I. INTRODUCTORY NOTES

Winter 2009, Mon-Wed 1-3pm, Angel G228

PS699: STATISTICAL METHODS II

A. Books &C: Shauman Drum; Amazon/Bean; Reserve

1. We will follow Greene strongly recommended; book you will want when done.

2. Kennedy also highly rec'd; intuitive explanations (not sufficient alone)

3. Crouch (or Kennedy) to bridge the two; most students initially prefer this

4. We will follow King & 2 Sage books (Allinach & Nelson; Menard) more

5. Other "Required" Materials:

directly than Green on qualitative/limited-dep-var models, esp. Logit/probit,

6. Additional Math and Probabilistic Resources:

(c) My lecture notes at Dollar Bill Copy. See also my web page.

(b) Some other articles and materials on reserve or to be made available throughout.

5. Additional Math and Probabilistic Resources:

(a) Math & Puzzles on Interactions.

4. Other "Required" Materials:

directly than Green on qualitative/limited-dep-var models, esp. Logit/probit,

3. Crouch (or Kennedy) to bridge the two; most students initially prefer this

2. Kennedy also highly rec'd; intuitive explanations (not sufficient alone)

1. We will follow Greene strongly recommended; book you will want when done.

A. Books &C: Shauman Drum; Amazon/Bean; Reserve

I. INTRODUCTORY NOTES

Winter 2009, Mon-Wed 1-3pm, Angel G228

PS699: STATISTICAL METHODS II
DATA: DATASETS OR INTEREST ACROSS POL-SCI SUBFIELDS

C.


2. Mostly from Lane, McKay, and Newton, Political Data Handbook of the OECD.

DATA: DATASETS OR INTEREST ACROSS POL-SCI SUBFIELDS

Other good idea you may have (check with me)

3. Replication & Extension (R&E Paper): 35%

B.

ASSIGNMENTS:

1. Periodic Problem Sets: 30%

2. Final Exam: 35%

3. Take-home, open-book/note...

Due to your GSI in manner we announce in class & by email, according to the schedule on page 2 of the syllabus, except as we may modify later & announce...

Due to your GSI in manner we announce in class & by email, according to the schedule on page 2 of the syllabus, except as we may modify later & announce...

All 10 on my web page now (may be changes this term...)
5. We may add others if we get to more / other material.

(b) Comparative Politics; Parliamentary Government Formation and Collapse

(a) Our Duration Model Example (if we get to it)

4. King, Al, Burns, and Laver, "Government Dissolution," APSR

(b) International Relations: Determinacy

(a) Our Dichotomous Dependent Variable Example

3. Health & Russian, "General Determinacy Between Enduring Rivals," APSR

(b) US Politics; Presidential Popularity

(a) Our Time-Series Example

2. Mackuen, Fiorina, and Stimson, "Presidents or Bankers?" APSR

(d) Socio-Economic-Political-READINESS: Size (Geog, Econ, Pops), Voter Part, Income Distribution, etc.

Governmental Fracationalization/Politicalization of Governments, etc.

(c) Government Characteristics: e.g., Duration/Number of Governments, Peace/Insurrection, etc.

Federal/Unitary; Levels of Govt, etc.

(b) Constitutional Readiness: e.g., # Seats in Legislature, Presidential/Parliamentary

(a) Comparative Politics & Our Main Workhorse
G. OTHER LOGISTICS & ADMINISTRATIVE PRELIMINARIES

3. Additionally, return of those make-up sessions?

2. Reschedule to Fri 1/16 & Fri 1/23 & Fri 1/30 & Fri 4/10? Time?

1. Away Mon & Wed 1/12 & 1/14; sessions Mon 1/19 & Wed 4/15 also missed.

E. SCHEDULING ADJUSTMENTS:

3. p-set turn-in procedures...

2. Office hours & office...

1. Section: Thursdays, 11:00-12:00, C325 Mason Hall

\text{Website:} \text{macmilian@umich.edu} (www.macmilian@umich.edu)

F. CSI CHEERS: BILL MACMILLAN

4. Office & Office Hours: Wed 3:30-5:00 (not today or next wk) & by apt.

3. My email: fransezep@umich.edu (don't call: I don't answer or check voice mail)

2. My web page: www.umich.edu/~franszee

I. Email by Wolverine Access course-list (so, if not officially registered...)

D. COMMUNICATION:
(1) Just a convenient way of referring to and manipulating a bunch of numbers at once.

2. Definition: A vertical array of numbers

2. (Column) Vector:

2. US population on 1/4/00 at noon, Number of countries in Africa on 1/12/97, etc.

(b) Standard Notation: Usually written in plain font, lower-case letters (Greek or Roman): a, b, c, d, e. Sometimes in italics: a, b, c.

(c) Examples:

[a, b, c; sometimes in italics: a, b, c]

1. Scalar:

1. Scalar: a plain-old, ordinary, every-day, regular number

A. TERMINOLOGY

DATA ANALYSIS

II. REFRESHER IN LINEAR ALGEBRA FOR STATISTICAL
NOTE: Since $\mathbf{q}_1 \mathbf{q}_2$ and the like are scalars, sometimes written italics, like $q_1 q_2$,

$\mathbf{q} = [0.035 \ 0.2 \ 0.4 \ 1] = \mathbf{q} \cdot [1 \ 0] = \mathbf{q}_1 \mathbf{q}_2 = [\mathbf{q}_1 \mathbf{q}_2] = [q_1 q_2]$. 

**Examples:**

- Standard Notation: As column vector, except w/ prime / prime / (, ) / (, )
- **Standard Notation:** a horizontal array of numbers

### 3. **Row Vector**

$\mathbf{q} = [0.035 \ 0.2 \ 0.4 \ 1] = \mathbf{q} \cdot [1 \ 0] = \mathbf{q}_1 \mathbf{q}_2 = [q_1 q_2]$. 

**Examples:**

- $q_1 q_2 q_3 q_4$
- $[q_1 q_2 q_3 q_4]$
- $\mathbf{q}$

**Usage:**

- **Usage:** 1. Usually written in bold, lower-case, Roman or Greek letters: $\mathbf{a}$, $\mathbf{b}$.
- **Usage:** 2. Sometimes written as a lower-case letter with a slight (~) or caret (^) under or over it. Particularly in settings where it's not obvious, e.g., proofs scrubbing on boards.
- **Usage:** 3. Underlined on board when we must clarify something is a vector. Ask if uncertain.
- **Usage:** 4. Either $\mathbf{bold}$ or sans serif or super scripts can get tedious, often simple letters or any format are used once it is unambiguous that the item in question is a (column) vector.

**Form:**

- **Form:** $\mathbf{q} = [q_1 q_2 q_3 q_4]$
4. **Matrix**

a) *Definition*: a rectangular array of numbers

(1) Thus, can view as set row vectors stacked vertically or column vectors horiz'ly *(concatenated)*
(2) Usually some substantive interpretation to rows and columns (e.g., see *data matrix* below).

b) *Standard Notation*:

(1) Usually denoted by upper-case letter, Greek or Roman, usually in bold: \( \mathbf{A}, \mathbf{\Sigma}, \mathbf{\Omega}, \mathbf{B} \), etc.;
(2) Can also refer to matrix as the set, \( \{ \cdot \} \), of its elements, e.g., \( A = \{ a_{ij} \} \), or by referencing a characteristic element of the matrix, e.g., \( A = [ a_{ij} ] \).
(3) Vertical, horizontal, and/or diagonal ellipses often used to denote generic ranges of elements or repeated elements of a particular defined matrix (see below).

c) *Examples*:

\[
\begin{align*}
\mathbf{A} &= \begin{bmatrix} a_{ij} \end{bmatrix} = \begin{bmatrix} a_i \end{bmatrix} = \begin{bmatrix} a_j \end{bmatrix}, \\
\mathbf{A} &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \\
\mathbf{\Omega} &= \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}, \\
\mathbf{B} &= \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \Rightarrow \mathbf{B}_{22} = \beta_{22} = 4
\end{align*}
\]

\[
\begin{bmatrix}
1 & \cdots & \cdots & \cdots & \cdots \\
1 & \cdots & \cdots & \cdots & \cdots \\
\vdots & \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \\
1 & \cdots & \cdots & \cdots & \cdots \\
1 & \cdots & \cdots & \cdots & \cdots \\
\end{bmatrix}
\begin{bmatrix}
\sigma^2 & 0 & \cdots & \cdots & 0 \\
0 & \ddots & 0 & \cdots & \vdots \\
\vdots & 0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 & \vdots \\
0 & \cdots & \cdots & 0 & \sigma^2
\end{bmatrix}
\begin{bmatrix}
\mathbf{V}(\mathbf{e}) = \mathbf{\Sigma}
\end{bmatrix}
\]
just above (upper first minor) the prime diagonal. Terms second minor and so on also exist.

(b) First minor diagonal: the elements on the diagonal just below (lower first minor) or

so term often used generically for any element of the (prime) diagonal.

(a) The off-diagonal runs bottom-left to top-right, less-often substantially important,

informally; those on the diagonal from top-left to bottom-right of the matrix.

7. (Prime) Diagonal of a Matrix: Formally, the set of elements $a_{ii}$, where $i = j$.

This might be said, "the observation or "the $i$th observation on $X_i".

for row $i$ or column $j$ or $a_{ij}$ (see above).

(d) Can refer to elements by their position such as "the $i$th element of $B_i"; in a data matrix,

as "the $i$th row, $j$th column" or $a_{ij}$ (see above).

(c) Elements’ positions or Coordinates in a matrix indexed by subscripts, usu. $\{i, j\}$; for row

b) Since each element is a scalar, we can write it as such, e.g.: $4, 0.36$, etc. (see above).

a) Thus, e.g., in a Data Matrix each element is an (1) observation on a (1) variable.

6. Elements of a matrix: the scalars that comprise a matrix.

(b) Each row is an observation on (each $x_k$ all of whose variables), usu. $i = 1 \ldots n$ or $i = 1 \ldots N$.

(a) Each column is a variable, usu. indexed $j = 1 \ldots p$, or $j = 0 \ldots \chi$ (see above).


\textbf{Note:} Scalars and vectors are subsets/special cases (1x1 and Ix1, or IxK or IxN of matrices.
\[
\begin{bmatrix}
1.67 & 0.75 & 3 \\
1.1 i & 5 & 1 \\
-0.1 i & -2 & 0 \\
1.97 & 4 & -5 \\
\end{bmatrix} = Z
\]

(a) **Square Matrix**: Matrix with \( n \times n \) rows and \( n \times n \) columns. E.g.: \( A, B, 0, Z \), etc.

(9) **Special Types of Matrices**: The types are ordered as nested special cases.

(c) **Examples**: From above, \( A, B \) are each \( 2 \times 2 \); \( X \) is \( 5 \times 5 \); \( Y \) is \( 3 \times 1 \); \( Q \) is \( 3 \times 1 \), and \( P \) is \( 1 \times 3 \).

(b) **Note**: Column vectors are \( n \times 1 \) matrices; row vectors \( 1 \times n \) matrices. (So, scalars are \( 1 \times 1 \) or \( n \times 1 \).

(a) **Dimensions of a Matrix**: Number of rows (usual) and columns (usual).
Diagonal Matrix: Formally, symmetric matrices with only its diagonal elements (possibly non-zero). In fact, any matrix such a matrix is symmetric (must be? why?). In fact, any matrix such a matrix is symmetric (must be? why?). Informally, matrix $A$ w/ elements above Frobenius norm $\|A\|_F$ (see Frobenius norm below); in fact, any matrix such a matrix is symmetric (must be? why?).
(1) Definition: Formally, a matrix \( \mathbf{a} \) is a scalar matrix \( \mathbf{W} \).

(2) Standard Notation: usn, written \( \mathbf{I} \), often w/ its dimension (symmetric so one dimension suffices) subscripted; \( \mathbf{I}^{n \times n} \).

(3) Identity matrix—a matrix with all elements equal to zero—rarely used.

Example: \( \mathbf{I}^{3 \times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \), \( \mathbf{I}^{4 \times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \).

Example: Homoskedastic Error-Var Covariance Matrix (scalar), \( \mathbf{W} = \mathbf{I}^{n \times n} \).

Scalar Matrices: Formally, diagonal matrices with \( n \) diagonal elements all the same number (scalar) \( \mathbf{W} = \mathbf{I}^{n \times n} \).

\[
\begin{bmatrix}
\lambda & 0 & 0 & \ldots & 0 \\
0 & \lambda & 0 & \ldots & 0 \\
0 & 0 & \lambda & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & \lambda
\end{bmatrix} = \begin{bmatrix} 16 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 5 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]
\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
5.5 & 0 & 0 & 0 \\
2 & 3.5 & 0 & 0 \\
1.5 & 1 & 0 & 0
\end{bmatrix} = \mathbf{L} = \begin{bmatrix}
4 & -2 & -1 & 12 \\
0 & 0 & 1 & 0 \\
0 & 0 & 2 & 2 \\
1.5 & 0 & 0 & 0
\end{bmatrix} \quad \text{(Upper-triangular)}
\]

\text{and you will often hear me call it that,}
\]

\text{\textbf{NOTE}: What is the multiplicative identity scalar? Is it simply also one?}

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix} = \mathbf{I}
\]

\text{\textbf{NOTE}: Is column-vector of ones? not to be confused with an identity vector of any sort. In fact, inner-product (see below) is the sum of the elements of } \mathbf{x} \text{, which is also often subscripted.}
\[ A \neq x, 0 \leq x, \iff \text{negative definite} \]

\[ A \neq x, 0 < x, \iff \text{positive definite} \]

\[ A \neq x, 0 < x, \iff \text{positive definite} \]

\[ \sum_{i=1}^{n} x_i = x, \iff \text{quadratic form} \]

1. **Positive and Negative Semi-definite**: See Cramer's Rule for more information. Including some related concepts for matrices only: positive and negative definite.

2. Positive (for all elements, else not). For example, $0 \leq 0 \iff \text{positive}$

3. Matrix positive $\iff 0$, negative $\iff 0$, weakly positive $\iff 0$, weakly negative $\iff 0$.

4. Greater than less than: Element-by-element comparison:

\[ A \neq B, \iff \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 6 & 3 \\ 2 & 1 \end{bmatrix} = B \]

5. **Algebraic Manipulation of Matrices**


   2. Algebraic Manipulation of Matrices:
\[
\begin{bmatrix}
6 \\
4 \\
2
\end{bmatrix} = \begin{bmatrix}
6 & 4 & 2
\end{bmatrix} = P \begin{bmatrix}
3 \\
5 \\
1
\end{bmatrix} = \begin{bmatrix}
3 \\
5 \\
1
\end{bmatrix} = V
\]

Formal Definition: \(B = AV\)

Standard Notation: \(B = AV\) for \(A\) a matrix (col. vectors transpose to row vectors). For \(A\) a matrix, transpose row (column one becomes row one, etc.) \(x\) each row into col.

2. Transposition: Intuitively, flips matrix along its axes, making each column into row.

Many regression quantities of interest have form \(x'Ax\). Examples:

- If \(A\) is positive/negative (semi-) definite, then \(x'Ax \geq 0\) and so \(x'Ax\) exists. What's \(x'Ax\)? So what?
- If \(A\) is positive/negative definite, \(x'Ax \neq 0\) and so \(x'Ax\) exists. What's \(x'Ax\)? So what?
- If none of the above: indefinite. So what?
\[ \mathbf{V} + \mathbf{B} = \begin{bmatrix} \mathbf{v}_1 + \mathbf{b}_1 \\ \mathbf{v}_2 + \mathbf{b}_2 \\ \mathbf{v}_3 + \mathbf{b}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 + \mathbf{b}_1 \\ \mathbf{v}_2 + \mathbf{b}_2 \\ \mathbf{v}_3 + \mathbf{b}_3 \end{bmatrix} = \mathbf{V} + \mathbf{B} \]

Proof: Matrix Addition Properties:

(1) Commutative: \( \mathbf{V} + \mathbf{B} = \mathbf{B} + \mathbf{V} \)

(2) Associative

Assuming conformability:

\( \mathbf{V} \land \mathbf{V} = \mathbf{0} + \mathbf{V} \)

Because \( \mathbf{0} \) is the additive identity matrix, written \( \mathbf{0} \) or \( \mathbf{0} \). 

(3) 

Technically, subtraction is defined as multiplying the latter matrix by -1, then adding.

\[ \begin{bmatrix} 6 \\ 4 \\ 2 \end{bmatrix} = \mathbf{V} + \mathbf{e} \quad \begin{bmatrix} 6 \\ 4 \\ 2 \end{bmatrix} = \mathbf{V} + \mathbf{e} \]

(4) 

Exception: \( \mathbf{Z} \) (non-square) and \( (\mathbf{1} \times \mathbf{m}) + (\mathbf{m} \times \mathbf{1}) \)

conformable:

\[ \begin{bmatrix} 4 \\ 2 \end{bmatrix} \]

such as:

\[ \begin{bmatrix} a + b \\ a + b \end{bmatrix} = \mathbf{V} + \mathbf{Z} \]

(5) 

Exception: I

Formally, adding or subtracting an element by element:

\[ \begin{bmatrix} \mathbf{a} + \mathbf{b} \\ \mathbf{a} + \mathbf{b} \end{bmatrix} = \mathbf{V} + \mathbf{B} \]

Examples:

For row-by-row or column-by-column:

\[ \begin{bmatrix} \mathbf{a} + \mathbf{b} \end{bmatrix} = \mathbf{V} + \mathbf{B} \]

Examples:

\[ \begin{bmatrix} \mathbf{a} + \mathbf{b} \end{bmatrix} = \mathbf{V} + \mathbf{B} \]

Examples:

\[ \begin{bmatrix} \mathbf{a} + \mathbf{b} \end{bmatrix} = \mathbf{V} + \mathbf{B} \]

(6) 

Exception: For addition A \& B conformable for addition: \( A \\& B \)

conformable for addition: \( A \\& B \)

\( \dim(\mathbf{V}) = \dim(\mathbf{B}) \)

\( \dim(\mathbf{V}) = \dim(\mathbf{B}) \)
Examples: BOARD (see notes)

(a) Inner-product multiplication is commutative: \( a \cdot b = b \cdot a \)

(b) This implies:

\[
\sum_{n} q_n^2 = q_1^2 + q_2^2 + \ldots + q_n^2 = q \cdot b
\]

(c) Written as \( a \cdot b \) or \( a \cdot b \\
Vector Multiplication: Inner Products

where the term originates: scaling up or down, 

A vector times a vector or matrix: just multiply every element by the scalar: this is probably

You already know all about multiplying two or more scalars.

You must be able to do this:

Scalar Multiplication:

(a) Just like in scalar multiplication, absence of an operation signifies multiplication.

Matrix Multiplication (and „Division“):

\[
(b + c) + d = [b + c + d] \cdot a = [b + c + d] = (b + c) + (d + a) = c + (b + a)
\]
Matrix Multiplication

Examples:

(a) Not Commutative: $AB$ does not necessarily equal $BA$ (may not even exist).

(b) $B$ post-multiplied by $A$ is $BA$.

(c) $B$ pre-multiplied by $A$ is $AB$.

(d) Pre-Multiplication and Post-Multiplication are different and may not both exist.

When $A \times C$ is multiplied times $B \times C$, you get an $A \times C$ matrix.

Properties and Facts (assuming conformability):

(1) Formal definition: $C = AB = \sum_{f} a_{B}^{c} p_{A}^{f}$

(2) For two matrices to be conformable for multiplication, the number of columns in the first must equal the number of rows in the second (e.g., $m \times n$).

(3) Informal Recipe for Solutions to Matrix Multiplication Problems

- Start by noting dimensions of the matrices to be multiplied.
- Draw an $(p \times q)(q \times r)$ matrix box for the answer.
- Then, the $f$th element of the solution is the inner product of the $f$th row of the first matrix and the $q$th column of the second.

(4) Examples: BOARD (see notes)

Column of the second, fill in the answer cell for column-by-column.

(5) When $A \times B = C$, if $E$ is the matrix solution will be $A \times B$. Draw an $A \times B$ matrix box for the answer.

(6) First, $p$ must equal $q$; if not, the matrices are not conformable and it can't be done so you're done.

(7) Recipes are just special cases of matrices, we already know that matrix multiplication must work something like inner products.

First, since vectors are just special cases of matrices, we
\[ u_3 + u_1 x_1 g + \cdots + u_0 x_0 g = u_0 \Lambda \]

\[ \cdots = \cdots \]

\[ v_3 + v_1 x_1 g + \cdots + v_0 x_0 g = v_0 \Lambda \]

\[ z_3 + z_1 x_1 g + \cdots + z_0 x_0 g = z_0 \Lambda \]

\[ l_3 + l_1 x_1 g + \cdots + l_0 x_0 g = l_0 \Lambda \]

 realmente, Einstein once noted all advancement in mathematics is advance in notation, (brief part of the Whole Point,)

Writing Sets of Equations in Matrix Notation

(8) Important: Matrix \( A \) is Idempotent if \( AA = A \) (implies that \( A \) must be square to be idempotent). If \( A \) is symmetric, then \( AA \) and \( AA^T \) are also \( A \) (since \( AA = A \)) and \( AA^T = A \).

(i) Transposition is Distributive in Reverse Order: \( AB^T = (BA)^T \)

may not even exist: EXAMPLES

(6) Distributive: \( A(B+C) = AB + AC \) (note not commutative, so order matters here; again, BA and CA

EXAMPLES: \( AB \neq BA \))
\[ 0 + qX = 3 + \gamma g + \delta g + \epsilon g = \lambda \]
I. Terminology

A. Scalar: \( a_i = 7 \)

B. Vectors:
\[
\begin{bmatrix}
1 & 3 & 5 & 7 \\
\end{bmatrix}
\]
\( a_2 = 3 \) or \( a_3 = 3 \)

C. Matrices:
\[
A = \begin{bmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{bmatrix}
\]
\( B = \begin{bmatrix}
1 & 4 \\
3 & 2
\end{bmatrix}
\)
\( X = \begin{bmatrix}
1 & | & | & | & | \\
V_1 & V_2 & \cdots & V_k
\end{bmatrix}
\]
\( \to \) All vectors & scalars are also special cases of matrices

D. Data Matrix: \( \text{rows} = \text{observations} \)
\( \text{cols} = \text{variables} \)

E. Elements:
\( A = \xi a_{ij}^3 \)
- Position: \( i^{th} \) row, \( j^{th} \) column
- The \( i^{th} \) observation on \( j^{th} \) column
- The \( i^{th} \) element of \( A \)

F. (Prime) Diagonal
\[
\begin{bmatrix}
a_{11} & a^{(1,2)} \\
a^{(2,1)} & a_{22}
\end{bmatrix}
\]

G. Dimensions:
- \( a \) is \( 1 \times 4 \)
- \( \overline{a} \) is \( 4 \times 1 \)
- \( A \) is \( 2 \times 2 \)
So is \( B \), \( X \) is \( n \times (k+1) \)

H. Symmetric Matrix:
\( \to \) Any \( A' A \) is symmetric; e.g., \( X' X \)

I. Var-Cov Matrices are \( \mathbf{cov} \) symmetric; e.g., in \( AR(1) \) case
\[
V(E_d) = \frac{\sigma_2}{\phi_2^2}
\begin{bmatrix}
1 & \phi_1 & \phi_2 \\
\phi_1 & 1 & \phi_2 \\
\phi_2 & \phi_1 & 1
\end{bmatrix}
\]
\( \sigma_1 \equiv \begin{bmatrix}
4 & 2 & 5 \\
2 & 3 & 1 \\
5 & 1 & 7
\end{bmatrix} \)

G. Diagonal Matrix:
- Symmetric & zero off-diagonal
- \( \eta \), \( \xi \) diagonal
- \( V(E) \) in the usual heteroskedastic case

H. Scalar Matrix:
Diagonal & the diagonal is same \( \neq \)

I. Identity:
\[
I_3 = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\]
\( \sigma_e \equiv \begin{bmatrix}
3000 \\
0030 \\
0003
\end{bmatrix} \)
\( V(E_h) = \begin{bmatrix}
\sigma_2^2 & 0 & 0 \\
0 & \sigma_2^2 & 0 \\
0 & 0 & \sigma_2^2
\end{bmatrix} \)

J. Trang:
\[
I_2 = \begin{bmatrix}
3 & 0 & 0 \\
1 & 5 & 0 \\
4 & 2 & 7
\end{bmatrix}
\]
\( \sigma_e \equiv \begin{bmatrix}
3 & 1 & 4 \\
1 & 5 & 2 \\
0 & 7 & 1
\end{bmatrix} \)

- Note the shorthand; this is used a lot
I. Algebraic Manipulation:

A. Equality:

\[
A = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}, \quad A \neq B
\]

\[
C = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, \quad C = D
\]

B. Transpose:

\[
A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}, \quad A' = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}, \quad a = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \quad a' = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}
\]

\[
B = \begin{bmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \end{bmatrix}, \quad B' = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 \end{bmatrix}
\]

C. Adding Subtracting:

\[
A = \begin{bmatrix} 3 & 5 \\ 7 & 9 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}, \quad C = A + B = \begin{bmatrix} 5 & 9 \\ 13 & 17 \end{bmatrix}
\]

\[
a = 1, \quad D = A + a = \begin{bmatrix} 4 \\ 6 \\ 8 \\ 10 \end{bmatrix}
\]

\[
\bar{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad E = A - \bar{a} = \begin{bmatrix} 4 \\ 6 \\ 9 \\ 11 \end{bmatrix}
\]

\[
a = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad F = A + a = \begin{bmatrix} 4 \\ 7 \\ 8 \\ 11 \end{bmatrix}
\]

D. Matrix Multiplication:

1. \(a'\ b\)

\[
a' = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \quad b = \begin{bmatrix} 3 & 5 \\ 7 & \end{bmatrix}
\]

\[a'\ b = 12 + 3\cdot4 + 5\cdot6 = 2 + 12 + 30 = \begin{bmatrix} 44 \\ 30 \end{bmatrix}
\]

2. \(b'\ a = \begin{bmatrix} 2 & 4 & 6 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} = 2.1 + 4.3 + 6.5 = \begin{bmatrix} 44 \end{bmatrix}
\]

3. \(a'\ b' = \begin{bmatrix} 1 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} 2 & 4 & 6 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} 2 & 14 \\ 18 & 28 \end{bmatrix}
\]

\[n.b.: \quad e'e = \Sigma e_i^2
\]

4. \(a = \begin{bmatrix} 7 \\ 1 \\ 3 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}
\]

\[a\ B = \begin{bmatrix} 7 \\ 21 \\ 35 \end{bmatrix}, \quad a\ B = \begin{bmatrix} 7 & 14 \\ 21 & 28 \end{bmatrix}
\]

Properties:

1) Commutative does not hold

2) \([A \times B] = B \times A\) if not square matrices

3) \(A\ B = B\ A\) if dimensions match

4) \(B\ A = \begin{bmatrix} 7 & 19 \\ 14 & 30 \end{bmatrix}\) if dimensions match
Further Examples (cont.)

II. D. Matrix Multiplication:

Properties:

b) Associative Holds: \((AB)C = A(BC)\)
\[
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
\end{bmatrix}
\begin{bmatrix}
5 & 7 \\
6 & 8 \\
\end{bmatrix}
\begin{bmatrix}
0 & 2 \\
1 & 0 \\
\end{bmatrix}
= 
\begin{bmatrix}
17 & 23 \\
36 & 53 \\
\end{bmatrix}
\begin{bmatrix}
0 & 2 \\
1 & 0 \\
\end{bmatrix}
= 
\begin{bmatrix}
23 & 34 \\
53 & 78 \\
\end{bmatrix}
\]
\[
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
\end{bmatrix}
\left( 
\begin{bmatrix}
5 & 7 \\
6 & 8 \\
\end{bmatrix}
\begin{bmatrix}
0 & 2 \\
1 & 0 \\
\end{bmatrix}
\right)
= 
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
\end{bmatrix}
\begin{bmatrix}
7 & 10 \\
8 & 12 \\
\end{bmatrix}
= 
\begin{bmatrix}
23 & 34 \\
53 & 78 \\
\end{bmatrix}
\]

c) Distributive: \(-A(B + C) = AB + AC\)
\[
\begin{bmatrix}
1 & 0 \\
0 & 2 \\
\end{bmatrix}
\left( 
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
\end{bmatrix}
+ 
\begin{bmatrix}
4 & 3 \\
2 & 1 \\
\end{bmatrix}
\right)
= 
\begin{bmatrix}
1 & 0 \\
0 & 2 \\
\end{bmatrix}
\begin{bmatrix}
5 & 5 \\
5 & 5 \\
\end{bmatrix}
= 
\begin{bmatrix}
5 & 5 \\
10 & 10 \\
\end{bmatrix}
\]
\[
\begin{bmatrix}
1 & 0 \\
0 & 2 \\
\end{bmatrix}
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
\end{bmatrix}
+ 
\begin{bmatrix}
1 & 0 \\
0 & 2 \\
\end{bmatrix}
\begin{bmatrix}
4 & 3 \\
2 & 1 \\
\end{bmatrix}
= 
\begin{bmatrix}
2 & 4 \\
4 & 2 \\
\end{bmatrix}
+ 
\begin{bmatrix}
4 & 2 \\
4 & 2 \\
\end{bmatrix}
= 
\begin{bmatrix}
5 & 5 \\
5 & 5 \\
\end{bmatrix}
\]

*Note: \(-A(B + C) \neq BA + CA\) because it doesn't hold.

d) Identity: \(I \cdot A = A\)
\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
5 & 6 \\
\end{bmatrix}
= 
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
5 & 6 \\
\end{bmatrix}
= 
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
5 & 6 \\
\end{bmatrix}
\]
\[
A \cdot I = A
\]
\[
\begin{bmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9 \\
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
\end{bmatrix}
= 
\begin{bmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9 \\
\end{bmatrix}
\]

C) \((AB)^T = B^T A^T\)

Term: \(A\) is idempotent \(\iff\) \(AA = A\)

E: Writing Sets of equations in Matrix Notation
\[
\begin{align*}
3x + 4y &= z_1 \\
5x + 6y &= z_2
\end{align*}
\]
\[
\begin{bmatrix}
3 & 4 \\
5 & 6 \\
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
\end{bmatrix}
= 
\begin{bmatrix}
z_1 \\
z_2 \\
\end{bmatrix}
\]
\[
\begin{align*}
2x &= z_1 \\
y_x + 2y &= z_2
\end{align*}
\]
\[
\begin{bmatrix}
2 & 0 \\
4 & 2 \\
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
\end{bmatrix}
= 
\begin{bmatrix}
z_1 \\
z_2 \\
\end{bmatrix}
\]

III. Geometry of Matrix Algebra

A. Thinking of it as coordinates: \(\bar{a} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}\)

1) Any \(\bar{c}\) in 2-D space can be written as a linear combo of \(\bar{a}\) and \(\bar{b}\):
\[
\bar{c} = \begin{bmatrix} 15 \\ 17 \end{bmatrix} = a_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + a_2 \begin{bmatrix} -7 \\ 3 \end{bmatrix}
\]
\[
\Rightarrow a_1 = -\frac{7}{3}, a_2 = \frac{54}{13}
\]

Zoinks! (How did I get these?)
(See Cramer's Rule)

2) Thus, usu. any two vectors "span" 2-D space. Two lines determine a plane. But not if the two vectors are prop. Consider \(\bar{a}\) and \(\bar{c}\):
\[
\begin{bmatrix} 4 \\ 2 \end{bmatrix}
\] Notice \(\bar{c} = 2\bar{a}\). XX

Oops: that draws just one line. You can draw \(\infty\) planes through that!
3) This generalizes: usually \( n \) vectors (\( n \) columns of a matrix) give you \( n \)-dimensions of information, but not if some are linearly dependent on others.

\[
\begin{bmatrix}
  \mathbf{a}
\end{bmatrix} = \begin{bmatrix}
  1 \\
  2
\end{bmatrix}, \quad
\begin{bmatrix}
  \mathbf{b}
\end{bmatrix} = \begin{bmatrix}
  1 \\
  4
\end{bmatrix}, \quad
\begin{bmatrix}
  \mathbf{c}
\end{bmatrix} = \begin{bmatrix}
  1 \\
  5
\end{bmatrix}
\]

\[
\begin{bmatrix}
  \mathbf{c}
\end{bmatrix} = \mathbf{a} + \mathbf{b}, \quad \text{since } \mathbf{c} = \mathbf{a} + \mathbf{b}
\]

you throw the third away; only 2 pieces contribute.

B. Rank: Rank of Matrix is simply # of indep. columns [Rules 1-4 regarding rank.]

1) But why do we care about linear indep of rank?

We are going to be interested in finding (estimating) \( \mathbf{b} \) in

\[
\mathbf{y} = \mathbf{X} \mathbf{b} + \mathbf{e}
\]

\[
\Rightarrow \begin{bmatrix}
  \mathbf{X}
\end{bmatrix} = \begin{bmatrix}
  \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_n
\end{bmatrix}
\]

so that \( \mathbf{X} \) has as many columns as there are variables.

But for this to work, we need columns of \( \mathbf{X} \) to be linearly independent.

Example: \( \mathbf{X} = \begin{bmatrix}
  1 & 1 & 1 \\
  1 & -1 & 1
\end{bmatrix} \)

Then

\[
\mathbf{y} = \begin{bmatrix}
  1 \\
  2
\end{bmatrix}
\]

The solution is

\[
\mathbf{b} = \begin{bmatrix}
  b_1 \\
  b_2 \\
  b_3
\end{bmatrix}
\]

2) If all \( \mathbf{X} \)'s columns are indep., we say it has "full (column) rank.

We want \( \mathbf{X} \)'s of full rank.

C) But how can you tell if some matrix is of full rank? Determinants.

a) \( \det \mathbf{A} \) = \( \sum_{i=1}^{n} a_{i1} a_{i2} \ldots a_{in} \)

b) \( \det \mathbf{A} = \begin{vmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{vmatrix} = a_{11} a_{22} a_{33} + a_{13} a_{22} a_{31} + a_{12} a_{32} a_{31} - a_{13} a_{23} a_{32} a_{21} a_{33} \\
\]

\[
\Rightarrow \frac{1}{\mathbf{X}} = \begin{bmatrix}
  \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_n
\end{bmatrix}
\]

\[
\Rightarrow \mathbf{X} = \begin{bmatrix}
  \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_n
\end{bmatrix}
\]

\[
\Rightarrow \begin{bmatrix}
  \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_n
\end{bmatrix} = \begin{bmatrix}
  \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_n
\end{bmatrix}
\]

\[
\begin{bmatrix}
  1 & 1 & 1 \\
  1 & -1 & 1
\end{bmatrix}
\]

\[
= 1 \cdot 3 \cdot 5
\]

C) \( \det \mathbf{A} = \sum_{i=1}^{n} a_{ij} (-1)^{i+j} \det \mathbf{A}_{ij} \) for any \( i \), where \( \mathbf{A}_{ij} \) is \( \mathbf{A} \) with row \( i \) and col \( j \) removed.

\[
\mathbf{A} = \begin{bmatrix}
  1 & 2 & 3 \\
  2 & 1 & 2 \\
  1 & 3 & 0
\end{bmatrix}
\]

\[
\Rightarrow \det \mathbf{A} = (\text{next page})
\]

\[
\Rightarrow \det \mathbf{A} = \begin{bmatrix}
  1 & 2 & 3 \\
  2 & 1 & 2 \\
  1 & 3 & 0
\end{bmatrix}
\]

\[
\Rightarrow \det \mathbf{A} = \begin{bmatrix}
  1 & 2 & 3 \\
  2 & 1 & 2 \\
  1 & 3 & 0
\end{bmatrix}
\]

\[
\Rightarrow \det \mathbf{A} = \begin{bmatrix}
  1 & 2 & 3 \\
  2 & 1 & 2 \\
  1 & 3 & 0
\end{bmatrix}
\]

\[
\Rightarrow \det \mathbf{A} = \begin{bmatrix}
  1 & 2 & 3 \\
  2 & 1 & 2 \\
  1 & 3 & 0
\end{bmatrix}
\]

\[
\Rightarrow \det \mathbf{A} = \begin{bmatrix}
  1 & 2 & 3 \\
  2 & 1 & 2 \\
  1 & 3 & 0
\end{bmatrix}
\]

\[
\Rightarrow \det \mathbf{A} = \begin{bmatrix}
  1 & 2 & 3 \\
  2 & 1 & 2 \\
  1 & 3 & 0
\end{bmatrix}
\]

\[
\Rightarrow \det \mathbf{A} = \begin{bmatrix}
  1 & 2 & 3 \\
  2 & 1 & 2 \\
  1 & 3 & 0
\end{bmatrix}
\]
Further Examples, Cont.

III.C.1. \[ \cdots = \begin{vmatrix} 1 & 0 & 2 & \frac{1}{3} \\ 0 & 1 & \frac{1}{2} & 0 \\ 1 & 0 & \frac{3}{8} & 0 \\ 0 & 1 & 1 & 0 \end{vmatrix} - 3 \begin{vmatrix} 1 & 0 & 2 & \frac{1}{3} \\ 0 & 1 & \frac{1}{2} & 0 \\ 1 & 0 & \frac{3}{8} & 0 \\ 0 & 1 & 1 & 0 \end{vmatrix} - 3 \begin{vmatrix} 1 & 0 & 2 & \frac{1}{3} \\ 0 & 1 & \frac{1}{2} & 0 \\ 1 & 0 & \frac{3}{8} & 0 \\ 0 & 1 & 1 & 0 \end{vmatrix} \]

\[ = (1+0+0-6-0-0) + 2 \left((0+1+1) - (0+2+6)\right) - 3 \left((0+1+0) - (0+0+0)\right) \]

\[ = -5 + 2 \cdot (1-8) - 3 \cdot (1-6) \]

\[ = -5 + 10 + 15 = 20 \]

So what? Remember if \( |A| = 0 \), then \( A \) "less than full rank" or "has linearly dependent columns".

\[ \begin{pmatrix} 1 & 2 & 4 \\ 5 & 6 & 18 \\ 3 & 10 & 14 \end{pmatrix} \]

\[ |A| = (1 \cdot 6 \cdot 14 + 2 \cdot 18 \cdot 3 + 4 \cdot 10 \cdot 5) - (4 \cdot 6 \cdot 3 + 2 \cdot 18 \cdot 14 + 1 \cdot 10 \cdot 18) \]

\[ = (84 + 108 + 200) - (72 + 140 + 180) \]

\[ = 392 - 392 = 0 \]

(!!!)

2) Why? (loose intuition):
- Note that \( A = \begin{pmatrix} 1 & 0 \end{pmatrix} \) can be taken to describe a rectangle.
- Area of a rectangle is \( \text{Area} = \ell w = a_{11}a_{22} \); more generally, any two independent columns will describe a parallelogram such as \( B = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \)
- Thus determinant is the area of the parallelogram described by the two column vectors. If, however, the columns are dependent, then as before they lie on a line.

\[ \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \]

For 3-dimensions, determinant is volume. Again, dependence implies we describe only a plane or a line, either of which have \( \emptyset \) volume.

- Analogously for more than 3D (if anyone can think that high)

D. Projections

\[ y = (4, 2) \]
\[ x = (2, 0) \]

Shortest gap from \( y \) \( (\text{which for now can be thought of as "what we'd like to know"}) \) and \( x \) will be achieved when \( e = y - xb \) is perpendicular to \( x \).

1) Orthogonality: a) Just the \( n \)-dimensional analogue to perpendicular

b) Why is \( a \cdot b = 0 \) when \( a \perp b \)?
or "orthogonal" they don't "go together" at all. They go in 4 directions.

2) So, to get the best $b$ here, we want to solve:

\[ (\overline{x} b)' e = 0 \]
\[ (\overline{x} b)' (\overline{y} - \overline{x} b) = 0 \]
\[ b \overline{x}' (\overline{y} - \overline{x} b) = 0 \]
\[ b \overline{x}' \overline{y} - b \overline{x}' \overline{x} b = 0 \]
\[ \overline{x}' \overline{x} b \implies b = (\overline{x}' \overline{x})^{-1} \overline{x}' \overline{y} \]

Least-Squares, if $\overline{x}' \overline{x}$ is a scalar, but what if it's not?

Least-Squares if $\overline{x}' \overline{x}$ (n x 1)

IV Solving a System of Equations:

A. Suppose Problem is $\overline{y} = \overline{x} b + e$; ie, have bunch of columns of info

1) Again, that means we want $(\overline{x} b)' e = 0$ for as we saw before:

2) Rewrite $e$:

$$(\overline{x} b)' (\overline{y} - \overline{x} b) = 0$$

Rearrange: $b' \overline{x}' \overline{y} - b' \overline{x}' \overline{x} b = 0$

$\overline{x}' \overline{y} - \overline{x}' \overline{x} b = 0$

$\overline{x}' \overline{x} b = \overline{x}' \overline{y}$

3) So, if we could divide by $(\overline{x}' \overline{x})$ we would finish as in the single-case, but $(\overline{x}' \overline{x})$ is $k \times k$, so we don't know how to divide by a matrix. In fact, you can't, but you can multiply by an inverse matrix.

B. Inverse Matrices: $BA = I \iff B = A^{-1}$

Two-by-Two Case: $\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\Rightarrow b_{11} a_{11} + b_{12} a_{21} = 1$
\[ @ b_{11} a_{12} + b_{12} a_{22} = 0 \]
\[ @ b_{21} a_{11} + b_{22} a_{21} = 0 \]
\[ @ b_{21} a_{12} + b_{22} a_{22} = 1 \]

So, $A^{-1} = \frac{1}{|A|} \cdot \text{Cofactor}(A)_{i,j}$

* In the 2x2 case, the cofactor is 1) swap the diagonal elements

2) put minus signs in front of off-diagonals

* More generally, $A^{-1} = \frac{1}{|A|} \cdot \begin{vmatrix} -1 \end{vmatrix} \begin{vmatrix} A \end{vmatrix}^{-1}$

Another BEAR to do manually Computers
(2) Noting that the coefficient is the same in each equation, we could collapse them to a scalar and rewrite each variable as a column vector like so:

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  \vdots \\
  y_n \\
\end{bmatrix} =
\begin{bmatrix}
  X_{10} \\
  X_{20} \\
  X_{30} \\
  \vdots \\
  X_{n0} \\
\end{bmatrix}
+ \begin{bmatrix}
  b_1 \\
  b_2 \\
  b_3 \\
  \vdots \\
  b_n \\
\end{bmatrix} + \begin{bmatrix}
  X_{11} \\
  X_{21} \\
  X_{31} \\
  \vdots \\
  X_{n1} \\
\end{bmatrix} + \begin{bmatrix}
  X_{12} \\
  X_{22} \\
  X_{32} \\
  \vdots \\
  X_{n2} \\
\end{bmatrix} + \begin{bmatrix}
  X_{13} \\
  X_{23} \\
  X_{33} \\
  \vdots \\
  X_{n3} \\
\end{bmatrix} + \cdots + \begin{bmatrix}
  X_{1k} \\
  X_{2k} \\
  X_{3k} \\
  \vdots \\
  X_{nk} \\
\end{bmatrix} + \begin{bmatrix}
  \varepsilon_1 \\
  \varepsilon_2 \\
  \varepsilon_3 \\
  \vdots \\
  \varepsilon_n \\
\end{bmatrix}
\]

...thus highlighting that the vector of observations on the dependent variable is a linear combination of the vectors of observations on the independent variables plus residuals.

f. Or we could write it like this which will often save us all a lot of hassle:

\[
y = X \cdot b + e
\]

\[(nx1) \quad (nxk) \quad (kx1) \quad (nx1)\]

C. The Geometry of Matrix Algebra

1. Think of the elements of an \(nx1\) vector as coordinates in \(n\)-dimensional space; e.g., the two-dimensional \(x-y\) graphs you are familiar with since grade school, or, equivalently, as a line segment from the origin to those coordinates (familiar from H.S.).

   a. Given two such vectors, \(a\) and \(b\), which are not on the same line, any third vector in the same dimensions, \(e\), can be expressed as a linear combination of them: any \(e = \alpha_1 a + \alpha_2 b\).

   b. The set of all possible third vectors that can be expressed as a linear combination of \(a\) and \(b\) would cover the whole plane. So we can draw the whole plane from just 2 "different" vectors: the old saw that two lines or three points define a plane (third point is the origin).

   (1) Important exception: if \(b\) strictly proportional to \(a\), i.e., if \(b = \delta a\) for some \(\delta\), then \(b\) isn’t really adding any new info to \(a\) – there’s only one unique vector here, the other lies right on top of it, \(b\) is just scaled up (down) version of \(a\). Thus \(a\) & \(b\) give only one line: not enough to define plane (rather you can draw any number of planes through a single line).

   (2) Example: suppose I gave you two vectors: one was \(a = \text{[Female Voters as % of total registered, Male Voters as % of total registered]}\) and the other was \(b = \text{[FV as % of Population, MV as % of Population]}\), and if Males and Females register to vote at the same rate \(\alpha\), then \(a\) is proportional to \(b\) since \(a\) will be \((1/\alpha)b\). Notice, if there is no difference in their registration rates, \(b\) adds no new information to \(a\) and vice versa.

   c. All of this generalizes to the more than two-dimensional cases. Usually \(n\) vectors will define \(n\)-dimensional space, but they must be \(n\) linearly independent vectors. Any vector that can be perfectly expressed as some linear function \((y = mx + b)\) of some other(s) adds no new information (and thus does not "span" another dimension when brought into the matrix).

2. Rank of a Matrix: the number of linearly independent columns (vectors) in that matrix.

   a. Rules:

   (1) Rule #1: Can’t have more dimensions (higher rank) than columns; recall, each column contains the coordinates of a point \(\Rightarrow\) can’t have more dimensions than points.

   (2) Rule #2: Can’t have more dim’s (higher rank) than rows; recall each row provides the
actual values for those coordinates \(\Rightarrow\) can’t have more dim’s than you give values to.

(3) Rules 1 & 2 \(\Rightarrow\) Rule #3: \(\text{Rank}(A) \leq \min\{\text{rows}(A), \text{cols}(A)\}\)

(4) Rule #4: \(\text{Rank}(AB) \leq \min\{\text{Rank}(A), \text{Rank}(B)\} \) : since \(AB\) is a linear combination of \(A\) and \(B\) it can’t create any new linearly independent information. Thus, it can span no more dimensions (provide no more information) than the least of those two. It could provide less information if some columns in \(A\) are linearly dependent on \(B\) and/or vice versa.

b. Why do we care? See Examples: Handout

3. **Determinants**: How can we tell if a matrix has **full rank**? Determinant \(\neq 0\).
   a. **Standard Notation**: “the determinant of \(A\)” is written \(|A|\) (only defined for square matrices)
   b. How do you calculate it?
      (1) 1x1 (scalar) case” \(|a| \equiv a\)
      (2) 2x2 case: \(|A| = a_1a_2 - a_1a_2\)
      (3) 3x3 case: \(|A| = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32})\)
      (4) General Case: using any row \(i\), \(|A| = \sum_j (-1)^{i+j} a_{ij} |A_{ij}|\) where \(A_{ij}\) is the matrix made by deleting row \(i\) and column \(j\) from the original matrix \(A\).
         a) Solving for anything larger than a 3x3 determinant without a computer should be done only once – in a mean prof’s first problem set – and never again!
         b) Hint: Where possible, choose a row \(i\) to delete that leaves as many zeros as possible, and then as many ones, to simplify (a little) calculating \(|A_{ij}|\).
   c. Examples: Handout
   d. **Why is determinant zero if matrix not full rank?** Coordinates of a pair of (2x1) vectors can be seen as describing a parallelogram (add each vector to the other’s end to draw the figure). The determinant then happens to be the formula for the area of this figure. Recall that if these two vectors are linearly dependent, then they lie on the same ray from origin. Accordingly, the “parallelogram” they draw collapses to a line, which has zero area. This generalizes. In 3-dimensional space, the vectors define a 3-d shape, the volume of which is given by the determinant. If one of the three vectors is a linear function of either or both the others, the shape collapses to a plane which has 0 volume. (If all three are linear functions of one vector, then it collapses to a line, which also has 0 volume.) This generalizes in some loose sense.

4. **Projections**: Suppose now we have 2 dimensions of information and we’d like to summarize it in one dimension. Alternatively, suppose we have some \(y\) and we’d like to express it as nearly as possible using some other vector of information \(x\) (perhaps \(x\) is more readily available), but we want to keep everything linear because curved lines gives us a headache.
   a. In other words, suppose we want to choose some scalar coefficient, \(b\), which when multiplied by \(x\) gets us as close as possible to \(y\): \(i.e., \) we want to minimize \(e\) in \(y = bx + e\). (Examples)
   b. Clearly, the thing to do is extend (scale up or down by \(b\)) \(x\) until the line from \(y\) to \(xb\), \(i.e., y-xb\) or \(e\), is perpendicular to \(xb\). (see graph in examples)
      (1) sub-point 1: \(e \perp xb\) is read “\(e\) is orthogonal to \(xb\)”; orthogonal is the general (n-dimensional) analog to perpendicular
      (2) sub-point 2: if \(a \perp e\), then \(a'c = 0\) (for why see examples)
c. so, we want to solve \((\mathbf{x}b)'\mathbf{e} = 0\) for \(b\).

\[
(1) \quad (\mathbf{x}b)'\mathbf{e} = (\mathbf{x}b)'(\mathbf{y} - \mathbf{x}b) = b\mathbf{x}'\mathbf{y} - b\mathbf{x}'\mathbf{x}b = \mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}b = 0
\]

(2) rearranging the last equation: \(\mathbf{x}'\mathbf{y} = \mathbf{x}'\mathbf{x}b\), or \(b = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x}} \approx (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}\)

d. Don’t look now, but we just derived the least-squares formula for \(b\) in the univariate case. Unfortunately, this particular solution only works when \(\mathbf{x}'\mathbf{x}\) is a scalar. (When will and won’t it be?) Suppose, instead, \(\mathbf{X}'\mathbf{X}\) is a matrix; how could we solve the analogous problem?

D. Solving a System of Equations:

1. Suppose that now we have \(k\) columns of information in \(\mathbf{X}\) (i.e., \(\mathbf{X}\) is of full rank \(k\)) and that we’d like to use all of this information to get as close as possible to the vector \(\mathbf{y}\).

a. First of all, notice that this is just an extension of the problem from “projections”. We have a bunch of equations like so:

\[
\begin{align*}
\mathbf{y}_1 &= b_0\mathbf{X}_{10} + b_1\mathbf{X}_{11} + b_2\mathbf{X}_{12} + \ldots + b_k\mathbf{X}_{1k} + \mathbf{e}_1 \\
\mathbf{y}_2 &= b_0\mathbf{X}_{20} + b_1\mathbf{X}_{21} + b_2\mathbf{X}_{22} + \ldots + b_k\mathbf{X}_{2k} + \mathbf{e}_2 \\
\mathbf{y}_3 &= b_0\mathbf{X}_{30} + b_1\mathbf{X}_{31} + b_2\mathbf{X}_{32} + \ldots + b_k\mathbf{X}_{3k} + \mathbf{e}_3 \\
&\vdots \\
\mathbf{y}_n &= b_0\mathbf{X}_{n0} + b_1\mathbf{X}_{n1} + b_2\mathbf{X}_{n2} + \ldots + b_k\mathbf{X}_{nk} + \mathbf{e}_n
\end{align*}
\]

(1) Which, as we noted in a previous section can be written as:

\[
\begin{bmatrix}
\mathbf{y}_1 \\
\mathbf{y}_2 \\
\mathbf{y}_3 \\
\vdots \\
\mathbf{y}_n
\end{bmatrix}
= b_0
\begin{bmatrix}
\mathbf{X}_{10} \\
\mathbf{X}_{20} \\
\mathbf{X}_{30} \\
\vdots \\
\mathbf{X}_{n0}
\end{bmatrix}
+ b_1
\begin{bmatrix}
\mathbf{X}_{11} \\
\mathbf{X}_{21} \\
\mathbf{X}_{31} \\
\vdots \\
\mathbf{X}_{n1}
\end{bmatrix}
+ b_2
\begin{bmatrix}
\mathbf{X}_{12} \\
\mathbf{X}_{22} \\
\mathbf{X}_{32} \\
\vdots \\
\mathbf{X}_{n2}
\end{bmatrix}
+ \ldots + b_k
\begin{bmatrix}
\mathbf{X}_{1k} \\
\mathbf{X}_{2k} \\
\mathbf{X}_{3k} \\
\vdots \\
\mathbf{X}_{nk}
\end{bmatrix}
+ \begin{bmatrix}
\mathbf{e}_1 \\
\mathbf{e}_2 \\
\mathbf{e}_3 \\
\vdots \\
\mathbf{e}_n
\end{bmatrix}
\]

(2) or as: \(\mathbf{y} = \mathbf{Xb} + \mathbf{e}\).

b. So, analogously to the preceding section, we will be interested in finding the vector of coefficients, \(\mathbf{b}\), on \(\mathbf{X}\) that brings us closest to \(\mathbf{y}\).

(1) (equivalently, we want to find the \(\mathbf{b}\) that minimizes \(\mathbf{e}'\mathbf{e}\), but that’s for later...)

(2) (...,or to find the \(\mathbf{b}\) that maximizes the likelihood of \((\mathbf{y},\mathbf{X})\) under normality assumptions).

(3) This, again analogously, means that we seek the vector \(\mathbf{b}\) such that \(\mathbf{Xb} \perp \mathbf{e}\) (orthogonal).

(4) Or, equivalently, we want to find \(\mathbf{b}\) such that \((\mathbf{Xb})'\mathbf{e} = 0\).

c. Rewrite using our previous rules (transposition & distributive) of matrix multiplication:

(1) Note that \(\mathbf{e} = \mathbf{y} - \mathbf{Xb}\), so

(2) \((\mathbf{Xb})'\mathbf{e} = \mathbf{b}'\mathbf{X}'\mathbf{e} = \mathbf{b}'\mathbf{X}'(\mathbf{y}-\mathbf{Xb}) = \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{Xb} = \mathbf{b}'(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Xb}) = 0\).

(3) Which implies we want to solve: \(\mathbf{X}'\mathbf{Xb} - \mathbf{X}'\mathbf{y} = 0\) [we don’t want the \(\mathbf{b} = 0\) “solution” because it’s trivial (vector of all 0’s is just the origin, we want a \(\mathbf{b}\) set of coords for \(\mathbf{b}\)].

(4) Which, rearranging, implies we want to solve: \(\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y}\)

(5) Two notes:
(a) If we could divide by \( X'X \), we could just divide both sides by it and be done as before: \( X'y/X'X \). But, this time, \( X'X \) is \( k \times k \), so we don’t know how to divide by that.

(b) Note the solution for \( b \) will involve only things we know: \( X, y \). That is, in regression terms, we can say: the estimate of the coefficient vector is “a function of the data.”

d. Before going further, note the problem’s dimensions: \( y=Xb+e \) is \( (n \times 1)=(n \times k)(k \times 1)+(n \times 1) \).

1. So, we have \( n \) observations on 1 (dependent) variable arranged in the \( (n \times 1) \) vector \( y \).

2. \( n \) observations on \( k \) (independent) variables arranged in the \( (n \times k) \) matrix \( X \).

3. since there are \( k \) (independent) variables in \( X \), we will have \( k \) coefficients thereupon and they will be arranged in the \((k \times 1)\) vector \( b \).

4. & there are \( n \) differences between \( Xb \) & \( y \) (residuals) arranged in the \((n \times 1)\) vector \( e \).

5. The dimensions of \( Xb \) (fitted values) are \((n \times k)(k \times 1)=(n \times 1)\).

6. So, the equation is correctly arranged to be conformable for addition & multiplication.

E. **Inverse Matrices**: the matrix analog to division

1. In scalar algebra, we know that \( x^{-1}x = 1 \), so thinking back to our solution in the bivariate analog to this problem we had \( x'bx = x'y \) before we divided both sides by \( x'x \) to get \( b = x'y/x'x \).

   a. We could equivalently have written \((x'x)^{-1}x'xb=(x'x)^{-1}x'y \) if we wanted, which would have reduced to \( 1^{-1}b = (x'x)^{-1}x'y \) or \( b = (x'x)^{-1}x'y \).

   b. So, rather than trying to divide by \( X'X \), we could try to find a matrix analog to the inverse of \( X'X \). In fact, we should because matrix division doesn’t exist while matrix inversion does.

2. **Inverse matrix**: So, in the general case, we want to define, for any matrix \( A \), another matrix \( B \) such that \( BA = (the\ matrix\ analog\ to\ 1) = I \).

   a. If \( B \) exists (it won’t always exist), we shall define it as \( B = A^{-1} \) (read “\( A \) inverse”)

   b. So, \( A^{-1} \) is that matrix such that \( AA^{-1} = I \)

   c. Two incidental points

      (1) pre-multiplying by \( A \) & post-multiplying by \( A^{-1} \), we find out \( AA^{-1} \) is idempotent & \( = I \):

         a. \((AA^{-1}) \) is idempotent:

            i) \( A(A^{-1}A^{-1}) = AIA^{-1} \)

            ii) \((AA^{-1})(AA^{-1}) = (AA^{-1}) \) (by associative & because \( I \) times anything is that)

         b. \((AA^{-1}) = I \)

            i) pre-multiplying by \((AA^{-1})^{-1} \Rightarrow (AA^{-1})^{-1} (AA^{-1})(AA^{-1}) = (AA^{-1})^{-1} (AA^{-1}) \)

            ii) \( = I (AA^{-1}) = I \Rightarrow (AA^{-1}) = I \)

   (2) Since \( A^{-1}A = I = AA^{-1} \), we know that \( A^{-1} \) can be multiplied on either side of \( A \).

      a. Thinking of the dim’s now, this means that both \((n \times k)(m \times p) \) and \((m \times p)(n \times k) \) exist.

      b. The first implies that \( k = m \), the second that \( p = n \).

      c. Either way, result is \( I \), which is square, so \( k = m \) and \( p = m \) as well.

      d. Putting it all together: \( n = k = m = p \), i.e. \( A \) and \( A^{-1} \) must be square & have same dim’s
d. Starting with the 2x2 case, then, the inverse, B, must satisfy:

\[ BA = I, \]

so:

\[
\begin{bmatrix}
  b_{11} & b_{12} \\
  b_{21} & b_{22}
\end{bmatrix}
\begin{bmatrix}
  a_{11} & a_{12} \\
  a_{21} & a_{22}
\end{bmatrix}
= \begin{bmatrix}
  1 & 0 \\
  0 & 1
\end{bmatrix}
\]

(2) which written long-hand as a set of equations is:

\[
\begin{align*}
    a_{11}b_{11} + a_{12}b_{21} &= 1 \\
    a_{11}b_{12} + a_{12}b_{22} &= 0 \\
    a_{21}b_{11} + a_{22}b_{21} &= 0 \\
    a_{21}b_{12} + a_{22}b_{22} &= 1
\end{align*}
\]

(3) which can be solved (by a lot of tedious substitution, or Kramer's rule) to:

\[
\begin{bmatrix}
  b_{11} & b_{12} \\
  b_{21} & b_{22}
\end{bmatrix}
= \frac{1}{a_{11}a_{22} - a_{12}a_{21}}
\begin{bmatrix}
  a_{22} & -a_{12} \\
  -a_{21} & a_{11}
\end{bmatrix}
\]

(4) which is \( B = [C_{ij}] / |A| \) where \( c_{ij} \) is the \( j^{th} \) cofactor of \( A \) (note the reversal of indices)

e. **Cofactor:** the cofactor of \( a_{ij} \), or equivalently, the \( j^{th} \) cofactor of \( A \) is calculated so:

(1) delete row \( i \) column \( j \) of \( A \)

(2) Then \((-1)^{i+j}\) times the determinant of the remaining matrix is the cofactor

(3) 2x2 easy, 3x3 bearable, 4x4 almost torture, & computer's nightmare (not yours) after that

3. To sum: \( A^{-1} = [C_{ij}] / |A| \)

a. **So, if the determinant is zero, i.e.** if there is linear dependence in the columns, the inverse is **undefined** implying that we can't solve our original problem

b. A matrix whose inverse does not exist is said to be **singular**; if it does exist, it's **nonsingular**

c. Some properties:

(1) \( |A^{-1}| = |A|^{-1} \) (i.e., \( 1/|A| \))

(2) \( (A^{-1})^{-1} = A \) (practically by definition: what matrix times \( A^{-1} \) gives \( I \) \( \begin{bmatrix} 
    1 & 0 \\
    0 & 1
\end{bmatrix} \) ? \( A \))

(3) \( (A^{-1})' = (A')^{-1} \)

(4) If \( A \) is symmetric, \( A^{-1} \) is symmetric (follows from 3; can you prove it given 3?)

(5) if both inverses exist, \( (AB)^{-1} = B^{-1}A^{-1} \). Note the order reversal and that this is extendable beyond two by applying associative property of matrix multiplication: \( (ABC)^{-1} = C^{-1}B^{-1}A^{-1} \)

F. **Back to the ALL-IMPORTANT SOLUTION:** So, if \( (X'X)^{-1} \) exists, we can get the solution by premultiplying both sides by it: \( (X'X)^{-1}X'Xb = (X'X)^{-1}X'y \rightleftharpoons \hat{b} = (X'X)^{-1}X'y \)

(Don't look now, but we just gave the general, matrix algebra formula, for ordinary-least-squares regression coefficient estimates! Know this equation; live it, learn it, love it. It is your mantra now...
I. Finishing up Matrix Addition:

- **Conformability of \( A \& B \) for Addition \( \iff \) \( \dim A = \dim B \)**

  - A pair of exceptional cases:
    - **Scalars**: a scalar can be added to any matrix
      \[ a + A = [a_{ij} + b] \quad \text{e.g.,} \quad 2 + A = [a_{i1} + 2 \quad a_{i2} + 2] \]
    - **Vectors**: 2 vectors must have same dims, but usually we allow:
      \[ \overline{a} + A = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad a = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \]
      \[ \overline{a} + A = \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} \quad a + A = \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix} \]

II. Starting back up from Matrix Multiplication:

A) Inner products: \( \overline{a} \cdot \overline{b} = \sum a_i b_i \) \quad e.g., \[ \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 9 \\ 5 \\ 6 \end{bmatrix} = 41 + 10 + 18 \]
\( \Rightarrow \overline{a} \cdot \overline{a} = \sum a_i^2 \) \quad e.g., \[ \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = 3^2 + 4 \quad b \cdot c = \sum b_i c_i \]

B) Matrix:
\[ C = AB \iff C_{ik} = \sum_j A_{ij} B_{jk} \]  
\[ \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 14 + 15 \\ 44 + 30 \\ 34 + 18 \\ 20 + 18 \end{bmatrix} = \begin{bmatrix} 29 \\ 74 \end{bmatrix} \]

Conformability: algebras of rows & cols:
\[ \frac{A}{n \times k} \cdot \frac{B}{k \times m} \Rightarrow \frac{C}{n \times m} \]  
recipe:
- 1) write dims  
- 2) draw box for resulting matrix  
- 3) fill in blanks

Assoc. \( (AB)C = A(BC) \checkmark \) (examples in handout; do some more if need practice)

Distrb. \( (A + B)C = AC + BC \checkmark \) \( \text{Note: the order matters be commutative} \)

Transposition \( (AB)' = B' A' \)

Idempotent: \( A \) is idempotent \( \iff AA = A \) \( \text{if symmetric, then } A A = A \text{ to}^2 \)

III. Writing sets of **Linear Forms**: (draw in vector form)
\[ \overline{O} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} \]  
\[ \overline{e} = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \]  
\[ \overline{y} = \overline{x} + \overline{e} \]  
\( \text{sometimes} \)
III. Geometry:
Handwritten notes are through solving for $A^{-1}$; so, returning to $(X'X)b = X'$, premultiplied: $(X'X)(X'Z)b = (X'X)\hat{y}$

\[ Q = (X'X) \quad \text{the products & cross-products of the indep. vars.} \]
\[ A = Q^{-1}X' \quad \text{the matrix that provides the coefficient vector} \hat{b}, \text{for} \]
\[ (\Rightarrow Ay = \hat{b}) \quad \text{fitting any} \ y \ \text{it premultiplies using} \ X \]
\[ N = XA \quad \text{the matrix that produces the fitted-value vector}, \ y'(using \hat{b}) \]
\[ (\Rightarrow Ny = XAy = X\hat{b} = \hat{y}) \]
\[ M = (I-N) \quad \text{the matrix that produces the fitted-residual vector}, \ e'(using} \ y' \]

Selected Properties:

1. $AX = I$: suppose you are using $X$ to predict $X$ (as LHS says), what coefficients would you want to, mult. $X$ by to get $X$? One of course, or some matrix $X$, of course.

2. $NN = N$: now you're fitting a thing made up of all of $X$'s to predict $N$ (made all of $X$'s). How well do you suppose you'd fit it? Perfectly of course, so $NN = N$

3. $NX = X$: same story, even more directly (fitted $X$ with itself, perfectly)

\[ (Q')' = Q^{-1} \quad (Q\text{'s symmetric}) \]

\[ MM = M \quad \text{& } NN = N \quad (\text{idempotent}) \]

\[ \text{for intuition ask after PSI} \]

\[ \text{what residuals do you suppose you'll get if you try to use the same X's to predict residuals already obtained using X to predict X?} \]
\[ \text{Well, you already used up everything X has to tell you in the first. } e = y - X\hat{b}, \text{shouldn't be any more info in e, so } MMy \text{ had better be zero. But } e \]
\[ \text{by assoc. property: } MMy = (MM)y \]
\[ \text{by above discussion: } e = (MM)y \]
\[ \text{definition of } M \quad e = My \]
\[ \text{comparing last two lines, } M = MM \]

IV. Partitions: Do look at Greene 2.6 (or relevant section in 3rd ed.), you should be perfectly capable of following it now.

Some Further Notes:

→ Main Point is Notational:

\[ A = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 4 & 7 \\ 2 & 6 \end{bmatrix} \quad \text{we can write } A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \]

→ For Block Diag: $| A | = \prod_{i=1}^{N} | A_{ii} |$

→ Multiplication also as before, but note: all submatrices $A_{ij}$ must be conformable with $B_{jk}$ (relevant) pairs (i.e.)

→ Addition works exactly as before

→ Kronecker Products useful especially in pooled Time-Series-Cross-Section for example in Systems of Equations
VI. Characteristic Equations, Roots, & Vectors

* Problem could be solved like so: \[ A \vec{c} = \lambda \vec{c} \]
  \[ \Rightarrow A \vec{c} - \lambda \vec{c} = 0 \Rightarrow [A - \lambda I] \vec{c} = 0 \]
  \[ \Rightarrow \text{so, we pick a particular solution, one where } \vec{c} \approx 1 \text{ (this is called renormalizing, to get } \vec{c} \approx 1) \]

* \( \vec{c} \) is not zero, so the solution must have:
  \[ |A - \lambda I| = 0 \]
  e.g.,
  \[ \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} - \begin{vmatrix} \lambda & 0 \\ 0 & \lambda \end{vmatrix} = 0 \]
  \[ a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} \]

* This "polynomial" is called the characteristic equation of \( A \).

* The set of \( \lambda \) that solve it are called the characteristic roots.

\[ \text{e.g.: } A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} \]
\[ A \vec{c} = \lambda \vec{c} \Rightarrow [A - \lambda I] \vec{c} = 0 \]
\[ |A - \lambda I| = 0 \Rightarrow \begin{vmatrix} 1 - \lambda & -1 \\ 0 & 2 - \lambda \end{vmatrix} = 0 \]
\[ (1 - \lambda)(2 - \lambda) = 0 \Rightarrow \lambda = 1 \text{ or } \lambda = 2 \]

* Putting lambda back in to orig. equations, we get:

\[ \begin{bmatrix} 1 - \lambda & -1 \\ 0 & 2 - \lambda \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = 0 \]
\[ \Rightarrow (1 - \lambda)C_1 + C_2 = 0 \]
\[ \Rightarrow (2 - \lambda)C_2 = 0 \]

* For \( \lambda = 1 \), the solution for \( C_2 \) is:
\[ (2 - \lambda)C_2 = 0 \]
\[ C_2 = 0 \]

* As advertised, \( \vec{c} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \) is not unique, any \( \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \) will do.

* Let's normalize so \( \vec{c} \approx 1 \)
\[ \Rightarrow C_1^2 + C_2^2 = 1 \]
\[ \Rightarrow C_1 = C_2 \]

* The normalized solution is \( \vec{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \), called a characteristic vector

* For \( \lambda = 2 \)
\[ C_1 = C_2 \]
\[ C_2 = \text{ anything } \]

* Again, any \( C_1 = C_2 \) works.
Continuing the example: \( C_1 = C_2, C_3 = \) anything for \( j=2 \)

- normalize: \( C_i^2 + C_j^2 = 1 \), substitute in \( C_i = C_2 \)
  \[ 2C_i^2 = 1 \quad \Rightarrow \quad C_i = \frac{\sqrt{2}}{2} \quad \text{or} \quad \frac{\sqrt{2}}{2} \quad \text{or} \quad \frac{2}{\sqrt{2}} \quad \text{or} \quad \ldots \]

- so the (normalized) characteristic vector associated with the characteristic root \( \lambda = \lambda_i \) is \( C = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \)

Some Results & Properties: (Try to prove them to yourself)

1. A \((k \times k)\) matrix \( A \) has, in general, \( k \) characteristic roots, (some of which may be identical to others) (that is, this \( k \) char. eq. is a \( n \)th order polynomial)

2. A \((k \times k)\) symmetric matrix \( A \) has \( k \) unique characteristic vectors (once normalized); the \( k \) characteristic roots need not be unique.

3. The characteristic vectors of a symmetric matrix are orthogonal.

\( \text{Aside: } (X'X) \text{ is symmetric, } \therefore \text{ its char. vectors are orthogonal} \)

We can (if it's frequently useful to) gather up the char. vectors of \( A \), i.e., the \( C_i \), and the char. roots that go with them, \( \lambda_i \), including repeats. We write:

\[ \Lambda = \begin{bmatrix} \lambda_1 & \lambda_2 & 0 \\ 0 & \lambda_k \end{bmatrix} \quad \text{(n.b., } \Lambda \text{ is capital } \Lambda) \]

\( d \quad C = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ C_1 & C_2 & \cdots & C_k \end{bmatrix} \)

More Properties: (Prove them to yourself)

- The full set of characteristic eqns is now \( AC = C\Lambda \)
- \( C'C = I \)
- \( C'C = C'C = C'C = C'C = I \)

All of this is useful because ... ?

- Makes finding rank of \( A \) easy:
  \[ \begin{align*}
  &\text{facts: } C'AC \text{ diagonalizes } A: C'AC = C'\Lambda = \Lambda C = \Lambda \\
  &\text{rank} (BAC) = \text{rank} (BA) = \text{rank} (A)
  \end{align*} \]

\( \forall \ B \text{ & } C \text{ non-singular (full rank)} \)

so \( \text{rank} (A) = \text{rank} (C'A') = \text{rank} (A') \)

\( \Delta \text{ is diagonal, so rank is just # of non-zero elements on its diagonal.} \)

\( \text{fact } \text{rank} (A) = \text{rank} (A'A), \text{ so this can be considerably easier than it first looks} \)

A measure of how close to colinear the columns of \( A \) are can be derived from the characteristic roots of \( A \)

(more on next page)
C & \Lambda^T can be used to find B such that BB^T or B^2 = A

B can be thought of as the "square-root matrix" of A.

More generally, for any "positive definite" A, \( A = C \Lambda C^T \).

Recall that charact. vectors are orthogonal if that in some sense orthogonal vectors "best" or "most stable" span a vector space.

It's no surprise then, that Spectral Decomposition, A = C \Lambda C^T, and much of this talk is closely related / very useful to factor analysis ... a process concerned, among other things in pulling out orthogonal vectors (variables) which summarize a (usually larger) set of correlated vectors (variables).

These concepts & the useful matrices C & \Lambda are also used in the "Cholesky Decomposition" or "Ch. Factorization" of matrices.

- Basically, you can take any symmetric, "positive definite" matrix and write it as a product of a lower-triangular & an upper-triangular matrix.

Useful in Vector-Auto-Regression (VAR)

→ A is ...

- **Positive Definite** \( \iff x'Ax > 0 \quad \forall x \neq 0 \)

- **Negative Definite** \( \iff x'Ax < 0 \quad \forall x \neq 0 \)

- **Positive (Negative) Semi-Definite** \( \iff \text{replace } > \text{ with } \geq \text{ or } \leq 0 \)

- **Positive Semi-Definite** \( \equiv \text{Non-Negative Definite} \)

- **Negative Semi-Definite** \( \equiv \text{Non-Positive Definite} \)

- **If it's none of these, then it's indefinite.**

  who cares? • among other things, if a positive definite, \( |A| \neq 0 \) so \( A^{-1} \) exists

  • A # of other useful properties which may become important because a lot of regression quantities are of the form: \( x'Ax \)
Say we want to solve the system

\[
\begin{align*}
3x + 2y + z - 2w &= 4 \\
2x - y + 2z - 5w &= 15 \\
4x + 2y - 5w &= 1 \\
3x - 2z - 4w &= 1.
\end{align*}
\]

The associated matrix is

\[
\begin{pmatrix}
3 & 2 & 1 & -2 \\
2 & -1 & 2 & -5 \\
4 & 2 & 0 & -1 \\
3 & 0 & -2 & -4
\end{pmatrix}
\]

whose determinant is \( \Delta = -65 \). Since the determinant is non-zero, we can use Cramer's rule. To obtain the value of the \( k \)-th variable, we replace the \( k \)-th column of the matrix above by the column vector

\[
\begin{pmatrix}
4 \\
15 \\
1 \\
1
\end{pmatrix}
\]

the determinant of the obtained matrix is divided by \( \Delta \) and the resulting value is the wanted solution.

So

\[
x = \frac{\Delta_1}{\Delta} = \frac{\begin{vmatrix}
4 & 2 & 1 & -2 \\
4 & 2 & 0 & -1 \\
15 & -1 & 2 & -5 \\
1 & 0 & -2 & -4
\end{vmatrix}}{-65} = \frac{-65}{-65} = 1
\]
\[ y = \frac{\Delta_2}{\Delta} = \frac{130}{-65} = 2 \]

\[ z = \frac{\Delta_3}{\Delta} = \frac{-195}{-65} = 3 \]

\[ w = \frac{\Delta_4}{\Delta} = \frac{65}{-65} = -1 \]

"example of Cramer's rule" is owned by drini.

(view preamble)

This object's parent.

Cross-references: column vector, column, Cramer's rule, determinant, matrix

This is version 3 of example of Cramer's rule, born on 2002-07-15, modified 2003-12-23. Object id is 3169, canonical name is ExampleOfCramersRule. Accessed 1489 times total.

Classification:
AMS MSC: 15A15 (Linear and multilinear algebra; matrix theory :: :: Determinants, permanents, other special matrix functions)

Pending Errata and Addenda
When we find simultaneous equations
\[
\begin{align*}
ax + by &= e \\
ax + dy &= f
\end{align*}
\]
we can find \(x\) and \(y\) by using the following determinants.
\[
\begin{array}{c|c}
\begin{bmatrix} e & b \\ f & d \\ a & b \\ c & d \end{bmatrix} & \begin{bmatrix} e & a \\ f & c \\ b & a \\ d & c \end{bmatrix}
\end{array}
\]
This determinant is Cramer's rule.

Alternate product \(\begin{pmatrix} a \\ c \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix}\) is the area of a parallelogram made by
\[
\begin{pmatrix} a \\ c \end{pmatrix} \text{ and } \begin{pmatrix} b \\ d \end{pmatrix}\] (This value of area has a sign.)
\[
\begin{pmatrix} a \\ c \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} = ad - bc\] \(\begin{pmatrix} a \\ c \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} = 0\).

Use alternate product to find the following simultaneous equation.
\[
\begin{align*}
ax + by &= e \\
ax + dy &= f
\end{align*}
\]
\[
\begin{align*}
x \begin{pmatrix} a \\ c \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix} &= \begin{pmatrix} e \\ f \end{pmatrix} \\
\Rightarrow x \begin{pmatrix} a \\ c \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} &= \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} \\
\Rightarrow x \begin{pmatrix} a \\ c \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} &= \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} \\
\Rightarrow x \begin{pmatrix} a \\ c \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} &= \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} \\
\Rightarrow x \begin{pmatrix} a \\ c \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} &= \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} \\
\Rightarrow x \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix} \\
\Rightarrow x \begin{pmatrix} a \\ c \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix}
\]
\[
\begin{align*}
\therefore x &= \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} b \\ d \end{pmatrix} \\
y &= \begin{pmatrix} e \\ f \end{pmatrix} \times \begin{pmatrix} a \\ c \end{pmatrix}
\end{align*}
\]
These are expressed by determinant as follow.
\[
\begin{array}{c|c}
\begin{bmatrix} e & b \\ f & d \\ a & b \\ c & d \end{bmatrix} & \begin{bmatrix} e & a \\ f & c \\ b & a \\ d & c \end{bmatrix}
\end{array}
\]
I. Where we were last:

\[ e = X - X\hat{b} \]

A. Regression as Projection Problem

1. Scalar/Vector

\[ x^T e = 0 \]

\[ b x^T e = 0 \]

\[ b^T (y - x\hat{b}) = 0 \]

\[ b^T y = b^T x\hat{b} = 0 \]

\[ x^T y = x^T x\hat{b} = 0 \]

\[ x^T b = x^T y \]

Notice: Suppose \( x^* = x - \bar{x} \),

\[ y^* = y - \bar{y} \],

then:

\[ x^T x^* = \sum_{i=0}^{n} (x_i - \bar{x})^2 = n \sigma^2 \]

\[ x^T y^* = \sum_{i=0}^{n} (x_i - \bar{x})(y_i - \bar{y}) \]

\[ b^* = \frac{x^T y^*}{x^T x^*} = \frac{\sum_{i=0}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n} (x_i - \bar{x})^2} \]

\[ b = \frac{\text{cov}(x,y)}{\text{var}(x)} \]

2. Vector/Matrix:

\[ y = X\hat{b} + \epsilon \]

\[ X\hat{b} \perp \epsilon \Rightarrow (X\hat{b})^T \epsilon = 0 \]

\[ (X\hat{b})^T (y - X\hat{b}) = 0 \]

\[ b^T X^T y - b^T X^T X\hat{b} = 0 \]

\[ \frac{b^T X^T y}{(1 \times b)} = \frac{b^T (X^T X)b}{(1 \times 1)} \]

\[ \Rightarrow b^T (X^T y - X^T X\hat{b}) = 0 \]

\[ \Rightarrow \text{one solution is } b = 0 \text{, but this is not optimal; one we want is} \]

\[ \hat{x}^T \hat{X}^T \hat{b} = \hat{x}^T y \]

\[ \Rightarrow \hat{b} = (X^T X)^{-1} X^T y \]
B. Further Topics:

1. Work with Determinants & Inverses -- "what's in them, intuitively?"

2. Trace: sum of diagonal elements of a matrix:

\[
\text{Trace}(e_i e_i'') = \sum_{i=1}^{n} e_i^2
\]

\[\rightarrow \text{some properties of use}\]

3. Regression Matrices of Importance:

\[
Q = (X'X)^{-1} \quad (V \text{-cov of } X \text{ mat})
\]

\[
A = Q^{-1} X' \quad (\text{coefficient maker})
\]

\[
N = XA \quad (\text{predicted-value maker})
\]

\[
M = (I - N) \quad (\text{residual maker})
\]

* Bunch of cool properties that you'll want to become intuitive to you. E.g.:  

\[
NX = X \quad MX = 0 \quad AX = I
\]

4. Partitions: often convenient to partition matrices like, e.g.:  

\[
A = \begin{bmatrix}
1 & 2 \\
3 & 4 \\
1 & 2 \\
3 & 4
\end{bmatrix}
\]

\[
\begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}
\]

→ Then: Add (Sub) & Mult as before (but relevant pairs must be conformable)

→ Block-Diagonal: \( A_{ij} = 0 \quad \forall \ i \neq j \)

→ For block diag, \( |A| = \prod_{i=1}^{n} |A_{ii}| \)

(very convenient)

5. Kronecker Product (\( \otimes \)): \( A \otimes B \) matrix \( A \) each element of \( B \)

6. Characteristic Equations, Vectors, Roots -- that they exist, that we use them in some configurations
Calculus (And Multivariable Calculus): The Bare Basics

- (scalar) function: $y$ is a (scalar) function of $x \iff$ for every $x$ there is a (single) $y$ that corresponds to it, (if more than one $y$ we can say $y$ is a "correspondence" of $x$

notation: $y = f(x)$, $y = g(x)$, $y = p(x)$

this means $y$ (maybe) a function of other variables too; left open what else though we might say in this case: $y$ is a function of $x$ & some other stuff

- terminology: $y$ is a function of $x$" or "$y$ equals $f$ of $x$" being a technical term for "$y$ equals $f$ of $x$" whatever the case may be, have no idea whether right now.

- if we don't want to specify everything for whatever reason, or just as shorthand once $y$ has already been specified

- examples: 1 perhaps simplest: $y = x$

   where is a set of

   $\{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\}$

   - if $x = 0$ (is not in the set)

   - if $x = 2$ (is in the set)

   - if $x = 7$ (is in the set)

   - if $x = 8$ (is in the set)

   - if $x = 9$ (is in the set)

   - if $x = 10$ (is in the set)

   - if $x = 11$ (is in the set)

   - if $x = 12$ (is in the set)

   - if $x = 13$ (is in the set)

   - if $x = 14$ (is in the set)

   - if $x = 15$ (is in the set)

- usually $y = 3x$ for example or $y = x$ is considered "$y$ is not a function of $x$"

   but note: $y = 3x + 2$ is same, so sometimes we torture the language a bit & say "$y$ is a zero or full function of $x$"

   this, sure, but a useful trivial function sometimes

- 4 $y = e^{-x^2}$

- 5 $y = 3x^2 + 4x + 2x - 17$

- Notice: just 44 $y$ is a function of $x$; doesn't mean it's not also a function of something else:

   - thus in example 6 we could (or even should) say $y$ is a function of $x$ & $z$

   - Notice: $y$ is a function of $x$ says nothing about which sort of function — can be as simple or as complicated as we like or need.

   - In example 6 we need to expand the definition a bit: $y$ is a function of $x$ & $z$ $\iff \forall x, z \exists$ a single $y$; we can do so ad nauseam. So for another example, we could have:

- 6 $y = f(x)$ where $x$ is a vector

   - this still a scalar function (by most definitions); here is left-hand side that defines the type of function

   - spelling out $f(x)$ here, $y$ could be, e.g.:

- 6 $y = 3x$

- 6 $y = 3x$

- going back to example 4, we could draw this all graphically:

   - now we might be interested in how $y$ changes as $x$ changes:

   $\Delta y \approx \frac{y - y_i}{x - x_i}$

   where $y_i$ is $y$ when $x = x_i$ & so on

   - this, as we all recognize, is just the same old "slope" of the line given by $y = x$. In this case, the slope is 1 at any point on the line, but what about:

- 7 $y = \sqrt{x}$

   - or 8 $y = ln(x)$ ("natural log of $x"")

   - These too can be graphed as 7 is in the figure,

   - but: 8 $\Delta y/\Delta x$ depends on where you start from

   - 9 $\Delta y/\Delta x$ depends on how big a $\Delta$ you use
**Derivatives:**

- Measure: "slopes"
- Measure how much \( y \) changes when \( x \) moves (an arbitrarily small bit)
- **can be thought of, then, as rate of change of \( y \) (with respect to \( x \))**
- Most basic: tells us in what relationship \( y \) moves with \( x \)

**Notation:**

- **given \( y = f(x) \)**
  - \( \frac{dy}{dx} \) or \( \frac{df}{dx} \) or \( f'(x) \) or \( \frac{dy}{dx} \)
- **given \( y = g(x, z, \ldots) \)**
  - \( \frac{\partial y}{\partial x} \) or \( \frac{\partial g}{\partial x} \) or \( g_x \)

All of above notations work, as does \( g_x \) or \( \frac{\partial y}{\partial x} \), which are like the \( f_x \) above, but now we just subscript the position (in this case, \( x \) of the variable the derivative is taken "with respect to"

*(n.b., derivatives are also called differentials if there are total or partial derivatives, don't worry too much about that for now)*

**Further Note:** Derivatives can only be taken at smooth points in the function.

That is, at \( C \)

\[
\lim_{x \to C} \frac{dy}{dx} \quad \text{OK} \quad \text{but at } \sqrt{3} \quad \text{is not OK.}
\]

*Turns out these points are generally true; only case they're not is basically when \( y = f(x) \) is a straight line, i.e., \( y \) is a linear function of \( x \)*

But this is a drag, b/c the slope tells us how \( y \) "responds" to \( x \), that is, how much and where \( y \) moves when \( x \) moves. We'd like some standard to answer that question & the slope is our ideal candidate. Returning to point (1) & (2), it turns out that (1) is unavoidable--nor should we avoid it; how much \( y \) moves does depend on where we start from. Point (3) is just as much a fact of \( y = f(x) \), but it's also of feature of how we asked the question. Suppose we rephrased it a little more precisely:

exactly at point \( C \), I want to know the slope of \( y \); i.e. I want to know the marginal (i.e. teeny-weeny) slope of \( y \) at \( C \).

Now, we could draw a tangent to \( y = f(x) \) at point \( C \):

That'll answer our question: slope at \( C \) is slope of tangent at \( C \)

But it's kinda hard to draw tangents, so time consuming to draw it all out, is there no other choice?

Thought experiment: start from \( C \) & go some \( \Delta x \) right & see how much \( \Delta y \) you get--for a small \( \Delta x \), this'll be a pretty close approximation to the tangent, won't it?

How about we just keep taking smaller \( \Delta x \)'s? In the limit, we'd be right on the tangent! But, unfortunately in limit \( \Delta x = 0 \), \( \frac{\Delta y}{\Delta x} \) wouldn't be defined. Hold on, though: the slope measured as \( \Delta x \) got smaller was getting closer to slope of tangent just before \( \Delta x \) became undefined. So, let's define:

\[
\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \quad \text{"limit of } \frac{\Delta y}{\Delta x} \text{ as } \Delta x \text{ goes to zero"}
\]

as what \( \frac{dy}{dx} \) is approaching as \( dx \) approaches but almost reached zero...

* This is (in the limit) the slope we want!
* It is also called the derivative of \( y \) with respect to \( x \)
### Table A1. Some Useful Differentiation Rules

<table>
<thead>
<tr>
<th>Expression</th>
<th>$\frac{\partial y}{\partial x}$</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = c$</td>
<td>$\frac{\partial c}{\partial x} = 0$</td>
<td>The derivative of a constant is zero.</td>
<td>$\frac{\partial 1}{\partial x} = 0$</td>
</tr>
<tr>
<td>$y = cz$</td>
<td>$\frac{\partial (cz)}{\partial x} = 0$</td>
<td>The derivative of a term that does not depend on $x$ is zero.</td>
<td>$\frac{\partial (3x)}{\partial x} = 0$</td>
</tr>
<tr>
<td>$y = cx$</td>
<td>$\frac{\partial (cx)}{\partial x} = c$</td>
<td>The derivative of a term involving a linear coefficient and $x$ is that coefficient.</td>
<td>$\frac{\partial (3x)}{\partial x} = 3$</td>
</tr>
<tr>
<td>$y = cx^a$</td>
<td>$\frac{\partial (cx^a)}{\partial x} = acx^{a-1}$</td>
<td>The derivative of a term involving a linear coefficient and $x$ raised to the $a$th power is the product of $a$, $c$, and $x$ raised to the $(a-1)$ power.</td>
<td>$\frac{\partial (3x^3)}{\partial x} = 15x^2$</td>
</tr>
<tr>
<td>$y = cxz$</td>
<td>$\frac{\partial (cxz)}{\partial x} = cz$</td>
<td>The derivative of a term involving a linear coefficient, $x$, and another variable $z$ is the product of the coefficient and the variable (we can treat the other variable as a constant with respect to $x$ here).</td>
<td>$\frac{\partial (3xz)}{\partial x} = 3z$</td>
</tr>
<tr>
<td>$y = cxz w$</td>
<td>$\frac{\partial (cxz w)}{\partial x} = cz w$</td>
<td>The result extends to higher-order interactions, where again variables that are not a function of the variable with respect to which one is differentiating are fixed.</td>
<td>$\frac{\partial (3xz w)}{\partial x} = 3zw$</td>
</tr>
<tr>
<td>$y = \ln(x)$</td>
<td>$\frac{\partial (\ln(x))}{\partial x} = \frac{1}{x}$</td>
<td>The derivative of a logged variable is the inverse of that variable.</td>
<td>$\frac{\partial (3\ln(x))}{\partial x} = \frac{3}{x}$</td>
</tr>
<tr>
<td>$y = e^x$</td>
<td>$\frac{\partial (e^x)}{\partial x} = e^x$</td>
<td>The derivative of base $e$ raised to a variable is base $e$ raised to that variable.</td>
<td>$\frac{\partial (3e^x)}{\partial x} = 3e^x$</td>
</tr>
<tr>
<td>$y = b_0 + b_1 x + b_2 z + b_3 x z$</td>
<td>$\frac{\partial b_0}{\partial x} + \frac{\partial (b_1 x)}{\partial x}$ $+ \frac{\partial (b_2 z)}{\partial x} + \frac{\partial (b_3 x z)}{\partial x}$ $= b_1 + b_3 z$</td>
<td>The derivative of some linear additive function equals the sum of the derivative of each of the terms.</td>
<td>$\frac{\partial (1 + 2x + 3z + 4xz)}{\partial x}$ $= 2 + 4z$</td>
</tr>
<tr>
<td>$y = f(x) \times g(x)$</td>
<td>$\frac{\partial (f(x) \times g(x))}{\partial x}$</td>
<td>The derivative of the product of two functions equals the sum of derivative of the first function, multiplied by the undifferentiated second function; plus the derivative of the second function, multiplied by the undifferentiated first function.</td>
<td>$\frac{\partial ((2x + 5) \times (3\ln(x)))}{\partial x}$ $= \frac{\partial ((2x + 5))}{\partial x} (3\ln(x))$ $+ \frac{\partial (3\ln(x))}{\partial x} (2x + 5)$ $= 2(3\ln(x)) + (3/x)(2x + 5)$</td>
</tr>
</tbody>
</table>
\[ y = f(g(x)) \]

\[ \left( \frac{df}{dg} \right) \times \left( \frac{dg}{dx} \right) \]

This is the chain rule for nested functions.

\[ \frac{\partial F(x)}{\partial x} = f(x) \]

The derivative of any cumulative probability function is the corresponding probability density function.

\[ \frac{\partial \left( (2(3\ln x) + 5) \right)}{\partial x} \]

\[ = \frac{\partial \left( 2(g) + 5 \right)}{\partial g} \times \frac{\partial g}{\partial x} \]

\[ = 2 \times \frac{3}{x} = \frac{6}{x} \]

\[ \frac{\partial \Phi(x)}{\partial x} = \phi(x) \]
Some Useful Derivatives:

- Derivative of a constant is zero \( \frac{d}{dx}c = 0 \) 
  e.g., \( \frac{d}{dx} 27 = 0 \)
- (liner)
  - Coefficient rule: \( \frac{d}{dx} (ax) = c \) 
  e.g., \( \frac{d}{dx} 3x = 3 \)
  - e.g., \( \frac{d}{dx} \frac{d}{dx} = x \)
- Power rule: \( \frac{d}{dx} (x^n) = n \cdot x^{n-1} \) 
  e.g., \( \frac{d}{dx} (x^3) = 12x^2 \)
- (power)
  - Exponential rule: \( \frac{d}{dx} e^{(x)} = e^{(x)} \cdot \frac{d}{dx} \) 
  e.g., \( \frac{d}{dx} e^{3x} = 3e^{3x} \)
- (exponential)
  - Log rule: \( \frac{d}{dx} \ln(x) = \frac{1}{x} \cdot \frac{d}{dx} \) 
  e.g., \( \frac{d}{dx} \ln(x) = \frac{1}{x} \cdot \frac{d}{dx} \)
- (log)
  - Addition rule: \( \frac{d}{dx} (f(x) + g(x)) = \frac{d}{dx} f(x) + \frac{d}{dx} g(x) \) 
  e.g., \( \frac{d}{dx} (f(x) + g(x)) = 14x + 3 \)
- (addition)
  - Multiplication rule: \( \frac{d}{dx} (f(x)g(x)) = \frac{d}{dx} f(x) + f(x) \) 
  e.g., \( \frac{d}{dx} (f(x)g(x)) = 14x \)
- (multiplication)
  - Chain Rule: \( \frac{d}{dx} f(g(x)) = \frac{d}{dx} f(g(x)) \cdot \frac{d}{dx} g(x) \) 
  e.g., \( \frac{d}{dx} (e^{x}) = e^{x} \cdot \frac{d}{dx} \) 

(a.k.a., There's a division rule, but multiplication rule + power rule is some things. In fact, many of these could be used; alternatively in various settings)

Some Things about \( \ln \) & \( e \) you should recall:

- \( \ln [x \cdot y \cdot z] = \ln x + \ln y + \ln z \) 
  \( \Rightarrow \ln \left( \prod_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} \ln x_i \) 

- \( \ln x^y = y \ln x \) 
  (follows from 0 actually)

- \( \ln \frac{1}{x} = \ln y - \ln x \) 
  (follows from 0 and 0 using \( y^{-1} \) instead of \( x \))

- \( e^{x+y} = e^{x} \cdot e^{y} \) 
  (e just a number, this just the way powers work)

- \( e^{y-x} = e^{x-y} \) 
  (follows from 0)

- \( \prod_{i=1}^{n} e^{x_i} = e^{x_1} \cdot e^{x_2} \cdot \ldots \cdot e^{x_n} \) 

- \( \ln e^{x} = x \) 
  (e = b cancel each other, this just the way powers & logs work)

- \( \ln 1 = 0 \) 
  & \( e^{0} = 1 \) 
  \( e^{\infty} = \infty \) 
  \( e^{-\infty} = 0 \) 

- These help solve integrals and whatnot.

- There are some more (I think rarely used in our line) derivative rules (mostly surrounding sine, cosine, tangent, secant & all their inverses). You may wish to know that exponential rule is special case of \( \frac{d}{dx} a^{x} = a^{x} \ln a \) 

- Combining the simple rules using (confusingly) the addition, multiplication, & chain rules, you should be able to differentiate anything (scalar)
Integration: suppose for some reason you wanted to find the area under a curve. Geometrically, this is easy if the curve is not too curvy:

\[ \text{Area} = b \cdot h + \frac{1}{2} ab \]

but what if \( y(x) \) could do this by approximation; just use a large number of narrow rectangles to approximate the irregular shape of the curve. In fact, this just what most computer programs do.

- This could be tedious if one summer some guy in England (Newton) & some guy in Germany (or was it Austria? I dunno. Anyway: Leibniz) were getting annoyed with all this in the context of some physics problem or another. They each knew a lot about derivatives or intuitively saw the slope of the tangent to this wacky set of curves \( y(x) \) at some point \( x_0 \). So it occurred to them: if I want a long infinitesimal step by infinitesimal step, colored in a line only infinitesimally wide, I then used the slope (derivative) to tell me where \( y(x) \) went next, repositioned at another width/distance, and so on, eventually I would have filled the whole thing, in the limit as the slices got smaller, perfectly!

So that's what they did (sort of)

Actually, they both noticed that if you could find some function \( g(x) \) such that \( \frac{dy}{dx} = g(x) \), then you could take a shortcut. Same higher dimensional thinking led them both to the conclusion that if you could find such a \( g(x) \), then the difference of \( g(x) \) at \( x \) & \( g(x) \) at zero gave you the whole area in one step (whoa!). I still don't really see the intuition, but it's true!

So, Area Under a Curve is the Antiderivative, evaluated at low point less some constant, evaluated at the low point. First, some notation/terminology.

- This is a Definite Integral:
  \[ A = \int_{a}^{b} f(x) \, dx \]
  area under curve
  lower bound
  upper bound
  curve being traced
  the dimension we're moving along as we're tracing
  continuous
  discrete \( \sum \)

- These are called definite integrals. We the \( a \) \& \( b \), we wish to span have been specified. Could be left off:
  \[ \int f(x) \, dx \]
  is an indefinite integral; it basically says: set the problem up so that I can solve it as soon as I decide what \( a \) \& \( b \) I care about.

So, anyway, the whole trick, as we have seen, is to find \( g(x) \) such that it's derivative will be the function we are trying to trace. In general this is still the dark given the derivative rules I gave you as a (guttering) candle.
But there are some easy ones: let's go back to the flat line of example 1.

The line \( y = f(x) = 3 \) gives us many possible functions. To find the area under that "curve," we have

\[
A = \int_0^5 f(x) \, dx = \int_0^5 3 \, dx
\]

For simplicity, let's try to find a function whose integral gives us this area. The simplest choice, if we think of a table of integrals, is \( f(x) = 3x \).

So now we evaluated \( 3x \) at \( x = 5 \) and \( x = 0 \). The difference, we write

\[
\left. g(x) \right|_0^5 \text{ or } \left[ g(x) \right]_0^5 = g(5) - g(0)
\]

which is confusing with matrices lying around in our minds, so use the other.

\( g(x) \) in this example is \( 3x \), so \( 3 \cdot 5 - 3 \cdot 0 = 15 \).

**Some Useful Integral Rules**

(anti-derivative)

As before, \( a \) a & \( n \) a constants

- \( x \) a variable

- \( f(x) \) a function

\( g(x) \) a function

Notice:

- Most of these are just the derivative rule read backwards.
- Just as in that case, easily obtaining the area under a variety of curves.

\[ \begin{align*}
\int a \, dx & = ax + C & \text{antiderivative of a constant is that constant times } x \\
\int af(x) \, dx & = a \int f(x) \, dx & \text{we can pull constants out of} \\
\int [f(x) + g(x)] \, dx & = \int f(x) \, dx + \int g(x) \, dx & \text{integral of sum is sum of} \\
\int x^a \, dx & = \frac{1}{a+1} x^{a+1} & (a \neq -1) \\
\int x^2 \, dx & = \frac{1}{3} x^3 \\
\int e^x \, dx & = e^x \\
\int a^x \, dx & = \frac{a^x}{\ln a} \\
\int b^x \, dx & = \frac{b^x}{\ln b} \\
\int \ln x \, dx & = x \ln x - x
\end{align*} \]

Examples:

\[ \begin{align*}
a) \int \frac{1}{x} \, dx & = \ln x \bigg|_1^7 = \ln 7 - \ln 1 = \ln 7 \\\n\int e^{-x} \, dx & = -e^{-x} \bigg|_0^\infty = -e^{-\infty} + e^0 = 0 + 1 = 1 \\
c) \int (x^3 + 2x^4 + 3x) \, dx & = \frac{1}{4} x^4 + \frac{2}{5} x^5 + \frac{3}{2} x^2 + C \\
& \left( \text{could do this step too if you like} \right)
\end{align*} \]
Gradients

Not hard to think of vector functions now that we have scalar functions down:

\[ y = f(x) \]

In fact, we saw already that we can write the whole set of equations:

\[ y = c_0 x^T + c_0 \]

\[ y = x^T + c_0 \]

*This was a vector-valued vector function.*

I already gave an example of a scalar-valued vector-function:

\[ y = x \]

Also, we already talked about case where.

*Just change the names of the variables in \( f() \) (incidentally, those are called the arguments of the function):*

\[ y = f(x_1, x_2, \ldots, x_n) \]

that we are back in vector (matrix) land

\[ y = f(x) \]

So if we can write \( y = f(x) \) it will probably happen that we want to know:

\[ \frac{\partial y}{\partial x} \quad \text{some time (i.e., the "slope", but n-dimensional, now, called the "gradient")}
\]

\[ \text{written } \nabla y = \nabla f(x) \quad \text{written } \nabla y = \nabla f(x) \]

This is the gradient vector. It's the n-dimensional slope of has its elements the partial derivatives of \( y \) w/r to the element numbers \( x_i: (\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \ldots, \frac{\partial y}{\partial x_n}) \).

Cross-Derivatives, Second- & Higher-Derivatives:

Also, if derivatives are just rates of change, we can think of rates of change of rates of change (second derivatives) of rates at which rates of change themselves change (higher derivatives) as we change another variable (cross-derivatives).

First Derivative:

\[ \frac{\partial y}{\partial x} \]

Second Derivative:

\[ \frac{\partial^2 y}{\partial x^2} \]

Cross-Derivative:

\[ \frac{\partial^2 y}{\partial x \partial z} \]

Usual written:

\[ \frac{\partial^2 y}{\partial x^2} \]

(Congratulate yourself if you can see why second derivative is just a special case of a cross-derivative. If you can't, pause & figure it out)

Hessians: the matrix of cross-derivatives

The Hessian of \( y = f(x) \) is given by

\[ \mathbf{H} = \begin{bmatrix}
\frac{\partial^2 y}{\partial x^2} & \frac{\partial^2 y}{\partial x \partial z} \\
\frac{\partial^2 y}{\partial x \partial z} & \frac{\partial^2 y}{\partial z^2}
\end{bmatrix} \]

Convince yourself \( \mathbf{H} \) is square & symmetric.
Matrix Derivative Rules:

1. \( \nabla = \frac{\partial^2}{\partial x \partial y} \) (just a definition)

2. \( \frac{\partial (A X)}{\partial X} = A \) (not \( A' \))

3. \( \frac{\partial A X}{\partial X} = A' \) (not \( A \))

4. if \( A \) symmetric,
   \[ \frac{\partial (X' A X)}{\partial X} = 2 A X \]

5. \( \frac{\partial (X' A X)}{\partial A} = X X' \)

Optimization:

One of the neatest things you can do with calculus is solve problems like maximize the volume/surface-area ratio of a rectangular box, or like minimize the sum of squared residuals.

Find the maximum of: \( f(x) \) here:

\[ f(x) = 1 + e^{-x} \]

maximun of \( f(x) \) occurs at slope = \( 0 \)

- Could be only "local max"
- Global max:
- Could also be a local or global min.

\( f(x) \) can be checked:

\( \frac{d^2 f(x)}{dx^2} \) at maxing its positive
\( \frac{d^2 f(x)}{dx^2} \) at minning its negative

\( (x, y) \) solution only one way
(e.g. \( y = x^2 \)), then globals
The Regression Examples:

0 scalar algebra

\[ \begin{align*}
\operatorname{Min} & \sum_{b} \varepsilon_i^2 \quad \text{where} \quad \varepsilon_i = y_i - x_i b \\
\operatorname{Min} & \sum \left( y_i - x_i b \right)^2 = \sum \left( y_i^2 - 2x_i y_i b + x_i^2 b^2 \right)
\end{align*} \]

First-Order Condition:

\[ \sum \left( \frac{\partial}{\partial b} y_i^2 - \frac{\partial}{\partial b} 2x_i y_i b + \frac{\partial}{\partial b} x_i^2 b^2 \right) = \phi \]

\[ \begin{align*}
\sum \left( -2x_i y_i + 2x_i^2 b \right) &= \phi \\
\sum x_i y_i &= \sum x_i^2 b \\
b^* &= \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i y_i \varepsilon_i}{\sum x_i^2} \\
\text{(by there it is again)}
\end{align*} \]

2) Matrix Algebra, Multivar Calc.

\[ \operatorname{Min} \sum_{b} \varepsilon_i \varepsilon_i \quad \text{where} \quad \varepsilon_i = y_i - X_i b \]

\[ \begin{align*}
\operatorname{Min} \left( y - X b \right)^T \left( y - X b \right)
\end{align*} \]

\[ \begin{align*}
\left( y - X b \right)^T = -b^T X^T y + \sum_{i} \varepsilon_i \left( X_i b \right) + b^T X^T X b
\end{align*} \]

\[ \begin{align*}
\sum_{i} \varepsilon_i \left( X_i b \right) &= \sum_{i} \varepsilon_i \left( X_i b \right) \\
- X^T y - X^T y + 2X^T b &= \phi \\
X^T b &= X^T y \\
b^* &= \left( X^T X \right)^{-1} X^T y
\end{align*} \]

Important Further Topics:

1. Constrained Optimization (Kuhn-Tucker, etc.)
2. Taylor Series Approximation
3. Differential Equations (Differential Equations)
4. Dynamic Optimization (Maximizing Integrals, Functions)
Constrained Optimization by Lagrangian Method

- \( \forall f(x) \text{ and } g_i(x), \ i = 1, \ldots, m \text{ differentiable} \)
  
  that map \( \mathbb{R}^n \rightarrow \mathbb{R}^m \) to solve:

  \[
  \min f(x) \quad \text{s.t. } g_i(x) = 0, \ i = 1, \ldots, m
  \]

- (prior theorem), an extremum implies

  \[
  \nabla f = \sum_{i=1}^{m} \lambda_i \nabla g_i,
  \]

  but this \( \Rightarrow \)

- \( \min f(x) \text{ s.t. } g_i(x) = 0 \equiv \)

  \[
  \min f(x) - \sum_{i=1}^{m} \lambda_i g_i(x)
  \]

  (without restriction)
TAYLOR SERIES (EXPANSIONS, APPROXIMATIONS, POLYNOMIALS, LINEARIZATIONS...)

\( \forall f(x) = \text{function defined on any open interval containing } x_0, \text{ differentiable at } x_0 \text{ at any order, } \)

\[ f(x) = T(x_0) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k \]

\[ = \frac{f(x_0)}{0!} (x-x_0)^0 + \frac{f'(x_0)}{1!} (x-x_0)^1 + \frac{f''(x_0)}{2!} (x-x_0)^2 + \ldots \]

\[ = f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2} f''(x_0)(x-x_0)^2 + \ldots \]

- \( N^{th} \) order(degree) Taylor Series expansion (approx)

\[ T_n(x_0) = \sum_{k=0}^{n-1} \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k \]

Remainder: \( R_n(x_0) = f(x_0) - T_n(x_0) \)

\[ \lim_{n \to \infty} R_n(x_0) = 0 \]

(Wow!)
Where we were last -- calculus wrap-up

Optimization:

\[ f(x) \]

- Extrema at \( x_0, x_1, x_2 \)
- Maxima at \( x_0, x_2 \)
- Global Minimum at \( x_3 \)

\[
\begin{align*}
\text{Maxima} @ & (x_0, x_2) \\
\text{Global Minimum} @ & (x_3) \\
\end{align*}
\]

\[ f'(x) = 0 \]

- Notice slope at extrema = 0

\[
\Rightarrow \text{First Order Condition (F.O.C.) of max/min:} \quad \frac{df(x)}{dx} = 0 \]

- But could be max or min

\[
\Rightarrow \text{Second Order Condition:} \\
\begin{align*}
\text{Max:} & \quad \frac{d^2f(x)}{dx^2} < 0 \\
\text{Min:} & \quad \frac{d^2f(x)}{dx^2} > 0 \\
\end{align*}
\]

- But still could be local or global extremum

\[ \Rightarrow \text{"Globality" hard to prove, but certain conditions on } f(x) \text{ will ensure just one max or just one min.} \]

Regression as a "Least Squared-Errors" Problem:

0 Scalar Algebra & Calc:

a) Bivariate regression: \( \sum_{i=1}^{n} E_i^2 \) where \( e_i = y_i - x_i b \)

\[
\begin{align*}
\text{Min } & \text{SSE} = \sum_{i=1}^{n} (y_i - x_i b)^2 \quad \Rightarrow \frac{\partial \text{SSE}}{\partial b} = 0 \\
\end{align*}
\]

\[
\frac{\partial \text{SSE}}{\partial b} = \sum_{i=1}^{n} 2(y_i - x_i b)(-x_i) = 0
\]

\[
\text{Check 2nd order cond.} \quad \Rightarrow \quad -2 \sum_{i=1}^{n} (y_i x_i - x_i b) = 0
\]

\[
\Rightarrow \quad b = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}
\]
b) Scalar Algebra & Calculus: Multivariate Regression

\[
\text{Min } \text{SSE}_b = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - x_{i1}b_1 - x_{i2}b_2 - \ldots - x_{ik}b_k)^2
\]

\[
\Rightarrow \frac{\partial \text{SSE}_b}{\partial b_k} = 0 \quad \forall k \in \{1, \ldots, K\}
\]

\[
\begin{align*}
0 &= \sum_{i=1}^n (y_i - x_{i1}b_1 - \ldots - x_{ik}b_k)(x_{ki}) = 0 \\
0 &= \sum_{i=1}^n (y_i - x_{i1}b_1 - \ldots - x_{ik}b_k)(x_{2i}) = 0 \\
0 &= \sum_{i=1}^n (y_i - x_{i1}b_1 - \ldots - x_{ik}b_k)(x_{ki}) = 0
\end{align*}
\]

... so, just plug and chug (plug & chug & plug & chug...)

OR:

2 Matrix Algebra & Multivariable Calculus

\[
\text{Min } \text{SSE}_b = \sum_{i=1}^n e_i^2 = (y_i - x_i b)^T (y_i - x_i b)
\]

\[
\Rightarrow \nabla_b (y - x b)^T (y - x b) = 0
\]

\[
\frac{\partial \text{SSE}_b}{\partial b_k} = \nabla_b (y - x b)^T (y - x b) = 0
\]

\[
\begin{align*}
\nabla_b (y - x b)^T (y - x b) &= 0 \\
\nabla_b (y' - b'X') (y - x b) &= 0 \\
\nabla_b (y' - \frac{1}{2} b'X'X b - y'X b + \frac{1}{2} b'X'X b) &= 0
\end{align*}
\]

\[
\Rightarrow \quad -X'X + 2X'X b = 0
\]

\[
\Rightarrow \quad X'X b = X'X
\]

\[
\Rightarrow \quad b^* = (X'X)^{-1} X'X
\]
Some other, final Math Notes in passing

1) Constrained Optimization by "Lagrangian Method"

\[ \min_{x} f(x) \quad \text{s.t.} \quad g_i(x) = 0 \]

\[ = \min_{x, \lambda} \mathcal{L}(x, \lambda) = f(x) - \sum \lambda_i g_i(x) \] \quad \text("Lagrangian multiplier")

\[ \Rightarrow \nabla_{x} \mathcal{L} = \nabla_{x} f(x) - \sum \lambda_i \nabla_{x} g_i(x) = 0 \]

\[ \nabla_{x} \mathcal{L} = g_i(x) = 0 \quad \text{This part always just gives you the constraints back} \]

In this part, two kinds of solutions may exist:

(a) Solutions with $\lambda \neq 0$ imply that the constraints "bind":

\[ (\min \mathcal{L}(x) \neq \min f(x) \quad \text{s.t.} \quad g_i(x) = 0) \quad \text("Interior solutions") \]

(b) Solutions with $\lambda = 0$ imply the constraints do not bind, "are not binding" ("corner solutions") so

\[ \min (k) = \min (k) \quad \text{s.t.} \quad g_i(k) = 0 \]

→ One strategy of statistical testing rests on comparing

\[ \frac{\partial f(x)}{\partial x} \quad \text{at a max/min constrained by imposition of the null hypothesis to} \quad \frac{\partial f(x)}{\partial x} \quad \text{at the unconstrained max/min. If non-binding, ...} \]

→ This usually achieved by clever auxiliary

(regression in which coefficients $\lambda_i$ are the $L$-multipliers

(so can test if zero(s)).
2. Taylor Series (Linear) Approximations

For any \( f(x) \) (continuous and continuously differentiable), the \( n \)-th derivative of \( f \) at some point \( X_0 \) is:

\[
    f(x + \Delta x) = f(x) + \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + \frac{1}{2!} f^{(2)}(x_0)(x - x_0)^2 + \ldots
\]

Point is:
- Can approximate by dropping higher orders if keeping \( \Delta x \) small.
- Could start with \( f(b) \) at a nonlinear point.
- Could approximate by

\[
    f(b) = f(a) + f'(a)(b-a)
\]

This is linear:

\[
    \Rightarrow V(f(b)) \approx V[f'(a)(b-a)]
\]

3. One more thing, while we're here: Infinite Sums Useful Too. Let \( p \) be

\[
    \frac{\Gamma(p)}{\Gamma(p+1)} = \frac{1}{p} \quad \forall 0 < p < 1
\]
(Univariate) Probability & Distribution Theory

- We can write the outcome of every aspect of the physical or social world as a product of some "random experiment."

- "Random" here means that prior to the outcome, there is some degree (from zero to infinite) of uncertainty about what the outcome will be.

  - Now, zero & infinite uncertainty are both, in some sense, a strain on the meaning of "uncertainty". We are usually uninterested in either case: zero uncertainty means there is nothing to talk about, & it just is. Infinite uncertainty is even a little hard to grasp (probably not empirically relevant in the vast majority of cases) -- we usually have some idea of range of possible outcomes if nothing else.

- % of the most texts exclude & define uncertainty from the definition of "random." This is fine, so long as you remember that to say something happens w/ probability 0 or 1, (which is zero uncertainty) is actually possible.

- Say let \( A \) represent the outcome of some random experiment as in: \( A \) is some "event," a possible outcome of the experiment.

  - Then let \( S \) be the set of all possible outcomes, called the Sample Space.

  - And let \( \Pr(A) \) be some function of \( A \) defined over the Sample Space.

\[ \Rightarrow \text{Axiomatic Definition of Probability Function (PF):} \]

\[ \text{Set Theory:} \begin{align*}
\text{1. } & \emptyset \leq \Pr(A) \leq 1 \quad \forall A \in S \\
\text{2. } & \Pr(S) = 1 \\
\text{3. } & \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \quad \forall (A,B) \in S
\end{align*} \]

Notes:

- On \( \text{1.} \) Notice that "nothing happens" is always an element of \( \emptyset \subseteq \emptyset \) (null set) always a subset of every other set. Including any sample space.

- On \( \text{3.} \) if \( A \neq B \) disjoint, that is \( A \cap B \) have "nothing in common" or "no overlap".

Now suppose for every possible outcome (i.e., \( \forall A \in S \)) we assign a numeric value, e.g., on a roll of a die, there are six possible faces that might come up, we could assign to each face (1,2,3,4,5,6) the numeric value that face.

Random Variables: are mathematically defined thusly \( X = \{ 1 \text{ if coin heads up, 0 if coin tails up} \} \)
If, like all of our examples so far, the values $X$ may take are finite or countably infinite, the $X$ is said to be a discrete random variable.

We define probability functions over random variables axiomaticall:

\[
f(X) = \text{Pr}(X = x)
\]

Axioms:

1. $\forall x \leq \sum \Pr(X = x_i) = 1$ or $\forall x \leq f(x) \leq 1$
2. $\sum_{x \in \Omega} \Pr(X = x_i) = 1$ or $\sum_{x \in \Omega} f(x) = 1$

Notice by these axioms that random variables are only defined over disjoint events. That is, for any outcome only one value is assigned to $X$ and for any value assigned to $X$ only one outcome corresponds. (There is a "one-to-one" correspondence, or, identically, a function.)

What if the set of outcomes is neither finite nor countably infinite? (An example might be the weight (exactly) of a randomly chosen person.) The set of outcomes is infinitely divisible (the only option left), we can still assign a value to all possible outcomes (as in our weight example), but there is no way we can sum the probabilities of such a random variable (since we can't count the possible outcomes).

$\Rightarrow$ resort to integration (calculus) if the axiomatic def.

becomes:

1. $\Pr(a \leq X \leq b) = \int_a^b f(x) \, dx \geq 0$
2. $\Pr(X = a) = \int_a^a f(x) \, dx = 0$ (by definition of integrals)
3. $\int_{-\infty}^{\infty} f(x) \, dx = 1$

Notes:

a) perfectly analogous to discrete case, except that probability at any specific point in the sample space is zero.

b) Thus, $f(x)$ is not a probability function (PF) but a probability density function (PDF)

c) Since $Pr(X = a) = 0$, $Pr(a \leq X \leq b) = Pr(a < X \leq b) = Pr(a \leq X < b) = Pr(a < X < b)$

d) Recall that integration is "scooping" up the area under a curve. So, to find the probability of some range of possible outcomes, say from $a$ to $b$, we "scoop" the area under the PDF, $f(x)$, from $a$ to $b$, i.e. we integrate from $a$ to $b$.

![Probability Function](image1.png)

![Probability Density Function](image2.png)
Some Common Distributions (p64f's):

Discrete:

1. Bernoulli (Binary Experiment) e.g. coin toss:
   distributions have pfs or pdfs
   these have arguments. They're called parameter(s) of the distribution. They're what you need to know about the specific example.

   Bernoulli: \( f(x) = \begin{cases} \pi x(1-\pi)^{1-x} & \text{for } x = 0, 1 \\ \emptyset & \text{for all else} \end{cases} \)

   Bernoulli's parameter \( \pi \in [0,1] \)

First Principles of Bernoulli:

- only two possible outcomes
- outcomes mutually exclusive (disjoint)
- but you can always redefine outcomes so that this is true

2. Discrete Uniform (equiprobable) Distribution e.g. coin toss:

   \( f(x) = \begin{cases} \frac{1}{N} & \text{for } x = 1, 2, \ldots, N \\ \emptyset & \text{for all else} \end{cases} \)

   parameter: \( N \)

First Principles:

- \( N \) disjoint events
- all equally likely

3. Binomial Distribution:

   \( f(x) = \begin{cases} \frac{N!}{x!(N-x)!} \pi^x (1-\pi)^{N-x} & \text{for } x = 0, 1, 2, \ldots, N \\ \emptyset & \text{for all else} \end{cases} \)

   parameters: \( N, \pi \)

First Principles:

- \( N \) Bernoulli trials
- Each trial has constant, same probability, \( \pi \)
- All trials are independent (which \( \Rightarrow \) trials are indp. but we haven't defined that yet)

Continuous:

1. Continuous Uniform Distribution \([a, b] \)

   \( f(x) = \begin{cases} \frac{1}{b-a} & \text{for all } a < x < b \\ \emptyset & \text{else} \end{cases} \)

   parameters: \( a, b \) -- the range of possible values

   First Principles:

   - fixed range of possible values (bounded)
   - nothing "more likely" than anything else in this range

2. Normal (Gaussian) Function:

   \( f(x) = \left( \frac{1}{2\pi \sigma^2} \right)^{-\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)

   parameters: \( \mu, \sigma^2 \) -- mean, variance

   "bell-shape" symmetric, "thin-tailed" empirical regularity
Cumulative Distribution Functions: \[ F(x) = \sum_{x \leq x} f(x) = \text{Pr}(X \leq x) \quad (\text{discrete}) \]
\[ \Rightarrow f(x_i) = F(x_i) - F(x_{i-1}) \]

Continuous:
\[ F(x) = \text{Pr}(X \leq x) = \int_{-\infty}^{x} f(t) \, dt \]
\[ \Rightarrow \frac{dF(x)}{dx} = f(x) \]

Further Implications:
1. \( 0 \leq F(x) \leq 1 \) (by def. of prob. density function)
2. If \( x > y \), \( F(x) \geq F(y) \) (by some def. \( x \) beyond \( y \), \( \text{incl.} \) \( y \), \( \text{Pr}(X \geq y) \))
3. \( F(\infty) = 1 \) (should be obvious)
4. \( F(-\infty) = 0 \) (this too)
5. \( \text{Pr}(a \leq X \leq b) = F(b) - F(a) \)

Measures of "central tendency": (trying to answer "where's this pdf or pdf center?", "what's the middle or center?"")

Mean of \( X \) = \( \mu_x = E(X) = \sum_{i} x_i \cdot \text{Pr}(X = x_i) = \sum_{i} x_i \cdot f(x_i) \quad \text{discrete} \)
\[ = \int_{-\infty}^{\infty} x \cdot f(x) \, dx \quad \text{continuous} \]

E.g. \( E(X) \) where \( X = \# \) dots on face of rolled die
\[ E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{21}{6} = 3\frac{3}{6} = 3.5 \]

E.g. \( E(X) \) where \( X \) continuous uniform on \( 0-1 \):
\[ E(X) = \int_{0}^{1} x \cdot \frac{1}{b-a} \, dx = \left[ \frac{1}{b-a} \cdot \frac{1}{2} x^2 \right]_{0}^{1} = \frac{1}{2} - 0 = \frac{1}{2} \]

(Actually cont. uniform is always \( \frac{1}{b-a} \))

Also median: "median of \( X \)" is \( x \) such that \( \text{Pr}(X \leq x) \geq \frac{1}{2} \) \( \Rightarrow \) it's point where \( \frac{1}{2} \) of pdf or pdf on either side

N.B. I say "mean of \( X \)" not "average of \( X \)". In loose terms, they are equivalent, but, more strictly, the mean is the true expectation of \( X \) in the pdf, the average is an estimate of the mean in some sample from that population. (More on this next week)
mode: value of $x$ where $f(x)$ is maximum; 
median: point with highest-probability "range" around it for discrete highest-probability outcome; uniform point with highest-probability "range" around it for a tricky def of range
median rarely used, mode almost never, so worry about mean.

Expectation of Functions of $X$:
1. $E(g(X)) = \sum x g(x) \cdot f(x)$
2. $E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) \, dx$

Proof:
$E(a + bX) = \int_{-\infty}^{\infty} (a + bx) f(x) \, dx$
$= \int_{-\infty}^{\infty} a f(x) \, dx + \int_{-\infty}^{\infty} b f(x) \, dx \cdot x$
$= a \int_{-\infty}^{\infty} f(x) \, dx + b \int_{-\infty}^{\infty} f(x) \, dx \cdot x$
$= a + bE(X)$

Hugely Important Special Case: of functions of $X$:

\[ \text{VARIANCE: } \text{Var}(X) = E[(X - \mu)^2] = \begin{cases} \sum (x - \mu)^2 \cdot P(x) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx & \text{continuous} \end{cases} \]

E.g. variance of die roll: (recall $E(X) = 3.5$)
\[ \frac{1}{6} (1-3.5)^2 + \frac{1}{6} (2-3.5)^2 + \frac{1}{6} (3-3.5)^2 + \frac{1}{6} (4-3.5)^2 + \frac{1}{6} (5-3.5)^2 + \frac{1}{6} (6-3.5)^2 \]
\[ = 2.916 \]

Variance is the most common measure of the "spread" of a distribution
Properties of Variances: \[ V(X) = \mathbb{E}(X - \mu)^2 = \mathbb{E}(X^2 - 2X\mu + \mu^2) = \mathbb{E}(X^2) - 2\mu \mathbb{E}(X) + \mu^2 \]

Thus:
1. \[ V(X) = \mathbb{E}(X^2) - \mu^2 \]
2. \[ \text{Var}(a + bX) = b^2 \text{Var}(X) \]
3. \[ \text{Var}(g(x)) = \int_a^b (g(x) - \mathbb{E}(g(x)))^2 dx \]

Mean-Squared-Error (MSE): MSE of \( X \) around some point \( c \)

\[ \text{MSE} = \mathbb{E}[(X - c)^2] \]

First, obviously, MSE of \( X \) about its mean \( \mu \) is \( 0 \).

More generally:

\[ \text{MSE}_{c \text{ about } c} = \mathbb{E}[(X - c)^2] = \sigma_X^2 + (c - \mu)^2 \]

- More on this next week.

Just some FYI definitions:

- **Skew**: Measure the symmetry of a distribution

\[ \text{skew} = \mathbb{E}[(X - \mu)^3] \]

- **Kurtosis**: Measures "how fat the tails are" (note: this is not necessarily a simple relationship to variance).

\[ \text{Kurtosis} = \mathbb{E}[(X - \mu)^4] \]

Moments: Notice that mean = \( \mathbb{E}(X) \), var = \( \mathbb{E}[(X - \mu)^2] \), skew = \( \mathbb{E}[(X - \mu)^3] \), etc.

Each of these is called a *moment* of the distribution. Mean is 1st-moment, Kurtosis is 4th-moment.

Moments are related to each other by derivatives: Anti-derivatives of the pdf (which is usually, but not necessarily, a simple relationship to variance.

As I said, just FYI: you may have heard of the "method of moments." This involves, basically, estimating relationships by variables only by estimating the moments of their joint distribution (to be defined) rather than spelling out the actual joint distribution.

Method of Moments (MOM) is new stuff (e.g., sci. stats). I don't think we know a lot of its properties well yet.
I. Wrapping up univariate distributions (for us)

A. Prob \( (X \geq a) \equiv \begin{cases} \sum_{X \geq a} p_{X_i}(X_i) & \text{discrete} \\ \int_{a}^{\infty} f(x) \, dx & \text{continuous} \end{cases} \)

- \( p_{X_i}(X_i) \), etc.
- \( \Pr(a \leq X \leq b) \), etc.

Examples: \( \Pr(X \geq 5) \) where \( X \) outcome of die roll = \( p_X(3) + p_X(4) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2} \)

\( X \sim N(0,1) \):

\[
\Pr(X > 2) = \int_{-\infty}^{2} f_X(x) \, dx + \int_{2}^{\infty} f_X(x) \, dx
\]

Std Norm cdf.

\( \text{Std Norm Symmetric, so} \)

\[
= 2 \int_{0}^{2} f_X(x) \, dx
\]

B. Properties of Expectations & Variances

1. \( \text{Expect: } E(X) \equiv \sum_{X} X_i \cdot p_{X_i}(X_i) \equiv \sum_{X} X_i \cdot f(X_i) \)

\[
\Rightarrow E(a + bX) = a + bE(X)
\]

\( E(\alpha) = \alpha \) \( \quad \text{(a)} \)

\( E(bX) = bE(X) \) \( \quad \text{(b)} \)

Law of Iterated Expectations: \( E(E(X)) = E(X) \)

Intuition: "you expect to expect" what you expect

\( E(X) \) will generally be some constant, e.g., \( E(X) = \mu \), \( E(x^2) = \sigma^2 \)

2. Var: \( \text{Var}(X) \equiv E[(X - \mu)^2] = E(x^2) - [E(x)]^2 \) \( \quad \text{(a)} \)

\( E(a + bX) = b^2 \text{Var}(X) \)

\( \Rightarrow E(x^2) = [E(x)]^2 + \text{Var}(x) \)

\( = \mu^2 + \sigma^2 \)

(2) \text{Var}(a) = 0 \quad \text{for any constant } a

(2) \text{Var}(bX) = b^2 \text{Var}(X) \quad \text{for any constant } b \text{ and random var. } X

\( \Theta \text{ Var}(g(X)) \) where \( g \) any old function

\[
\approx \frac{d^2}{dx^2} g'(\mu)^2 \text{Var}(X)
\]

evaluated at \( \mu \)

C. Mean-Squared-Error of \( \hat{\theta} \) around \( \theta \): usually we will be interested in MSE of some "estimator" around the "parameter"

\[
E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2
\]

\( = \sigma^2 + \text{bias}^2 \) where \( \theta = \text{true value of parameter} \)

\( \text{MSE supposes to estimate} \)

\( \theta \)

\( e.g., \hat{X} \text{ around } \theta \)
D. $\chi^2$, $t$, and $F$ Distributions:

1. Chi-Squared Distribution
   - if $Z \sim N(0,1)$, then, letting $X = Z^2$, $X \sim \chi^2$
     ("$X$ is distributed Chi-Squared with 1 degree of freedom")
   - $E(X)$ for any Chi-Squared r.v. $X$ is its degrees of freedom
   
   Aside: for any $X \sim N(\mu, \sigma^2)$, $Z = \frac{(X-\mu)}{\sigma} \sim N(0,1)$
   - if $Z_1, Z_2, \ldots, Z_n$ each dist.
     independently $N(0,1)$, then $\left(\sum_{i=1}^n Z_i^2\right) \sim \chi^2_n$
   - if $X_1, X_2, \ldots, X_n$ each dist.
     independently $\chi^2_k$, then $\left(\sum_{i=1}^n X_i^2\right) \sim \chi^2_n$.
   
   Example: if the $e_i$ from $Y_i = \beta_0 + \beta_1 x_i + e_i$ distributed independently" normal with
   
   $\left(\frac{\sum e_i^2}{\sigma^2}\right) \sim \chi^2_n$

2. F Distribution: if $X_1$ and $X_2$ are "independent" chi-squared r.v.s with $n_1$ and $n_2$ degrees of freedom respectively, then

   $F = \left(\frac{X_1/n_1}{X_2/n_2}\right) \sim F_{n_1, n_2}$ ("$F$ is distributed F with $n_1$, $n_2$ degrees of freedom")

   Example: if $e_i$ from $Y_i = \beta_0 + \beta_1 x_i + e_i$ "independent" Normal $N(0, \sigma^2)$ "independent"

   and $e_i^*$ from $Y_i^* = \beta_0 + \beta_1 x_i + e_i^*$ "independent" Normal $N(0, \sigma^2)$ denote

   Then $\left\{ \frac{\sum (e_i^* - \bar{e})^2}{\sum (e_i - \bar{e})^2} \right\}/n_2 \sim F_{n_1, n_2}$

3. t-Distribution (also called "student" $t$):

   - if $Z \sim N(0,1)$ and $X \sim \chi^2_n$ is "independent" of $Z$, then

   \[ t = \frac{Z}{\sqrt{\chi^2/n}} \sim \chi^2_n \] ("$t$ is distributed F with $n$ degree of freedom")

   Example: if $\hat{\beta}$ from $Y_i = \beta_0 + \beta_1 x_i + e_i$ is $N(0,1)$ and $e_i$ is "independent" of $\hat{\beta}$, also Normal $N(0, \sigma^2)$

   then $\sqrt{\frac{\hat{\beta}^2}{\hat{\sigma}^2}} \sim t_n$

   - it can be shown that $\left(\frac{\hat{\beta}^2}{\hat{\sigma}^2}\right) \sim F_{1, n}$ for any r.v. distributed $F_{n_1, n_2}$

E. Some limiting distributions (FYI):

1) $t_n$ becomes more like $N(0,1)$ as $n \to \infty$; beyond $n=100$, the difference will (almost) never matter practically; in fact, considerably sooner than that, say 50 or so, the difference is already small enough that it's not much worth sweating.

2) $X \sim \chi^2_n$ for $n > 30$ or so, is well approximated by noting that

   $Z = \sqrt{\frac{X}{\chi^2/n}}$ is approximately $N(0,1)$

3) $n_{\text{F}}, n_{\text{F}}$ becomes $\chi^2_n$ as $n \to \infty$; i.e. as $n$ becomes really large relative to $n$.

These can be useful in two ways: 0) they simplify things if tables must be used or

1) if you're looking at some $t$-statistic and you know $N$ is large, you can just think of $N(0,1)$, the critical values of which we might have better memorized.

However, any computer stats program has $X^2, t, F$ built in (as do the better spreadsheets), so really, you shouldn't ever need to approximate even a little. (Or if you do, that program "reasonably" approx. for you)
II. JOINT, BIVARIATE DISTRIBUTIONS:

A. Suppose you have two R.V.'s, $X_1$ & $X_2$ (e.g., $X_1 = \text{Gender}$ & $X_2 = \text{Height}$).

Defining Joint (Bivariate) Distribution:

$$\text{Prob}(a \leq x \leq b, c \leq x \leq d) = \begin{cases} \sum \sum f(x_1, x_2) & \text{discrete} \\ \iint_{[a,b] \times [c,d]} f(x_1, x_2) \ dx_1 \ dx_2 & \text{continuous} \end{cases}$$

Axiomatic Definition:

1. $f(x_1, x_2) \geq 0$
2. $\sum \sum f(x_1, x_2) = 1$

Cumulative Distribution Function (Joint C.D.F.):

$$F(x_1, x_2) = \text{Prob}(X_1 \leq x_1, X_2 \leq x_2) = \begin{cases} \sum \sum f(x_1, x_2) & \text{discrete} \\ \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(x_1, x_2) \ dx_2 \ dx_1 & \text{continuous} \end{cases}$$

Intuition: A Snow-Covered Driveway

Consider a driveway covered with snow. Dimension:

- think of the length of the driveway as the $X_1$ dimension (length dimension)
- the width of the driveway as the $X_2$ dimension (width dimension)

Then, the height of the snow at any point on the driveway reflects the probability density that any randomly chosen snowflake will have fallen at that $(X_1, X_2)$ coordinate. In general, the height could vary from spot to spot on course.

First of all, we're defining snowflakes that fall on the driveway as the only relevant ones, so by definition, the "probability density" of any point on the driveway is zero (no snow) or positive (some snow).

$$f(x_1, x_2) \geq 0$$

Since we've defined the driveway as the whole universe of relevant snow (seems that way modeling it), any relevant snowflake must have fallen on the driveway somewhere.

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) \ dx_2 \ dx_1 = 1$$

integrating over the snow-covered area (snow plow) over length & width of driveway.
So, if we took a shovel or a plow & plowed all the snow down to a line at the end or side of the driveway (that’s the first integral), then turned the plow & plowed all the snow on that line into a point in the corner (that’s the second integral after having done the first), we would have all the snow on that point. In probability, we call all that the probability, or 1.

\[
\int_{-\infty}^{\infty} f(x_1, x_2) \, dx_1 \, dx_2 = 1
\]

A discrete example:

<table>
<thead>
<tr>
<th>Population</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>18%</td>
<td>10%</td>
</tr>
<tr>
<td>sex</td>
<td>30%</td>
<td>42%</td>
</tr>
<tr>
<td>height</td>
<td>48%</td>
<td>52%</td>
</tr>
<tr>
<td>total</td>
<td>58%</td>
<td>62%</td>
</tr>
</tbody>
</table>

Marginal Distributions: (i.e., getting the univariate distributions back out of the bivariate distribution)

\[
f_y(x) = \int_{\mathbb{R}} f(x_1, x_2) \, dx_1
\]

\[
f(x) = \int_{\mathbb{R}} f(x_1, x_2) \, dx_2
\]

Back to the driveway analogy for intuition: in general, snow falls all over the driveway, \((x_1, x_2)\) coordinates all over. Suppose you wanted to know the probability (or proportionate amount) of snow falling at \(x_2\), not caring where it fell or relative to the \(x_1\) axis? Say, proportionate amount in line with some tree next to the driveway. Well, get out your snow plow and plow the whole driveway (parallel to the street) so that it keeps up in a line along the side of the driveway (the line is perpendicular to the line from the tree of interest across the driveway). Wherever the snow was generally higher across the driveway will now be the highest point along the line you just plowed, won’t it? (rhetorical)

That height of that line of snow now represents the marginal (i.e., univariate) distribution of \(x_1\) as it gives the “probability density” that \(x_1\) coordinate was that point what ever \(x_2\) was. (i.e., the “accumulated” height of snow along the line at \(x_1\)).

Example: in the sex & height table above, height = \{ over six foot \(p = 28\%\) \}

is the marginal distribution of height. It’s obtained by summing rows, i.e., summing over sexes,

Expectations From Joint Distributions:

\[
\mathbb{E}[g(x_i)] = \int \sum g(x_i) f_{X_i}(x_i) \, dx_i
\]
Expectations from Joint p.d.f.'s & p.f.'s: from the definition just given, we can show a whole bunch of other useful things.

1) \[ E[x] = \sum_{x} x \cdot f(x) \]
   \[ = \sum_{x_1} \sum_{x_2} x \cdot f(x, x_2) \]
   \[ = \sum_{x_1} \sum_{x_2} f(x, x_2) \]
   - from the definition of the marginal dist., \( f_i \)
   - \( x_i \) is not changed as we sum over \( x_2 \), i.e., relative to summing over \( x_2 \), it is a constant, thus we can move it in or out of summing over \( x_2 \) as will.

The Continuous Analogue is

\[ E(x) = \int_{-\infty}^{\infty} x \cdot f(x, x_2) \, dx \, dx_1 \]

(Notational note: I may from time to time write it this way if both (or all) integrals over some range. If \( I \) ever leave it off, or write \( \int_{-\infty}^{\infty} I \), it means \(-\infty \) to \( \infty \) (not indefinite integral) or over the range of \( x \).)

2) \[ \nu(x) = E[|x - \mu|^2] \]
   \[ = \sum_{x_1} (x - \mu)^2 \cdot f(x) \]
   \[ = \sum_{x_1} \sum_{x_2} (x - \mu)^2 \cdot f(x, x_2) \]
   \[ = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x, x_2) \, dx \, dx_1 \]

- logic for this exactly as above
- by analogy

Covariance & Correlation:

\[ \text{Covariance}(x, x_2) = \text{Cov}(x, x_2) = C(x, x_2) = E(x, x_2) - E(x) \cdot E(x_2) \]

\[ = E(x_1, x_2) - E(x_1) \cdot E(x_2) \]

\[ = E(x_1, x_2) - E(x_1) \cdot \mu_2 + \mu_1 \mu_2 \]

\[ = E(x_1, x_2) - \mu_1 \mu_2 + \mu_1 \mu_2 \]

\[ = E(x_1, x_2) - \mu_1 \mu_2 \]

\[ \text{Correlation}(x, x_2) = \text{Corr}(x, x_2) = \frac{\text{Covariance}(x, x_2)}{\sigma_1 \sigma_2} \]

(Where \( \sigma_1 \) is std. dev. \( x \), etc.)

Both Covariance & Correlation tell us whether (zero or not zero) \( d \) is in what direction (+ directly, - inversely) two R.V.'s go together, how much they go together (larger, in general, means more). From standpoint of answering the "how much" question, Cov is problematic in its magnitude depends not only on "how much" they go together but how large (scale) they each are to begin with. Corr. "scales" this problem by dividing by two quantities related to covariance \( d \) to \( x_1 \) & \( x_2 \) that similar reflect how large-scale \( x_1 \) and \( x_2 \) are to begin with. So by dividing, the "scale" information cancels out & we get a "pure" measure of how much the \( x_1 \) and \( x_2 \) go together. Conveniently, the "scale" information cancels out just perfectly so that:

\[ -1 \leq \rho \leq 1 \]

\[ \rho = 0 \Rightarrow \text{zero or "no" correlation} \]

\[ \rho \rightarrow \pm 1 \Rightarrow \text{perfect (positive) correlation} \]

\[ x_1 \text{ and } x_2 \text{ go together perfectly, as } x_1 \text{ goes up or down, } x_2 \text{ also goes up or down \& by some exact fixed amount} \]
Conditional Distributions

\[ f(x_a \mid x_1) \equiv \frac{f(x_1, x_a)}{f(x_1)} \]

(definition works for discrete or continuous; in fact also for set theory analogy; also for set theory analog)

\[ P(A \mid B) = \frac{P(\cap AB)}{P(B)} \]

\[ P(A \mid B) = \frac{P(\cap AB)}{P(B)} \]

\[ \text{So, within any joint distribution, } f(x_1, x_2), \text{ there are 2 marginal distributions, } f_1(x_1) \text{ and } f_2(x_2), \text{ which are the} \]

\[ \text{conditional distributions, } f(x_1 \mid x_2) \text{ and } f(x_2 \mid x_1), \text{ or in the multivariate case.} \]

**Conditional Expectations**

We may very well be interested in the expectation of \( x_a \) given \( x_1 \), or rephrasing the variables for obvious reasons, \( y \) given \( x \).

\[ E(y \mid x) = \sum_{y} y \cdot f(y \mid x) \]

(\( E(y \mid x) \) sometimes written \( E(y \mid x) \))

Greene also calls this "regression of \( y \) on \( x \)"

Example: Go back to our sex & height table.

- The condition distribution of (sex | height \( \geq 6' \) is

\[ f(y \mid x) = \frac{P(\text{male} \& y \mid 6')}{P(y \mid 6')} = 0.19/0.8 = \frac{19}{8} \]

- The conditional expectation, calling female = 1, male = 0,

\[ E(y \mid x) = \sum_{y} y \cdot f(y \mid x) = 0 \cdot \frac{19}{8} + 1 \cdot \frac{51}{8} = \frac{51}{8} \]

Harder to draw for continuous, but suppose

**Conditional Variance:**

Analogous to the usual partitioning, get \( E(y \mid x) \) by

\[ \sum_{y} g(y) \cdot f(y \mid x) \]

\[ \int_{-\infty}^{\infty} g(y) \cdot f(y \mid x) \, dy \]

So, the most important example, the expectation of squared deviation, or variance, is

\[ V(y \mid x) = E[(y - E(y \mid x))^2 \mid x] = \sum_{y} (y - E(y \mid x))^2 \cdot f(y \mid x) \]

writing it out, we can simplify

\[ V(y \mid x) = E[(y - E(y \mid x))^2 \mid x] = E[y^2 - 2y \cdot E(y \mid x) + (E(y \mid x))^2 \mid x] \]

(Expect of sum is sum of expects)

\[ = E[y^2 \mid x] - 2 \cdot E[y \cdot E(y \mid x) \mid x] + E[(E(y \mid x))^2 \mid x] \]

\[ = E[y^2 \mid x] - 2 \cdot E[y \cdot \mu \mid x] + E[\mu^2 \mid x] \]

\[ \Rightarrow V(y \mid x) = E(y^2 \mid x) - [E(y \mid x)]^2 \]

\[ \text{These are hills coming out of the page, darker is higher, higher represents more probable, the slice of the hill at } x_0 \text{ would be 2 dimensional & could be drawn in the second chart. That's } E(y \mid x) \text{ is just the mean of that slice.} \]
Conditional Variance:  
\[ V(y|x) = E[(y - E(y|x))^2 | x] = \sum_y (y - E(y|x))^2 f(y|x) = \int_{y|x} (y - E(y|x))^2 f(y|x) dy \]

- Returning to the (beautiful) contour map of the "hills" & the slice from that map: the conditional variance of \( y \), given \( X = X_0 \) is the variance of \( f(y|x) \) which is, the variance of that slice of the "hill". A measure of how "spread" the distribution is, how "broad" or "fat" the "hill" is at that point. In general, the conditional variance depends on which \( x \) you give. The cross-sectional (conditional) shape of the hill is going to differ by where you slice it; unless the hill is very regularly shaped. No "true hill" world; for example, have identically shaped cross-sections.

FYI: The conditional variance function is called the skedastic function. A joint distribution with identical \( V(y|x) \) \( \forall x \) is called homoskedastic. If \( V(y|x) \) varies by \( x \) we give, then the (joint) distribution is heteroskedastic.

N.B.: Expectations of \( E(y|x), V(y|x) \) are generally going to be functions of \( x \) or simply constants. Thus, they "useless to describe" (give central tendency & spread of \( y \).)

FYI: As thinking about the hill-crosssection ontology may make clear, there are conditional skewness, conditional kurtosis, etc. (i.e., conditional moments).

- Alternative Analogy for Conditional Distributions: Think of a sheet hovering in a room over air jets blowing at various strengths. The strength with which a particular part of the sheet is being "blown" is actually the force & frequency with which it is hit by particles in the air stream. Thus the height of the sheet is actually a pretty good analogy for a joint pdf. The coordinates \( x, y \) give the layout of the room, & the height of the sheet \( z \), \( (x,y) \) is \( f(x,y) \), the pdf! A strength with which it's being blown.

The conditional distribution \( f(y|x = x_0) \), say, would be by the thread running through the sheet at \( x_0 \).

\( f(x|y = y_0) \) is analogously the thread running the other way \( a + y_0 \).

Some Important Properties & Notes to Wrap Up: (here \( a, b, c \) always constants, \( x \) \& \( y \) R.V.'s)

1. \( E[\sum a_i + \sum b_i X_i] = a + \sum b_i E(X_i) \)
2. \( V(a + bX + cY) = b^2 V(X) + 2bc Cov(X, Y) + c^2 V(Y) \)
3. \( Z_1 = (a + bX + cY) \quad Z_2 = (a + bX + cY) \quad Z_1, Z_2 \)
   \( \Rightarrow Cov(Z_1, Z_2) = b_1 b_2 Cov(X, Y) + b_2 Cov(X, Y) + b_2 Cov(X, Y) + c_2 Cov(Y) \)
4. \( E[(a + bX + cY) | x] = a + bx + c E(Y|x) \)
5. \( V(x + y) = V(x) + V(y) + 2 Cov(x, y) \)
6. \( V(x - y) = V(x) + V(y) - 2 Cov(x, y) \)
7. \( E(y) = E_x[E(y|x)] \quad \text{where } E_x \text{ "take the expectation over all } x \" \text{ thus somewhat self-evidently the inverse of \( \text{giving (a particular) } x \). This is also called, the law of iterated expectations.}

8. \( \text{ Cov}(X, y) = Cov(X, E_y(x)) \)
Some More Properties:

\[ \sigma^2_y = \mathbb{E}_y \left( \mathbb{V}(y|X) \right) + \mathbb{V}_y \left[ \mathbb{E}(y|X) \right] \]
(provisional analog to \( TSS = RSS + ESS \))

\[ \mathbb{E}(XY) = \mathbb{E}_x \left[ X \cdot \mathbb{E}(Y|X) \right] \]

...why again, part of \( Y \) that can't be predicted by \( X, Y-E(Y|X) \), is uncorrelated with \( X \), i.e. orthogonal so

\[ \mathbb{E}(XY) = \mathbb{E}(X \cdot (E(Y|X)+E)) = \mathbb{E}(X \cdot E(Y|X)) + E(E) \]

Still More Properties

Define \( \varepsilon = Y - E(Y|X) \)

(i) \( \mathbb{E}(\varepsilon|X) = E(Y|X) - E(E(Y|X)|X) = E(Y|X) - E(Y|X) = 0 \)

(ii) \( \mathbb{E}(EE|Y) = E(Y|Y) - E(E(Y|X)|Y) = Y - Y = 0 \)

(iii) \( \mathbb{E}(\varepsilon) = E(Y) - E(E(Y|X)) = E(Y) - E(Y) = 0 \)

(iv) \( \text{Cov}(X, \varepsilon) = E(\varepsilon X) - E(\varepsilon)E(X) = E(X(Y - E(Y|X))) - E(X)E(Y) = E(XY) - E(X)E(Y) = 0 \)

(v) \( \mathbb{V}(\varepsilon) = \mathbb{V}[E(Y|X)] \)

(vi) for any function \( h(X) \), \( \text{Cov}(h(X), \varepsilon) = 0 \)

In Special case where \( E(Y|X) \) is linear: \( E(Y|X) = a + bX \)

\[ \begin{cases} 
\text{1. } a = E(Y) - bE(X), \\
\text{2. } b = \text{Cov}(X,Y) / \text{Var}(X)
\end{cases} \]

\[ \Rightarrow a = E(Y) - bE(X) \]

\[ \Rightarrow a = E(Y) - bE(X) \]

\[ \Rightarrow b = \text{Cov}(X,Y) / \mathbb{V}(X) \]

Aside: suppose, then you have (or have estimated) \( E(Y|X) \) and found: \( E(Y|X) = bX + E(Y|X) \), what's \( \sqrt{b^2 \mathbb{V}(X)} \)?

Decomposition of Variance: we already saw that \( \mathbb{V}(Y) = \mathbb{V}_x \left( \mathbb{V}(Y|X) \right) + \mathbb{V}_y \left( \mathbb{E}(Y|X) \right) \)

in linear case, as we hinted above, this has important substantive meaning:

- Variation in \( Y \) comes from two sources:
  1. Variation due to the fact that \( E(Y|X) \) varies with \( X \), (the variance of the expectation, or "explained variance")

\[ \text{regression variance} = \mathbb{V}_x \left( E(Y|X) \right) \]

(\( = b^2 \mathbb{V}(X) \) in bivariate case)

2. Variation due to the fact that \( Y \) varies around the conditional expectation (\( Y \) varies around the prediction):

\[ \text{residual variance} = \mathbb{V}_x \left( V(Y|X) \right) \]

Provided \( \mathbb{V}(\varepsilon) \) constant (homoscedasticity, then variance decomposition implies

\[ \mathbb{V}(E(Y|X)) \leq \mathbb{V}(Y) \]
Coefficient of Determination = \text{COD} = \rho^2 = \text{Explained Variance} / \text{Total Variance}

\[ \Rightarrow \text{in Multivariate setting: } R^2 \]

*measure of linear association between X \& Y*

(we'll obviously have a lot more to say about \( R^2 \)
when we come to OLS)

**INDEPENDENCE:**

There are (at least) three types of (in)dependence that appear from time-to-time in considerations of how to analyze relationships between various R.V.s.

1. **Stochastic (or Statistical) Independence** (this is strongest): Stochastic independence is the opposite of correlation, it means having non-zero variance; refers to R.V.s in general:

\[
\begin{align*}
\text{y stochastically independent of } x & \iff f(y|x) = f(y) \\
\iff f(x,y) = f(x)f(y) & \iff f(x|y) = f(x)
\end{align*}
\]

*proof of identically of these definitions:

\[
\begin{align*}
0. f(y|x) = \frac{f(x,y)}{f(x)} & \text{ by definition of } f(y|x) \\
0. f(x|y) = \frac{f(x,y)}{f(y)} & \text{ by definition of } f(x|y) \\
0. f(x,y) = f(x)f(y) \text{ then } f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(x)}{f(y)} & \text{ but it also } \equiv f(y|x). \\
\text{so } \Rightarrow \Rightarrow \Rightarrow \\
\Rightarrow f(x,y) = f(x) \text{ but it also } \equiv f(x|y). \\
\text{so } \Rightarrow \Rightarrow \Rightarrow \\
\text{finally, taking } 0 \text{ or } 0 \text{ (it doesn't matter)}
\end{align*}
\]

\[
\begin{align*}
f(y|x) = f(y) & \text{ by } f(f(x,y)|x) = f(x|y) \\
\Rightarrow f(x,y) = f(x) & \text{ by rewriting} \\
\Rightarrow f(x,y) = f(x)f(y)
\end{align*}
\]

*meaning of this definition: knowing X tells us nothing about Y, it gives us no better guess of E(Y) or \( f(Y) \) or anything, X is on every basis \& from all directions a meaningless piece of info as it relates to the distribution of Y.*

2. **Mean Independence** \( \iff E(Y|X) = E(Y) \)

*corollary: if Y mean-independent, X then Y mean-independent (X)

\[
\begin{align*}
\text{E}(Y|X) = E(Y) \iff E(x|y) = E(x) \\
\text{i.e., mean independence is symmetric, not stochastic.}
\end{align*}
\]

3. **"Linear" Independence:** (Uncorrelatedness)

(\text{Def. } X \& Y \text{ "linearly independent" (in the statistical sense, or maybe perfectly linearly independent)} \iff \text{Cov}(X,Y) = 0 \implies p(x,y) = 0)
**Multivariate Extensions:** Let $X_i$ be a vector of random variables. $X_i$ is not the matrix of expectations of $X_i$, but rather a vector of expectations of $X_i$.

- **Linear independence:** If $X_i$ is independent of $X_j$ for all $i 
eq j$, then the covariance matrix is diagonal, and $X_i$ and $X_j$ are said to be independent.

- **Strong dependence:** If $X_i$ and $X_j$ are not independent, then they are said to be dependent.

**Stochastic Interdependence:**

- $X$ is mean independent.
- $X$ is mean uncorrelated.
- $X$ is mean dependent.
- $X$ is mean correlated.

**Variance-covariance matrix:**

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

**Maximum likelihood estimation:**

- For the case of $n$ observations, $\hat{\beta}$ is the maximum likelihood estimator of $\beta$.
- The likelihood function is $L(\beta|x) = \prod_{i=1}^{n} f(x_i|\beta)$.

**Errors:**

$$\epsilon_i = x_i - \hat{x}_i$$

**Residuals:**

$$\hat{\epsilon}_i = x_i - \hat{x}_i$$

**Homoscedasticity:**

- The variance of the errors is constant across observations.
- The errors are homoscedastic.

**Normality:**

- The errors are normally distributed.
- The errors follow a normal distribution.

**Independence:**

- The errors are independent across observations.
- The errors are independent of each other.

**Random effects:**

- The errors are random variables.
- The errors are stochastic variables.

**Fixed effects:**

- The errors are fixed variables.
- The errors are non-stochastic variables.

**Estimation:**

- $\hat{\beta}$ is the maximum likelihood estimator of $\beta$.
- The maximum likelihood estimator is $\hat{\beta}$.

**Hypothesis testing:**

- The hypothesis that $\beta = 0$ is tested.
- The hypothesis that $\beta = 0$ is rejected.

**Confidence intervals:**

- The confidence interval for $\beta$ is $\hat{\beta} \pm t_{\alpha/2} \sqrt{\text{var}(\beta)}$.
- The confidence interval for $\beta$ is $\hat{\beta} \pm t_{\alpha/2} \sqrt{\text{var}(\beta)}$.

**Standard errors:**

- The standard error of $\hat{\beta}$ is $\sqrt{\text{var}(\hat{\beta})}$.
- The standard error of $\hat{\beta}$ is $\sqrt{\text{var}(\hat{\beta})}$.

**Goodness of fit:**

- The goodness of fit is measured by the $R^2$ statistic.
- The goodness of fit is measured by the $R^2$ statistic.

**Diagnostic checks:**

- The diagnostic checks include checking for autocorrelation, heteroscedasticity, and influential observations.
- The diagnostic checks include checking for autocorrelation, heteroscedasticity, and influential observations.

**Model specification:**

- The model specification is $y = \beta_0 + \beta_1 x + \epsilon$.
- The model specification is $y = \beta_0 + \beta_1 x + \epsilon$.

**Example:**

- In a regression analysis, the relationship between $y$ and $x$ is estimated.
- In a regression analysis, the relationship between $y$ and $x$ is estimated.
Conditional Distributions, Expectations, and Variances:

1. Joint pdf is like a contour map
   - Conditional distribution is a slice of that contour.

2. Some properties of note:
   a) $V(a + bX + cY) = b^2V(X) + 2bc\text{Cov}(X,Y) + c^2V(Y)$
      
      \[ E[(a + bX + cY)^2] = E(a^2 + 2abX + 2acY + b^2X^2 + 2bcXY + c^2Y^2) = [a + bE(X) + cE(Y)]^2 \]
      \[
      = a^2 + 2abE(X) + 2acE(Y) + b^2E(X^2) + 2bcE(XY) + c^2E(Y^2) \\
      = b^2E(X^2) + 2bcE(XY) + c^2E(Y^2) \\
      = b^2V(X) + 2bc\text{Cov}(X,Y) + c^2V(Y) \\
      \]
   b) Any RV can be written $Y = E(Y|X) + (Y - E(Y|X)) = E(Y|X) + e$
   c) $E[E(Y|X)] = E(Y)$
   d) $\text{Cov}(X, E(Y|X)) = \text{Cov}(X, Y)$
      
      \[
      \text{Cov}(X, Y) = \text{Cov}(X, E(Y|X) + e) \\
      = \text{Cov}(X, E(Y|X)) + \text{Cov}(X, e) \\
      = \text{Cov}(X, E(Y|X)) + \varnothing \\
      \text{intuition: unrelated part of } Y \text{ doesn't covary with } X \text{ by definition of "unrelated"}
      \]
   e) $V(Y) = E(V(Y|X)) + V(E(Y|X))$
      
      \[
      V(Y) = V(E(Y|X)) + \sum_{x} \left[ E(Y|X=x) - E(Y|X) \right]^2 = \sum_{x} \left[ E(Y|X=x) - E(Y|X) \right]^2 \\
      \text{TSS} = \text{RSS} + \text{ESS} \Rightarrow V(Y) \leq V(Y) \text{ why?}
      \]
   f) $E(XY) = E(X \cdot E(Y|X))$
      
      \[
      Y = E(Y|X) + \varepsilon \quad \text{by construction}
      \]
(More Properties) (for \( E[Y - E(Y|x)] \))

\[
\begin{align*}
g) \quad E(\epsilon | x) &= E(Y - E(Y|x) | x) = E(Y | x) - E[E(Y|x) | x] = E(Y | x) - E(Y) \\
E(\epsilon | y) &= 0 = E(Y | y) - E[E(Y|x) | y] = E(Y) - E(Y | y) = 0 \\
E(\epsilon) &= 0 = E(Y - E(Y|x)) = E(Y) - E(Y) = 0 \\
\Rightarrow \text{You're not supposed to know anything about your residuals (may know its stochastic features).}
\end{align*}
\]

h) \( \text{cov}(x, \epsilon) = E(x\epsilon) - E(x)E(\epsilon) = E(xy - xE(Y|x)) - \mu_x \cdot \sigma_x \)

\[
\text{cov}(x, y) = E(xy) - E(x)E(y) = 0
\]

i) \( \text{V} (\epsilon) = E(\text{V}(Y|x)) \)

\[
\text{This is the thing about which } \text{the variance-of-residuals is talking.}
\]

(More Properties) for case where \( E(Y|x) \) linear:

\[
\begin{align*}
E(Y) &= a + bX \\
E(Y|x) &= a + bX
\end{align*}
\]

j) \( E(Y) = a + bE(X) \) i.e., \( \bar{Y} = \bar{a} + \bar{b} \bar{X} \) (regression line passes through the mean)

k) \( b = \frac{\text{cov}(x,y)}{\text{V}(x)} \) Proofs: \( \phi = \frac{\text{cov}(x, E(Y|x))}{\text{V}(x)} = \frac{\text{cov}(x, a + bX)}{\text{V}(x)} \)

\[
(\text{V}(x)^2 + \text{E}(x)^2) \phi = \frac{\text{cov}(x,y)}{\text{V}(x)} \]

Notice: \( \text{if } E(Y) = a + bX \) \( \text{then } \phi = \frac{\text{cov}(x,y)}{\text{V}(x)} \)

\[
[\text{cov}(x,y)^2] = \frac{\text{cov}(x,y)}{\text{V}(x)} = \frac{\text{V}(y)}{\text{V}(x)} = b
\]

\[
(b_{xy} = \frac{\text{cov}(x,y)}{\text{V}(x)} = \frac{\text{cov}(x,y)}{\text{V}(x)} \cdot \frac{\text{cov}(x,y)}{\text{V}(y)} = \frac{[\text{cov}(x,y)]^2}{\text{V}(x) \text{V}(y)}
\]

\[
\Rightarrow \text{correlation coefficient: } \rho_{xy} = \frac{b_{xy}}{\sqrt{b_{xx} \cdot b_{yy}}}
\]

\[
\rho_{xy} = \frac{(x,y \in X \in Y) \text{ var } x \text{ var } y}{\sqrt{\text{var } x \text{ var } y}}
\]

\[
\text{cov}(x,y) \geq 0 \iff \rho_{xy} + \text{ cov } x \text{ cov } y \\
\text{cov } x \text{ cov } y = \frac{\text{V}(y)}{\text{V}(x)} = b
\]

\[
\Sigma = (x,y) \in X \in Y \quad \text{csc } x \text{ csc } y
\]

\[
\Sigma = (x,y) \in X \in Y \quad \text{csc } x \text{ csc } y
\]
Independence:

1. Stochastic (Statistical) Independence
   \[ f(x, y) = f(x) f(y) \]
   \[ f(y|x) = f(y) \]
   \[ f(x|y) = f(x) \]
   \[ f(x) = \frac{f(x, y)}{f(y)} \]
   familiarity:
   knowing \( x \) doesn't tell you one damn thing about the distribution of \( y \) \& v.v.

2. Mean Independence:
   \[ E(Y|X) = E(Y) \]
   \[ E(Y|X) = E(Y) \rightleftharpoons E(X|Y) = E(X) \]
   knowing \( x \) tells you \( E(Y|X) \), the mean of \( X \);
   may not tell you \( V(Y|X) \) or other stuff about the distribution of \( Y \)

3. "Linear Independence" or Uncorrelatedness
   \[ \text{cov}(X, Y) = 0 \rightleftharpoons \text{linear independence} \]
   \[ \text{cov}(X, Y) = 0 \rightleftharpoons \text{linear independence} \]

   \[ E(Y|X) \]
   \[ E(Y|X) \]
   similarly, proportional from other direction also

   \[ f(y|x) = f(y) \]
   \[ f(x|y) = f(x) \]

   \[ f(x) = \frac{f(x, y)}{f(y)} \]
   \[ f(y|x) = f(y) \]
   \[ f(x|y) = f(x) \]
\[ E(X) = \begin{bmatrix} E(X_1) & \cdots & E(X_n) \\ \vdots & \ddots & \vdots \\ E(X_n) & \cdots & E(X_{2n}) \end{bmatrix} \]

\[ V(X) = E[(X - \mu)^2] = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \]

\[ V(Y) = E[(Y - \mu)^2] = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \]

\[ V(b) = \begin{bmatrix} V(b_1) & \cdots & V(b_n) \\ \vdots & \ddots & \vdots \\ V(b_n) & \cdots & V(b_{2n}) \end{bmatrix} \]

Special Case: \( X_1, X_2, \ldots, X_n \) multivariate normal, then linear \( \Rightarrow \) mean \( \Rightarrow \) stochastic.
I. Greene's Sections: What to Do/Know
A. Focus on Sections 4.1-4.3.1 (Intro through Point Estimation)
B. Section 4.3.2, Efficient Unbiased Estimation, gets hard when the
Cramer-Rao lower bound is introduced—struggle with that, but by sure
you understand the point about what the C-R lower bound tells us.
C. Section 4.4, Large-Sample Distribution Theory is key, again especially a
level of Conceptual understanding. Don't worry about 4.4.1-4.4.5 through for
the exact focus.
D. Do try to grasp 4.5, Maximum Likelihood Estimation, especially the first
estimation variance of the MLE is always harder. Be sure you at least
know what the properties are and what they mean in 4.5.1-4.5.2,
just do the best you can to see what's going on—we'll return to this in
more applied settings perhaps later, so don't kill yourself on 4.5.2.
E. Section 4.6, Consistent Estimation: The Method of Moments
is a bit beyond this class. You may (should?) have looked just
as far as what it's about; FYI.
F. Section 4.7 on Interval Estimation, & 4.8 on Hypothesis Testing, are
essential.

G. In sum: 4.1-4.3.1, 4.4, 1st part of 4.5, 4.7-4.8 are the key; rest
(possibly except 4.6) could be attempted though

Notational Note (Definition): iid means independently and identically distributed

II. We will think of outcomes or events in the world as having been generated
by some underlying (and usually unobservable) probabilistic process.
A. This process is often characterized as having been generated by
some probability function or probability density function (pdf).
1. Eg., we might think of election outcomes as having been
probable by some Bernoulli probability function.
2. Eg., we might similarly think of how much the two candidates
spent in the campaign as coming from pdf's.
3. The sum of their spending might be considered Normal(µ, σ²)
4. Each candidate's spending might be considered to have
been drawn from a bounded normal N(µ, σ²)
B. This does not mean that we think of the world as
random in the sense of completely unpredictable or "having nothing
systematic about it"
1. For one thing, the statement that the outcome in
question comes from some particular type of pdf or pdf,
is itself a statement of probabilistic regularity.
2. For another thing, we usually also go on to say that we know
or think we know something about what determines
the parameters of the distribution
a) We may think, for example, that Th., probability
the incumbent wins is some function of the amount
of pork she brought to her district.
b) Total spending we might think to be a function
of how close to 50-50 the partisanship of
voters in the district, for example.
C. We might, then, wish to use the outcomes that actually occurred
in some real-world sample of like events to infer something
about the (parameters of the underlying distribution.

D. This is the whole point: we are interested in inferring from what
we observe, how the world actually works; the underlaying systematic.
elements of the political $, more broadly, the socio-politico-economic
world.
1. This course is about how best to use the information available
from what has happened.
2. "Best" here refers to the spare of accurately inferring; it is
decidedly an applied science as we are interested in it.
III. Samples, Sampling Theory, & Simple Statistics

A. (Simple) Random Sample (SRS): n observations, \( X_1, X_2, X_3, \ldots, X_n \) drawn independently from the same population (i.e., from the universe of possibilities as described by a p.f. or p.d.f.) is a SRS.

\[ f(X_j, \Theta) \]

may be an observation on a single p.f. or on a random vector of even a random matrix

\[ \text{e.g. } X_1, \ldots, X_n \text{ a SRS from } N(\mu, \sigma^2) \rightarrow \Theta \]

\[ \Rightarrow \text{ each } X_i \text{ a drawn from an independent } N(\mu, \sigma^2) \]

1) In practice, our sample will be in the form of a

d) cross-section: a set of observations on various units (CS) (individuals, states, counties, governments, etc.) at some time, t.

b) time-series: a set of observations on a single unit across some number of (usually evenly spaced) time periods, \( t=1, 2, 3, \ldots, T \)

c) time-series-cross-sections: the combination of (a) & (b) -- a set of cross-sectional observations across some number of time periods, or equivalently, a set of time-series observations across some cross-section of units.

FTR: i) a TSCS of data with (many) more CS than TS is usually called panel data

(ii) a TSCS of data with (many) more TS than CS is usually called pooled data

this nomenclature is not all that terribly adhered to though

2) Since an SRS, each \( X_i \sim \text{ independently } f(X_i, \Theta) \), the joint distribution of the \( X_i \)'s is

\[ f(X, \Theta) = f(X_1, \Theta)f(X_2, \Theta)\cdots f(X_n, \Theta) \]

vector of the \( X_i \)'s or matrix of the vectors \( X_i \)

B) (Sample) Descriptive Statistics

1) A statistic is any function computed from the data.

of course, we usually restrict our attention to functions of the data that are informative about the distribution of the data (i.e., hopefully \( \hat{\Theta} \), by some inference process informative about the underlying population).  

2) Some common "descriptive" statistics (i.e., they purport to help describe or summarize the entire sample distribution, \( \hat{\Theta} \), again, by inference, the population distribution)
III. B. 2. Descriptive Statistics

a) First, generically, a statistic \( y \) is some function of the data \( y = g(x_1, x_2, \ldots, x_n) \).

b) Examples:

i) The Sample Mean or Average

(a) The mean, i.e., \( \mu \) or \( \mathbb{E}(X) \), is a population feature; the sample mean or average is an estimate of it.

(b) Def. \( \bar{x} = g(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \).

(c) Other measures of the sample's central tendency, the sample median, mode, midrange, etc., can be defined analogously to their population cousins as well.

ii) The Sample Standard Deviation:

(a) Usually, defined \( s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \).

(b) Occasionally, we use (for reasons to be explained) \( s_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \) instead or divide by some other quantity (related to \( n \)) for other reasons.

(b) Other measures of the sample spread, sample mean absolute deviation, or sample range, etc., can again be defined similarly to their population analogs.

iii) Sample Relationships

(a) Sample covariance:

\[ s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \]

(b) Sample correlation:

\[ r_{xy} = \frac{s_{xy}}{s_x s_y} \]

(c) It is often useful to organize and arrange sample variances and covariances for correlations in matrices:

(1) Sample Covariance Matrix

\[
S_x = \begin{bmatrix}
S_{xx} & \cdots & S_{x1} \\
\vdots & \ddots & \vdots \\
S_{1x} & \cdots & S_{11}
\end{bmatrix}
\]

(2) Sample Correlation Matrix

\[
R_x = \begin{bmatrix}
1 & \cdots & r_{12} \\
\vdots & \ddots & \vdots \\
r_{21} & \cdots & 1
\end{bmatrix}
\]

n.b.: Diagonal elements of \( S_x \) are variances of \( x_i \), right?

n.b.: Diagonal elements of \( R_x \) are all one (1), right?

1) \( s_x^2 = \text{sample variance} = \frac{1}{n-1} \left[ \left( \frac{n}{\sum_i} \right) - \frac{n \overline{x}}{n} \right] = \sum_i \frac{1}{n-1} x_i^2 - \frac{n}{n-1} \overline{x}^2 \)

2) \( s_{xy} = \text{sample cov. of } x,y = \frac{1}{n-1} \left[ \left( \sum_i x_i ight) \cdot \left( \sum_i y_i \right) - n \overline{x} \overline{y} \right] \)

3) \( r(a \cdot x, b \cdot y) = \frac{a \cdot b}{|a| \cdot |b|} \cdot r_{xy} \quad \text{[e.g. } r_{xy} = 0.5, \text{ } \Rightarrow r(2 \cdot x, 1 \cdot y) = -0.5 \]}

4) \( s(a \cdot b \cdot x_c + d \cdot y) = b \cdot d \cdot s_{xy} \quad \text{[e.g. } s(2+3x, 4+2y) = 12 \cdot s_{xy} \]}

\[ s^2(a \cdot b \cdot x) = b^2 \cdot s_x^2 \quad \Rightarrow s(a \cdot b \cdot x) = |b| \cdot s_x \]

in short, all the useful properties of expectations, variances, and covariances we used for population RV's last week work exactly analogously for their sample statistic analogs.

III. C. Sampling Distributions (of Simple Statistics)

1) The random sample itself can be seen as presenting a "sample distribution." This is not to be confused with the sampling distribution of a statistic.

2) Since each observation is a draw from a probability distribution, if we draw another sample, the observations would be different. Thus, across repeated samples, sample statistics calculated from those samples will vary in some way. The sampling distribution of a statistic is the p.f. or p.d.f. which describes how the statistic varies across these (usually hypothetical) samples.

3) Thus, we can speak, for example, of \( \overline{X} \sim f(\overline{X}, \theta) \), i.e. of the sample average being distributed across repeated samples, according to \( f(\overline{X}, \theta) \). We might, furthermore, be interested in the expected value of variance, etc., of \( \overline{X} \) or whatever other statistic across these repeated samples:

\[ E(\overline{X}) = E\left[ \frac{1}{n} \sum X_i \right] \quad \text{by definition of } \overline{X} \]

\[ = \frac{1}{n} E(\sum X_i) \quad \text{by } \frac{1}{n} \text{ can pull constants out of } E(\cdot) \]

\[ = \frac{1}{n} (E(X_1) + E(X_2) + E(X_3) + \ldots) \quad \text{writing out the sum} \]

\[ = \frac{1}{n} (n \cdot \mu) \quad \text{each } X_i \text{ is a SRS, from population with same mean } \mu \]

\[ = \mu \]

\[ V(\overline{X}) = V\left( \frac{1}{n} \sum X_i \right) = \left( \frac{1}{n} \right)^2 V(\sum X_i) = \left( \frac{1}{n} \right)^2 \sum_i V(X_i) \]

\[ = \frac{1}{n^2} \sum_i \sigma^2 \quad \text{by in this step, since } X_i \text{ independent, } \text{Cov}(X_i, X_j) = 0 \]
III. C. 3. Mean & Variance of the Sampling Distribution of \( \bar{X} \)

\[
E(\bar{X}) = \mu
\]

\[
\text{Var}(\bar{X}) = \frac{1}{n} \sigma^2
\]

(N.B. \( \text{Var}(\bar{X}) \) is independent of the random variable \( X \) being averaged, \( \mu \) is constant) 

**Notes:**

a) This is true for \( X \)'s drawn from any p.f. or p.d.f. with sample size \( n \) being averaged.

b) Variance of averages across repeated samples is \( \frac{1}{n} \) (the variance of the underlying random variable)

\[
\Rightarrow \text{"averaging reduces variance" in this sense if by a factor of \( \frac{1}{n} \) where \( n \) is the sample size being averaged.}
\]

An example: If \( X_i \sim N(\mu, \sigma^2) \) \( \forall i \) (i.e. the \( X_i \) are independent)

\[
\bar{X} \sim N(E(\bar{X}), \text{Var}(\bar{X}))
\]

\[
\text{because sum of normals is normal}
\]

\[
\bar{X} \sim N(\mu, \frac{\sigma^2}{n})
\]

4. We will be interested in how a statistic behaves across repeated samples as those (hypothetically repeated) samples get larger and larger:

c) \( \lim_{n \to \infty} \text{Var}(\bar{X}) = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0 \)

b) \( \lim_{n \to \infty} E(\bar{X}) = \lim_{n \to \infty} \mu = \mu \) (constant doesn't vary with \( n \))

so, as \( n \to \infty \), the distribution of \( \bar{X} \) "goes to a spike" over \( \mu \)

more on this in a bit.

IV. (Point) Estimation of Parameters

A. First some Definitions

1. **Point Estimate**: a statistic that gives a single value estimate or guess for a parameter \( \Theta \) or a set of parameters, \( \Theta \)

2. **Standard Error**: the standard deviation of the sampling distribution of some point estimate.

- the sampling standard deviation is the standard error; the square of that is the sampling variance.

3. **Interval Estimate**: a statistic that gives some range of values as a guess of \( \Theta \) or \( \Theta \). Usually, the interval is constructed in an appropriate (i.e. useful) manner from a point estimate and its standard error.

- e.g. "Margin of Error": 90% favor Clinton +/\( \pm \) 2; \( \Theta \) = true %

- Favoring Clinton: 90% is the point estimate, \( \hat{\Theta} \) = the standard error, "Margin of Error" \( \equiv 2 \hat{\sigma} \) (so here, \( \hat{\sigma} \) must be 1) if the interval estimate is, \( 48\% \pm 2 \) or \( 48\% \pm 2 \) or \( 46\% \pm 2 \%

4. **Estimator**: an estimator is a rule or a strategy for guessing at a parameter; i.e., it is a function that produces an estimate:

\[
g(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

is an estimator for \( \mu \) (that we often call \( \bar{X} \))

5. **Finite (or "small") Sample**: any fixed sample size less than \( \infty \)

Asymptotic (or "large" or "infinite") Sample: a sample of infinite or "going toward infinite" size.

1. Unbiasedness: An estimator of \( \Theta \), call it \( \hat{\Theta} \), is unbiased iff \( E(\hat{\Theta}) = \Theta \)

   a. Example: \( \bar{X} \) is an unbiased estimator of \( \mu \)
      \[ E(\bar{X}) = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n} \sum E(X_i) = \frac{1}{n} \sum \mu = \mu \]

   b. \( \hat{\Theta} = \text{ith observation of } X_i, X_i \) is an unbiased estimator of \( \theta \)
      \[ E(\hat{\Theta}) = E(X_i) = \mu \]

2. Efficiency: An unbiased estimator, \( \hat{\Theta} \), is more efficient than another unbiased estimator, \( \hat{\Theta}_2 \), iff \( V(\hat{\Theta}) < V(\hat{\Theta}_2) \)

   e.g. \( \hat{\Theta} = \bar{X} \) is more efficient than \( \hat{\Theta} = X_i \):
   \[ V(\bar{X}) = \frac{\sigma^2}{n} < V(X_i) = \sigma^2 \]

3. Mean-Squared Error (MSE):
   \[ \text{MSE}(\hat{\Theta}) = E[(\hat{\Theta} - \Theta)^2] = \text{Var}(\hat{\Theta}) + (\text{bias}(\hat{\Theta}))^2 \]

   a. \[ E(\hat{\Theta}^2) - 2E(\hat{\Theta} \cdot \Theta) + \Theta^2 \]
   b. \[ [E(\hat{\Theta})]^2 + V(\hat{\Theta}) - 2E(\hat{\Theta}) \cdot \Theta + \Theta^2 \]

   for any RV, \( E(X^2) = \mu^2 + \sigma^2 \)

   c. \[ \text{Bias}(\hat{\Theta}) = E(\hat{\Theta}) - \Theta \]
      \[ [\text{bias}(\hat{\Theta})]^2 = [E(\hat{\Theta})]^2 - 2E(\hat{\Theta}) \cdot \Theta + \Theta^2 \]

   - Another criteria, then, for comparing 2 estimators or judging
     one is minimum MSE.

4. More generally, we can assume/or it is appropriate to apply some other function reflecting the costs of "mistakes".

   e.g. Min. MSE assumes/appplies costs of mistakes according to the function \( (\Theta - \hat{\Theta})^2 \). We could imagine scenarios where costs of mistakes are zero for small mistakes & constant for any large error e.g.
   \[ L(\Theta) = \begin{cases} 0 & \Theta = \Theta_0 \in C \\ 4 & \Theta = \Theta_0 \notin C \end{cases} \]
   or any old loss function we like.
Two Ways to Show $E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2$

1. $E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2$

\[
E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) = E[(\hat{\theta} - E(\hat{\theta}))^2 + [E(\theta) - \theta]^2]
\]

\[
E(\hat{\theta}^2) - 2\theta^2 + \theta^2 = E(\hat{\theta}^2 - 2E(\hat{\theta})\hat{\theta} + [E(\hat{\theta})]^2) + E(\theta)^2 - 2E(\theta)\theta + \theta^2
\]

\[
\begin{bmatrix}
\hat{\theta} = \theta + (\hat{\theta} - \theta)
\end{bmatrix}
\]

\[
\begin{bmatrix}
\theta & [\theta + (\hat{\theta} - \theta)]
\end{bmatrix}
\]

\[
\hat{\theta} + \theta = \theta
\]

\[
E(\hat{\theta}^2) - \theta^2 = E(\hat{\theta}^2) - \theta^2
\]

\[
E[(\hat{\theta} - \theta)^2] = E[ E[(\hat{\theta} - E(\hat{\theta})) + (E(\theta) - \theta)]^2]
\]

\[
= E\left\{ (\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\theta) - \theta) \right\}
\]

\[
= V(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2 + 2E(\hat{\theta} - E(\hat{\theta}) - \theta E(\theta))^2 + E(\theta)\theta + 2\left[ (E(\hat{\theta})^2 - E(\theta)\theta - [E(\hat{\theta})]^2 + E(\theta)\theta) \right]
\]

\[
= V(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2
\]

**Efficiency** in the multivariate case:

$\hat{\theta}_1$ is more efficient than $\hat{\theta}_2 \iff V(\hat{\theta}_2) - V(\hat{\theta}_1)$ is positive definite;

i.e., $x^T[V(\hat{\theta}_2) - V(\hat{\theta}_1)]x > 0$, $\forall x$. 

Page 83 of 364
5. Minimum Variance Unbiased Estimator (One which is unbiased and has the lowest variance of any estimator that is likewise unbiased.

Minimum Variance Linear Unbiased Estimator or "Best" Linear Unbiased Estimator (BLUE):

One which is unbiased and linear and has the lowest variance of any estimator that is likewise linear and unbiased.

In general, we cannot know if some estimator is MVUE or MVLUE/BLUE. Similarly, we do not generally know if some estimator is Minimum MSE. However, Cramér-Rao proved an interesting result:

Cramér-Rao Lower Bound for any "well-behaved" p.d.f., any unbiased estimator of parameter \( \theta \), will have variance at least as large as:

\[
[I(\theta)]^{-1} = \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]\right)^{-1}
\]

Here, \( L(\theta) \) is the "likelihood function" of the data, i.e., loosely, it is:

\[L(\theta) = \prod_{i=1}^{n} f(X_i, \theta)\]

E.g., suppose \( X_i \) drawn from exponential distribution:

\[X_i \sim f(x_i, \theta) = \theta e^{-\theta x_i}\]

Then, \( L(\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum x_i}\)

Thus, \( \ln L(\theta) = n \ln \theta - \theta \sum x_i \)

\[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = \frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta} \left( \frac{n}{\theta} - \theta \sum x_i \right) = -\frac{n}{\theta^2}
\]

So, the C-R lower bound = \((-E[\frac{n}{\theta}]\) = \([E[\frac{n}{\theta}]\])^{-1} = \left[\frac{n}{\theta^2}\right]^{-1}

Thus, C-R tells us that any estimator of the parameter \( \theta \) in an exponential distribution must have a variance (across repeated samples) of at least \([\frac{\theta^2}{n}]\). That implies, for example, that if we can find an estimator with that variance, we can stop looking: no other unbiased estimator will do any better.

6. Large-Sample Criteria for Estimators:

a. Convergence in Probability

1. \( \lim_{n \to \infty} X_n \) converges in probability to a constant \( C \), written \( X_n \to C \)

\[\lim_{n \to \infty} \text{Prob}(|X_n - C| > \epsilon) = 0 \text{ for any positive } \epsilon \]

2. We also write this \( \text{plim} X_n = C \)

ii. \( \lim_{n \to \infty} X_n \) converges in [mean square]

\[\text{plim} X_n = C \text{ if } \lim_{n \to \infty} \text{Var} X_n = 0 \]

iii. 1 & ii are not quite the same (2 is stricter), but for our purposes don't sweat the difference.
IV.B.6.b. Consistency: an estimator, $\hat{\theta}$, of a parameter, $\theta$, is consistent if

$$\lim_{n \to \infty} \hat{\theta} = \theta$$

- The point: a consistent estimator goes to exactly the right answer (i.e., exactly the parameter you are estimating) as the sample size goes to infinity.

- Example: $\bar{X}$ is a consistent estimator of $\mu$
  1. $E(\bar{X}) = \mu$
  2. $V(\bar{X}) = \sigma^2/n \Rightarrow \lim_{n \to \infty} E(\bar{X}) = \mu$
  3. $\lim_{n \to \infty} V(\bar{X}) = 0$

- $\bar{X}$ converges in quadratic mean to $\mu$, which is stricter than and therefore $\Rightarrow \lim \bar{X} = \mu$

- Extensions: $\lim_{n \to \infty} \frac{1}{n} \sum g(X_i) = E(g(X))$
  - i.e., the average of any function evaluated at each observation is a consistent estimator of the mean of that function.

\[ \hat{\theta} = \frac{1}{n} \sum X_i \]

$\Rightarrow \hat{\theta}$ is consistent.

- (Slutsky Theorem) for any continuous $g(X_i)$, $\lim g(X_n) = g(\lim X_n)$

Examples:

- $\lim 3x^2 = 3(\lim x)^2$
- $\lim \frac{x^2}{3} = \frac{(\lim x)^2}{3(\lim s^2)} = \frac{\mu^2}{3}$

C. Convergence in Distribution & Limiting Distribution

i. Def: $X_n$ converges in distribution to a random variable $X$ with cumulative probability function $F(x) \iff \lim_{n \to \infty} P_n(X_n \leq x) = P(X \leq x)$

- (written $X_n \xrightarrow{d} X$

- i.e., as $n \to \infty$, the distribution of $X_n$ becomes ever more like, in the limit exactly equal to, some (other) distribution.

- The distribution being "approached" in this way is called the limiting or asymptotic distribution of $X_n$

ii. The most common & most important example: $X_n \xrightarrow{d} N(\theta, \sigma)$

iii. Useful Combinations of the Two Convergences:

- $X_n \xrightarrow{d} X$ & $\lim Y_n = C \Rightarrow X_n + Y_n \xrightarrow{d} X + C$
- $\lim Y_n = C \Rightarrow X_n Y_n \xrightarrow{d} X/C$ (C ≠ 0)

iv. Another Useful Example:

- $\frac{1}{n} \xrightarrow{d} N(\theta, \sigma)$

v. Convergence in Distribution is also sometimes written

$X \xrightarrow{d} f(x)$ "$X$ is asymptotically distributed f($x$)"
IV. B. G. a. Best (i.e., Minimum Variance) Asymptotic Normal Estimator

\[ \hat{\theta} \stackrel{d}{\to} N(\theta, \phi^2) \]

for any alternative estimator which converges in distribution to \( N(\theta, \phi^2/n) \)

V. The Law of Large Numbers (LLN) & Central Limit Theorem(s) (CLT)

A. The Law of Large Numbers: \( \bar{X} \) is sample mean of random sampling from any population with \( \mu \) & \( \sigma^2 \), average \( \bar{X} \to \mu \) (i.e., \( \lim_{n \to \infty} \frac{1}{n} \sum (\text{die rolls}) = E(\text{die roll}) \))

B. Central Limit Theorem: in S.R.S. from any population w/mean \( \mu \) variance \( \sigma^2 \),

\[ \left\{ Z = \frac{\bar{X} - \mu}{\sigma} \right\} \stackrel{d}{\to} N(0,1) \]

1. This is the most useful form of the CLT; notice also, though, that the left-hand side can be rewritten:

\[ \sqrt{n} \cdot (\frac{\bar{X} - \mu}{\sigma}) \stackrel{d}{\to} N(0,1) \]

2. Given our rule that \( V(bX) = b^2 V(X) \), we can also rewrite it:

\[ \sqrt{n} \cdot (\bar{X} - \mu) \stackrel{d}{\to} N(0, \sigma^2) \quad \text{as Greene does} \]

3. Given our rule that \( E(bX) = b \cdot E(X) \), we can also rewrite it:

\[ \bar{X} \stackrel{d}{\to} N(\mu, \sigma^2/n) \]

4. It can be shown further, but Greene does not, we won’t either, that if \( X \) is drawn from population w/mean \( \mu \), but variance which is different for every observation, \( \sigma^2 \), then

\[ \sqrt{n} \cdot (\bar{X} - \mu) \stackrel{d}{\to} N(0, \sigma^2) \]

\( \sigma^2 \) is the “average variance,” \( \frac{1}{n} \cong \sigma^2 \)

C. Multivariate Extensions of the CLT:

1. If \( X_1, \ldots, X_n \) are a S.R.S. from a joint distribution with mean \( \mu \)

and variance \( \Sigma \), then \( \sqrt{n} (\bar{X} - \mu) \to N(0, \Sigma) \)

\( \Sigma \) is a positive definite matrix.

2. If \( X_1, \ldots, X_n \) S.R.S. with mean \( \mu_i \), def variance \( \sigma_i^2 \), then \( \sqrt{n} (\bar{X}_i - \mu_i) \stackrel{d}{\to} N(0, \sigma_i^2) \)

\[ \bar{X}_i = \frac{1}{n} \sum_{i=1}^{n} X_i \]

where \( \bar{X}_i = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \)

\( Q = \lim_{n \to \infty} \sigma_i^2 \)

3. That is, the CLT generalizes in a more or less transparent way to the multivariate case.
V.D. Asymptotic Distributions

The reason we care about the CLT and limiting distributions is that we (more or less, speaking a bit loosely) "use" them to talk about the asymptotic distributions of statistics we care about.

\[ X_n \sim N(\mu, \sigma^2/n) \]

\[ Z=(X_n-\mu)/\sigma\sqrt{n} \sim N(0,1) \]

- \( Z \) is not exactly distributed \( N(0,1) \), but
  1. As \( n \) (sample size gets larger) it is more and more like it
  2. The closer \( X \) (i.e. the actual, exact distribution of \( X \)) is to normal to begin with, the sooner \& quicker the distribution of \( Z \) goes to \( N(0,1) \)

Consider, for example, the statistic:

\[ \frac{\hat{\beta} - \beta_0}{\text{S.E.}(\hat{\beta})} = t \]

i.e., the statistic we use to test the hypothesis \( H_0 : \beta = \beta_0 \). The CLT & Asymptotic Distribution theory tell us that if:
  a) \( \hat{\beta} \) is an unbiased estimator of \( \beta \)
  b) S.E. (\( \hat{\beta} \)) is a consistent estimator of \( \text{S.E.}(\beta) \), then

\[ t \sim N(0,1) \]

Now, if, as in fact we often will, we know a bit more about \( \beta \) & S.E. (\( \hat{\beta} \)) we might know their actual distribution (e.g., under OLS assumptions, \( \hat{\beta} \) is a S.N. \( N(\beta, \sigma^2/n) \) but we don't actually need anything more than "good" estimate of \( \beta \) \& S.E. (\( \hat{\beta} \)) + the CLT to get "asymptotic normality" for the "t" statistic.

VI. Maximum Likelihood Estimation

A. We now actually have more or less all the tools we need for Maximum Likelihood Estimation (MLE).

   1. We view outcomes as being produced by some p.f. or p.d.f. \( f(x; \theta) \)
   2. If each outcome is i.i.d. (i.e., is from a S.R.S.), then their joint distribution is the product of their marginal distributions:

\[ f(x_1, x_2, x_3, \ldots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) f(x_3; \theta) \ldots f(x_n; \theta) \]

   3. What we do now is choose \( \theta \) (i.e. estimate \( \theta \)) so that it maximizes the "probability" or likelihood of our having observed the data we actually have observed. (\( \hat{\theta} \) model \( \hat{f}(\theta) \))

B. That is, given the data, \( x \), define the (joint) conditional distribution of \( \theta \) given \( x \). Maximize this with respect to \( \theta \).

C. Examples:
   1. Each \( x_i \) ~ Poisson (\( \Theta \)) : \( f(x_i; \Theta) = \frac{e^{-\Theta} \Theta^{x_i}}{x_i!} \) (A Poisson Distribution describes a count (RV) of events happening at average rate \( \Theta \) per fixed period)
   a. \( \hat{\Theta} = \frac{1}{n} \sum_{i=1}^{n} x_i \)
   b. \( \hat{\Theta} = \text{Max} \sum_{i=1}^{n} \left[ x_i \ln(\Theta) - \Theta - \ln(x_i!) \right] \)

   C. Condition \( \frac{\partial}{\partial \Theta} \hat{L}(\Theta) = 0 \) (continued on page 11)
Maximizing Log-Likelihood of Poisson Model

\[ L(\theta) = \left(\frac{\theta}{e}\right)^n \frac{e^{-\theta}}{x!} \]

F.O.C.

\[ \frac{dL(\theta)}{d\theta} = n \left(\frac{1}{e}\right)^n \frac{e^{-\theta}}{x!} - \left(\frac{\theta}{e}\right)^n \frac{e^{-\theta}}{x!} \ln x \]

\[ \Rightarrow \theta^* = \frac{1}{n} \sum x_i = \bar{x} \]

Second-Order Condition:

\[ \theta > \frac{d^2 L(\theta)}{d\theta^2} \]

\[ \frac{d^2 L(\theta)}{d\theta^2} = -n \left(\frac{1}{e}\right)^n \frac{e^{-\theta}}{x!} \ln 2 \]

Thus, MLE for Poisson parameter \( \theta \), is \( \hat{\theta} = \bar{x} \).

MLE Example 2: MLE for Normal

\[ f(x, \theta) = \left(\frac{1}{2\pi \sigma^2}\right)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\[ L(\theta) = \prod_{i=1}^n \left(\frac{1}{2\pi \sigma^2}\right)^{-1/2} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \]

\[ \ln L(\theta) = \sum_{i=1}^n \left[ -\frac{1}{2} \ln (2\pi \sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} \right] \]

\[ \Rightarrow \sum_{i=1}^n \left[ -\frac{1}{2} \ln (2\pi \sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} \right] \]

\[ \Rightarrow \sum_{i=1}^n \left[ -\frac{1}{2} \ln (2\pi \sigma^2) - \frac{1}{2} (x_i-\mu)^2 \right] \]

\[ \Rightarrow \sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \mu)^2 \]

\[ \Rightarrow \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{n} \sum x_i = \bar{x} \]

\[ \Rightarrow \frac{1}{n} \sum (x_i - \bar{x})^2 = \sigma^2 \]

We note (in passing for now) that Maximum Likelihood Estimation is especially useful when we have some substantive model for the parameter \( \theta \), such as \( \theta = a + bX \) or something. It's especially useful when the probability function underlying the outcomes are linear.

Properties of Max Like Estimators:

1) Consistent (\( \lim_{n \to \infty} \hat{\theta}_n = \theta \))

2) Asymptotically Normal (\( \hat{\theta}_n \approx N(\theta, \sigma^2 \{H(\theta)\}) \))

3) Achieves Cramer-Rao Lower Bound for Consistent Estimators (i.e. is "best")

(eg. in normal example above, MLE for \( \sigma \) is square root of MLE for \( \sigma^2 \))
A. Some Preliminary Facts: We use a lot in Interval Estimation Hypothesis Testing:

1. As we know, in sampling from a normal distribution, with known mean \( \mu \) and variance \( \sigma^2 \),
   \[ X - \mu \sim N(0, \sigma^2) \]
   This is exact, not an approximate distribution.

2. When taking a sample average from a known \( N(\mu, \sigma^2) \):
   \[ \frac{X - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \]
   Also exact.

3. When sampling from an unknown distribution with known variance,
   \[ \frac{X - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \]
   Not exact, but converges to Normal.

4. When sampling from a normal distribution with unknown variance:
   \[ \frac{X - \mu}{S/\sqrt{n}} \sim t_{n-1} \]
   \[ \frac{(X - \mu)}{(S/\sqrt{n})} \sim t_{n-1} \overset{d}{\rightarrow} N(0, 1) \]

5. When sampling from an unknown distribution with unknown parameters
   \[ \frac{X - \mu}{S/\sqrt{n}} \overset{d}{\rightarrow} N(0, 1) \]
   \[ \frac{X - \mu}{S/\sqrt{n}} \sim N(0, 1) \]

6. \( V = \sum_{i=1}^{n} X_i^2 \) when each \( X_i \sim N(0, 1) \) \( \Rightarrow V \sim \chi^2_n \)
   \( \text{if each } X_i \sim N(0, 1) \text{ then } V \sim \chi^2_n \)

7. \( W = \frac{V_1}{V_2} \) where \( V_1 \sim \chi^2_{n_1}, V_2 \sim \chi^2_{n_2} \)
   \( \Rightarrow W \sim F_{n_1, n_2} \)
   \( \text{if each } V \sim \chi^2_n, \text{ then } W \sim F_{n_1, n_2} \)

   \( \text{e.g.: } R^2 = 1 - \frac{\text{ms error}}{\text{ms total}} \)
   \( \overset{\text{under certain conditions, } R^2 \sim F}{} \)

B. Constructing Confidence Intervals:

1. We start with some estimate (statistic) which we "standardize" or "normalize" (i.e. subtract mean and divide by standard deviation):
   \[ \text{estimate} \rightarrow X \rightarrow \left( \frac{X - \mu}{S/\sqrt{n}} \right) \rightarrow \text{standardization} \]
   S.D. of \( X \) (estimated)

2. Then we ask what "bounds" would contain this with some probability, given its distribution (here \( t_{n-1} \) by (4) above); i.e., we look for \( z \) such that satisfy
   \[ \text{Prob} \left( -z < \frac{X - \mu}{S/\sqrt{n}} < z \right) = (1 - \alpha) \]
VII. B. Constructing Confidence Intervals (C.I.'s)

2. \( \text{Prob}\left(-Z < \frac{x - \mu}{\sigma/\sqrt{n}} < Z\right) = (1 - \alpha) \) 
   \( \Rightarrow \) confidence level, e.g. \( \alpha = 0.10 \)
   \( \Rightarrow 90\% \) confidence interval

3. More usually, we are interested in an interval anchored on our estimate \( \overline{x} \) here, i.e., going \( \pm t \) on either side of that so that the interval might encompass the parameter we're trying to estimate.
   So we can rearrange (2.):
   \[ \text{Prob}\left(\overline{x} - \frac{ts}{\sqrt{n}} < \mu < \overline{x} + \frac{ts}{\sqrt{n}}\right) = (1 - \alpha) \]

4. Notes:
   a. We're implicitly assuming that we won't a symmetric confidence interval.
      i) if distribution at the statistic is symmetric, then the symmetric C.I. is also the smallest C.I. from lower to upper bound.
      ii) if statistic is not symmetric, distributed may be reasonable to get the smallest bound-range instead; this is complicated (rarely done because of that, but we may try it later).

5. Mechanics (Example)
   \( \overline{x} \) taken from \( x_i \sim \text{i.i.d.} \ N(\mu, \sigma^2) \) \( i = \{1, \ldots, 10\} \)
   \( \overline{x} = 0.02 \), \( s = 0.1 \) from our sample estimates.
   \[ \text{Pr}\left(\overline{x} - \frac{ts}{\sqrt{n}} < \mu < \overline{x} + \frac{ts}{\sqrt{n}}\right) = (1 - \alpha) \]
   \[ \text{Pr}\left(0.02 - \frac{0.1}{10} < \mu < 0.02 + \frac{0.1}{10}\right) = (1 - \alpha) \]

6. Final Note: Often when the distribution of the sample statistic is asymptotically normal, we use the \( T \) distribution rather than \( Z \) right to using the normal as an approximation. The \( T \) is clearly more conservative than \( N(0,1) \), but we can't actually say this is warranted nonetheless, except in the limit (asymptotically).
VII.C. Hypothesis Testing

(First we will stress the formal, rigid process of hypothesis testing, its interpretation; Then, throughout the course, I will press a more intuitive approach to interpretation of hypothesis test results.)

1. The broadly described steps:
   a. Formally, the testing procedure begins by specifying a null hypothesis, $H_0$, (i.e., the default hypothesis is usually the thing we would like to "disprove", and an alternative hypothesis, $H_1$ (the thing we would like to "prove") or $H_2$.

   b. Then a "size", or "α-level", or "p-level" for the test is pre-determined.

   $H_0: \theta = \theta_0$

   $H_1: \theta \neq \theta_0$

   c. Then a test statistic is computed. The test-stat, in some transformation of the estimate of the parameter, is distributed in such a way that we know or can approximate under the condition that the null is true. The test-stat is then compared to this distribution against the pre-determined $\alpha$.

2. Type I & Type II Errors, Size, Power, Bias & Consistency of Tests

   a. Type I Error: Rejection of null hypothesis when it is, in fact, true.

   b. Type II Error: Failure to reject the null hypothesis, when it is, in fact, false.

   (Type II Error: confusing a Type I with a Type II error.)

   c. The probability of a Type II error is the size $(\alpha$, level, p-level, significance level) of a test.

   d. The probability that a test will correctly reject the null, i.e., $1 - P(\text{Type II Error})$, is the power of a test.

   e. In general, if we are using all available info, correctly constructed the test to make best use of this information 

   f. Example: if the criminal justice system is efficiently using available evidence, there is no way to make conviction of a criminal more likely without also making conviction of the innocent more likely. Or, v.v., there is no way to make conviction of the innocent less likely without also making conviction of guilt less likely.

   Similarly, for fiscal efficiency->bearing changes in efficiency, Miranda Rights, for example, lead to more innocent being free and more guilty going free.

   g. A test is Uniformly Most Powerful if it has higher power than any other test of the same size for any admissible value of the true parameter.

   h. A test is consistent when its power $\rightarrow 1$ as the sample size increases. I.e., as sample size $n \rightarrow \infty$, the test always rejects a false null.

   i. In general, if the estimate being used in the test is consistent, then the test is consistent. Why? Consistent estimate $\hat{\Theta} \Rightarrow \lim \hat{\Theta} = \Theta$, so in the limit $\hat{\Theta}$ converges to the true parameter and we can see directly whether $H_0$ is true or not.
Mechanics of Conducting Hypothesis Tests:

1. Null Hypothesis: $H_0: \theta = \theta_0$
   - Example: $H_0: \beta = 0$

2. Alternative Hypothesis: $H_1: \theta \neq \theta_0$
   - Example: $H_1: \beta \neq 0$

b. Convert estimate, $\hat{\theta}$, to standardized form as in confidence intervals.
   - $T = \left\{ \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})} \right\} \sim \mathcal{N}(0, 1)$
   - Example: $T = \frac{\hat{\beta} - \beta}{\text{se}(\hat{\beta})} \sim \mathcal{N}(0, 1)$

   or $T = \frac{\hat{\theta}}{\text{se}(\hat{\theta})}$

C. Check that test statistic, $T$, against its distribution or approximate distribution, and record the probability of $T$ being more extreme (greater in absolute value) than your actual $T$.

   - Example: $T = 1.76$, distribution $t_{1000}$
   - Prob (1.76 ≤ $t_{1000}$) = 0.05

   - This is your p-level. It is the probability of having observed a test statistic, here $T$, this far or farther from the null, here $\theta$, if the null were in fact true.

   - If the p-level is ≤ the predetermined size or α-level of the test, we reject the null (in favor of the alternative).

D. Interpretational Note:
   - The p-level does not tell us the probability the null is true (in non-Bayesian analysis). The null is either true (with probability 1) or not true (with probability 1). The p-level tells us the probability we would have observed a test statistic far or farther from the null if the null were true. From this we infer that the null was likely untrue: or not; the facts, however, are that the null either is or is not true (with p=1).

E. Rob's Rules: The standard procedure is to set the size of the test, α-level, to .10 or .05 or .01. If then select anything that makes this or lower p-level. That's fine, but with modern tech, it's simple to know the 'exact' p-level. This is more info, we have the info, therefore, conditional distribution, give it, assumptions/approximations

F. The Broad (Better by Rob's Rules) Interpretation:
   - The question is, "Is the estimated quantity, e.g., the relationship between X and Y, far from the null, e.g., 0.5?"
   - "Far" is defined in standard deviation units, how "significant" any increment of. For ex. is, how far has the straining of credibility given the evidence, i.e., how "long" a standard deviation is, is measured in increments of likelihood.

   - Thus, p-levels tell us just how far we're straining credibility given the evidence of what has happened; if we cling to the null.

   - The lower the p-level, the harder is the stretch of imagination required to continue believing the null into swallow.
VII. D. Three Common Types of Tests in Regression Analysis.

1. Think of hypotheses regarding (a) parameter(s), \( \Theta \), such as
\[
H_0: c(\Theta) = 0
\]
\[
H_1: c(\Theta) \neq 0
\]

2. Likelihood-Ratio (LR) Tests:

(a) If \( c(\Theta) = q \) is true, then imposing it to begin with prior to estimation should make little difference to the likelihood of the data (or fit of the regression),
\[
L(\Theta) = L(\Theta) \quad \text{imposed}
\]

(b) \( \therefore \) we ought to be able to base a test on
\[
\ln L(\Theta) - \ln L_R(\Theta) \quad \text{just applying ln rules}
\]

(c) In fact, under certain conditions, we won't get into
\[
\ln L(\Theta) - \ln L_R(\Theta) \sim \chi^2_n
\]

(d) So, we can reject \( c(\Theta) = q \) if LR larger than \( \chi^2_n \) for some predetermined \( \alpha \)
    \* or we can report the p-level of the test as before.

3. Wald Tests:

(a) If \( c(\Theta) = q \) true, then \( c(\Theta) \) should not be too far from \( q \)

(b) \( \therefore \) we ought to be able to find a test based on \( c(\Theta) - q \)

(c) In fact, under certain conditions, we won't get into norm
\[
W = (c(\hat{\Theta}) - q)^T \left( \text{Var}(c(\hat{\Theta}) - q) \right)^{-1} (c(\hat{\Theta}) - q)
\]
\[
W \sim \chi^2_n \quad \text{for} \quad n = \text{number of conditions given by} \quad c(\Theta) = q
\]

(d) Reject or not, or report p-level from \( \chi^2_n \) exactly as before.

4. Lagrange Multiplier Tests:

(a) If \( c(\Theta) = q \) true, then imposing constraint should not have much effect in our optimizing the likelihood.
    In particular, the first-order condition that
\[
\frac{\partial L(\Theta)}{\partial \Theta} = 0
\]
should not be "far" from holding
\[
\frac{\partial \ln L(\Theta)}{\partial \Theta} = 0
\]

(b) \( \therefore \) should be able to base a test on
\[
L_M = \left[ \frac{\partial \ln L(\Theta)}{\partial \Theta} \right] \left[ I(\hat{\Theta}) \right]^{-1} \left[ \frac{\partial \ln L(\Theta)}{\partial \Theta} \right]
\]
\[
L_M \sim \chi^2_n \quad \text{number of conditions again}
\]

(c) In fact, under conditions we're not going into, etc.

5. We will here more of these later, but be sure you get the (a)(b) parts of each type of test.
\[ E(\mathbf{X}) = \begin{bmatrix} E(X_1) & \cdots & E(X_n) \end{bmatrix} \]

\[ V(\mathbf{\varepsilon}) = E[(\mathbf{\varepsilon} - \mathbf{\mu})(\mathbf{\varepsilon} - \mathbf{\mu})'] = \begin{bmatrix} \sigma^2 & \epsilon_{12} & \cdots & \epsilon_{1n} \\
\epsilon_{12} & \sigma^2 & \cdots & \epsilon_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\epsilon_{1n} & \epsilon_{2n} & \cdots & \sigma^2 \end{bmatrix} \]

\[ = \sum_{\mathbf{\varepsilon} \in \Omega} \mathbf{\varepsilon} \mathbf{\varepsilon}' \]

\[ = \sum_{\mathbf{\varepsilon} \in \Omega} \mathbf{\varepsilon} \mathbf{\varepsilon}' \]

\[ = \mathbf{\Gamma} \]

\[ = \mathbf{O} = \begin{bmatrix} V(\varepsilon_1) & \cdots & C(\varepsilon_1, \varepsilon_n) \\
\vdots & \ddots & \vdots \\
C(\varepsilon_1, \varepsilon_n) & \cdots & V(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & \epsilon_{12} & \cdots & \epsilon_{1n} \\
\epsilon_{12} & \sigma^2 & \cdots & \epsilon_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\epsilon_{1n} & \epsilon_{2n} & \cdots & \sigma^2 \end{bmatrix} \]

Special Case: If \( \mathbf{X} = (X_1, X_2, \ldots, X_n) \) are \( \mathbf{\varepsilon} \)-multivariate normal, then linear \( \Rightarrow \) mean \( \Rightarrow \) stochastic.
1. Properties of Expectations, (Co)Variances, and Conditioning

A. Expectations

1) Laws of Iterated Expectations
- \( X, Y, Z \) = R.V.s
- \( a, b, c \) = constants

a) \( E(E(X)) = E(X) \)
b) \( E(E(Y|X)) = E(Y) \)

2) Expectation of a sum is sum of the expectations

\[ E(\sum_{i=1}^{n} g(x)) = \sum_{i=1}^{n} E(g(x)) \]

3) Expectation of a constant is the constant

a) \( E(a) = a \)
b) \( E(X|X) = X \)

4) Rob's Rule of Expectations:

"\( E \) slides on through constants, snagging around random variables"

a) \( E(a + bX + cY) = a + bE(X) + cE(Y) \)
b) \( E(XX' - X[Ee']X(X'X)^{-1}) = (X'X)^{-1}X'E[e'e']X(X'X)^{-1} \)

B. Conditioning

1) Conditioning is distributive:

\[ E(a + bX + cY|X) = E(a|X) + bE(X|X) + cE(Y|X) = a + bX + cE(Y|X) \]

2) Conditioning More than Once is Redundant:

\[ E[E(Y|X)|X] = E(Y|X) \] (on same info!)

3) Conditioning on \( X \) renders \( X \) constant (see A3b)

4) Conditioning & Expectation sort of undo each other (see A3b)
© (Co)Variances:

1) Rob's Rules of (Co)Variances:
   a) Constants neither vary nor co-vary.
      \[ V(c) = 0 \quad C(c, X) = 0 \]
   b) "Variance slides through sums, eliminating constants, smashing on random variables
      and squaring their coefficients, while splitting out 2 times all the covariances
      times their associated coefficients."
      \[
      V(a + bX + cY) = b^2 V(X) + c^2 V(Y) + 2bc Cov(X, Y)
      \]
      Notice that in Matrix notation, with \( X \) a constant vector and \( b \) a random vector, this becomes just:
      \[
      V(X^T b) = X^T V(b) X
      \]
      "X squared" "in matrix land"
   c) The Analogous for Covariances becomes a kind of FOIL:
      \[
      Cov(a + bX + cY, d + eW + fV) = \]
      just be Cov(X, W) + bf Cov(X, V) + ce Cov(Y, W) + de Cov(Y, Y)

(Selected)

D) Properties of Note Combining the Above:

1) Any R.V. \( Y \) can be written: \( Y = E(Y | X) + (Y - E(Y | X)) \)

2) Properties that follow:
   a) \( \text{Cov}(X, Y) = \text{Cov}(X, E(Y | X)) \)
   b) \( V(Y) = E(V(Y | X)) + V(E(Y | X)) \)
      \[
      = \sigma^2 + V(Y) \quad \text{error, regression}\n      \Rightarrow TSS = RSS + ESS \quad \text{explained, residual}\n      \]
      "variance decomposition"

Note: \( \Rightarrow V(Y) \leq V(Y) \) (and a so \( + V(\hat{Y}) \leq V(Y) \) though that less notable)
I.D.2. (cont) Properties that follow from \( E(Y) = E(Y|X) + (Y - E(Y|X)) \)

c) \( E(XY) = E(X \cdot E(Y|X)) \)
d) \( E(\epsilon) = E(\epsilon|X) = E(\epsilon|Y) = 0 \)
e) \( E(\epsilon) = E(Y - E(Y|X)) = E(Y) - E(E(Y|X)) = E(Y) - E(Y) = 0 \)

\[ E(\epsilon|X) = E(Y - E(Y|X)|X) = E(Y|X) - E(E(Y|X)|X) = E(Y|X) - E(Y|X) = 0 \]

\[ E(\epsilon|Y) = E(Y - E(Y|X)|Y) = E(Y|Y) - E(E(Y|X)|Y) = Y - Y = 0 \]

f) \( \text{Cov}(X, \epsilon) = 0 \); in fact, \( \text{Cov}(h(X), \epsilon) = 0 \forall h(h) \)

\[ E(X\epsilon) - E(X)E(\epsilon) = E(X(Y - E(Y|X))) = E(X)E(\epsilon) \]
\[ = E(X\epsilon) = 0 \]

I.D.3. For Special Case where \( E(Y|X) \) is linear

\( E(Y|X) = a + bX \)

a) \( a = E(Y) - bE(X) \)

b) \( b = \frac{\text{Cov}(X,Y)}{\text{V}(X)} \)

proof: \( \text{Cov}(X,Y) = \text{Cov}(X, E(Y|X)) \)
\[ = \text{Cov}(X, a + bX) \]
\[ = b \text{Cov}(X,Y) \]
\[ \Rightarrow b = \frac{\text{Cov}(X,Y)}{\text{V}(X)} \]

c) \( \text{V}(Y) = \text{V}(E(Y|X)) + \text{V}(E(Y|X|X)) \)
\[ = E(\text{V}(Y - a - bX)) + V(a + bX) \]
\[ = E(V(E(X))) + b^2 \text{V}(X) \]

d) \( \rho_{xy} = \sqrt{\frac{b_{xy}}{\text{V}(X) \text{V}(X)}} = \sqrt{\frac{\text{Cov}(X,Y)^2}{\text{V}(X) \text{V}(X)}} = \frac{\text{Cov}(X,Y)^2}{\text{V}(X) \text{V}(X)} \)

\[ \text{V}(Y) = \text{b} \text{sd}(X) \text{sd}(\bar{Y}) \]

e) \( E(Y) = a + bE(X) \); i.e., \( \bar{Y} = a + b \bar{X} \)

f) "Coefficient of Determination" \( R^2 \equiv \frac{\text{Cov}(X,Y)^2}{\text{sd}(X) \text{sd}(\bar{Y})} \)
Ch. 24: The Classical (Normal) Linear Regression Model
I. Preliminaries: Recall Greene Ups (King 1-4, esp. 2) is here "intro" to MLE perspective, Karlton, Cipriani, et al.
A. Theory: for example, Political Science theory, most typically makes statements regarding observable relationships between variables. It is common in science, and all but the universal norm in social science, that these theories describe not exact, deterministic, relationships, but rather probabilistic, random, or stochastic (these are synonyms in this context) relationships.

1. i.e., we theorize: \[ y = f(X, \varepsilon) \]

   a. \( y \) is a variable (vector or matrix of variables) which is the output of some process about which we are theorizing: The Dependent Variable(s).

   b. \( f(\cdot) \) is the function which describes the theory, i.e., it tells us the theorized (hypothetical) process producing outcomes \( y \): The functional form.

   c. \( X \) is some (set of) variable(s) which are hypothesized to be related to \( y \), i.e., to "produce" \( y \) in some manner described by \( f(\cdot) \). \( X \) is not, in general, stochastic or random in any consequential way, thus the part of the process defined by \( Y = f(X, \varepsilon) \) is generally the systematic part of the world: The Independent Variable(s).

   d. \( \varepsilon \) is some (set of) random variable(s) which describe the non-systematic, random, or stochastic (synonyms) part of the process \( f(\cdot) \) producing \( y \).

2. Examples: \[ y = f(X, \varepsilon) \]

   \[ \text{(vote-share of incumbent)} = f(\text{the socioeconomic world, competitive characteristics, chance}) \]

   \[ \text{(duration of Parliamentary Govt)} = g(\text{number parties in govt, polarization of parties in govt, scandals, chance}) \]

B. "A Language of Inference" (King, Unifying Political Methodology, 1-2)
1. Social System: The object under study (e.g., a neighborhood, a government, a market, a ward, etc.)
   a. It's what the researcher wishes to know about
   b. Contains features that are known or estimable features which are unobservable & some unobservable

2. Explanatory Variables: measures on features of the social system which are hypothesized to be related to (to produce) the output(s) of interest

3. Output (Dependent Variable(s)): consequences of the social system which can be observed and measured (anything that can be observed, can be measured at least as present (1) not-present (0)).

4. Random Variables: operationally, an assignment of numbers to events or features (e.g., 1 = Republican wins, 0 = Democrat wins)
5. **Experiment** when parts of the social system are under control of the researcher and manipulated, to observe their effects on the output. This is generally not possible in Social Science, but we frequently consider hypothetical experiments where they are theoretically possible.

6. **Model:** a model is a (mathematical or logical argumentative) simplification of an approximation to a more complex system or social system. It is not even intended to replicate the social system in all its detail but to focus on certain aspects.

7. **Statistical Model:** a formal representation of the process by which the social system produces output (S). They have systematic and stochastic components.

II. **The Classical (Normal) Linear Regression Model** is then a particular example of the broader class of statistical models.

**A. Standard Econometric Approach** vs. (Identical) Standard Statistical Approach

1. The Model
   a) $Y_i = X_i \beta + \varepsilon_i$
   
   (linearity)
   b) $\varepsilon_i \sim N(0, \sigma^2)$
   
   (normal distribution)

2. Zero Mean of Stochastic Component
   
   $E(\varepsilon_i) = 0 \iff E(Y_i - X_i \beta) = 0$

3. Homoskedasticity (Constant Variance) of Stochastic Component
   
   $V(\varepsilon_i) = \sigma^2 \iff V(Y_i | X_i) = \sigma^2$

4. Non-correlation of Stochastic Component (Weak Independence)
   
   $Cov(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j \iff Cov(Y_i, Y_j | X_i, X_j) = 0$

5. Uncorrelatedness of Regressors and Stochastic Component (Regressors Dependent Variable are, Parametrically, but not Stochastically Dependent)
   
   $Cov(X_i, \varepsilon_j) = 0 \forall i \neq j \iff \beta_j$ depends on $X_i$ only through its mean, determined by $X_i \beta$

6. Nonstochastic Regressors:
   
   $X_i$ are fixed in repeated samples, $0 < \gamma^2(x, -x)^2 < \infty$

7. Normality:
   
   $\varepsilon_i \sim N(0, \sigma^2) \quad \beta \sim N(0, \sigma^2)$

**B. The Linearity Assumption** (The model is linear in its parameters ($\beta$))

1. Non-linear in Variables (Examples: Non-linearity in $X = \exp X$)
   
   $y_i = a \exp b x_i + \varepsilon_i \iff y_i = a \exp (b x_i + \varepsilon_i)$

   i. => Really considerably more flexible than it first appears

   ii. => Even when "true" not linear, we get "best" linear approximation (BLS) instead

2. Non-linear in Parameter(s) (Examples: Non-linear in $\beta = \exp$ non-linearly)
   
   $y_i = \frac{\alpha}{\beta + x_i} + \varepsilon_i \iff y_i = (1 + e^{-x_i})^{-\frac{\alpha}{\beta + x_i}} [1 - (1 + e^{-x_i})^{-\frac{\alpha}{\beta + x_i}}]^{-\frac{\alpha}{\beta + x_i}}$
8.2. Elasticity Models: an important special case

a. Consider: \( \ln y_i = \beta \ln x_i + \varepsilon_i \) (assume for the moment that the intercept is zero)

b. What is "effect" of \( x_0 \) on \( y_i \)? (2. total differential: I neglected to go into these in our calculus course)

\[
\frac{d \ln y_i}{d x_i} = \beta \frac{d \ln x_i}{d x_i}
\]

\[
\Rightarrow \frac{d y_i}{y_i} = \beta \frac{1}{x_i}
\]

\[
\beta = \frac{d y_i / y_i}{d x_i / x_i}
\]

- Which in discrete notation/terms/form is like

\[
\beta = \frac{\Delta y_i / y_i}{\Delta x_i / x_i} \quad \text{% change of } y
\]

\[
\text{% change of } x
\]

- Thus, in a log-log model, \( \beta \) is the % change in \( y \) brought about by a 1% change in \( x \). In economics this is called the "elasticity of \( y \) with respect to \( x \)."

- More generally, the concept of things \( (x, y) \) being related in terms of % changes rather than absolute changes (i.e., proportional changes) is an important option to consider. Which does your theory imply holds about \( x \& y \)? (If either)

C. Zero-Mean For Stochastic Component: \( (E(\varepsilon_i) = 0 \) or \( E(y - x, \varepsilon) = 0 \)

1. First of all, the stochastic component is supposed to be non-systematic.

- E.g., you know nothing about \( \varepsilon \) beyond possibly the class of probability function it is drawn from. If \( E(\varepsilon_i) \neq 0 \), then you know something put it in the systematic part.

2. Second of all, we do this in practice anyway! We model

\[
y_i = \beta x_i + \varepsilon_i \]

- Suppose \( E(\varepsilon_i) \neq 0 \), then we could write \( \xi = \varepsilon_i + E(\varepsilon_i) \) with

\[
\xi \sim E(\xi) = 0
\]

\( \Rightarrow \) assumption of zero mean for \( \varepsilon \) utterly means that either

- a) we are willing to assume intercept is zero & evaluate everything given that assumption, or

- b) we include a constant (intercept) \( \beta \) that "soaks up" \( E(\varepsilon) \)

\( \Rightarrow \) This assumption is of little consequence unless we are particularly interested in \( \beta \) as opposed to \( \beta + E(\varepsilon_i) \) for some reason. We rarely if ever are.

D. Homoskedasticity: Uncorrelated Errors: these assumptions have more bite if we will spend considerable time on their violation.

1. Homoskedasticity: \( \sigma(y_i | x_i) = \sigma(\varepsilon) = \sigma^2 \) i.e. the variance of the distribution from which your observations are drawn is constant.
2. Non-Correlation:

-> basic notion: knowing any $E_i$ wouldn't not help us predict what some other $E_j$ might be.

I.e., given $X_i$, there is no further information in

$E_i = E(y_i|X_i) = y_i - x_i \beta = E_i$,

which would help us predict $E_j$ (some other observations residual).

Typical Violations:
- Observations are ordered in time if something is "sticky," so that if $E_i$ is large then $E_{i+1}$ is more likely to be large, etc. (serial correlation)
- Observations are ordered in space (e.g., geographic proximity of observational units so that if $E_i$ is large then $E_j$ from an adjacent unit is likely to be large or small (contagion)

E. Uncorrelatedness of $X \& E$ (Parametric but not Stochastic Dependence of $Y$ on $X$

1. The 2nd Version of this says that $f(y_i|x_i)$ depends on $X_i$ only in that $E(y_i|x_i) = x_i \beta$. That is, only the mean of $y_i$ depends on $x_i$; the rest of the shape of $f(y_i|x_i)$ (see pictures above) does not depend on $x_i$. i.e., it is constant across $X_i$. This includes, for example, that $V(y_i|x_i) = \sigma^2$ which is no masked elasticity. It is more general though: none of the rest of the distribution should depend on $X_i$ either.

2. Basically, if you have the causality right, you'll be OK here.

I.e., high $X$ cannot be associated with high $E$ through $Y_i$.

Most typical Violations:
- $Y_i \rightarrow X_i \& X_i \rightarrow Y_i$ ("endogeneity")
- "simultaneity"
- "reverse causality"

which leads to
Illustration of Mutual Causality, \( y \Leftrightarrow x \), implying Covariance of Regressor with Residual

\[
\begin{align*}
y &= \beta x + \gamma z + \varepsilon_y \\
x &= \theta y + \lambda w + \varepsilon_x
\end{align*}
\] 

\[
\begin{align*}
\text{Cov}(x, \varepsilon_y) &= \text{Cov}(\varepsilon_y, \theta y + \lambda w + \varepsilon_x) = \text{Cov}(\varepsilon_y, \theta y) \\
&= \text{Cov}(\varepsilon_y, \theta(\beta x + \gamma z + \varepsilon_y)) = \text{Cov}(\varepsilon_y, \theta \varepsilon_y) = \theta \text{Var}(\varepsilon_y) \\
\text{Cov}(y, \varepsilon_x) &= \text{Cov}(\varepsilon_x, \beta x + \gamma z + \varepsilon_y) = \text{Cov}(\varepsilon_x, \beta x) \\
&= \text{Cov}(\varepsilon_x, \beta \varepsilon_x) = \beta \text{Var}(\varepsilon_y)
\end{align*}
\]
\[ y_i \sim N(\mu_i, \sigma^2) \quad ; \quad \mu_i = \beta x_i \]

\[ f(y_i; \mu_i, \sigma^2) = \left(\frac{1}{2\pi \sigma^2}\right)^{\frac{n}{2}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}} \]

\[ L(\mu_i; \sigma^2 | x) = \frac{1}{\Gamma(n/2)} \left(\frac{2\pi \sigma^2}{2}\right)^{\frac{n}{2}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}} \]

\[ \ln L(\mu_i; \sigma^2 | x) = -\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu_i)^2 / \sigma^2 \]

\[ = -\frac{n}{2} \ln 2 - \frac{n}{2} \ln \pi - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 (\sigma^2)^{-1} \]

\[ \frac{\partial \ln L}{\partial \beta} = -\left(\sigma^2\right)^{-1} \sum_{i=1}^{n} (y_i - \beta x_i) x_i = 0 \]

\[ \sum_{i=1}^{n} (y_i - x_i \hat{\beta}) x_i = 0 \]

\[ \Sigma \hat{\beta} x_i^2 = \Sigma y_i x_i \]

\[ \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \]

\[ \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2} \left(\sigma^2\right)^{-2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 = 0 \]

\[ -n + \left(\sigma^2\right)^{-1} \sum_{i=1}^{n} (y_i - \beta x_i)^2 = 0 \]

\[ \left(\sigma^2\right)^{-1} \sum_{i=1}^{n} (y_i - \beta x_i)^2 = n \]

\[ \frac{n}{\sigma^2} = \sum_{i=1}^{n} (y_i - \beta x_i)^2 \]

\[ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \]

\[ \frac{\sum_{i=1}^{n} (y_i - \beta x_i)^2}{\sigma^2} < 0 \]
A. Least-Squares Estimation

Let's consider estimating \( \alpha \) and \( \beta \) such that \( y_i = \alpha + \beta x_i + e_i \).

Perhaps, following from prior weeks, we could consider merely "fitting" the line \( \alpha + \beta x_i \) so that we minimize the total (sum) distance from the observed \( y_i \) to that fitted line. This is equivalent to minimizing the sum of squared \( e_i \) with respect to \( \alpha \) and \( \beta \).

1. i.e., \( \min_{\alpha, \beta} \sum e_i^2 \) \( \equiv \min_{\alpha, \beta} \sum \varepsilon_i^2 \) \( \equiv \min_{\alpha, \beta} (y_i - x_i \beta)'(y_i - x_i \beta) \)

Matrix of independent variables, here:

\[
X = \begin{bmatrix}
1 & x_1 & x_2 & \cdots & x_n \\
1 & x_1 & x_2 & \cdots & x_n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_1 & x_2 & \cdots & x_n \\
\end{bmatrix}
\]

Column vector of \( n \) observations on \( X \)

Column vector of \( n \) "ones" = "the constant"

\( \Rightarrow \) \( \beta \) = "the intercept"

\( \Rightarrow \) First-order Condition:

\[
\frac{\partial \sum e_i^2}{\partial \beta} = -x'Y - x'Y + 2x'XB = 0
\]

\[
\Rightarrow 2(x'X)B = 2x'Y
\]

\[
B = (x'X)^{-1}(x'Y)
\]

2. Scalar Equivalent:

\[
\min_{\alpha, \beta} \sum e_i^2 = \min_{\alpha, \beta} \sum (y_i - \alpha - \beta x_i)^2
\]

F.o.C.:

\[
\frac{\partial \sum e_i^2}{\partial \beta} = \sum_{i=1}^{n} 2 (y_i - \alpha - \beta x_i) (-x_i) = 0
\]

\[
-2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(x_i) = 0
\]

\[
\Rightarrow \sum_{i=1}^{n} e_i x_i = 0
\]

\[
\Rightarrow \sum_{i=1}^{n} e_i x_i = 0
\]

\[
\sum_{i=1}^{n} e_i = 0
\]
III. A. 2 Scalar First-Order Conditions for Least Squares

\[ \frac{\partial SSE}{\partial b} = 0 \quad \Rightarrow \quad \sum \left( y_i - a - bx_i \right) = 0 \]

\[ \frac{\partial SSE}{\partial a} = 0 \quad \Rightarrow \quad \sum y_i = n a + b \sum x_i \]

\[ \frac{\partial}{\partial a} \sum \left( y_i - a - bx_i \right) = 0 \quad \Rightarrow \quad \sum y_i = na + b \sum x_i \]

\[ \frac{\partial}{\partial b} \sum \left( y_i - a - bx_i \right) = 0 \quad \Rightarrow \quad \sum x_i y_i = a \sum x_i + b \sum x_i^2 \]

(a) and (b) are the so-called "normal equations," but this is entirely misleading so forget Greene told you this, just continue to solve.

*dividing both sides of (a) by \( n \) \( \Rightarrow \quad \bar{y} = a + b \bar{x} \)

**Conclusion 1:** Least-Squares Regression Line \( \hat{y} = a + b \bar{x} \) passes through the point \((\bar{x}, \bar{y})\)

**Conclusion 2:** Least-Squares Estimate of \( a = \bar{y} - b \bar{x} \)

*Using conclusion 2 in equation (b)*

\[ \sum x_i y_i = (\bar{y} - b \bar{x}) \sum x_i + b \sum x_i^2 \]

Note: \( \sum x_i = n \bar{x} \)  
(why? \( \bar{x} \equiv \frac{1}{n} \sum x_i \), so \( n \bar{x} = n \cdot \frac{1}{n} \sum x_i = \sum x_i \))

\[ \Rightarrow \sum x_i y_i = (\bar{y} - b \bar{x}) n \bar{x} + b \sum x_i^2 \]

\[ \Rightarrow \sum x_i y_i = \bar{y} n \bar{x} - b n \bar{x}^2 + b \sum x_i^2 \]

\[ \Rightarrow \sum x_i y_i - n \bar{x} \bar{y} = b (\sum x_i^2 - n \bar{x}^2) \]

(iii) \( b = \frac{\sum (x_i y_i - n \bar{x} \bar{y})}{\sum (x_i^2 - \bar{x}^2)} \)

**Conclusion 3:** \( b = \frac{\sum (x_i \bar{y} - \bar{x} \bar{y})}{\sum (x_i^2 - \bar{x}^2)} = \frac{S_{xy}}{S_{xx}} \)

Aside: How do we get from (ii) to (iii)

\[ \sum (x_i^2 - \bar{x}^2)(y_i - \bar{y})^2 = \sum (x_i y_i - \bar{x} \bar{y})^2 \]

\[ = \sum (x_i y_i - \bar{x} \bar{y})^2 \]

\[ = \sum (x_i y_i^2 - \bar{x} \bar{y} y_i - n \bar{x} \bar{y} y_i + n \bar{x} \bar{y}) \]

\[ = \sum (x_i y_i^2 - n \bar{x} \bar{y} - n \bar{x} \bar{y} x + n \bar{x} \bar{y} \bar{x}) \]

\[ = \sum (x_i y_i^2 - n \bar{x} \bar{y}) \]
III. B. Given Assumptions 1-6: (linearity, \(E(e)=0\), \(\text{Cov}(e_i, e_j)=0\), \(\text{Var}(e)=\sigma^2\)) and no endogeneity \(x^*\)'s fixed

**Statistical Properties of the LS Estimator**

\[
b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}
\]

where \(c_i = \frac{x_i - \bar{x}}{S_{xx}}\)

\[
= \frac{\sum (x_i - \bar{x})(\alpha + \beta x_i + \varepsilon_i - \bar{y})}{S_{xx}}
\]

\[
= -\frac{\alpha \Sigma (x_i - \bar{x})}{S_{xx}} + \frac{\beta \Sigma (x_i - \bar{x})^2}{S_{xx}} + \frac{\Sigma (x_i - \bar{x}) \varepsilon_i^2}{S_{xx}}
\]

so \(b\) differs from true parameter \(\beta\) because \(\varepsilon_i\) are stochastic.

**Property 1:** \(E(b) = E(\beta + \sum \frac{n}{S_{xx}} C_i \varepsilon_i)\)

\(\Rightarrow b = \beta + \sum \frac{n}{S_{xx}} C_i \varepsilon_i\)

\(b_{LS}\) is an unbiased estimate of \(\beta\)

**Property 2:** \(\text{Var}(b_{LS}) = \text{Var}(\beta + \sum \frac{n}{S_{xx}} C_i \varepsilon_i)\)

\(\Rightarrow \text{Var}(b) = \text{Var}(\beta) + \sum \text{Cov}(\beta, \frac{n}{S_{xx}} C_i \varepsilon_i) + \sum \text{Var}(\frac{n}{S_{xx}} C_i \varepsilon_i)\)

\(\Rightarrow \text{Var}(b) = 0\)

\[
\text{Var}(b_{LS}) = \frac{\sigma^2}{S_{xx}}
\]

\(\text{Cov}(\varepsilon_i, \varepsilon_j) = 0\) (by GLM assumptions)

\(\text{Var}(\varepsilon_i) = \frac{\Sigma C_i^2 \sigma^2}{(S_{xx})^2}\)

\(\text{Var}(\varepsilon_i) = \frac{\Sigma C_i^2 S_{xx}}{(S_{xx})^2}\)
III.8.3. Given Assumptions 1-6, we have established that

\[ \beta \] \iff \text{unbiased est. of } \beta

\[ \sigma^2_{\hat{y}} \]

Now, \( E(a_{LS}) = \alpha \) \iff \( a \) is also an unbiased est. of \( \alpha \).

\text{Proof: } E(a_{LS}) = E(\hat{y} - b\bar{x}) \quad \text{by def. of } a_{LS}

\[ = E\left(\frac{1}{n} \sum_{i=1}^{n} y_i - b\bar{x}\right) \quad \text{by def. of } \bar{x}

\[ = E\left(\frac{1}{n} \sum_{i=1}^{n} (\alpha + \beta x_i + \epsilon_i) - b\bar{x}\right) \quad \text{by } a_{LS} \text{ est. of } \alpha

\[ = \frac{1}{n} \sum_{i=1}^{n} \alpha + \frac{1}{n} \beta \sum_{i=1}^{n} x_i + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i - b\bar{x} \quad \text{working the sum out}

\[ = \alpha + \frac{\beta}{\frac{1}{n}} + 0 - \frac{\beta}{\frac{1}{n}} = \alpha \]

\[ V(a_{LS}) = V(\hat{y} - b\bar{x}) \]

\[ = V(\hat{y}) - 2 \text{Cov}(\hat{y}, b\bar{x}) + V(b\bar{x}) \]

\[ = \frac{\sigma^2}{n} - 2 \text{Cov}(a_{LS}, b\bar{x}) + \bar{x}^2 V(b) \quad \text{by Cov. rules}

\[ = \frac{\sigma^2}{n} - 2 \text{Cov}(a_{LS}, b\bar{x}) + \bar{x}^2 V(b) \quad \text{by Cov. rules}

\[ = \frac{\sigma^2}{n} - 2 \bar{x} \text{Cov}(a_{LS}, b\bar{x}) + \bar{x}^2 V(b) \quad \text{by Cov. rules}

\[ = \frac{\sigma^2}{n} - 2 \bar{x} \text{Cov}(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} b\bar{x}) + \bar{x}^2 V(b) \quad \text{by the intercept,} \alpha \text{ when the slope, } b \text{ is estimated}

\[ = \frac{\sigma^2}{n} - 2 \bar{x} \text{Cov}(\hat{y}, b\bar{x}) + \bar{x}^2 V(b) \]

\[ = \frac{\sigma^2}{n} - 2 \bar{x} \text{Cov}(\sum_{i=1}^{n} y_i - b\bar{x}, b\bar{x}) + \bar{x}^2 V(b) \]

\[ = \frac{\sigma^2}{n} - 2 \bar{x} \text{Cov}(\hat{y} - b\bar{x}, b\bar{x}) + \bar{x}^2 V(b) \]

\[ = \frac{\sigma^2}{n} - 2 \bar{x} \text{Cov}(\hat{y}, b\bar{x}) + \bar{x}^2 V(b) \quad \text{by the intercept,} \alpha \text{ when the slope, } b \text{ is estimated}

\[ = \frac{\sigma^2}{n} - 2 \bar{x} \text{Cov}(\hat{y}, b\bar{x}) + \bar{x}^2 V(b) \quad \text{by the intercept,} \alpha \text{ when the slope, } b \text{ is estimated}

\[ = \frac{\sigma^2}{n} + \bar{x}^2 \quad \text{(so far)}

\[ V(a_{LS}) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2} \right] \]

\( \Rightarrow V(B_{LS}) = V(a_{LS}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \]

\( \Rightarrow 4. \text{ Three Facts About } V(B_{LS}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \]

\( \sigma^2 \) \text{ is higher the larger } \sum_{i=1}^{n} x_i^2 \text{ is (the more different the observations)}

\( \beta \) \text{ is lower the larger } \sum_{i=1}^{n} x_i^2 \text{ is (the greater the variation of } x \text{ in your sample)}

\( a_{LS} \) \text{ is lower as } \bar{x} \to 0 \text{ (as sample average of } x \text{ tends to zero)}
III.6. Given Assumptions (1-6) (all but Normality), **GAUSS-MARKOV**

**OLS is BLUE.** (best (min.variance) linear-unbiased estimator)

**Proof:** Consider some alternative to $b_{ls}$ (recalling $b_{ls} = \sum_{i=1}^{n} c_i x_i$). The $i$th part of the $Y_i$ vector.

To be a valid estimator, it must be a function of some function of the data. To be unbiased, it must be a function of known quantities, $X_i, \beta$ (not unknowns, $\varepsilon$).

To be linear, we must have $b' = \sum_{i=1}^{n} q_i y_i$. Here $q_i$ must be a function of $X_i$'s and $\varepsilon$,

$$b' = \alpha \sum_{i=1}^{n} q_i + \beta \sum_{i=1}^{n} q_i x_i + \sum_{i=1}^{n} q_i \varepsilon_i$$

Now, for $b'$ to be unbiased, $E(b')$ must equal $\beta$.

$$\Rightarrow E(b') = \alpha \sum_{i=1}^{n} E(q_i) + \beta \sum_{i=1}^{n} E(q_i x_i) + \sum_{i=1}^{n} E(q_i \varepsilon_i) = \beta$$

$$\Rightarrow q_i's$$ are functions of $X_i$'s which are fixed, so $E(q_i) = q_i$ and $E(q_i x_i) = q_i x_i$.

So, we must have

$$\sum_{i=1}^{n} q_i = 0 \text{ and } \sum_{i=1}^{n} q_i x_i = 0$$

for $E(b') = \beta$ to be true. So $b'$ is a permissible alternative to $b_{ls}$.

So $V(b_{ls}) = \sigma^2 \sum c_i^2$ and note $\sum c_i^2 = E q_i^2 = 0$ from

$$V(b') = \sigma^2 \sum q_i^2$$

Let $\nu_i = q_i - c_i$, so that $q_i = c_i + \nu_i$.

Then $V(b') = \sigma^2 \left( \sum (c_i^2 + \sum \nu_i^2 + 2 \sum c_i \nu_i) \right)$

= $0$ by orthogonality since $\nu_i = q_i - c_i$ the net effect was already

This is the Big Bang, so let me re-iterate:

1. $y_i = \alpha + \beta x_i + \varepsilon_i$ (i.e. $y$ a linear function of $x$)
2. $E(\varepsilon_i) = 0$ (trivial really)
3. $V(\varepsilon_i) = \sigma^2$ (constant variance)
4. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j$ (no covariance of $\varepsilon_i$'s)
5. $\text{Cov}(X_i, \varepsilon_i) = 0 \forall i \neq j$ (Y_i depends on X_i only in its mean)
6. $X_i$ is nonstochastic (fixed in repeated samples)

Then: $\beta = (X'X)^{-1} X'Y = [\alpha_{ls} \beta_{ls}]$ is the BLUE estimator of $\beta$. $\beta$ is a linear unbiased estimator.

*Notice that no assumptions about the exact shape of $f(y)$ need to be made, just statements (2-5) at this point.*
IV. Statistical Inference under the CLR Model:

- Gauss-Markov is true w/o normality; however, if we also have normality, then Assumption 1: That $y = \alpha + \beta x_i + \varepsilon_i$, i.e., linearity is unnecessary.

Under Assumptions 2-7 (normality included, linearity unnecessary), the Best Unbiased Estimator (BUE) happens to be linear.

For moving on to statistical inferences regarding LS estimates, it is useful to begin with the normality assumption included.

Then, $a$ and $b$, which are linear functions of non-stochastic $x$, if $x$ is normally distributed, must also be normally distributed. A property of normal R.V.'s, we noted before.

A. So, if $\varepsilon \sim N(0, \sigma^2)$ or equivalently, $f(y|x)$ is normal mean $\kappa \beta$, variance $\sigma^2$ then:

- $a \sim N(\alpha, \sigma^2(\frac{1}{n} + \frac{\hat{\beta}^2}{\hat{\beta}_{\text{Var}}}))$
- $b \sim N(\beta, \sigma^2/\hat{\beta}_{\text{Var}})$

or, more completely:

$$\begin{bmatrix} a \\ b \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \sigma^2(\frac{1}{n} + \frac{\hat{\beta}^2}{\hat{\beta}_{\text{Var}}}) & -\frac{\hat{\beta}}{\hat{\beta}_{\text{Var}}} \\ -\frac{\hat{\beta}}{\hat{\beta}_{\text{Var}}} & \sigma^2/\hat{\beta}_{\text{Var}} \end{bmatrix} \right)$$

B. So, if we were willing to maintain the normality assumption, we're almost all set to test hypotheses, construct confidence intervals, etc.

1. Suppose, for example, we wanted to test $H_0: \beta = 1$ against $H_1: \beta \neq 1$.

Then, we construct:

$$T = \frac{b-\beta}{\sigma/\hat{\beta}_{\text{Var}}} = (b-1) \frac{1}{\sigma/\hat{\beta}_{\text{Var}}} \sim N(0, 1)$$

under the null.

2. But we don't know $\sigma^2$ (the true variance of the stochastic component).

⇒ use an estimate.

One possible estimate: $E(\varepsilon_i) = 0$, so $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2$ seems a logical estimate of $\sigma^2$.

Problem: exactly analogous to previous results.

$$E(\sigma^2) = \frac{\sigma^2}{n} \text{ or } \frac{\sigma^2}{n} \text{ is biased by factor of } \left(\frac{n-2}{n}\right)$$

where does bias come from?

$$E_i = y_i - \bar{a} - \hat{\beta} \bar{X}_i$$

⇒ 2 degrees of freedom used up already.

⇒ Just as before, $S^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$ is an unbiased estimate

⇒ Just as before, $\left(\frac{b-\beta}{S^2/\hat{\beta}_{\text{Var}}}ight) \sim t_{n-2}$ i.e., $(b-\beta)/\text{estimated std error of } b \sim t_{n-2}$

(i.e. $(b-\beta)/\text{estimated std error of } b \sim t_{n-2}$)

why? $(b-\beta) \sim N(0, \sigma^2)$ $S^2/\hat{\beta}_{\text{Var}}$ is $\chi^2$ divided by its degrees of freedom ($\chi^2$, sum of squared normals, $\varepsilon_i \sim N(0, \sigma^2)$.}$
I. C. Statistical Inference in CLRM: Confidence Intervals

We want to construct an interval \([b - t \cdot s(b), b + t \cdot s(b)]\)

Such that

\[
(1) \quad \Pr \left[ b - t \cdot s(b) \leq \beta \leq b + t \cdot s(b) \right] = 1 - \alpha
\]

\[
\Rightarrow \text{read } t_{n-2} \text{ for (1-}\alpha) \text{ level } \frac{\alpha}{2}
\]

Then substitute your estimates for \(b \text{ and } s(b)\) into (1) if you're done.

2. What about a confidence interval (test) for \(\sigma^2\) ?

From before \(\frac{s^2}{\sigma^2} \sim \chi^2_n\) where \(n = \text{degrees of freedom of the estimate } \sigma^2\) is distributed \(\chi^2_n\)

\[
\Rightarrow \frac{s^2}{(\sigma^2/\sigma^2)} \sim \chi^2_n
\]

\[
\Rightarrow \alpha \text{ 95\% confidence interval: } \frac{(n-2) \cdot s^2}{\chi^2_{n-2}(0.975)} \text{ to } \frac{(n-2) \cdot s^2}{\chi^2_{n-2}(0.025)}
\]

Value from \(\chi^2_{n-2}\) which has 97.5% probability to the right of it.

\[
\Rightarrow \text{ Hypoth. Test: } T = \frac{(n-2) \cdot s^2}{\sigma^2} \text{, check } T \text{ on a } \chi^2_{n-2} \text{ table (or computer program)}
\]

V. The "Fit" of a Regression:

A. Variance Decomposition:

1) Total Sum of Squares: \(SS T = \sum_{i=1}^{n} (y_i - \bar{y})^2\) \{the sample {variation in } \(y\)\}

\[
a) \quad y_i - \bar{y} = (y_i + e_i) - \bar{y} = x_i b + e_i - (\bar{x} b) + a - a
\]

\[
\Rightarrow \quad \frac{SS T}{n-1} = b \left( x_i - \bar{x} \right) + e_i
\]

\[
b) \quad \sum (y_i - \bar{y})^2 = \sum [\left( y_i - \bar{y} \right) + e_i]^2 = \sum (y_i - \bar{y})^2 + \sum e_i^2
\]

\[
= \sum (b(x_i - \bar{x}))^2 + \sum e_i^2
\]

2. \(SS T = SSR + SSE\)

B. \(R^2 = \frac{\text{Variation in } y \text{ fit by the Regression}}{\text{Total Variation in } y} = \frac{SSR}{SS T} = 1 - \frac{SSE}{SS T}\)
IV-B. \( R^2 \) (Continued)

2. \( R^2 = \frac{SSR}{SST} = \frac{b^2 Sxx}{Syy} = \frac{(\bar{Y}/\bar{X})^2 Sxx}{Syy} = \frac{Sxx^2}{Syy Sxx} = r^2 \) 

by def of \( Syy \)

3. More generally: \( R^2 = (r_{xy})^2 \)

C. i) \( R^2 \) F-Tests:

\[ SSR = b^2 Sxx \text{ has } 1^0 \text{ of freedom (only } b \text{ varies)} \]
\[ SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \text{ has } (n-2)^0 \text{ of freedom (each } e \text{ varies across samples but estimating } \alpha \text{ and } \beta \text{ ties two of them down)} \]
\[ SST = Syy \text{ has } (n-1)^0 \text{ of freedom (each } Y \text{ varies across samples, but estimating } \beta \text{ ties one of them down)} \]

\[ \frac{SSR}{(n-2) MSE} \sim F_{1, n-2} \]

\[ \frac{R^2}{(1-R^2) n} = \frac{SSR}{SST} \frac{MSE}{\text{MSE} n} = \frac{SSR}{\text{MSE} n} \sim F_{1, n-2} \]

is a test of the regression model, \( Y = \alpha + bX + e \) against the alternative: \( Y = \alpha + e \)

\[ \Rightarrow \text{ in the bivariate case, exactly equivalent to a } t \text{-test on } \beta = 0 \]

D. So, since \( R^2 \) is the proportion of variation in \( Y \) "explained" (or "accounted for") by variation in \( X \), what's a good \( R^2 \)? Shouldn't we try to maximize it? \( \boxed{\text{No}} \)

1. \( R^2 = \frac{SSR}{SST} \) is therefore an estimate of the proportion of the world (simplified & approximated in your model) that is systematic in the X's. The world is not 100% systematic, even if it were, it'd be too cumbersome a model to carry around that would account for all of it. That is, some portion of the world is stochastic, so we don't know how much \( \Rightarrow R^2 \) is something you estimate, not something you maximize.

2. \( S = \left( \frac{1}{n-2} \sum e_i^2 \right)^{1/2} \) is also called the "standard error of the estimate" or "\( s_e \) of the regression" (SEE or SER) & it is an estimate of \( \sigma \), the true error of the regression model.

Obviously, this too is something you estimate, not something you minimize. Notice:

\[ R^2 \sim 1 - \frac{s_e^2}{SST} \left( \bar{R}^2 = 1 - \frac{s^2}{SST} \right) \]
1. First of all, $\frac{SSR}{SST}$ as proportion of the world given the model that is systematic is definitely something we'd like an estimate of (in general, though, we are usually even more or almost exclusively interested in our estimates of $\beta$ (the relationships $b$ the $x$'s $d$ the dependent variable(s)).

2. Second, as we've just seen, we can test hypotheses about coefficients (it's more useful when the hypothesis involves more than one coefficient) using it.

3. Within a single data set, it is meaningful to compare two models by their $R^2$: (in a manner we shall see next time), alternative.

VI. All of these tests, confidence intervals, etc. assumed normality, what if $f(y|x)$ is not normal (i.e., $E[y|x] = N(0,\sigma^2)$)?

Answer: Central Limit Theorem

- $b$ is linear combo of (large) numbers of random variables $(y_i's)$, so...
- $b \sim N(\beta, \sigma^2/n)$
- Whatever the actual distribution of $y_i$:

$\Rightarrow$ All of our discussion about "how large is large enough?", "does the limit kick in?" or "why do we use a t even though all we know is asymptotic normality"... all of that from last week now applies to OLS.

VII. Prediction: Sometimes, we use OLS to make predictions about $x_i$'s not observed (yet). In these cases, we would like to know what uncertainty to attach to our predictions just as we would like to know what uncertainties to attach to our coefficient estimates ($\alpha$ and $b$).

A. We have estimated $\hat{y}_i = \hat{a} + \hat{b}x_i$; now we get some new data $x$, we want to use it "predict" the $Y_i$th observation.

1. What's the prediction error? $e_i = Y_i - \hat{Y}_i = Y_i - \alpha - bX_i$

$e_i = (\alpha + \beta X_i + \epsilon_i) - \alpha - bX_i$

$= (\alpha - \alpha) + (\beta - b)X_i + \epsilon_i$

2. Thus two sources of error:

a) One because $\alpha$ and $\beta$ are only estimates of $\alpha$ and $\beta$

b) One because there is a natural stochastic component to the world

B. Prediction Intervals: $\sqrt{\varepsilon_i^2} = \sqrt{\varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2}$

$\Rightarrow \varepsilon_i = \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n$

$\Rightarrow \text{cov}(\varepsilon_1, \varepsilon_2) + \text{cov}(\varepsilon_1, \varepsilon_2) = 0$

$\Rightarrow \text{var}(\varepsilon_1, \varepsilon_2) + \text{var}(\varepsilon_1, \varepsilon_2) = 0$

$\Rightarrow \text{var}(\varepsilon_1, \varepsilon_2) + \text{var}(\varepsilon_1, \varepsilon_2) = 0$

$\Rightarrow \text{var}(\varepsilon_1, \varepsilon_2) + \text{var}(\varepsilon_1, \varepsilon_2) = 0$

$\Rightarrow \text{var}(\varepsilon_1, \varepsilon_2) + \text{var}(\varepsilon_1, \varepsilon_2) = 0$
VII. B. (cont.) Forecast Error

So, we can test statements such as "given our model, the outcome in this new x observation was 'surprising'" where that last is taken to mean beyond some specified confidence interval.

Conf. Intervals:

\[ \text{for } y = (a + bx) \]

\[ (a + bx) \pm t_{\alpha/2} \cdot \sqrt{\frac{s_x^2}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}} \]

\[ \text{we don't know } \sigma^2, \text{ so we use its estimate} \]

n.b.: \( y \)

\[ E(y|1|x) \]

\[ \text{the forecast:} \]

\[ \text{confidence interval bounds} \]

\( \text{eteris paribus}, \) \( \text{n.b.: Just like with coefficients} \)

\[ \uparrow S_{xx} \text{ (sample variation in } x) \]

\[ \Rightarrow \downarrow \text{forecast error variance} \]

\[ \uparrow S^2 \text{ (sample variance of } e_i) \]

\[ \Rightarrow \uparrow \text{forecast error variance} \]

\[ \uparrow n \text{ (sample size used to estimate } a \text{ and } b) \]

\[ \Rightarrow \downarrow \text{forecast error variance} \]

\[ \text{one new thing (which we'll see again)} \]

\[ \uparrow (x_i - \bar{x}) \Rightarrow \uparrow \text{forecast error variance} \]

\[ \text{(Further the new observation is from the mean of those used to estimate a and b)} \]

VIII. Some Remaining Thoughts on Bivariate Regression Analysis

A. Theory should drive your econometric modelling. That is, we do not specify models to fit lines to data (generally), we estimate relationships between variables of interest.

*Empirical Tests must be theoretically informed so that the results can be theoretically informing.*

B. Despite plethora of restrictive assumptions, OLS often does better than it has a right to do. Why? In general b/c of its simplicity: it is the best linear approximation & the CLT helps. I we often simply do not know enough to go beyond linear approximation & bounded (<=oo) variance.

*Still sometimes we do know more, sometimes a lot more & sometimes even the best linear approximation is still a crummy one. We the relationship is very non-linear.*

C. When does OLS do poorly?

\[ \left\{ \begin{array}{l}
1. \text{ Since } \min (y - x^2) \text{ is sensitive to "outliers". I.e., when things not normal in a particular way: when tails too fat. We'll elaborate on this later.} \\
2. \text{ When things are highly non-linear} \\
3. \text{ When assumptions 1 & 2 of these violated} 
\end{array} \right. \]
1. Gauss-Markov: OLS is BLUE

\[ b_{ls} = \sum C_i \tilde{y}_i \]
(recall the \(-\hat{\beta}\) drops out)

consider some other linear estimated from the data that is also unbiased

\[ b' = \sum q_i \tilde{y}_i \]
(linear \(q_i\) s just coefficients)

Now, if CLRM assumptions hold, \( \tilde{y}_i = \alpha + \beta x_i + \varepsilon_i \)

\[ b' = \alpha \sum q_i + \beta \sum q_i x_i + \sum q_i \varepsilon_i \]
(n.d. no N)

**Unbiased** \( b' \)

\[ \Rightarrow E(b') = \alpha \sum E(q_i) + \beta \sum E(q_i x_i) + \sum E(q_i \varepsilon_i) = \beta \]

\( q_i = f(x_i), \) d \( x_i \) is fixed, so \( E(q_i) = q_i \)

\[ E(b') = \alpha \sum q_i + \beta \sum q_i x_i + \sum q_i E(\varepsilon_i) \]

\( \cdots \)

\( \sum q_i \varepsilon_i = 0 \) if \( \sum q_i x_i = 1 \)

\[ V(b_{ls}) = \sigma^2 \sum C_i^2 \quad \rightarrow \quad \sum C_i \varepsilon_i = 0 \]

\[ V(b') = \sigma^2 \sum q_i^2 \quad \Rightarrow \quad V(b') = \sigma^2 (\sum C_i^2 + \sum V_i C_i + \sum V_i^2) \]

\( \Rightarrow \) by definition of \( V_i, V_i \) and \( C_i \) are orthogonal

\[ V(b') = \sigma^2 \left( \sum C_i^2 + \sum V_i C_i + \sum V_i^2 \right) \]

so \( \sum V_i C_i = 0 \)

\[ \Rightarrow V(b') \geq V(b_{ls}) \quad \forall b' \text{ linear and unbiased} \]

& CLRM assumptions hold

\[ \Rightarrow \text{restate Assumptions, emphasize that } f(y) \text{ or } N(3) \text{ not specified beyond assumption 2-5} \]
I. Statistical Properties of the LS estimator

\[ A_i = b = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]

Let \[ C_i = \frac{(X_i - \bar{X})}{S_{XX}} \Rightarrow b = \frac{\sum C_i \cdot (Y_i - \bar{Y})}{\sum C_i \cdot Y_i} \]

\[ b = \frac{\left( \sum (X_i - \bar{X}) \right) (\alpha + \beta X_i + \varepsilon_i - \bar{Y})^2}{S_{XX}} \]

but \[ \sum (X_i - \bar{X}) = \sum X_i - n \cdot \bar{X} = n \cdot \bar{X}_i - n \cdot \bar{X} = 0 \]

so \[ b = 0 + \frac{\beta \sum (X_i - \bar{X}) X_i^2}{S_{XX}} + \frac{\sum (X_i - \bar{X}) \varepsilon_i^2}{S_{XX}} = 0 \]

but \[ \sum (X_i - \bar{X})^2 = \sum X_i^2 - 2 \bar{X}X_i + \bar{X}^2 \]

so \[ S_{XX} = \sum X_i^2 - \sum X_i \bar{X}_i = \sum (X_i - \bar{X}) X_i^2 \]

so \[ b = \beta \frac{\bar{X}}{S_{XX}} + \frac{\sum (X_i - \bar{X}) \varepsilon_i^2}{S_{XX}} = \beta + \sum C_i \varepsilon_i \]

unbiased \[ E(b) = \beta + \sum C_i E(\varepsilon_i) = \beta \]

\[ V(b) = V(\beta + \sum C_i \varepsilon_i) \]

\[ = V(\beta) + \sum C_i \sum C_j V(\varepsilon_i \varepsilon_j) \]

\[ = \sum C_i^2 V(\varepsilon_i) \]

\[ = \sum \frac{\varepsilon_i (X_i - \bar{X})^2}{S_{XX}} \]

\[ = \frac{\sum \left( \frac{\varepsilon_i}{S_{XX}} \right)^2 (X_i - \bar{X})^2}{(S_{XX})^2} \sigma^2 = \frac{S_{XX}}{(S_{XX})^2} \sigma^2 = \frac{\sigma^2}{S_{XX}} \]

\[ V(b) = \frac{1}{S_{XX}} \sigma^2 = \frac{1}{\sum (X_i - \bar{X})^2} \sigma^2 \]

For a true \[ E(\alpha) = \alpha \]

unbiased \[ V(\alpha) = \frac{\sigma^2}{S_{XX}} \]

\[ V(b) = \frac{1}{S_{XX}} \Sigma \left[ \begin{array}{c} V(\alpha) \vspace{1cm} \frac{-\bar{X}}{S_{XX}} \end{array} \right] \]

Things to Note:

- \[ V(b) = \frac{1}{S_{XX}} \sum (\text{Sample Variation in } X_i, \sigma^2) \]
- \[ V(b) = \frac{1}{S_{XX}} \sum (\text{Sample Variance in } X_i, \sigma^2) \]
- \[ V(b) = \frac{1}{S_{XX}} \sum (\text{Sample Variance in } X_i, \sigma^2) \]

estimated \[ V(b) = \frac{1}{S_{XX}} \sum (\text{Sample Variance in } X_i, \sigma^2) \]

\[ \Rightarrow g(N) \]

\[ \Rightarrow g(N) \]

\[ \Rightarrow g(N) \]
The Classical (Normal) Linear-Regression Model

Assumptions:

1. Linearity of the dependent variable is a linear-additive function of the regressors, with an additively separable stochastic component:

   \[ y = X\beta + \varepsilon \quad \text{or} \quad \varepsilon \sim (\mu, \sigma^2 I) \quad \text{or} \quad E(\varepsilon) = \mu = X\beta \quad \text{(n.b.} \; \varepsilon \text{)} \]

2. Zero mean of the stochastic component:

   \[ E(\varepsilon) = 0 \quad \text{(n.b.} \; \varepsilon \text{)} \]

3. Homoskedasticity (i.e., constant conditional variance) of the stochastic component:

   \[ \forall (\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I \quad = \sqrt{\varepsilon} \]

4. Zero (conditional) covariance of the stochastic component:

   \[ \text{Cov}(X, \varepsilon) = 0 \quad \text{i.e.} \quad E(X'\varepsilon) = 0 \]

5. The regressors are (weakly) exogenous, i.e., the regressors do not covary with the stochastic components:

   \[ \text{Cov}(X, \varepsilon) = 0 \]

6. \textit{X} is non-stochastic, i.e. fixed across repeated (hypothetical) samples

7. Normality:

   \[ \varepsilon \sim N(0, \sigma^2 I) \quad \text{or} \quad y \sim N(X\beta, I) \]

Issues/Challenges/Violations/Opportunities

common in multilevel contexts:

1. The implicit parameter-homogeneity assumption in \( \text{1.} \& \text{2.} \)
2. The spherical error-structure assumption in \( \text{3.} \& \text{4.} \)
3. (In some contexts, the \( E(X'\varepsilon) = 0 \) assumption by interdep.)
OLS: An Estimator for the $C(N)$ LRM

CLRM: $\chi = X_\beta + \varepsilon$, $\varepsilon \sim (\varnothing, \sigma^2 I)$

1. Regression as a Projection Problem in Geometry:

\[ \Rightarrow (\chi - X\beta) \perp \chi \]
\[ \Leftrightarrow \chi' (\chi - X\beta) = 0 \]

2. Regression as a Minimization Problem in Calculus:

\[ \Rightarrow \chi'\chi - \chi'X\beta = 0 \]
\[ \Rightarrow b^* = (X'X)^{-1}X'\chi \]
\[ \Rightarrow \min_{\beta} \chi'X - b'X\chi - \chi'X\beta - b'X\chi X\beta \equiv \mathcal{L} \]
\[ \nabla_b \mathcal{L} = -X'\chi + X'\chi X\beta - X'X\beta - X'X\beta = 0 \]
\[ \Rightarrow X'X\beta = X'\chi \Rightarrow b^* = (X'X)^{-1}X'\chi \]

3. Regression as a Maximum-Likelihood Problem in Statistics

\[ \chi_i \sim N(\mu_i, \sigma^2) \]
\[ \Rightarrow f(\chi_i | X_i \beta) = (2\pi \sigma^2)^{-\frac{1}{2}} e^{-\frac{(\chi_i - X_i \beta)^2}{2\sigma^2}} \]

\[ \text{Maximize the likelihood of observing this set of } \chi \text{ given the model} \]
\[ \mathcal{L}(Y) = \prod_{i=1}^n (2\pi \sigma^2)^{-\frac{1}{2}} e^{-\frac{(\chi_i - X_i \beta)^2}{2\sigma^2}} \]
\[ \ln \mathcal{L} = \sum_{i=1}^n \left[ -\frac{1}{2} \ln(2\pi \sigma^2) - \frac{1}{2} \ln \sigma^2 - (\chi_i - X_i \beta)^2 / 2\sigma^2 \right] \]

\[ \frac{\partial \ln \mathcal{L}}{\partial \beta} = -\frac{1}{\sigma^2} \sum_{i=1}^n (\chi_i - X_i \beta) X_i \]
\[ \text{set to } 0 \Rightarrow \sum_{i=1}^n \chi_i \beta_i = \sum_{i=1}^n \chi_i X_i \beta \]
**Properties of OLS for the CLRM:**

**UNBIASED:** \( E(\hat{\beta}_{OLS}) = \beta \)

\[
E[(X'X)^{-1}X'\varepsilon] = E[(X'X)^{-1}X'(X\beta + \varepsilon)] = (X'X)^{-1}X'X\beta + E[(X'X)^{-1}X'\varepsilon] = \beta + (X'X)^{-1}E(X'\varepsilon) = \beta
\]

- Expectation of a sum is the sum of the expectations.
- "Expectation operator slides through constants & sticks to R.V's"

**VARIANCE:** \( V(\hat{\beta}_{OLS}) = V[(X'X)^{-1}X'(X\beta + \varepsilon)] = V[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon] \)

- Constants neither vary nor covary.
- "Variance operators eliminate isolated constants, slide through constant coefficients, squaring them on the way by, snagging on R.V's, & spitting out 2x all the Covs"

\[
V(\hat{\beta}_{OLS}) = 0 + V[(X'X)^{-1}X'\varepsilon] = (X'X)^{-1}X'V(\varepsilon)X(X'X)^{-1}
\]

**EFFICIENT:** (means "no unused useful information")

\[
\rightarrow \text{suffice to say } \sigma^2(X'X)^{-1} \text{ hits C-R lower bound}
\]

[go to "CLRM Crash Course"]
Multivariate Regression: Estimating & Reporting Certainty, Hypotheses Tests

I. The C(N)LRM:

A. Core Assumptions:

1. \( y = X\beta + \epsilon \)
2. \( E(\epsilon) = 0 \)
3. \( V(\epsilon) = \sigma^2 \epsilon I_n \)
4. \( E(\epsilon | X) = 0 \)
5. \( X \) of full-column rank.

B. Convenience Assumptions (Unnecessary):

1. \( X \) non-stochastic.
2. \( \epsilon \sim N(0, \sigma^2 \epsilon I_n) \). (If not: CLT & asymptopia!)

II. Properties of OLS Estimator under CLRM:

\[ \hat{\beta}_{LS} \equiv b_{LS} = (X'X)^{-1} X'y \equiv Ay \]

A. \[ = A(X\beta + \epsilon) = \beta + A\epsilon \]
\[ E\left( \hat{\beta}_{LS} \right) = E\left( \beta + A\epsilon \right) = \beta + E(A\epsilon) \]
\[ = \beta + AE(\epsilon) = \beta, \text{ if } X \text{ non-stoch.} \]
\[ = \beta + E(A\epsilon) = \beta, \text{ if } X \text{ stoch.} \& E(\epsilon \mid X) = 0 \]

\[ V\left( \hat{\beta}_{LS} \right) = V\left( \beta + A\epsilon \right) = AV(\epsilon)A' \]
\[ = (X'X)^{-1}X'V(\epsilon)X(X'X)^{-1} \]
\[ = (X'X)^{-1}X'\sigma_\epsilon^2 I_n X(X'X)^{-1} \]
\[ = \sigma_\epsilon^2 (X'X)^{-1}X'X(X'X)^{-1} \]
\[ = \begin{cases} 
\sigma_\epsilon^2 (X'X)^{-1}, \text{ if } X \text{ non-stoch.} \\
\sigma_\epsilon^2 E\left\{(X'X)^{-1}\right\}, \text{ if } X \text{ stoch.} 
\end{cases} \]

\[ \Rightarrow V\left( b_k \right) = f\left( \sigma_\epsilon^2, V(X), R_{x_k,-x_k}^2 \right) \]

D. EFFICIENT: "using all useful information".

\[ V\left( \hat{\beta}_{LS} \right) = \sigma_\epsilon^2 (X'X)^{-1} \text{ is the Cramer-Rao lower-bound;} \]

i.e., \( \hat{\beta}_{LS} \) is efficient; i.e., \( \hat{\beta}_{LS} \) is BLUE.

1. In fact, if \( \epsilon \sim MVN \), then \( \hat{\beta}_{LS} \) is BUE.
E. If \( \epsilon \sim \text{MVN}(0, \sigma_\epsilon^2 I_n) \), then:

\[
\mathbf{b}_{LS} = \beta + \mathbf{A} \epsilon = \text{constant + linear-sum of normals}
\]

\[
\Rightarrow \mathbf{b}_{LS} \sim \text{MVN}(\beta, \sigma_\epsilon^2 (X'X)^{-1})
\]

if not, then

\[
\mathbf{b}_{LS} = \beta + \mathbf{A} \epsilon = \text{constant + linear-sum of } ?
\]

\[
\Rightarrow \mathbf{b}_{LS} \sim^A \text{MVN}(\beta, \sigma_\epsilon^2 (X'X)^{-1})
\]

III. Estimating \( \sigma_\epsilon^2 \) using \( \frac{1}{n-k} \mathbf{e}'\mathbf{e} \):

\[
\mathbf{e} = \mathbf{M}(X\beta + \epsilon) = \mathbf{MX}\beta + \mathbf{M}\epsilon = 0\beta + \mathbf{M}\epsilon = \mathbf{M}\epsilon
\]

\[\Rightarrow E(\mathbf{e}'\mathbf{e}) = E\{(\mathbf{M}\epsilon)'\mathbf{M}\epsilon\} = E\{\epsilon'\mathbf{M}'\mathbf{M}\epsilon\} = E\{\epsilon'\mathbf{M}\epsilon\}
\]

\[= E\{trace(\epsilon'\mathbf{M}\epsilon)\} = E\{trace(M\epsilon\epsilon')\}
\]

\[= trace\{\mathbf{M}\epsilon\epsilon'\} = trace\{\mathbf{M}\sigma_\epsilon^2 I_n\}
\]

\[= \sigma_\epsilon^2 \times trace\{I_n - \mathbf{N}\} = \sigma_\epsilon^2 \times trace\{I_n\} - trace\{\mathbf{N}\}
\]

\[= \sigma_\epsilon^2 \times \{n - trace\{X(X'X)^{-1}X'\}\}
\]

\[= \sigma_\epsilon^2 \times \{n - trace\{(X'X)^{-1}X'X\} = \sigma_\epsilon^2 \left[ n - trace(I_k) \right]\}
\]

\[= \sigma_\epsilon^2 (n-k)
\]

\[\Rightarrow E\left(\frac{\mathbf{e}'\mathbf{e}}{n-k}\right) = \sigma_\epsilon^2 \Rightarrow \text{use } s_\epsilon^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} \text{ as LS (unbiased) est. } \sigma_\epsilon^2.
\]
A. NOTE: $\frac{e'e}{n-k}$ is sum of squared (asympt.) normals, divided by deg freedom, so $s_e^2$ is (asym.) $\chi_{n-k}^2/n-k$.

B. Therefore, std **Wald** $t$-tests & conf. ints. by:

1. 

$$T = \frac{b_j - c_0}{s.e.(b_j)} = \frac{b_j - c_0}{\sqrt{s_e^2 \{(X'X)^{-1}\}_{jj}}} \sim \text{(A)} t_{n-k}$$

2. 

$$b_j \pm t_{n-k}^\alpha \times s.e.(b_j) = b_j \pm t_{n-k}^\alpha \times \sqrt{s_e^2 \{(X'X)^{-1}\}_{jj}}$$

$$\Rightarrow \{1 - \alpha\} \% \text{ (asympt.) conf. int.}$$

3. Tests of linear restrictions (by **Wald strategy**):

a) For instance, $H_0: \beta_1 + \beta_2 = 1$. If null-hypothesis true, then estimate not far (in std. err. units) from null:

$$T = \frac{(b_1 + b_2) - 1}{s.e.(b_1 + b_2)} = \frac{(b_1 + b_2) - 1}{\sqrt{\hat{V}\left(\hat{b}_1 + \hat{b}_2\right)}}$$

$$= \frac{(b_1 + b_2) - 1}{\sqrt{\hat{V}\left(\hat{b}_1\right) + \hat{V}\left(\hat{b}_2\right) + 2 \times \hat{C}\left(\hat{b}_1, \hat{b}_2\right)}} \sim A t_{n-k}$$

b) But note how we could use matrix algebra to generalize:
\( H_0: \beta_1 + \beta_2 = 1 \Rightarrow r\beta = q; \) for instance, if \( \beta \) is \((4 \times 1)\), then:

\[ r' = [0 \ 1 \ 1 \ 0] \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \]

\[ \Rightarrow H_0: r\beta = q; \quad \text{e.g.,} \quad H_0: r\beta = 1 \Rightarrow T = \frac{r'b - 1}{\sqrt{\hat{V}(r'b)}} = \frac{r'b - 1}{\sqrt{r'\hat{V}(b)r}} \sim A^{-1} t_{n-k} \]

Notice how \( r' \ldots r \) plucks \( \hat{V}(b_1), \hat{V}(b_2), \hat{C}(b_1, b_2), \hat{C}(b_1, b_2) \), just as in scalar!

4. Test of joint hypotheses (by Wald Strategy):

\( H_0: \beta_1 = q_1, \beta_2 = q_2 \Rightarrow R\beta = q; \)

for instance, \( H_0: \beta_1 = 0, \beta_2 = 1 \) and \( \beta \) is \((4 \times 1)\), then:

\[ R = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \text{and} \quad q = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \]

\[ \Rightarrow H_0: R\beta = q \Rightarrow C = (Rb - q)' \left[ \hat{V}(Rb - q) \right]^{-1} (Rb - q) \]

is a ratio of chi-squares. "Denominator" has n-k deg. free, and has n-k in its denom. "Numerator" is square 2\times1 normals. So, C/2, or, more generally, C/J where J=rows(R) is \( F_{j,n-k} \)!

\[ F = (Rb - q)' \left[ \hat{V}(Rb - q) \right]^{-1} (Rb - q) / J \sim A^{-1} F_{j,n-k} \]
C. “Degradation of Fit” Strategy for Testing:

1. Logic: If null hypothesis were true, then imposing it as true rather than estimating its parameters should result in “little” loss of fit.

2. Strategy: Measure fit-loss, then determine how that measure or some function of it would be distributed under the null hypothesis, so we can determine how likely this much fit-loss is to have occurred by chance.

3. The “change-in-R²” or “delta-R²” or “ΔR²” Test:

   a) Determine how to “impose the null hypothesis”. Example:

   \[ H_o: \beta_j = \beta \ \forall j \implies \begin{cases} H_o: y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \\ H_1: \beta_j \neq \beta \ \forall j \implies H_1: y_{ij} = \beta_0^j + \beta_1^j x_{ij} + \varepsilon_{ij} \end{cases} \]

   b) Measure loss of explanatory power relative to gap from big-model explanatory power to one, and divide each numerator and denominator by its degrees of freedom:

   loss of fit: \( \Delta R^2 = R_1^2 - R_0^2 = \left(1 - \frac{SSE_{n-k_1}}{SST_{n-1}}\right) - \left(1 - \frac{SSE_{n-k_0}}{SST_{n-1}}\right) = \frac{SSE_{n-k_0} - SSE_{n-k_1}}{SST_{n-1}} \)

   fit-gap: \( 1 - R_1^2 = \frac{SSE_{n-k_1}}{SST_{n-1}} \implies \frac{SSE_{n-k_0} - SSE_{n-k_1}}{SSE_{n-k_1}} \)

   \( \implies \frac{\chi^2_{n-k_0} - \chi^2_{n-k_1}}{\chi^2_{n-k_1}} \implies \text{free:} \quad \frac{k_1 - k_0}{n - k_1} = \frac{\Delta k}{n - k_1} = \frac{Jk - k}{n - Jk} \)
So: \[ F = \frac{\frac{\Delta R^2}{\Delta k}}{(1 - R_1^2)/(n - k_1)} \sim ^{(A)} F_{\Delta k, n-k_1} \]

4. Tests using other measures of fit, \( s_e^2 \) or \( \ln(L) \), also:

   a) **Likelihood-Ratio test**: \[ 2[\ln(L_0) - \ln(L_1)] \sim \chi^2_r . \]
   
   \[ \left( e' e_{\text{pooled}} - \sum_{j=1}^{J} e' e_j \right) / \Delta k_{jk-k} \]
   
   \[ \sim ^{A} F_{jk-k, n-\Sigma j k_j} \]

   b) **Chow test**: \[ \sum_{j=1}^{J} e' e_j / (n - \sum_{j=1}^{J} k_j) \]

5. Third logic, **Lagrange-Multiplier Tests**: if null hypoth were true, then to impose it as constraint on max \( \ln(L) \) or min \( SSE \) should not bind, implying that Lagrange multipliers, \( \lambda \), and \( \Delta \ln(L)/\lambda = 0 \) or \( \Delta SSE/\lambda = 0 \).
**Assumptions:**

1. \( Y = X \beta + \epsilon \), \( \epsilon \sim \mathcal{N}(0, \sigma^2 \Omega) \)
2. \( \Omega \) is some \( n \times n \) matrix
3. \( E(X' \epsilon) = 0 \), \( X \) full rank

For what did we use \( V(\epsilon) = \sigma^2 I \) from the CLRM?

1. Estimating \( \sigma^2 \); To show OLS efficient, & what its standard errors are.
2. Distribution of \( \hat{\beta}_{OLS} \); also for hypothesis tests, etc.
   - i.e., to show \( V(\hat{\beta}_{OLS}) = \sigma^2 (X'X)^{-1} \)

We did not use it to show:

3. \( \hat{\beta}_{OLS} \) unbiased, consistent, or asymptotically normal

What does \( \sigma^2 \Omega \) look like in multilevel data?

\[
V(\epsilon) = \begin{pmatrix}
\sigma_{y}^2 & \sigma_{y,1}^2 & \cdots & \sigma_{y,n_{i}}^2 \\
\sigma_{y,1}^2 & \sigma_{y,1}^2 + \sigma_{1}^2 & \cdots & \sigma_{y,1,\ldots,n_{i}}^2 \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{y,n_{i}}^2 & \sigma_{y,n_{i},\ldots,n_{j}}^2 & \cdots & \sigma_{y,n_{i,\ldots,n_{j}}^2}
\end{pmatrix}
\]
Properties of OLS under G(N)LRM:

1. **Unbiased:**
   \[ b_{\text{LS}} = \hat{\beta} = \hat{\beta} + \epsilon + A \epsilon \]
   \[ E(b_{\text{LS}}) = \beta \]

2. **Variance:**
   \[ V(b_{\text{LS}}) = V(\hat{\beta}) = A V(\epsilon) A' = A \sigma^{-2} \Omega A' \]
   \[ = \sigma^{-2} (X'X)^{-1} \left[ X' \Omega X \right] (X'X)^{-1} \]
   \[ \neq \sigma^{-2} I \]

   \[ \Rightarrow \text{OLS standard errors wrong (biased, inconsistent, inefficient)} \]

   OLS coefficients inefficient (unequal, unexploited)

Sandwich Estimators:

- Since OLS is "only" inefficient, could fix s.e.'s:
  - **Difference OLS \( \hat{\beta} \)** from CLRM to GLRM: In \( \Omega \)
    \[ [X' \sigma^{-2} I X] \quad \text{vs.} \quad [X' \sigma^{-2} \Omega X] \]
    \[ \uparrow \quad \Omega \quad \uparrow \Omega^* \]
    \[ \Omega^* \text{ is } k \times k, \text{ where as } \Omega \text{ is } N \times N \]

    \[ \Omega^* = \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_i \epsilon_j X_i X'_j \]

    \[ \Rightarrow \text{a correction need concern itself only with how } \sigma_i^2 \text{ and } \sigma_j^2 \text{ 's relate to } X_i X'_j \]
Sandwich Estimators: First one offered: "Pure Heteroskedastic."

\[ V(\varepsilon) = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_n^2 \end{bmatrix} \]

\[ Q^* = X' \Omega X = \begin{bmatrix} -X_1 & \cdots & -X_K \end{bmatrix} \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \]

\[ = \sum_{i=1}^n \omega_{ii} X_i X_i' \]

\[ \Rightarrow \hat{Q}^* = \sum_{i=1}^n \epsilon_i \omega_{ii} X_i X_i' \]

\[ \Rightarrow \sqrt{n}(\hat{\beta}_L) = (X'X)^{-1} X' \hat{Q}^* X (X'X)^{-1} \]

Many extensions since, essentially versions of:

\[ \hat{Q} = \sum_j \epsilon_i \epsilon_j X_i X_j' \]

but not all necessarily "well behaved" so 'tricks' added

\[ \Rightarrow \hat{Q}^* \text{ converges to constant as } N \rightarrow \infty \]

\[ \Rightarrow \hat{Q}^* \text{ positive definite} \]

Properties OLS + Sandwich under GLRM:

\[ \Rightarrow \hat{\beta}_L \text{ unbiased, consistent, inefficient} \]

\[ \Rightarrow \sqrt{n}(\hat{\beta}_L) \text{ consistent, inefficient} \]
(Feasibly) Efficient Estimation: Need use the wasted info!

Stipulate: \( \exists \) matrix \( \Omega^{-\frac{1}{2}} \), "square root of the inverse of \( \Omega \)"

Then: \( \Omega^{-\frac{1}{2}} y = \Omega^{-\frac{1}{2}} x \beta + \Omega^{-\frac{1}{2}} \varepsilon \) under GLRM follows the CLRM, so OLS on this BLUE

\[
\varepsilon^* = \Omega^{-\frac{1}{2}} \varepsilon \Rightarrow \text{V}(\varepsilon^*) = \text{E}(\varepsilon^* \varepsilon^\top) = \text{E}(\varepsilon \varepsilon^\top) \Omega^{-\frac{1}{2}}
\]

\[
= \Omega^{-\frac{1}{2}} \text{V} \Omega^{-\frac{1}{2}} = \sigma^2 \Omega \Omega^{-1} = \sigma^2 I
\]

"Pure Het" Example:

Multilevel Context \( \text{V}(\varepsilon) = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_k^2 \end{bmatrix} \)

\( \bar{\varepsilon} = \Omega^{-\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\sigma_k} \end{bmatrix} \)

\( \text{V}(\varepsilon) = \begin{bmatrix} 2 & & \\ & 6 & \\ & & 3 \end{bmatrix} \)

\( \Rightarrow \Omega^{-\frac{1}{2}} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\sigma_k} \end{bmatrix} \)

\( \Rightarrow \text{Using the wasted info:} \) More reliable units weighed more heavily

(see \( \varepsilon = \rho \bar{\varepsilon} + \gamma \) example if time)
Linear Regression in Matrix Notation

FGLS: Another Example

\[ y_t = x_t \beta + \epsilon_t \quad , \quad \epsilon_t = \rho \epsilon_{t-1} + \nu_t \]

\[ \begin{bmatrix} 
1 & p^2 & \ldots \\
1 & p & \ldots \\
1 & \ldots & \ldots 
\end{bmatrix} \Rightarrow \Omega^{-\frac{1}{2}} = \begin{bmatrix}
\sqrt{1-p^2} \\
0 \\
0 \\
\end{bmatrix} \begin{bmatrix}
0 \\
\frac{-p}{\sqrt{1-p^2}} \\
0 \\
\end{bmatrix} \begin{bmatrix}
0 \\
0 \\
1 \\
\end{bmatrix} \]

\[ \Omega^{-\frac{1}{2}} \gamma \Rightarrow \begin{bmatrix}
(1-p^2)\gamma_1 \\
y_2^2 \gamma_1 \\
\vdots \\
y_n^2 \gamma_1 \\
\gamma_n \gamma_n \\
\end{bmatrix} \]

Recall: issue with \( \bar{\epsilon}_{ij} \neq 0 \) was that each obs not wholly new info. This effects a weighing of only the "new" info.

Interactions: do some nonlinear examples

\[ y = \ldots + b_0 \text{age} + b_2 \text{age}^2 + b_{edu} + b_{edu} \text{age} + b_{edu} \text{age}^2 + \ldots \]

[\( \frac{\partial y}{\partial \text{age}} \) = \ldots \( \frac{\partial y}{\partial \text{edu}} \) = \ldots]

Types of Multilevel / Panel, Longitudinal & TSCS data
With a binary outcome, linear reg many probles

1. $\hat{p}$, $\hat{p}(y=1)$ can be $>1 \Rightarrow <0$
   - interpret "Non-sense predictions"
2. Heteroskedasticity: $V(y) = \pi (1-\pi)$
   - Fvus $V(\hat{p}) = V(p) = \hat{p} (1-\hat{p})$
3. Non-normality: $\hat{e} = y - x\beta = \{y - x\beta\}$
   - so bounded & related in particular way to $x\beta$ \Rightarrow non-normal

4. **Non-linearity:** Substance strongly suggests
   - $\lim_{x\beta \to \infty} p(y=1) = 1$ & $\lim_{x\beta \to \infty} = 0$, & smoothly
   - \Rightarrow want $p = s$-shaped $f(x\beta)$, not linear, smoothly bounded 0:

   **Logit:** $\frac{e^{x\beta}}{1+e^{x\beta}} = p$
   **Probit:** $\Phi(x\beta) = p$

**MaxLike Estimation Logit/Probit:**

$$f(y|x) = \pi^y (1-\pi)^{1-y}$$

$$L : f(y|x = \hat{f}(x\beta)) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

$$\ln L = \sum_{i=1}^{n} \{y_i \ln F(x_i\beta) + (1-y_i) \ln [1-F(x_i\beta)]\}$$

$$\frac{\partial (\ln L)}{\partial \beta} = 0 \Rightarrow \sum_{i=1}^{n} \{y_i \frac{F(x\beta)}{F(x_i\beta)} X + (1-y_i) \frac{F(x\beta)}{F(x \beta)} X\}$$

$$= \sum_{i=1}^{n} \{y_i \frac{F(x\beta)}{F(x\beta)} - (1-y_i) \frac{F(x\beta)}{F(x\beta)} \} X$$
Interpreting Logit/Probit:

**Logit**:
\[
\hat{p} = \frac{e^{\hat{\beta} x}}{1 + e^{\hat{\beta} x}} = \left[1 + e^{\hat{\beta} x}\right]^{-1}
\]

- Marginal Effects:
  \[
  \frac{\partial \hat{p}}{\partial x_2} = \frac{e^{-\hat{\beta} x}}{1 + e^{\hat{\beta} x}} \cdot \hat{\beta}_2 \\
  = \frac{1}{1 + e^{\hat{\beta} x}} \cdot \frac{e^{-\hat{\beta} x}}{1 + e^{\hat{\beta} x}} \cdot \hat{\beta}_2 \\
  = \hat{p} \cdot (1 - \hat{p}) \cdot \hat{\beta}_2
  
\]

**Probit**:
\[
\hat{p} = \Phi(\hat{\beta} x) \\
\Rightarrow \frac{\partial \hat{p}}{\partial x_2} = \phi(\hat{\beta} x) \cdot \hat{\beta}_2
\]

- Probit:
  - Effect → 0 as \( x \hat{\beta} \rightarrow \infty \) or \(-\infty \)
  - Effects vary
- Logit:
  - Effect → max as \( x \hat{\beta} \rightarrow 0 \)
  - So with all \( x \)

\[
\hat{p}, -\hat{p}_0 \text{ for given } x_0 \text{ and } x_1
\]

(First diffs)

Standard Error of Effects:

Delta Method:
\[
V(f(\hat{\beta})) \approx \left[f'(\hat{\beta})\right]^T \cdot V(\beta) \cdot \left[f'(\hat{\beta})\right]
\]

- e.g. Probit:
  \[
  \left[\nabla_\beta (\Phi(\hat{\beta} x) \hat{\beta})\right]^T \cdot V(\beta) \cdot \left[\nabla_\beta (\Phi(\hat{\beta} x) \hat{\beta})\right]
  
\]

Or simulate:
\text{gking.harvard.edu/docs/clarify.html}
I. General Background

As always, our goal is to use history (what has happened) as data, a database from which to infer the utility of (partial) validity of our theories or models about relationships among variables.

B. As an interesting aside, we might note that practically all empirical analysis in the social sciences, quantitative or qualitative, has this same goal. (One relatively rare exception is the occasional experimental analysis.)

1. We might note that anything that can be observed can be measured -- at least on a "present/not present" basis.

2. Finally, we might go still further and point out that the issues regarding this measurement & the inference from history to our theories remain the same whether we address them formally or not, whether we acknowledge them or not, whether we realize that any time we make any statement regarding evidence relating to our theory, we have resolved, often by assumption, implicit or explicit, all the logical issues of inference, or all the empirical issues regarding measurement, that relate to that statement.

Statistical inference, then, has the advantage of being greater explicitness about these necessary assumptions of grounding consistent proven, logically immutable rules, for making evidential inference given the data.

A large part of its "weakness" is often said to be that it provides, in its rules of the "over-simplistic" or its usually maintained assumptions, but...

a) "rigidity" in the rules of inference (4a) in parallel, in the rules of evidence, is paramount to objectivity itself, and is vital if evidence are to be used. We may suggest different sets of rules, but no one could seriously propose a set of rules that will not be followed... rather, this is said to be the problem of rigidity, coupled with overly simple assumptions, but...

b) it does not follow from the statement that one set of assumptions is too simple that we need not make assumptions. In fact, we are unwilling to assume that each observation or event, in some slice of history, is independent; it does not follow that we can drop that assumption without replacement. If we use several events as evidence for or against some theory, whether quantitatively or qualitatively, we are making some assumption about the representativeness of these events of about their weight relative to each other & to other evidence. In statistical inference, these assumptions are known (often simple) -- in qualitative inference, these assumptions may be explicit or implicit -- they are, in either case, usually more simplistic, (e.g., of the form: this case is representative & conclusive -- the latter implying no other case is representative, & has any weight of evidence).

4. More correctly, then, the limitation in statistical inference is not at all in its use of evidence, but rather, in its careless application and/or in the potential for its users to forget that any amount of evidence is useless without accompanying interpretation. Correctly establishing on empirical relationship that may help us understand our world only when coupled with a theory which interprets why it has held. A corollary, then...
is that all empirical work is both quantitative and qualitative. To the degree that we make use of evidence, we are at least implicitly applying statistical inference (I am unaware of any elaborated alternative). To the degree we are interpreting that evidence as something not only with raw historical but rather also social theoretical meaning, we apply qualitative analysis.

Upshots:

1. There is no quantitative/qualitative divide (may be a positivist/interpretivist divide); properly viewed, they are complementary.

2. Anything that can be observed can be measured. There is nothing positive which cannot, at least in principle, be subjected to quantitative analysis. (It may not always be tremendously useful.)

3. Avoiding quantitative analysis does not avoid any of the issues of inference or measurement involved in evidentiary analysis.

4. Inference itself is only half the exercise -- the other half is substantive interpretation of the inferential relationships.

5. We see then that this course fits into our broader training at any number of points.

C. Multiple Regression is simply the straightforward extension of the C(NLRM) in the bivariate case (last week) to the case where several independent variables (X's) are related to a dependent variable (Y).

1. As before, our model of the world begins with

\[ Y = f(X, \varepsilon) \]

where outcome \( Y \) is some function of some set of random elements the stochastic part of the world and some set of variables the systematic part of the explanatory variables of the world.

2. As before, we then proceed to restrict our attention (for now) to a class of functional relationships, \( Y = f(\cdot) \) which satisfy certain simplifying restrictions or assumptions. (The C(N)LRM)

3. We can then show, under these assumptions, how to estimate certain parameters of interest in the model using OLS, b) that these OLS estimates are BLUE, c) how to make statistical inferences using these estimates, and d) (most important) how to interpret these estimates -- what do they mean?

4. We then "clean up" with a few remaining issues.
II. The (Multivariate) Classical (Normal) Linear Regression Model

A. Assumptions (all are perfectly analogous to last week's) CLRM

1. \( Y = X \beta + \epsilon \)

2. \( E(\epsilon) = 0 \)

3. \( V(\epsilon) = \sigma^2 I_n \)

4. \( E(\epsilon | X) = 0 \)

5. \( X \) is non-stochastic (full column-rank)

3. \( Y \) is not perfectly linearly dependent

B. What do they mean? (\( n = \# \) of observations, \( n \times n \) ; \( k = \# \) of values, \( k \times k \))

1. We are assuming that \( Y \) is distributed somehow, \( Y = f(X, \epsilon) \) so that the expected value of \( Y \) (the best prediction) given \( X \) is a linear function

   \[
   E(Y | X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon
   \]

   or equivalently, \( E(Y | X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + E(\epsilon) \)

2. \( E(Y | X) = X\beta \) equivalently \( E(\epsilon) = 0 \)

   - the stochastic term, \( \epsilon \), has mean zero (for all its elements)

   \[
   E(\epsilon) = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0
   \]

3. \( V(\epsilon) = \begin{bmatrix} Cov(\epsilon_i, \epsilon_i) & \cdots & Cov(\epsilon_i, \epsilon_n) \\ \vdots & \ddots & \vdots \\ Cov(\epsilon_n, \epsilon_i) & \cdots & Cov(\epsilon_n, \epsilon_n) \end{bmatrix} = \sigma^2 I_n \)

   \( \Rightarrow \) Variance of each \( \epsilon_i \) (of each \( i \)) is some constant \( \sigma^2 \)

   (homoskedasticity)

   \( \Rightarrow Cov(\epsilon_i, \epsilon_j) = 0 \) \( \forall i \neq j \)

   (b.e.s.i. no \( \epsilon_i \) tells us anything about \( \epsilon_j \)

   or given \( x_i \) & \( x_j \) no \( \epsilon_i \) tells us anything

   (i.e. \( \epsilon_i \) & \( \epsilon_j \) more about \( y_i \) (b.e.s.i. b.s.i.)

   (uncorrelatedness)
The CLRM Assumptions' Meanings

4. $E(\varepsilon|X) = 0$ : once the systematic relationship between $E(y)$ if $X$ is netted out, $X$ has nothing more to tell us about $\varepsilon$.

5. $X$ is non-stochastic (this can be relaxed) if no column of $X$ is a perfect linear combination of the others.

### C. Typical Violations & Some Brief Intuition of Why We Need the Assumptions

1. $Y = X\beta + \varepsilon$ : relationship of $Y$ to $X$'s is non-linear in the parameters $\beta$.
   - Recall that we can include in $X$ d or transform $Y$ in a large manner of ways to make the equation non-linear in $X$.
   - We shall be able to apply the CLRM. It is $\beta$ which must enter the equation linearly.

2. $E(\varepsilon) = 0$ : If we mis-specify the $E(Y|X)$ part, it is possible that $E(\varepsilon) \neq 0$. But, if linearity holds, then including a constant assures us that $E(\varepsilon) = 0$. This is not problematic unless $(\varepsilon)$ is or unless we fail to include a constant.

3. $V(\varepsilon) = \sigma^2I_n$ : heteroskedasticity, $\sigma^2$ is not a constant.
   - usual data is plagued by the shadow of the past.
   - correlated residuals.

4. $E(\varepsilon|X) = 0$ : the most typical violations involve
   - Self-selection of observations into the sample.
   - Reverse Causality (Endogeneity, Simultaneity) (disjoint sets).
   - Some omitted factor causes both $Y$ and $X$.

   Basically, the assumption is violated whenever $X$ contains information about the distribution of $Y$ which goes beyond information about $E(Y)$.

5. $X$ non-stochastic, $X$ is the function of $X_{n \times p}$, non-stochastic $X$ means we treat it as fixed in repeated samples.

   No variable, $X$ (column of $X$) is an exact linear function of others.

   Why not?

   Suppose $X = 3X_2$, then suppose $b_1 = 1$ and $b_2 = 0$ is an optimal solution, so too then is $b_0 = 3$, $b_2 = 0$, and in fact, so is any pair of $b_i = b_2$ along the entire line ($-\infty$ to $+\infty$) given by the points $(0, 1)$ and $(3, 0)$.

   Also (why not?) (Same thing by geometry of Matrix Algebra):

   If some $X_0$ is an exact linear function of $X$, then $\text{rank}(X) < k \Rightarrow \text{rank}(X'X) < k \Rightarrow \det(X'X) = 0$.

   $\Rightarrow (X'X)^{-1}$ does not exist.

   - Most common violations (all are researcher dumb mistakes, type stuff, etc., around $Y$):
     - Type same variable name twice when giving computer program independent list.
     - Some set of $X$'s sum to 1 (e.g., set of dummies, % breakdowns of something)
     - $X$'s are perfectly collinear with the constant.

   - Resolution: leave one out (more than a couple weeks).
II. GLS Estimation of the CLRM in Multivariate Case (Again)

A. As always, we want to estimate \( \beta \) (the relationship between \( X \)'s and \( Y \)) by finding \( \hat{\beta} \) that minimizes the distance from the hyperplanes (or planes only in many dimensions) given by \( X \) to the lines given by \( \hat{\beta} \). (This is "projection".) Equivalently, we find the \( \hat{\beta} \) that minimizes the sum of squared distances from the \( X_i \) 's to the \( \hat{Y}_i \) 's:

\[
\min_{\hat{\beta}} \sum \hat{e}_i^2 \quad \text{where} \quad e_i = Y_i - X_i \hat{\beta}
\]

By definition of least squares:

\[\min_{\hat{\beta}} \mathbf{e}^T \mathbf{e} \quad \text{where} \quad \mathbf{e} \text{ is the } n \times 1 \text{ vector of the } e_i \text{'s}
\]

Writing out:

\[
\min_{\hat{\beta}} \left( Y - X \hat{\beta} \right)^T \left( Y - X \hat{\beta} \right) = \min_{\hat{\beta}} \left( Y - Y \hat{\beta} - X \hat{\beta} + X \hat{\beta} \right)^T \left( Y - Y \hat{\beta} - X \hat{\beta} + X \hat{\beta} \right)
\]

\[
= \min_{\hat{\beta}} \left( \sum (Y - X \hat{\beta})^T (Y - X \hat{\beta}) \right) = \min \left( \sum \hat{e}_i^2 \right)
\]

Multiplying out,

\[
= \min_{\hat{\beta}} \left\{ \sum (Y - X \hat{\beta})^T (Y - X \hat{\beta}) \right\} = \min \left( \sum \hat{e}_i^2 \right)
\]

First-order condition:

\[
\frac{\partial}{\partial \hat{\beta}} \hat{e}_i^2 = \frac{\partial}{\partial \hat{\beta}} \left( \sum (Y - X \hat{\beta})^T (Y - X \hat{\beta}) \right) = 0
\]

\[
-2 \sum \left( (X'Y)_{i} - (X'X)_{ii} \hat{\beta} + (X'X)_{ii} \hat{\beta} \right) = 0
\]

\[
-2X'Y + 2X'X \hat{\beta} = 0 \Rightarrow -X'Y + X'X \hat{\beta} = 0
\]

Put \( \hat{\beta} = (X'X)^{-1} X'Y \)

B. What did we just do? We had some function of \( \hat{\beta} \) named \( e_i \) and the same question as \( \frac{\partial e_i}{\partial \hat{\beta}} \leq 0 \) specifically, does the curve up or down at \( \hat{\beta} \) and

\[
\text{We found some } \hat{\beta}^* \text{ for which } e_i \text{ is the sum of squared residual (squared) had slope (derivative) equal to zero, but this could be a maximum of min for } e_i \text{ (see figure) depending on curvature.}
\]
III. A. Estimating $b^*$

$$\frac{\partial^2 \mathcal{L}}{\partial b \partial b^*} < 0$$

B. but is it a maximum or minimum?

$$\frac{\partial^2 \mathcal{L}}{\partial b^* \partial b^*} < 0 \Rightarrow \text{maximum}$$

rate of change of the rate of change, evaluated at $b^*$

$$\frac{\partial^2 \mathcal{L}}{\partial b \partial b^*} > 0 \Rightarrow \text{minimum}$$

$$\frac{\partial \mathcal{L}}{\partial b^*} = -X'Y + (X'X)b^*$$

$$\Rightarrow \frac{\partial^2 \mathcal{L}}{\partial b \partial b^*} = \frac{\partial}{\partial b^*} \left( \frac{\partial \mathcal{L}}{\partial b} \right) = (X'X)$$

rotational definition

Is this "$> 0$"? In matrices that translates to is $(X'X)$ positive definite? Which means is $v'(X'X)v > 0$ for any conformable vector, $v$?

Let $q \equiv Xv$, then this can be written $q'q$ which equals $\sum q_i^2$ which is greater than zero for any $q$.

Thus for any $X$ (of full rank), $X'X$ is positive definite, so among a lot of other things

$$\frac{\partial^2 \mathcal{L}}{\partial b \partial b^*} |_{b^*} > 0$$

so

$$b_{LS} = (X'X)^{-1}X'Y$$

minimizes $\sum e_i^2$ (the sum of squared residuals).

C. As per usual, we could solve all this out from the "normal equations":

$$(X'X)_b = X'Y$$

what are elements of $X'X$?

$$\begin{bmatrix} X'X_1 & X'X_2 & \cdots & X'X_k \\ X_1'X_1 & X_1'X_2 & \cdots & X_1'X_k \\ \vdots & \vdots & \ddots & \vdots \\ X_k'X_1 & X_k'X_2 & \cdots & X_k'X_k \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} X_1'Y \\ X_2'Y \\ \vdots \\ X_k'Y \end{bmatrix}$$

Each of these is a column of $X$ (a vector).

$$(k \times k)$$

$(k \times 1)$
III. Calculating out the Scalar Representations of the $b_{LS}$ estimator:

considering the 3 variable case:

\[
\begin{bmatrix}
X_1 & X_2 & X_3 \\
X_2^T & X_2 & X_3 \\
X_3^T & X_2 & X_3
\end{bmatrix}
\begin{bmatrix}
b_1 \\
b_2 \\
b_3
\end{bmatrix}
= \begin{bmatrix}
X_1' \cdot X_1 \\
X_2' \cdot X_2 \\
X_3' \cdot X_3
\end{bmatrix}
\begin{bmatrix}
y_1 \\
y_2 \\
y_3
\end{bmatrix}
\]

(3x3)  
(3x1)  
(3x1)

Multiply this out:

\[\Rightarrow X_1' \cdot X_1 \cdot b_1 + X_1' \cdot X_2 \cdot b_2 + X_1' \cdot X_3 \cdot b_3 = X_1' \cdot y\]

\[\Rightarrow X_2' \cdot X_1 \cdot b_1 + X_2' \cdot X_2 \cdot b_2 + X_2' \cdot X_3 \cdot b_3 = X_2' \cdot y\]

\[\Rightarrow X_3' \cdot X_1 \cdot b_1 + X_3' \cdot X_2 \cdot b_2 + X_3' \cdot X_3 \cdot b_3 = X_3' \cdot y\]

Now, in almost all cases, to illustrate here, $X_1$ is the constant, a vector of ones, so $X_1' \cdot X_1 = N$, $X_2' \cdot X_1 = X_1^2$, etc.

\[\Rightarrow n \cdot b_1 + b_2 \cdot \Sigma X_2 + b_3 \cdot \Sigma X_3 = \Sigma y\]

\[\Rightarrow b_1 = \bar{y} - b_2 \cdot \bar{X}_2 - b_3 \cdot \bar{X}_3\]

\[\Rightarrow b_2 \cdot \Sigma X_2 + b_3 \cdot \Sigma X_3 = \Sigma (X_2 \cdot y)\]

\[\Rightarrow b_2 \cdot \Sigma X_2 + b_3 \cdot \Sigma (X_2 \cdot y) = \Sigma (X_2 \cdot y)\]

Now, putting 0 into 0 and 0:

\[\Rightarrow \Sigma X_2 - b_2 \cdot \bar{X}_2 \cdot \Sigma X_2 = \Sigma X_2 - \Sigma X_2 - b_3 \cdot \Sigma X_3 = \Sigma X_3 - \Sigma X_3\]

\[\Rightarrow \Sigma X_2 - \Sigma X_2 = \Sigma X_3 - \Sigma X_3\]

Similarly for 0:

\[\Rightarrow \Sigma X_2 - \Sigma X_2 = \Sigma X_3 - \Sigma X_3\]

One more step, putting 0 into 0 or 0:

First define $X_2^*$ as $X_2 - \bar{X}_2$ and so on (deviations from sample mean).

\[b_2 = \frac{(\Sigma X_2 \cdot y^*)}{\Sigma (X_2^*)} = \frac{(\Sigma X_2 \cdot y^*)}{\Sigma (X_2^*)} - \frac{(\Sigma X_2 \cdot y^*)}{\Sigma (X_2^*)}\]

\[b_3 = \frac{(\Sigma X_2 \cdot y^*)}{\Sigma (X_2^*)} = \frac{(\Sigma X_2 \cdot y^*)}{\Sigma (X_2^*)} - \frac{(\Sigma X_2 \cdot y^*)}{\Sigma (X_2^*)}\]

So, noticing that $X_2^* \cdot y^*$ is in times the sample covariance of $X_2 \cdot y$, etc.

\[b_2 = \frac{S_{x_2 \cdot y}}{S_{x_2}^2} - \frac{S_{x_2 \cdot y} \cdot S_{x_2 \cdot x_3}}{S_{x_2}^2 \cdot S_{x_3}^2} - \frac{(S_{x_2 \cdot y})^2}{S_{x_2}^2 \cdot S_{x_3}^2} - \frac{(S_{y \cdot x_2 \cdot x_3})}{S_{x_2}^2 \cdot S_{x_3}^2}\]

\[b_3 = \frac{S_{x_2 \cdot y}}{S_{x_2}^2} - \frac{S_{x_2 \cdot y} \cdot S_{x_3}}{S_{x_2}^2 \cdot S_{x_3}^2} - \frac{(S_{y \cdot x_2 \cdot x_3})}{S_{x_2}^2 \cdot S_{x_3}^2}\]

\[\Rightarrow 5 \text{ factors matter for our estimation of } b_2 \text{ and } b_3: \]

\[S_{x_2}^2, S_{x_3}^2, \frac{v(x_2 \cdot y)}{v(x_2)}, \frac{v(x_2 \cdot y)}{v(x_3)}\]

\[S_{x_2} \cdot S_{x_3}, \frac{\text{Cov}(X_2, y)}{\text{Cov}(X_2, y)}\]

\[\& \frac{S_{x_2 \cdot x_3}}{\text{Cov}(X_2, X_3)}\]
### III.C. Hand Calculations of OLS in Matrix Algebra

Recall that the OLS of $\mathbf{Y}$ on $X_1, X_2 \Rightarrow b_2 = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_2)}$

1. **Bivariate regression** on $X_1$ gives:
   \[ b_{1x_1} = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)} \]

2. **Bivariate regression** on $X_2$ gives:
   \[ b_{2x_2} = \frac{\text{Cov}(X_2, Y)}{\text{Var}(X_2)} \]

Then:

\[ b_{2,x_2,x_1} = \frac{b_{2x_2} - b_{1x_1} \cdot b_{1x_1}}{1 - b_{1x_1}^2} \]

### D. Some More Biproducts of OLS Estimators

1. $(X'X)b - X'\mathbf{y} = -X'(\mathbf{y} - Xb) = -X'e = O$

   \[ \Rightarrow X'e = O \quad \forall \text{columns of } X \]

   (i.e., this is true for every variable)

   Since $X_i$ is constant, this $\Rightarrow \mathbf{e}'e = 0 \Rightarrow \Sigma e_i^2 = 0$

2. $\bar{Y} = \bar{X}'b$, i.e., the regression line passes through the sample means of the $X$'s and $\bar{Y}$

   \[ \Rightarrow \bar{Y} = \bar{Y} \quad \text{i.e., the average predicted value equals the average observed value} \]
Calculations of $b_{LS}$ without (much) Matrix Algebra

$$\min_{\alpha, \beta_1, ..., \beta_k} \sum_{i} e_i^2 \equiv \min_{\beta} e'e = \min_{\beta} (y - X\hat{\beta})' (y - X\hat{\beta})$$

$$\Rightarrow \frac{\partial e'e}{\partial \hat{\beta}} = 0 \Rightarrow X'X\hat{\beta}^* = X'y$$

now, notice that, for $X$ 3x3, for example, we have:

$$X'X = \begin{bmatrix}
    x'_1 x_1 & x'_1 x_2 & x'_1 x_3 \\
x'_2 x_1 & x'_2 x_2 & x'_2 x_3 \\
x'_3 x_1 & x'_3 x_2 & x'_3 x_3
\end{bmatrix},$$

so, we can rewrite:

$$\begin{bmatrix}
x'_1 x_1 & x'_1 x_2 & x'_1 x_3 \\
x'_2 x_1 & x'_2 x_2 & x'_2 x_3 \\
x'_3 x_1 & x'_3 x_2 & x'_3 x_3
\end{bmatrix}\begin{bmatrix}
    \hat{\beta}_1^* \\
    \hat{\beta}_2^* \\
    \hat{\beta}_3^*
\end{bmatrix} = \begin{bmatrix}
x'_1 y \\
x'_2 y \\
x'_3 y
\end{bmatrix},$$

which in scalar notation is:

(1) \hspace{1cm} x'_1 x_1 \hat{\beta}_1^* + x'_1 x_2 \hat{\beta}_2^* + x'_1 x_3 \hat{\beta}_3^* = x'_1 y

(2) \hspace{1cm} x'_2 x_1 \hat{\beta}_1^* + x'_2 x_2 \hat{\beta}_2^* + x'_2 x_3 \hat{\beta}_3^* = x'_2 y

(3) \hspace{1cm} x'_3 x_1 \hat{\beta}_1^* + x'_3 x_2 \hat{\beta}_2^* + x'_3 x_3 \hat{\beta}_3^* = x'_3 y
Now, usually first column (variable) is $\mathbf{x}_1 = [\cdots 1 \cdots]$, which implies:

\[ x'_1 x_1 = \sum_{i=1}^n (1 \times 1) = n, \text{ and } x'_1 x_2 = \sum_{i=1}^n (1 \times x_{2i}) = \sum_{i=1}^n x_2 \text{ and } x'_1 x_3 = \sum_{i=1}^n (1 \times x_{3i}) = \sum_{i=1}^n x_3 \]

Using this, we can simplify (1), (2), (3) as…

\[
(4) \quad n \hat{\beta}_1^* + \hat{\beta}_2^* \sum x_2 + \hat{\beta}_3^* \sum x_3 = \sum y \\
(5) \quad \hat{\beta}_1^* \sum x_2 + x'_2 x_2 \hat{\beta}_2^* + x'_2 x_3 \hat{\beta}_3^* = x'_2 y \\
(6) \quad \hat{\beta}_1^* \sum x_3 + x'_3 x_2 \hat{\beta}_2^* + x'_3 x_3 \hat{\beta}_3^* = x'_3 y
\]

Now, let’s finish converting to scalar notation:

\[
(7) \quad n \hat{\beta}_1^* + \hat{\beta}_2^* \sum x_2 + \hat{\beta}_3^* \sum x_3 = \sum y \\
(8) \quad \hat{\beta}_1^* \sum x_2 + \hat{\beta}_2^* \sum x_2^2 + \hat{\beta}_3^* \sum x_2 x_3 = \sum x_2 y \\
(9) \quad \hat{\beta}_1^* \sum x_3 + \hat{\beta}_2^* \sum x_2 x_3 + \hat{\beta}_3^* \sum x_3^2 = \sum x_3 y
\]

Now, divide (7) through by $n$ to get the solution for $\hat{\beta}_1$:

\[
(10) \quad \hat{\beta}_1^* = \bar{y} - \hat{\beta}_2^* \bar{x}_2 - \hat{\beta}_3^* \bar{x}_3
\]

**Upshots:**

The intercept, $\hat{\beta}_1^*$, is the conditional (sample) mean of $y$.

The regression line passes through the (multidimensional) mean of the data (which is called the centroid of the data).
Now, substitute (10) into (8) or (9), let’s use (8). ⇒ (11) substituting:

\[
\left(\bar{y} - \hat{\beta}_2^* \bar{x}_2 - \hat{\beta}_3^* \bar{x}_3\right) \sum x_2 + \hat{\beta}_2^* \sum x_2^2 + \hat{\beta}_3^* \sum x_2 x_3 = \sum x_2 y
\]

(multiplying) ⇒ \[
\left(\bar{y} \sum x_2 - \hat{\beta}_2^* \bar{x}_2 \sum x_2 - \hat{\beta}_3^* \bar{x}_3 \sum x_2\right) + \hat{\beta}_2^* \sum x_2^2 + \hat{\beta}_3^* \sum x_2 x_3 = \sum x_2 y
\]

(since \(\sum x_2 = n \bar{x}_2\) ) ⇒ \[
\left(\bar{y} n \bar{x}_2 - \hat{\beta}_2^* n \bar{x}_2 - \hat{\beta}_3^* n \bar{x}_3 \right) + \hat{\beta}_2^* \sum x_2^2 + \hat{\beta}_3^* \sum x_2 x_3 = \sum x_2 y
\]

(rearranging) ⇒ \[
\hat{\beta}_2^* \left(\sum x_2^2 - n \bar{x}_2^2\right) + \hat{\beta}_3^* \left(\sum x_2 x_3 - n \bar{x}_3 \bar{x}_2\right) = \sum x_2 y - n \bar{x}_2 \bar{y}
\]

⇒ \[
\hat{\beta}_2^* \times S_{x_2}^2 + \hat{\beta}_3^* \times S_{x_2 x_3} = S_{x_2 y}
\]

...by definitions of sample variations & covariations

(solving for \(\hat{\beta}_2^*\)) ⇒ \[
\hat{\beta}_2^* = \frac{S_{x_2 y} - \hat{\beta}_3^* \times S_{x_2 x_3}}{S_{x_2}^2}
\]

Which can also be rewritten (note the change in subscripting notation!):

\[
\Rightarrow \hat{\beta}_{2,3}^* = \frac{S_{x_2 y}}{S_{x_2}^2} - \frac{\hat{\beta}_{3,2}^* \times S_{x_2 x_3}}{S_{x_2}^2} = \hat{\beta}_2 - \hat{\beta}_{3,2}^* \frac{S_{x_2 x_3}}{S_{x_2}^2} = \hat{\beta}_2 - \hat{\beta}_{3,2}^* \times \hat{\beta}_{x_2 x_3}
\]

And, by analogy: (12) \[
\hat{\beta}_{3,2} = \hat{\beta}_3 - \hat{\beta}_{2,3}^* \times \hat{\beta}_{x_2 x_3}.
\]

The notation \(2,3\), etc., serves to underscore that these are the optimal estimates of the conditional or partial coefficients on \(x_2\) from a regression including, i.e., controlling for, \(x_3\), and *vice versa*. The notation \(x_2 x_3\), etc., means from a regression of \(x_2\) on \(x_3\). Finally, the lack of a * on the bivariate coefficients is to stress that they are from bivariate (unconditional) regression equations, not from this multivariate regression.
Working now from the more-elaborate expressions (1)-(3) in your notes:

1. \[ b_1^* = \bar{y} - b_2^* \bar{x}_2 - b_3^* \bar{x}_3 \]

2. \[ b_2^* = \frac{(\Sigma x_2^* y^*) (\Sigma (x_3^*)^2) - (\Sigma x_3^* y^*) (\Sigma x_2^* x_3^*)}{(\Sigma (x_2^*)^2)(\Sigma (x_3^*)^2) - (\Sigma x_2^* x_3^*)^2} = \frac{S_{x_2 y} S_{x_2}^2 - S_{x_3 y} S_{x_2 x_3}}{S_{x_2}^2 S_{x_3} - (S_{x_2 x_3})^2} \]

3. \[ b_3^* = \frac{(\Sigma x_3^* y^*) (\Sigma (x_2^*)^2) - (\Sigma x_2^* y^*) (\Sigma x_2^* x_3^*)}{(\Sigma (x_3^*)^2)(\Sigma (x_2^*)^2) - (\Sigma x_2^* x_3^*)^2} = \frac{S_{x_3 y} S_{x_2}^2 - S_{x_2 y} S_{x_2 x_3}}{S_{x_2}^2 S_{x_3} - (S_{x_2 x_3})^2} \]

First, the notational shift here from \( \hat{\beta} \) to \( b \) signifies nothing; it serves here only to match the notation from the notes for the versions of the expressions for the same quantities from which we are beginning.

To get to these expressions from (11) and (12), you must solve out the right-hand-side \( \hat{\beta} \)’s in those expressions. Doing so, you get:
\[ \hat{\beta}^{*}_{2.3} = \frac{S_{x_2 y}}{S_{x_2}^2} - \frac{\hat{\beta}_{3.2} \times S_{x_2 x_3}}{S_{x_2}^2} = \frac{S_{x_2 y}}{S_{x_2}^2} - \frac{\left( \hat{\beta}_{3} - \hat{\beta}^{*}_{2.3} \times \hat{\beta}_{x_2 x_3} \right) \times S_{x_2 x_3}}{S_{x_2}^2} \]

\[ \Rightarrow S_{x_2}^2 \times \hat{\beta}^{*}_{2.3} = S_{x_2 y} - \left( \frac{S_{x_3 y}}{S_{x_3}^2} - \hat{\beta}^{*}_{2.3} \times \frac{S_{x_2 x_3}}{S_{x_3}^2} \right) \times S_{x_2 x_3} \]

\[ \Rightarrow S_{x_2}^2 \times \hat{\beta}^{*}_{2.3} = S_{x_2 y} - \left( \frac{S_{x_3 y}}{S_{x_3}^2} \frac{S_{x_2 x_3}}{S_{x_3}^2} - \hat{\beta}^{*}_{2.3} \times \frac{\left( S_{x_2 x_3} \right)^2}{S_{x_3}^2} \right) \times S_{x_2 x_3} \]

\[ \Rightarrow S_{x_3}^2 \times S_{x_2}^2 \times \hat{\beta}^{*}_{2.3} = S_{x_3}^2 \times S_{x_2 y} - S_{x_3 y} \frac{S_{x_2 x_3}}{S_{x_3}^2} + \hat{\beta}^{*}_{2.3} \times \left( S_{x_2 x_3} \right)^2 \]

\[ \Rightarrow \left\{ S_{x_3}^2 \frac{S_{x_2}^2}{S_{x_2}} - \left( S_{x_2 x_3} \right)^2 \right\} \times \hat{\beta}^{*}_{2.3} = S_{x_3}^2 \times \frac{S_{x_2 y}}{S_{x_2}} - S_{x_3 y} \frac{S_{x_2 x_3}}{S_{x_2}} \]

\[ \Rightarrow \hat{\beta}^{*}_{2.3} = \frac{S_{x_3}^2 \frac{S_{x_2 y}}{S_{x_2}} - S_{x_3 y} \frac{S_{x_2 x_3}}{S_{x_2}}}{S_{x_3}^2 \frac{S_{x_2}^2}{S_{x_2}} - \left( S_{x_2 x_3} \right)^2} \]

Which was (2) as we had it before. Now, recall:

\[ b_2 = \frac{\text{Cov}(X_2, Y)}{\text{Var}(X_2)} = \frac{S_{x_2 y}}{S_{x_2}^2} \]
So, \( \div \) numerator & denom of (2) & (3) by \( S_{x_2}^2 \) and \( S_{x_3}^2 \):

\[
\hat{\beta}_{2.3}^* = \frac{S_{x_3}^2 S_{x_2 y} / S_{x_3}^2 S_{x_2}^2 - S_{x_3 y} S_{x_2 x_3} / S_{x_3}^2 S_{x_2}^2}{S_{x_2}^2 S_{x_3}^2 / S_{x_3}^2 S_{x_2}^2 - (S_{x_2 x_3} / S_{x_3}^2 S_{x_2}^2)^2} = \frac{S_{x_2 y} / S_{x_2}^2 - S_{x_3 y} S_{x_2 x_3} / S_{x_3}^2 S_{x_2}^2}{1-r_{x_2 x_3}^2}
\]

(2') \( b_{2.3}^* = \frac{b_2 - b_3 b_{x_3 x_2}}{1-r_{x_2 x_3}^2} \) \{ \text{where } \begin{align*}
& \ b_{2.3} \text{ is coeff reg } y \text{ on } x_2 \text{ in presence } x_3; \\
& \ b_{x_2 x_3} \text{ is coeff reg } x_2 \text{ on } x_3; \\
& \text{and } b_2 \text{ is bivariate reg coeff}.
\end{align*} \}

(3') \( b_{3.2}^* = \frac{b_3 - b_2 b_{x_3 x_3}}{1-r_{x_2 x_3}^2} \)

(We return to the original \( b \) notation here at the end to reflect from where we began these calculations.)

**Upshot:** whether by (11) & (12) or by (2') & (3'), **the multivariate, conditional coefficients equal the bivariate coefficients if either (i) \( x_2 \) & \( x_3 \) do not covary or (ii) \( x_3 \) does not affect \( y \).**

You might notice here, though, that the expressions for \( \hat{\beta}_{2.3}^* \) & \( b_{2.3}^* \) **seem** to differ & contradict, even though they are the same entity: the partial coefficient on \( x_2 \):

\[
b_{2.3}^* = \frac{b_2 - b_3 b_{x_3 x_2}}{1-r_{x_2 x_3}^2} \text{ vs. } \hat{\beta}_{2.3}^* = \hat{\beta}_2 - \hat{\beta}_{3.2}^* \times \hat{\beta}_{x_3 x_2}
\]

Latter says multivariate, partial coefficient equals the
bivariate (unconditional) coefficient if either (i) $x_2 \& x_3$ do not covary or (ii) $x_3$ does not affect $y$. The former seems to agree regarding (i) but not (ii). Even if $x_3$ does not affect $y$, non-zero $r_{x_2 x_3}^2$ could still imply $b_{2.3}^* \neq b_2$. The seeming contradiction, though, is actually not one. Note the different second (numerator) term, $b_3 \text{ vs. } \hat{\beta}_{3,2}^{\ast}$; the different denominators, $1-r_{x_2 x_3}^2 \text{ vs. } 1$, exactly cancel this difference in the numerators.

So, what the apparent difference in the expressions does is clarify and underscore how we should understand the verbal descriptions of the conditions under which unconditional and partial coefficients equate. Namely:

(i) $x_2 \& x_3$ do not covary means they do not covary, unconditionally; whereas:

(ii) $x_3$ does not affect $y$ means $x_3$ does not covary with $y$, conditionally; i.e., the partial coefficient on $x_3$ is (would be) zero; i.e., $x_3$ does (would) not matter, controlling for $x_2$. 

Showing the equivalence of the two expressions:

\[ b_{2.3}^* = \frac{b_2 - b_3 b_{x_3 x_2}}{1 - r_{x_2 x_3}^2} = \hat{\beta}_{2.3}^* = \hat{\beta}_2 - \hat{\beta}_{3.2}^* \times \hat{\beta}_{x_3 x_2} \]

...first unify the notation:

\[ \hat{\beta}_{2.3}^* = \frac{\hat{\beta}_2 - \hat{\beta}_3 \hat{\beta}_{x_3 x_2}}{1 - r_{x_2 x_3}^2} = \hat{\beta}_2 - \hat{\beta}_{3.2}^* \times \hat{\beta}_{x_3 x_2} \]

...then substitute analogous formula for \( \hat{\beta}_{3.2}^* \):

\[ \hat{\beta}_{2.3}^* = \hat{\beta}_2 - \left( \hat{\beta}_3 - \hat{\beta}_{2.3}^* \times \hat{\beta}_{x_2 x_3} \right) \times \hat{\beta}_{x_3 x_2} \]

...multiply through:

\[ \hat{\beta}_{2.3}^* = \hat{\beta}_2 - \hat{\beta}_3 \times \hat{\beta}_{x_2 x_3} + \hat{\beta}_{2.3}^* \times \hat{\beta}_{x_2 x_3} \times \hat{\beta}_{x_3 x_2} \]

...gather terms:

\[ \hat{\beta}_{2.3}^* \left( 1 - \hat{\beta}_{x_2 x_3} \times \hat{\beta}_{x_3 x_2} \right) = \hat{\beta}_2 - \hat{\beta}_3 \times \hat{\beta}_{x_3 x_2} \]

...solve & simplify:

\[ \hat{\beta}_{2.3}^* = \frac{\hat{\beta}_2 - \hat{\beta}_3 \times \hat{\beta}_{x_2 x_3}}{1 - \frac{\text{Cov}(x_2, x_3)}{\text{Var}(x_3)} \times \frac{\text{Cov}(x_3, x_2)}{\text{Var}(x_2)}} = \frac{\hat{\beta}_2 - \hat{\beta}_3 \times \hat{\beta}_{x_2 x_3}}{1 - \frac{[\text{Cov}(x_2, x_3)]^2}{\text{V}(x_3) \text{V}(x_2)}} = \frac{\hat{\beta}_2 - \hat{\beta}_3 \times \hat{\beta}_{x_3 x_2}}{1 - r_{x_2, x_3}^2} \]
IV. Some Regression Matrix Algebra:

A. Defined. $Q \equiv (X'X)$
   1. $A \equiv Q^{-1}X'$
      - the least-squares coefficient-estimator matrix
   2. $N \equiv XA$
      - the projection matrix (least-squares prediction)
   3. $M \equiv I_n - N$
      - the residual-estimator matrix

B. Then,
   1. $A_y = \hat{b}_{LS}$
      - $A$ turns what it pre-multiplies into coeff. estimates
   2. $N_y = \hat{\gamma}_{LS}$
      - $N$ turns """" into fitted values
   3. $M_y = \hat{e}_{LS}$
      - $M$ """" into estimated residuals

C. Some Useful Facts:
   1a. $Q$ is symmetric ($\Longleftrightarrow Q^{-1}$ is symmetric)
   1b. $Q$ is positive definite ($\Longleftrightarrow Q^{-1}$ is positive definite)
   2a. $M$ is symmetric (so is $N$)
   2b. $M$ is idempotent ($MM = M$) (so is $N$)

There are a large number of other useful facts that follow from these definitions, but these will do for now.

D. Note: if $X$ is in (simple) mean-deviation form ($x^* = x - \bar{x}$), then $X'X$ is (in time) the variance-covariance matrix of the $x$'s. (If not, then it only "kinda" is, it also includes sum of $x^2$'s & sum of $x_i x_j$ terms)

V. Partial Regression & "Controlling for..."

A. $y = \alpha + \beta x_1 + \gamma x_2 + \epsilon$

   See Kennedy, Ch. 3

   The circles represent the total variation of $y$ & $x_1, x_2$

   - The variation of $x_1$ on $y$ which is not also attributable to $x_2$ is "blue"
   - The variation of $x_2$ on $y$ which is not also attributable to $x_1$ is "green"

   red area is impossible to distinguish whether to be attributed to $x_1$ or $x_2$

   in some mix of $x_1$ & $x_2$, share this variation fits relation to variation in $y$

   Multiple regression coefficients measure effects corresponding to "blue" and "green"; i.e. $\hat{b}_i$ estimates $b_i$ using only variation in $x_i$ relating to variation in $y$ which is not also related by $x_j$'s shared variation w/ $y$.

   * $R^2$ is "(red+blue+green)" whole circle; $\hat{R}^2$ is yellow area;
     - $V(x_1)$ is whole $x_1$ circle; $\text{Cov}(x_1, x_2)$ is red+orange; $\text{Cov}(x_1, y)$ is blue+red;
     - $V(x_2)$ is whole $x_2$ circle; $\text{Cov}(x_2, y)$ is red+green.
B. Another way to see what partial coefficients mean:

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

\[ \beta_1 = \frac{\partial Y}{\partial X_1}, \quad \text{i.e. change in } Y \text{ produced by a change in } X_1 \]

\[ \beta_2 = \frac{\partial Y}{\partial X_2}, \quad \text{analogous} \]

holding all else constant (ceteris paribus).

C. What's so important & powerful about multiple regression is that we can estimate the "ceteris paribus" coefficients, i.e. these effects holding all other X's constant even though these usually are no observations where one X changes and all others do not. Looking back at the "Ballentine", we can do this because we can find variance in \( x_1 \& x_2 \) that is shared (covariance) of base our estimates of effects on \( y \) on variance remaining after setting out this shared variance.

D.究竟 how it's done:

1. Break \( X \) into two sets of variables, \( X_1 & X_2 \):

\[ Y = X_1 \beta_1 + X_2 \beta_2 + \epsilon \]

\( X_1 \) is variables (columns) 1-\( i \) of \( X \)

\( X_2 \) are the remaining \( (i+1)-k \) columns (variables)

2. Recall the "normal equations":

\[ (X'X)_n = X'y \]

Now these can be "broken up" into:

\[ X_1'X_1 \beta_1 = X_1'y \]

\[ X_2'X_2 \beta_2 = X_2'y \]

3. By rules of inverting partitioned matrices (chapter 7):

\[ b_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2b_2 \]

\[ b_2 = \frac{X_2'y - A_1X_2b_2}{A_2} \]

(loos regression of \( y \) on \( X_1 \) only times coefficients on \( X_2 \) in whole regression

Subst. this into equation (b) \& you get:

\[ X_2'X_2 N_1 y - X_2'N_2 X_2b_2 + X_2'X_2 b_2 = X_2'y \]

\[ (\text{but } XA = N_1) \]

\[ X_2'X_2 N_1 y - X_2'N_2 X_2b_2 + X_2'X_2 b_2 = X_2'y \]

\[ (\text{gathering up } b_2 \text{ terms on LHS & } y \text{ term on RHS}) \]

\[ X_2'(I - N_1)X_2 b_2 = X_2'(I - N_1) y \]

\[ X_2'(M_1X_2) b_2 = X_2'M_1 y \]

\[ b_2 = (X_2'M_1X_2)X_2'M_1 y \]
5. Dr. "Partial" Regression Coefficients
\[ b_{a2} = \left( X_{2}^\top M_{2} X_{2} \right)^{-1} X_{2}^\top M_{2} Y \]

- Now, \( M_{2} \) is symmetric & idempotent so: \( M_{2} M_{2} = M_{2} \); replace \( M_{2} \) in above by \( M_{2}^2 \):
\[ b_{a2} = \left( X_{2}^\top M_{2}^2 M_{2} X_{2} \right)^{-1} X_{2}^\top M_{2}^2 M_{2} Y \]

(\( \mathbf{C} \)) \[ b_{a2} = \left( (M_{2} X_{2})^\top (M_{2} X_{2}) \right)^{-1} (M_{2} X_{2})^\top M_{2} Y \]

- \( M_{2} \) is "residual-maker" for regression of thing it multiplies on \( X_{2} \),
- and \( b_{a2} \) is coefficient from regression of \( (M_{2} Y) \)
on \( (M_{2} X_{2}) \) (equation \( \mathbf{C} \) is in \((Z'Z)^{-1}Z'Y\)
form, right? so it's an \( A \) matrix
but \( M_{2} Y \) are the residuals from \( Y \) regressed on \( X_{2} \).
- \( M_{2} X_{2} \) are the residuals from (each of) the \( X_{2}'s \) regressed on (all of) the \( X_{2}'s \).

Thus, \( b_{a2} \), the coefficients on \( X_{2} \), are also the coefficients on \( X_{2}^* \) when regressed on \( Y^* \), where the *'s mean that all (linear) relationships with \( X_{1} \) has been netted out. This is what it means to say "the effect(s) of \( X_{2} \), controlling for \( X_{1} \)."

Thus, in the simple \( Y = a + b_{1} X_{1} + b_{2} X_{2} + e \) case,

- \( b_{2} \) is the effect of \( X_{2} \) only, netting out all linear relationship(s) of \( Y \) with \( X_{1} \)
- (the constant) & the latter (two) with \( X_{2} \)

The analysis is completely symmetric, so

- \( b_{1} \) is the effect of \( X_{1} \) on \( Y \), netting out all linear relationships of \( Y \) with \( X_{2} \)
- (the constant) & the latter (two)

Most Generally, \( b_{j} \) is the effect of \( X_{j} \) on \( Y \), netting out the relationships of all \( X_{notj} \) with \( Y \) and \( X_{j} \)

E. It is straightforward from here that if we want partial correlations of \( X_{j} \) & \( Y \) (controlling for \( X_{notj} \)) we simply net out the linear relationships of \( X_{notj} \) with \( X_{j} \) & \( Y \) & calculate the correlations. Two shortcuts exist:

1. \( r_{yk}^2 = t_{k}^2 / (k + degrees of freedom) \)

- where \( t_{k} \) is t-stat from test of hypoth. that coeff on \( X_{k} \) = 0

2. regress \( Y \) on all \( X \), regress \( X_{k} \) on \( Y \) and all \( X_{notk} \)

- Product of coeff on \( X_{k} \) in 1st reg & on \( Y \) in second is partial correlation squared.
IV. Partial Coefficients—One Important Example

Suppose we consider $X_1$ to be just the constant (the vector of ones) and $X_2$ to be all the other variables. Then, just as before:

$$b_{\gamma} = \left[(M'X_2)(M'X_2)^{-1}M'y\right].$$

* $M$, here is the "residual maker" regressing what it multiplies on the constant only. What do you get regressing any vector $Z$ on a constant? A coefficient of $Z$. Why?

We did this in linear regression:

$$\min_{b} \sum_{i=1}^{n} (z_i - b^Tc)^2 = \min_{b} \sum_{i=1}^{n} (z_i - \bar{z})^2$$

So, then $M$, here, nets out the mean of $y$ and of all $X_2$ columns. Thus $b_{\gamma}$ in OLS multiple regression including a constant, nets out the means of all $X$'s and $Y$'s.

V. Statistical Properties of the OLS Estimates

A. Just as before, $b_{\gamma} = A_{\gamma} = A(X\beta + \varepsilon) = (X'X)^{-1}X'(X\beta + \varepsilon)$

$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon$$

$$= \beta + A\varepsilon$$

So $b_{\gamma}$ differs from $\beta$ only because of stochastic component, $\varepsilon$.

B. $b_{\text{ols}}$ unbiased: $E(b_{\gamma}) = E(\beta) + AE(\varepsilon) = \beta + A\cdot 0 = \beta$

C. $\text{Var}(b_{\gamma}) = \text{Var}(\beta + A\varepsilon)$

$$= A\text{Var}(\varepsilon)A'$$

$$= (X'X)^{-1}X'\text{Var}(\varepsilon)X(X'X)^{-1}$$

$$= \sigma^2 Q^{-1}Q Q^{-1}$$

$$= \sigma^2 Q^{-1}$$

Under CLRM assumptions:

$$\text{Var}(b_{\text{ols}}) = \sigma^2 (X'X)^{-1}$$

Under CLRM:

D. $b_{\text{ols}}$ is BLUE: $\text{Var}(b_{\text{ols}}) \leq \text{Var}(b^*)$ for any linear, unbiased $b^*$

* each $b_j$ is BLUE estimate of $\beta_j$

* for any $X' \beta$ where $W$ fixed set of linear coefficients, $W' b_{\text{ols}}$ is BLUE

* for any $X' \beta$ where $W$ fixed set of linear coefficients, $X' W b_{\text{ols}}$ is the BLUE linear combination of $b$'s is the BLUE

Proofs for all this in Greene; they are analogous to bivariate proofs.

One difference: for matrices, $A' \preceq C$ means $C - A$ positive definite.

Thus $\text{Var}(b_{\text{ols}}) \leq \text{Var}(b^*)$ means $\text{Var}(b^*) - \text{Var}(b_{\text{ols}})$ is positive semidefinite.
VI. Statistical Properties of OLS under CLRM Assumptions

E. Relaxing the fixed $X$ assumption:

1) $\hat{b} = \hat{\beta} + A \varepsilon \sim N(0, \sigma^2 T^{-1})$ still, only now $A = (X'X)^{-1}X'$ made up of stochastic $X'$

2) $E(b_{i|x}) = \beta + \text{E}[A \varepsilon | X] = \beta + \text{E}(X\varepsilon) = \beta$ (given the $X$'s are fixed)

3) $\text{Var}(b_{i|x}) = 0 \iff \sigma^2 = \text{Cov}(X \varepsilon)$ (giving the $X$'s are fixed)

4) $V(b_{i|x}) = \text{Var}(b_{i|x})$ is linear & unbiased, $\forall b_i \text{ linear & unbiased}$

VII. Statistical Inference from OLS estimates of CLRM

A. Introduce Assumption $6: \varepsilon \sim N(0, \sigma^2 I)$, equivalently $E(\varepsilon | X)$ is the $N(\mu, \sigma^2 I)$

2. Mean is $E(b_{i|x}) = \beta$, as we've already shown

3. Variance is $V(b_{i|x}) = 0 \iff \text{Cov}(X \varepsilon) = 0^2 E(X \varepsilon')$ (fixed $X$) or $0^2 E(X \varepsilon')$ for stochastic $X$

$\Rightarrow \frac{b_{i|x}}{V(b_{i|x})}$ the $i$th coefficient

B. Hypothesis Tests: $H_0: \beta = \gamma \iff \hat{b} = \beta$

1. Test Statistic (as always): $\frac{(b_i - \gamma)}{\sigma \varepsilon \varepsilon'_i} \sim \chi^2(1)$

2. We don't know $\sigma^2 \Rightarrow$ estimate it.

$\hat{\sigma}^2 = V(\varepsilon_i)$ $\forall \varepsilon_i$ good unbiased, consistent, efficient estimate of $\sigma^2$, so sample variance of $\varepsilon_i$ a "good" (some properties) estimate of $\sigma^2$ (under CNLRM)

3. $\hat{\varepsilon} = \hat{\varepsilon} = M(\hat{X} \beta + \hat{\varepsilon}) = \hat{\varepsilon} = \hat{\varepsilon}$, residuals from $\beta$ regressed on $X$, but don't know $\varepsilon$, so can't reduce this any further.

$M \hat{\varepsilon} = M(X \beta) = 0 \iff \beta = 0$ (under CNLRM)
VII. B. Hypoth. Tests on Single Coefficients:

3. b. $\hat{e} = M \hat{e}$ \Rightarrow $E(\hat{e}' M') = E(\hat{e}' M' M \hat{e})$

(M symmetric & idempotent) \Rightarrow $E(\hat{e}' M' M \hat{e})$

$= E[\text{trace}(M' M \hat{e} \hat{e}')]$ (by a rule above - think of the covariance)

$= E[\text{trace}(M' M \hat{e} \hat{e}')]$ (by a rule of traces, namely that elements cyclic shift to back)

$= \text{trace}(M' M \hat{e} \hat{e}')$ (because $M$ fixed, or conditional on $X$'s will)

$= \text{trace}(M' 0 M \hat{e} \hat{e}')$ (a constant brought out of a sum where the sum is the trace)

$= 0^{-2} \text{trace}[I_n - N] \quad \text{(trace}(I_n) = 0^{n} I_{n} = n)$

$= 0^{2} (\text{trace}(I_n) - \text{trace}(N)) = 0^{-2} (n - \text{trace}(x'(x'x)^{-1}x'))$

$= 0^{-2} (n - \text{trace}[x'(x'x)^{-1}x'x])$ (cycling the $X$ from front to back inside the trace)

$= 0^{-2} (n - \text{trace}(I_k))$ (as long as $x'(x'x)^{-1}x' \Rightarrow I_k$

$\Rightarrow 4. \text{Unbiased Estimate of } \sigma^2 \text{ is } \frac{(e'e)}{n-k}$

i.e., $\frac{1}{n-k} \sum_{i=1}^{n} \bar{e} e_i$

so, as always, we estimated $k$ coefficients, checking up $k$ degrees of freedom, implying the unbiased estimate divides by $(n-k)$

$\hat{\sigma} = \sqrt{(e'e)/(n-k)}$ is the "standard error of the regression" (or, of the estimate)

$\hat{\theta}$ is the unbiased estimate of $\sqrt{V(\hat{e})} = \sqrt{\sigma^2} = 0$

Back to our test statistic: $(b_i - c) / \sqrt{\hat{\sigma}^2 (x'x)^{-1}} \sim t_{n-k}$

C. Mechanics of Conducting t-tests

1. t-tests: a) You have coeff. est., $b_j$, from $j^{th}$ element of $A\hat{y}$

b) You have a null hypothesis about $\beta$ (that you'd like to reject) ($\beta = c$)

c) You need an estimate of std. err. of $b_j$

i) You have $V(b_j) = \sigma^2 (x'x)^{-1} = \begin{bmatrix} V(b_1) & \ldots & \text{Cov}(b_1, b_2) \\ \vdots & \ddots & \vdots \\ \text{Cov}(b_2, b_1) & \ldots & V(b_k) \end{bmatrix}$

ii) You estimate this by using $\hat{\sigma}^2 = (e'e)/(n-k)$ to estimate $\sigma^2$

iii) the $j^{th}$ element ($j^{th}$ diagonal element) of estimated $V(b_j)$ is thus your estimate of $V(b_j)$

iv) estimated std. error is $\hat{\sigma}(b_j)$

$\Rightarrow d) \text{t-statistic: } \frac{(b_j - c)}{\hat{\sigma}^2 (x'x)^{-1}} \sim t_{n-k}$

$\Rightarrow e) \text{t}_{n-k} \text{ distribution}$

Shaded area = $P(t_{n-k} \geq T)$ (under the null being true)

The less area out here, the less likely you were to have estimated a $T$-stat this far or farther from zero if $\beta$ really were $C$. This is same as, the less area out here, the less likely you were to have found a $b_j$ this far or farther from $C$ if $\beta$ really were $C$. 
VII. C. Mechanics of \(t\)-Tests & Confidence Intervals

1. \(t\)-Tests:

\(t\)-Tests:

(i) Thus, \(i\) the p-level of your test statistic is the probability of a test statistic (coefficient estimate) this extreme or more extreme if the null hypothesis were true.

(ii) from low p-levels we infer that the null is not well supported by the evidence.

\(ii\): the p-level may alternatively be viewed as the \(\alpha\)-level, which, had we chosen that level, our test statistic would have just enabled us to reject the null.

\[ b = 1.96 \]
\[ \text{s.e.}(b) = 1 \]
\[ H_0: \beta = 0 \quad H_1: \beta \neq 0 \]
\[ N = 5\text{ million} \]
\[ T = \frac{(b - \beta)}{\text{s.e.}(b)} = \frac{1.96 - 0}{1} = 1.96 \]
\[ T \sim t_{5\text{Million}} \]

\[ \Rightarrow T\text{-stat} = 1.96 \]
\[ p\text{-level} = .05 = \alpha\text{-level at which a} T \text{ of 1.96 would just barely allow us to reject} H_0. \]

2. Confidence Intervals:

\[ \Pr(\beta - t_{\text{level}} \cdot \text{s.e.}(b) \leq \beta \leq \beta + t_{\text{level}} \cdot \text{s.e.}(b)) = C\% \]

(a) The c.i. level \( \equiv \) confidence interval level you choose to report, e.g. 95%.

(b) Find the \( t \)-distribution corresponding to your degrees of freedom, On that distribution find the \( T\)-Stat you want such that \( P(T > T\text{-Stat}) = 1 - (\text{c.i. level}) \), e.g. c.i. level = 95% \( \Rightarrow 1 - (\text{c.i. level}) = .05 \)

\[ T\text{-stat} \text{ you want is one that puts } \frac{1}{2} \cdot (1 - \text{c.i. level}) \text{ on either side here.} \]

(c) Plug in to get c.i. bounds:

\[ b = 4 \]
\[ T\text{-Stat} = 1.96 \text{ (from above: } 2\alpha - 2b) \]
\[ \text{s.e.}(b) = 2 \]

\[ 4 \pm 1.96 \cdot 2 = 4 \pm 3.92 = [0.08, 7.92] \]

D. More Generally, What can we say about estimated std. errors of \( b \)?

1. \[ \text{s.e.}(b) = \sqrt{\frac{\text{SSE}}{(n-1)(n-2)}} \]

That is:

\[ \text{s.e.}(b) = \sqrt{\frac{\text{SSE}}{(n-1)(n-2)}} \]

\[ = \sqrt{\frac{\text{Sample Covariance of \( X_j \) with all the \( X_{not\ j} \) \text{ specifically \( R^2 \) from \( X_j \) on all \( X_{not\ j} \)}}}{\text{Sample Variation of \( X_j \)}}} \]

\[ + \text{Sample Covariance of \( X_j \) with all the \( X_{not\ j} \) \text{ specifically \( R^2 \) from \( X_j \) on all \( X_{not\ j} \)}} \]
VII. D. Useful Facts about Estimated s.e. (b_i)

1. s.e. (b_i) = \sqrt{\frac{S_x^2}{ \sum \left( x_i - \bar{x} \right)^2 + R_x^2} \cdot \frac{1}{n}}

   a) s.e. is bigger when estimate of stochastic component bigger
   b) smaller when (meaningful) sample variation in X is larger
   c) larger when X_i covaries more with the other X's

2. Graphical Intuition

   a) Why does fact that E varies imply that b varies?

   Suppose instead of being 1 std. dev. above regression line, observation 3 were two standard deviations below at obs. 3'.

   Then regression line would be at the dotted line case.

   Thus, movements in E_i change our estimated slopes. The larger is V(E_i), the more b will move for a "typical" movement in E_i by the larger such a "typical" move will be.

   Thus, s.e. (b) which measures how b estimates change across repeated samples correctly increases as V(E_i) increases.

   Unfortunately, we only have an estimate of V(E_i) so our estimate of s.e. (b) can only respond to this estimate.

   b) Why does ↑ S_x \Rightarrow ↓ s.e. (b)?

   These two examples are similar in that b = \frac{S_y}{S_x} in 1st case in that obs 1, 2, 3, 4 are 2 units from the regression line & 3' is 3 units below the original line. Only difference is the X's are more spread out (S_x higher in second case), larger sample variation in X produces less variance in estimates across samples for a given V(E_i) (represented by the equal movements from 3 to 3').

   c) Why does ↑ sample covariance among X's increase variance of b_i?

   b_i measures \Delta Y/ \Delta X_i holding all other X's fixed, i.e., netting out covariance of X_i with other X's. When X_i more highly correlated with other X's, the more is netted out & the less information we have left to estimate the effect of X_i per se. Less Info = More Variance.
VIII. Tests of Joint Hypotheses

A. A test regarding a single coefficient, e.g. that \( \beta = 0 \), is conducted holding all other coefficients fixed. For example:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

Assumes \( b_0, b_1, b_2 \) estimates remain fixed.

The hypothesis then, is equivalent to testing \( Y = b_0 + b_1 X_1 + \epsilon \) as a restriction (restricting \( b_2 \) to be zero) on the previous model. Not infrequently, however, our hypotheses will involve more than one coefficient in some way. For example, we may be interested in the above model in whether \( b_1, b_2 \) are both zero or that they sum to one or whatever.

B. Testing a Linear Restriction:

1) \[ Y = X^\prime b + \epsilon = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k + \epsilon \]

\( H_0 : \beta_1 + \beta_2 + \cdots + \beta_k = 0 \)  

examples: \( \beta_1 + \beta_2 = 0 \)

\[ \sum \beta_k = \frac{1}{2} \beta_k \] (n.b. \( \beta_k = \frac{1}{2} \epsilon_k = 0 \))

2) As per usual, if the null hypothesis were true, then \( \beta \) will not differ (or, rather, will differ with low probability) from \( \hat{\beta} \), the estimate.

3) In symbols, call \( \tilde{\beta} = \hat{\beta} \), \( T = (\hat{\beta} - \bar{\beta}) \times \text{. s.e.}(\hat{\beta}) \) will be distributed

4) Need an estimate of s.e. \( \hat{\beta} \):

\[ V(\hat{\beta}) = \sigma^2 (X^\prime X)^{-1} \]

5) Example:

\[ Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \epsilon \]

\( H_0 : \beta_1 + \beta_2 = 1 \)

\[ \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ b_1 \\ b_2 \end{bmatrix} \quad q = 1 \]

\[ \sqrt{V(\hat{\beta})} = \begin{bmatrix} b_0/b_2 \\ b_1/b_2 \end{bmatrix} \]

\[ T = ((b_1 + b_2) - 1)/\text{. s.e.}(b_1 + b_2) \]

(especially)

\[ \text{s.e.}(b_1 + b_2) = \sqrt{V(b_1) + V(b_2) + 2 \text{ Cov}(b_1, b_2)} \]

get these from your output of \( V(\hat{\beta}) \)

get whole \( V(\hat{\beta}) \) matrix of your output

6. Sets of Linear Restrictions: \[ Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \epsilon \]

\( H_0 : d_1 = 0 \)

\( d_2 = 0 \)

\[ \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} \]

More examples of how to construct \( \hat{\beta} \) are in Greene.
Once you have a suitable $R$ (restriction) matrix, then the statistic
\[ F = \frac{(R\hat{\beta} - q)'[R V(C)]^{-1}(R\hat{\beta} - q)}{J} \]
where $J$ is the number of restrictions.

- Examples given in Greene

Note the basic principles in this sort of test:

- If null $(R\hat{\beta} - q) = 0$ were true, then our estimate $(R\hat{\beta} - q)$ of it
  should not be "far" from zero. "Far," as always is measured in units of standard-deviations or variances, thus we always take some measure of distance $(R\hat{\beta} - q)'[R V(C)]^{-1}(R\hat{\beta} - q)$ and divide by the standard deviation of that estimate (s.e.($\hat{\beta}$)) or by the variance $[R V(C)]^{-1}$ to get our test stat.

Then it's a matter of figuring out how the test stat,

- is distributed if the null is true (which we leave to the experts for this class).

C. Confidence Intervals For More than one Coefficient

1. A confidence interval for a single coefficient is based on a set of values for the $T$-statistic (for which we would not reject

   i.e.,

   distribution of $\hat{\beta}$ under null that $\beta = \beta$

   $b$-T-sec($b$) interval

   some critical level for a $T$-stat corresponding to $(1-\alpha)$

   alpha-level

   E.g., $N = 5$ million, so $\hat{\beta} \sim$ approx. $N(\beta, V(\hat{\beta}))$

   $T$ for 95% conf. interval $= T$ for $\alpha = .05\approx 1.96$ approx.

   So, 95% conf. int. range amounts to the range of $\hat{\beta}$ we might

   propose as null hypotheses & be unable to reject at the .05

   level.

2. So... we can do something for sets of coefficients.

   Namely, find some range of values for $(\beta_1, \beta_2, \ldots)$

   we could propose as null & not reject on an F-test

   at whatever level.

3. How do we do it? a) Pick the coefficients of interest, say $b_1$, $b_2$, $b_3$.

   b) Find $F$-stat on $\frac{(R\hat{\beta} - q)'[R V(C)]^{-1}(R\hat{\beta} - q)}{J}$

   c) Find the set of $(\beta_1, \beta_2, \beta_3)$ that satisfy

   \[ \frac{1}{J} \left[ \begin{array}{c} b_1 - \beta_1 \\ b_2 - \beta_2 \\ b_3 - \beta_3 \end{array} \right] \left[ \begin{array}{c} V(b_1) \\ V(b_2) \\ V(b_3) \end{array} \right]^{-1} \left[ \begin{array}{c} b_1 - \beta_1 \\ b_2 - \beta_2 \\ b_3 - \beta_3 \end{array} \right] \leq F \]

   d) Multiply this out and solve for the $(\beta_1, \beta_2, \beta_3)$ that satisfy...
C.3. Joint Confidence Intervals.

We solve

\[ f \left( b - \beta \right) \left[ V(b) \right]^{-1} \left( b - \beta \right) = F_{n-k, \alpha} \]

for the set of \((\beta, \beta)\) pairs that solve this.

Generally, they look like this:

![Confidence Interval Diagram]


a) Consider \((1-\alpha/2)\%\) c.i. for \(b_1\) & \((1-\alpha)\%\) c.i. for \(b_2\).

Then, these give a rectangle in \(b_1, b_2\) space:

![Rectangle Diagram]

This rectangle gives a confidence region with at least \((1-\alpha, \alpha)\%\) probability.

E.g., 90\% c.i. for \(b_1\) & 95\% c.i. for \(b_2\) draw a rectangular c.i. for \((b_1, b_2)\) with at least 85\% confidence in it.

⇒ This is not the smallest area 85\% c.i.⇒ It may have more than 85\% c.i. in it, but it's a convenient shortcut sometimes.

when is it really bad? Look back to points 3a-3d. It is really bad when c.i. is very thin, i.e., when \(\text{Cov}(b_1, b_2)\) is high. Larger is \(|\text{Cov}(b_1, b_2)|\) the more misleading the shortcut (in b, it is never right).

IX. Goodness of Fit:

A. Std. Err. of Est.: \(\sqrt{\frac{e^2}{n-k}} = \frac{S_e}{n-k}\) is an estimate of the standard deviation of the stochastic component of the world.

1) \(S_e\) is in the same units as the dependent variable (e.g., if dependent in dollars, \(S_e\) is in dollars).

2) It's "kinda" a measure of a typical mistake or your "average" mistake magnitude. (Only "kinda" b/c it's exactly the square root of the average squared mistake which isn't quite the same as your average absolute-value mistake).

(Seek models to)

3) Just as we don't maximize \(R^2\) but rather estimate it, we don't minimize \(S_e^2\) or \(S_e^2\) but rather estimate it. (Seek models to)
IX. A. Std. Err. of Regression, Std. Err. of Estimate, \( s_e \) (same thing)

4. \( s_e \) can be used to compare models estimated in the same sample.

Some claim it can be used to compare across samples if the dependent variable in the samples are measures in the same units of the same thing. (E.g. $\$s$ in 1950-70 compared to $\$s$ in 1971-1990).

This is an open question. I think you're better off not making or sticking to this claim. Rather, you might use \( s_e \) for different periods. This can be used to compare the models.

\[ R^2 = \frac{SSR}{SST} = \frac{\Sigma(y-\bar{y})^2}{\Sigma(y-\bar{y})^2} = 1 - \frac{SSE}{SST} = 1 - \frac{\Sigma e_i^2}{\Sigma(y-\bar{y})^2} \]

\[ R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma(y-\bar{y})^2} \]

The ratio of linear "explained" variation to total variation in y.

\[ \text{Adj. } R^2 \text{ or } \bar{R}^2 \]

\[ R^2 = \frac{1 - \frac{SSE}{(n-k)}/SST}{1-(k/n)} \]

2. See Email Message about \( R^2 \) and \( \bar{R}^2 \)

3. \( R^2 \) and \( \bar{R}^2 \) should be used not used basically in the same scenarios as \( s_e^2 \). One Key Exception:

There is absolutely no doubt that \( \bar{R}^2 \) are not comparable across samples.

C. Basic Rules:

- if it's not the same sample, we will want to avoid using \( R^2 \), \( \bar{R}^2 \), or \( s_e^2 \) to choose between models.

- if it is the same sample, we may use \( R^2 \), \( \bar{R}^2 \), \( s_e^2 \) to compare models; however, we will want some way of accounting for degrees of freedom (\( R^2 \) and \( s_e^2 \) insufficiently so).

D. Tests Based on \( R^2 \):

1. Thinking back to all our previous tests, notice that hypotheses that \( \beta = 0 \) or some \( \beta = 0 \) are equivalent to the hypotheses that we can drop these variables from a regression. E.g.

\[ y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e \]

\[ H_0: \beta = 0 \text{ is equivalent to } H_0: \text{Model is } y = b_0 + b_3 x_3 + e \]

\[ H_0: \beta = 0, \beta = 0 \text{ is equivalent to } H_0: \text{Model is } y = b_0 + b_3 x_3 + e \]

\[ \Rightarrow \text{If these nulls were true, then the restricted model } y = b_0 + b_3 x_3 + e \]

\[ \Rightarrow \text{not significantly worse than the unrestricted model } y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e \]

\[ \Rightarrow \text{If these nulls were true, then the restricted model } y = b_0 + b_3 x_3 + e \]

\[ \Rightarrow \text{not significantly worse than the unrestricted model } y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e \]

\[ \Rightarrow \text{If these nulls were true, then the restricted model } y = b_0 + b_3 x_3 + e \]

\[ \Rightarrow \text{not significantly worse than the unrestricted model } y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e \]

2. A common hypothesis is \( H_0: \text{all } \beta \text{ except the constant (intercept) are zero.} \]

That is, the model is no improvement on \( y = \bar{y} \) for all \( y \). I.e., the \( X \)'s are no help in predicting \( y \).
Tests based on $R^2$

2. Test that $\beta_i = 0 \forall i \neq 0$ (where $\beta_i$ is intercept): or, more generally, test of any restricted model against an unrestricted (i.e. less restricted) one

$$\frac{(\Delta R^2 / \Delta k)}{\left(1-R^2_{big}/n-k_{big}\right)} = F \sim F_{\Delta k, n-k_{big}}$$

a) $\Delta R^2$ is the change in $R^2$ from the restricted to the unrestricted model.

(N.B. $R^2 = 1 - \frac{\text{sum of squared residuals}}{\text{sum of squared total}}$)

b) $\Delta k$ is the number of restrictions which is most typically equal to the $\Delta$ in the number of X's ($= \Delta k$).

(N.B. $\Delta k$ is also the degrees of freedom of $\Delta R^2$, which is one $R^2$ minus another which yields still another $R^2$ with deg. free. equal to diff. of deg. free. in initial $R^2$).

$\Rightarrow$ Numerator is an $R^2$ divided by its degrees of freedom

c) We express the change in $R^2$ from restricted to unrestricted as a fraction of (unrestricted $R^2$) Thus divide by $R^2_{big}$ from the "big" i.e. unrestricted regression, $R_{big}$

as already noted, this $R^2_{big}$ is a $R^2$, it has degrees of freedom equal to $n-k_{big}$ (sample size - #Xs in big regression)

So whole thing is $\left(\frac{\Delta R^2/\text{free}}{\Delta R^2/\text{free}}\right)$ which is an $F$ with an $F$ distribution

Specifically $F_{\Delta k, n-k_{big}}$

Example: Return to $H_0: \beta_i = 0$ except intercept.

Note first that $R^2$ with just an intercept $= 0$ why? $\gamma$ with just an intercept $= \gamma$ as we showed way back

thus $R^2 = \frac{\sum (Y-\gamma)^2}{\sum (Y)^2}$

so $\Delta R^2 = R^2_{small}$ from the regression

$\Delta k = k-1$ the intercept

$R^2_{big}$ sample size - k including intercept.

So for this particular hypothesis we only need to run one regression $H_0$ we know the restricted regression has $R^2 = 0$.

More generally, the restrictions will reduce $R^2$ but not to zero:

E.g. $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$

$R^2 = 0.80$

$Y = b_0 + b_1X_1 + e$

$R^2 = 0.55$

$\Delta R^2 = 0.25$ $\Delta k = 2 \Rightarrow F$-stat. if this "big", then restriction really bit.

Check $p$-level in the tail (for right)

1. Note that $t$-test is equivalent to either sort of $F$-test (first sort was Wald, one we just did was $R^2$-ratio $F$-test), when hypothesis that one variable is zero.

2. In fact, the Wald Test, the $R^2$-test, & the likelihood ratio test—rather than compare $R^2$'s we could compare log likelihoods—are all asymptotically equivalent.

3. The $t$-Test: take $T$ stat & square it. That new statistic is distributed $F_{1,n-k}$ can be used to test the same hypothesis.

4. So why learn them all? They are only asymptotically equivalent—they have different small sample properties.

5. They are based on different estimated quantities, their validities depend on those different things being correctly estimated. In some cases, we have better confidence (or may even know for a fact) that certain of these things are estimated better than others.

X. Prediction: You can look over this on your own. We are less interested in predicting & forecasting sense than in evaluating theory which is about relationships between variables (i.e., coefficients), generally.

Still, note:

A. Predicting some $Y_j$ using the vector of explans $X_j$:

$$\hat{Y}_j = X_j^T \hat{B}$$

$\hat{B}$ is generally called estimated & on some sample (not including $j$ usually).

Then $e_j = Y_j - \hat{Y}_j$ is the prediction error

$$V(e_j) = V(Y_j - \hat{Y}_j) = \sigma^2 + V(\hat{Y}_j)$$

$$= \sigma^2 + X_j \hat{\Sigma} X_j'$$

(again, from the sample used to predict $j$)

Forecast interval for $Y_j$:

$$\hat{Y}_j \pm T_{a/2} \cdot \text{s.e.}(e_j) = \hat{Y}_j \pm T_{a/2} \left( \sigma^2 + X_j \hat{\Sigma} X_j' \right)^{1/2} X_j$$

That tells us the critical value from $t_{n-k}$ in sample as usual.

An example is in Greene.
SOME CRUCIAL REGRESSION MATRICES:

\[ Q \equiv X'X \] : matrix of products & cross-products of columns of \( X \); i.e., roughly variation-covariation of \( X \)

\[ A \equiv Q^{-1}X' = (X'X)^{-1}X' \] : that particular linear-additive-parititude combination of the \( X \)s, that comp of the info in \( X \), that produces the coefficients on \( X \) for regression of \( Y \) of whatever \( A \) multiplies on

\[ N \equiv XA = XQ^{-1}X' = (X'X)^{-1}X \]

\[ N_Y = \hat{Y}_s \]

\[ N = I - N = I - X(X'X)^{-1}X' \] : residual maker

\[ M_Y = e_{ls} \]

PARTIAL-REGRESSION COEFFICIENTS

\[ (X'X)_{k} = X'Y \]

\[
\begin{bmatrix}
    x_{1}'x_{1} & x_{1}'x_{2} \\
    x_{2}'x_{1} & x_{2}'x_{2}
\end{bmatrix}
\begin{bmatrix}
    b_{1} \\
    b_{2}
\end{bmatrix}
= \begin{bmatrix}
    x_{1}'y \\
    x_{2}'y
\end{bmatrix}
\]

\[ \Rightarrow b_{1} = (X'X)^{-1}X'Y - (X'X)^{-1}X'X_{2}b_{2} \]

\[ b_{1} = A_{1}Y - A_{1}X_{2}b_{2} \]

\[ = \hat{b}_{1} \text{ from Reg } Y \text{ on } X_{1} - (\text{Reg } X_{2} \text{ on } X_{1}) (\text{Reg } Y \text{ on } X_{2} \text{ conditionally for } X_{1}) \]

\[ X_{2}'X_{1} (A_{1}Y - A_{1}X_{2}b_{2}) + X_{2}'X_{2}b_{2} = X_{2}'Y \]

\[ X_{2}'N_{1}Y - X_{2}'N_{1}X_{2}b_{2} + X_{2}'X_{2}b_{2} = X_{2}'Y \]

\[ (X_{2}'X_{2} - X_{2}'N_{1}X_{2})b_{2} = X_{2}'Y - X_{2}'N_{1}Y \]

\[ X_{2}'(I - N_{1})X_{2}b_{2} = X_{2}'(I - N_{1})Y \]
\[ x_2' M_1 x_2 b_2 = x_2' M_1 y \]

\[ b_2 = (x_2' M_1 x_2)^{-1} x_2' M_1 y \]

*\text{M symmetric & idempotent}*

\[ \Rightarrow M_1 = M_1'M_1' \]

\[ b_2 = (x_2' M_1'M_1 x_2)^{-1} x_2' M_1'M_1 y \]

\[ b_2 = [(M_1' M_1 x_2)' (M_1' x_2)]^{-1} (M_1' x_2)' M_1 y \]

i.e., \( b_2 \) - partial reg. coeffs on \( x_2 \), are the coeffs from reg. of

Residual (\( y \) reg on \( x_1 \))

reg on

Residual (\( x_2 \) reg on \( x_1 \)).

\[ \Rightarrow \text{Partial coefficients use only the unique variation & covariation of } y \text{ & each } x. \]
Multivariate Regression: Estimating & Reporting Certainty, Hypotheses Tests

I. The C(N)LRM:

A. Core Assumptions:
   1. \( y = X\beta + \varepsilon \)
   2. \( E(\varepsilon) = 0 \)
   3. \( V(\varepsilon) = \sigma^2 \varepsilon \)
   4. \( E(\varepsilon | X) = 0 \)
   5. \( X \) of full-column rank.

B. Convenience Assumptions (Unnecessary):
   1. \( X \) non-stochastic. (Unnec. to unbiasedness; relaxed to “conditioning on \( X \)” for variance estimation.)
   2. \( \varepsilon \sim MVN(0, \sigma^2 \varepsilon I_n) \). (\& if not: CLT \& asymptopia!)

II. Properties of OLS Estimator under CLRM:

\[ \hat{\beta}_{LS} \equiv b_{LS} = (X'X)^{-1} X'y \equiv Ay \]

A. \[ = A(X\beta + \varepsilon) = \beta + A\varepsilon \]
\[ E\left( \hat{\beta}_{LS} \right) = E\left( \beta + A\varepsilon \right) = \beta + E(A\varepsilon) \]

\[ = \beta + AE(\varepsilon) = \beta, \text{ if } X \text{ non-stoch.} \]

\[ = \beta + E(A\varepsilon) = \beta, \text{ if } X \text{ stoch. & } E(\varepsilon | X) = 0 \]

\[ V\left( \hat{\beta}_{LS} \right) = V\left( \beta + A\varepsilon \right) = AV(\varepsilon)A' \]

\[ = (X'X)^{-1}X'V(\varepsilon)X(X'X)^{-1} \]

\[ = (X'X)^{-1}X'\sigma^2_{\varepsilon} I_n X(X'X)^{-1} \]

\[ = \sigma^2_{\varepsilon}(X'X)^{-1}X'X(X'X)^{-1} \]

\[ = \begin{cases} 
\sigma^2_{\varepsilon}(X'X)^{-1}, & \text{if } X \text{ non-stoch.} \\
\sigma^2_{\varepsilon}E\left\{(X'X)^{-1}\right\}, & \text{if } X \text{ stoch.} 
\end{cases} \]

C.

\[ \Rightarrow V\left( b_k \right) = f\left( \sigma^2_{\varepsilon}, V\left( X \right), R^2_{x_k,x_{-k}} \right) \]

D. **NOTES:**

1. \[ \overline{V\left( b_k \right)} = f\left( \sigma^2_{\varepsilon}, \overline{V\left( X \right)}, \overline{R^2_{x_k,x_{-k}}} \right) \]

\[ V\left( \hat{\beta}_{LS} \right) = \sigma^2_{\varepsilon}(X'X)^{-1} \text{ is the Cramer-Rao lower-bound;} \]

i.e., \( \hat{\beta}_{LS} \) is efficient; i.e., \( \hat{\beta}_{LS} \) is BLUE.

2. In fact, if \( \varepsilon \sim MVN \), then \( \hat{\beta}_{LS} \) is BUE.
E. If \( \epsilon \sim MVN(0, \sigma^2_\epsilon I_n) \), then:

\[ b_{LS} = \beta + A\epsilon = \text{constant + linear-sum of normals} \]

\[ \Rightarrow b_{LS} \sim MVN \left( \beta, \sigma^2_\epsilon (X'X)^{-1} \right) \]

if not, then

\[ b_{LS} = \beta + A\epsilon = \text{constant + linear-sum of ?} \]

\[ \Rightarrow b_{LS} \sim^A MVN \left( \beta, \sigma^2_\epsilon (X'X)^{-1} \right) . \]

III. Estimating \( \sigma^2_\epsilon \) using \( \frac{1}{n-k} \epsilon'\epsilon \):

\[ e = M(X\beta + \epsilon) = MX\beta + M\epsilon = 0\beta + M\epsilon = M\epsilon \]

\[ \Rightarrow E(e'e) = E \{ (M\epsilon)'M\epsilon \} = E \{ \epsilon'M'M\epsilon \} = E \{ \epsilon'M\epsilon \} \]

\[ = E \{ \text{trace}(\epsilon'M\epsilon) \} = E \{ \text{trace}(M\epsilon\epsilon') \} \]

\[ = \text{trace} \{ ME(\epsilon\epsilon') \} = \text{trace} \{ M\sigma^2_\epsilon I_n \} \]

\[ = \sigma^2_\epsilon \times \text{trace} \{ I_n - N \} = \sigma^2_\epsilon \times \text{trace} \{ I_n \} - \text{trace} \{ N \} \]

\[ = \sigma^2_\epsilon \times \{ n - \text{trace} \{ X(X'X)^{-1}X' \} \} \]

\[ = \sigma^2_\epsilon \times \{ n - \text{trace} \{ (X'X)^{-1}X'X \} \}

\[ = \sigma^2_\epsilon \times \{ n - \text{trace} \{ I_k \} \}

\[ = \sigma^2_\epsilon (n - k) \]

\[ \Rightarrow E \left( \frac{\epsilon'\epsilon}{n-k} \right) = \sigma^2_\epsilon \Rightarrow \text{use } s^2_\epsilon = \frac{\epsilon'\epsilon}{n-k} \text{ as LS (unbiased) est. } \sigma^2_\epsilon. \]
A. NOTE: $\frac{e'e}{n-k}$ is sum of squared (asympt.) normals, divided by degrees freedom, so $s_e^2$ is (asympt.) $\frac{\chi^2_{n-k}}{n-k}$.

B. Therefore, std Wald $t$-tests & conf. ints. by:

$$T = \frac{b_j - c_0}{\text{s.e.}(b_j)} = \frac{b_j - c_0}{\sqrt{s_e^2 \{(X'X)^{-1}\}_{jj}}} \sim^{(A)} t_{n-k}$$

1. $b_j \pm t_{n-k}^\alpha \times \text{s.e.}(b_j) = b_j \pm t_{n-k}^\alpha \times \sqrt{s_e^2 \{(X'X)^{-1}\}_{jj}}$

2. $\Rightarrow \{1 - \alpha\} \%$ (asympt.) conf. int.

3. Tests of linear restrictions (by Wald strategy):

   a) For instance, $H_0: \beta_1 + \beta_2 = 1$. If null-hypoth. true, then estimate not far (in std. err. units) from null:

   $$T = \frac{(b_1 + b_2) - 1}{\text{s.e.}(b_1 + b_2)} = \frac{(b_1 + b_2) - 1}{\sqrt{V(b_1 + b_2)}} = \frac{(b_1 + b_2) - 1}{\sqrt{V(b_1) + V(b_2) + 2 \times C(b_1, b_2)}} \sim^A t_{n-k}$$

   b) But note how we could use matrix algebra to generalize:
\[ H_o: \beta_1 = 0, \beta_2 = 1 \Rightarrow r' \beta = q; \text{ for instance, if } \beta \text{ is } (4 \times 1), \text{ then:} \]

\[ r' = [0 \ 1 \ 1 \ 0] \text{ and } \beta = \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix} \]

\[ \Rightarrow H_o: r' \beta = q; \text{ e.g., } H_o: r' \beta = 1 \Rightarrow T = \frac{r'b - 1}{\sqrt{V(r'b)}} = \frac{r'b - 1}{\sqrt{r'V(b)r}} \sim ^A t_{n-k} \]

Notice how \( r' \ldots r \) plucks \( \sqrt{V(b_1)}, \sqrt{V(b_2)}, C(b_1, b_2), C(b_2, b_1) \), just as in scalar!

4. Test of joint hypotheses (by Wald strategy):

\[ H_o: \beta_1 = q_1, \beta_2 = q_2 \Rightarrow R\beta = q; \]

for instance, \( H_o: \beta_1 = 0, \beta_2 = 1 \) and \( \beta \) is \( (4 \times 1) \), then:

\[ R = \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}, \quad \beta = \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix}, \quad \text{and } q = \begin{bmatrix}
0 \\
1
\end{bmatrix} \]

\[ \Rightarrow H_o: R\beta = q \Rightarrow C = (Rb - q)' \left[ \sqrt{V(Rb - q)} \right]^{-1} (Rb - q) \]

is a ratio of chi-squares. "Denominator" has \( n-k \) deg. free, and has \( n-k \) in its denom. "Numerator" is square \( 2 \times 1 \) normals. So, \( C/2 \), or, more generally, \( C/J \) where \( J = \text{rows}(R) \) is \( F_{j,n-k} \)!

\[ F = (Rb - q)' \left[ \sqrt{V(Rb - q)} \right]^{-1} (Rb - q) / J \sim ^A F_{j,n-k} \]
C. Confidence Intervals & Confidence Regions

1. Recall the simple formula for confidence intervals:

\[(1 - \alpha)\% \text{ confidence interval (c.i.) for } \beta:\]

\[\hat{\beta} \pm T \times \text{s.e.}(\hat{\beta})\]

2. Recall/Note, too, 1-for-1 correspondence b/w hypothesis test @ level \(\alpha\) and the \((1-\alpha)\%\) c.i.:
   a) c.i. overlaps \(0 \Leftrightarrow\) fail-to-reject; 0 lies outside c.i. \(\Leftrightarrow\) reject
   b) The \((1-\alpha)\%\) c.i. also corresponds to set of hypothesized \(\beta\) that one would fail-to-reject at level \(\alpha\) given estimated \(b\).

3. By latter understanding, can see how might construct a \((1-\alpha)\%\) confidence region for \((\beta_1, \beta_2)\) as set of \((\beta_1, \beta_2)\) that would fail-to-reject at \(\alpha\) given estimates \((b_1, b_2)\):

\[
\frac{1}{J} \left[ \begin{bmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \end{bmatrix} \right]' \left[ \begin{bmatrix} V(b_1) & C(b_1, b_2) \\ C(b_1, b_2) & V(b_2) \end{bmatrix} \right]^{-1} \begin{bmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \end{bmatrix} \leq F_{J, n-k}
\]
a) Using $F_{J,n-k}$ critical value for $\alpha$ from desired $(1-\alpha)\%$ c.i.,
b) $J$ is the number of estimates at issue (2 here) & so $J$ is also the dimensionality of the resulting confidence region.
c) Multiply & solve for $(\beta_1, \beta_2)$ that just-satisfy inequality.

4. They generally look like this:

![Graph showing confidence region]

5. Are properties of conf. reg. intuitive to you?
   a) Ellipsoidal and centered on $(b_1, b_2)$.
   b) Go top-left to bottom-right if $C(b_1, b_2)<0$, and this will be when (partial) $C(x_1, x_2)>0$; go from bottom-right to top-left if $C(b_1, b_2)>0$, and this will be when (partial) $C(x_1, x_2)<0$.
   c) Appear thinner as $|C(b_1, b_2)|$ &/or $|V(b_1)-V(b_2)|$ ‘greater’.
   d) Circular if $C(b_1, b_2)=0$ and $V(b_1)=V(b_2)$.

6. Short-cut Approximation: Rectangular region given by set of $k (1-\alpha_k)\%$ c.i.’s contains at least $(1-\Sigma \alpha_k)\%$: 
a) Ex: two 95% c.i.’s ⇒ region w/ min. (1-.05-.05)% = 90%

b) Worse approx. (area too big) the more ‘slanted’ and thinner the actual confidence region. Never “right” area (ultimately, b/c rectangular not ellipsoidal).

D. Measures of Fit (“Goodness of Fit” Statistics)

1. Std. Err. Est./Reg. (S.E.E., S.E.R., s.e.e., s.e.r.):

\[ s_e = \sqrt{s_e^2} = \sqrt{\frac{e'e}{n-k}} \]

Note: If \( \epsilon \sim^A \text{MVN} \), then \( s_e \sim^A \frac{\chi^2_{n-k}}{n-k} \)

a) Sometimes also denoted \( \sigma \) or \( \hat{\sigma} \), w/ or w/o subscript \( e \) or \( \epsilon \), but best to reserve for ML est. and use hat & \( e \):

\[ \hat{\sigma}_e = \sqrt{\sigma_e^2} = \sqrt{\frac{e'e}{n}} \]

b) Notes:

(1) Kinda measure of typical or avg error or mistake. (Act’ly, measures square root of average squared mistake…)

Page 8 of 11
(2) In same units as dep var. E.g., if dep var in $\$, s.e.e. in $\$

(3) Not construct models to min $s_e$ any more than to max $R^2$.

2. $R^2$: share of the variation in $y$ ‘explained’ (linearly accounted) by the model $(X\beta)$.

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{\sum [(y - e) - \bar{y}]^2}{\sum (y - \bar{y})^2} = \frac{\sum [(y - \bar{y}) - e]^2}{\sum (y - \bar{y})^2}$$

$$= \frac{\sum (y - \bar{y})^2 - \sum (2(y - \bar{y})e) + \sum e^2}{\sum (y - \bar{y})^2}$$

$$= 1 - \frac{2\sum (\hat{y}e + e^2 - \bar{y}e) - \sum e^2}{\sum (y - \bar{y})^2}$$

$$= 1 - \frac{2\sum e^2 - \sum e^2}{\sum (y - \bar{y})^2} = 1 - \frac{\sum e^2}{\sum (y - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

a) $R^2$ is also the square of the correlation of $y$ & $\hat{y}$, i.e., $r_{y,\hat{y}}^2$.

b) If $\varepsilon \sim \mathcal{MVN}$, then $R^2 \sim \mathcal{F} \left( \frac{\chi^2}{\chi^2} \right) = F$.

3. Adjusted $R^2$, Adj. $R^2$, R-bar squared:

a) Can always increase $R^2$ just by adding variables. Want some penalty for lack parsimony. Common adj. to $R^2$ is to replace numerator & denominator w/ unbiased estimates.
\[
\bar{R}^2 = 1 - \frac{\text{unbiased}(SSE)}{\text{unbiased}(SST)} = 1 - \frac{\sum e^2 / (n-k)}{\sum (y - \bar{y})^2 / n - 1} = 1 - \frac{s_e^2}{s_y^2}
\]

b) Weak penalty. Can show that adding variable w/ coeff. having \( t > 1 \) increases \( \bar{R}^2 \). Alternative adj.’s with stronger penalties, based on “Information Criterion”: A\textsuperscript{K}\text{a}i\text{k}e IC, B\text{ay}e\text{s}i\text{n}/S\text{chw}a\text{r}t\text{z} IC, ... \text{http://en.wikipedia.org/wiki/Akaike\%27s_information_criterion}.

c) Although not directly used (that I’m aware), notice that:

If \( \varepsilon \sim^A \text{MVN} \), then \( \bar{R}^2 \sim^A \frac{\chi^2_{n-k} / (n-k)}{\chi^2_{n-1} / (n-1)} = F_{n-k, n-1} \)

4. (Log) Likelihood from ML est. is also measure of fit.

5. Use & Abuse of Fit Statistics/Measures:

a) Can use to compare model performance \textit{in same sample}; use to compare across samples only w/ great attention and care how much V(y) to explain varies across samples.

b) At end, (relative) fit of model more something to estimate given a model, than something to model to maximize.

E. “Degradation of Fit” Strategy for Testing:

1. \textit{Logic}: If null hypothesis were true, then imposing it as true rather than estimating its parameters should result in “little” loss of fit.

2. \textit{Strategy}: Measure fit-loss, then determine how that measure or some function of it would be distributed
under the null hypothesis, so we can determine how likely this much fit-loss is to have occurred by chance.

3. The “change-in-\(R^2\)” or “delta-\(R^2\)” or “\(\Delta R^2\)” Test:

a) Determine how to “impose the null hypothesis”. Example:

\[
H_0: \beta_3 = \beta_4 = 0 \quad \Rightarrow \quad \begin{cases} \ H_0: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ H_1: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \end{cases}
\]

b) Measure loss of explanatory power relative to gap from big-model explanatory power to one, and divide each numerator and denominator by its degrees of freedom:

loss of fit: \(\Delta R^2 = R_1^2 - R_0^2 = \left(1 - \frac{SSE_{n-k_1}^1}{SST_{n-1}}\right) - \left(1 - \frac{SSE_{n-k_0}^0}{SST_{n-1}}\right) = \frac{SSE_{n-k_0}^0 - SSE_{n-k_1}^1}{SST_{n-1}}\)

fit-gap: \(1 - R_1^2 = \frac{SSE_{n-k_1}^1}{SST_{n-1}} \quad \Rightarrow \quad \text{ratio:} \quad \frac{SSE_{n-k_0}^0 - SSE_{n-k_1}^1}{SSE_{n-k_1}^1}\)

\[
\Rightarrow \frac{\chi^2_{n-k_0} - \chi^2_{n-k_1}}{\chi^2_{n-k_1}} \quad \Rightarrow \quad \text{\(\chi^2\) free:} \quad \frac{(n-k_0) - (n-k_1)}{n-k_1} = \frac{k_1 - k_0}{n-k_1} = \frac{\Delta k}{n-k_1}
\]

So: \(F = \frac{\Delta R^2 / \Delta k}{(1 - R_1^2) / (n-k_1)} \sim (\Delta) F_{\Delta k, n-k_1}\)

4. Tests using other measures of fit, \(s_e^2\) or \(ln(L)\), also...

5. Third logic, Lagrange-Multiplier Tests: if null hypoth true, then impose it as constraint on max \(ln(L)\) or min \(SSE\) should not bind, implying: Lagrange multipliers, \(\lambda=0\), and \(\partial ln(L)/\partial \beta\) at \(\beta_{\text{null}}=0\) or \(\partial SSE/\partial \beta\) at \(\beta_{\text{null}}=0\).
Topics in Linear Regression: Hypothesis Testing

These notes are based on Green (ed al, Chp 7) and are on Hypothesis Testing. The main goal of this chapter is to transform your substantive hypotheses into statistically testable statements, they imply.

Much of the "regressive regression," "non-linear restrictions," "non-parametric testing" is technically complicated but substantively transparent. Be sure you have it, the latter level.

I. Approaches to Hypothesis Testing

1. We estimate coefficients and then ask if those estimates are "far" from their hypothesized levels.

   A. For "far" is measured in standard-deviation or variance units.

   B. The counter-factual question is: "how likely is it that, if the true coefficient were B, I would have estimated a coefficient this far or further from B just by chance?" If this extreme estimate was unlikely to have occurred by chance, we tend to conclude that the true coefficient is not, in fact, B.

2. An alternative (but equivalent at least asymptotically) approach is to impose the hypothesized level of B rather than estimate it. 

   See if the fit worsens by more than is likely to have occurred just by chance.

3. Most generally, the basic question is, given our model, do given the variability in estimates in the stochastic component of the world, what is our best guess of the relationships between variables of interest? how precisely does that relationship appear to have held in our (historical) data? i.e., what have we found and how "certain" are we of that finding?

B. The most common hypothesis:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \]

estimated by 0.25 as:

\[ Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k + \varepsilon \]

Example: 

- \[ \beta_2 = 0 \]
- \[ \frac{b_2}{s_c(b_2)} = 80 \]
- \[ \Rightarrow T = 1.6 \]
- \[ n-k = 170 \]
- \[ (\text{our estimate is } 1.6 \text{ standard deviations from } \theta, \text{ thus...}) \]

1) \[ P (|T_{n-k}| > 1.6) = p\text{-level} \approx 0.124 \]

   i.e., if \( \beta_2 \) were truly zero, \( \varepsilon \) is \( \varepsilon \) and \( \varepsilon \sim N(0, \sigma^2) \) or the central limit theorem applies well enough for the normal approximation to be appropriate, then we would estimate \( \beta_2 = .8 \) or greater or \( -.8 \) or lower in about 12.4% of random samples.

2) 90% confidence interval: \[ b_2 \pm T_{.95} \cdot s_c(b_2) \]

   \( T_{.95} \) is such that .05 probability above \( T \) and .05 below -\( T \).

   \[ N(.8 \pm 1.65(.5)) = .8 \pm .825 = [-.025, 1.625] \]

   in 95% of random samples (if normally or CLT), an interval around the estimate, \( b_2 \), constructed like this would contain the true \( \beta_2 \).

3) Most generally, we have estimated that \( \frac{b_2}{s_c(b_2)} = 8 \) and that if the \( \varepsilon_i \) are independent, then the standard deviation of this estimate is \( \sigma \). Further, just to gauge how big a variability of .5 is, we note that if the \( \varepsilon_i \) were iid \( \varepsilon \) normal, then we have \( \frac{b_2}{s_c(b_2)} \approx N(0, .5) \) (90% C.I. \( -0.25, 1.625 \)).
I. B. Testing whether \( \beta \) coefficient estimate(s) is(are) for from some hypothesized level(s).

1. (Sets of) Linear Restrictions:
   \[ \begin{align*}
   R \beta &= q \\
   \text{examples:} & \quad a) \beta = 0 \\
   & \quad b) \beta_4 = 7 \\
   & \quad c) \beta_2 = 3 \beta_4 \\
   & \quad d) \beta_3 = 2 \beta_4 + 3 \\
   \\
   R \beta &= q & \text{states your hypotheses:} & \quad e) \beta = 0, \beta_2 = 3 \beta_4, \beta_3 = -\beta_4
   \end{align*} \]

2. As we saw last week, you construct \( R \beta = q \) so that:
   - a) \( R = \begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = 0 \Rightarrow q = 0 \\
   - b) \( R = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & \ldots \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = 0 \Rightarrow q = 0 \\
   - c) \( R = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & \ldots \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = 3 \Rightarrow q = 3 \\
   - d) \( R = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & -2 & 0 & 0 & \ldots \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = 3 \Rightarrow q = 3 \Rightarrow \beta_2 - 2 \beta_3 = 3 \Rightarrow \beta_3 = 2 \beta_4 + 3 \sqrt{\text{which is what we wanted}} \\
   - e) \( R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = 0 \Rightarrow q = 0 \\

3. Check to see whether \( (R \beta - q) \) differs a "lot" from \( (R \beta - q) \)

   where "a lot" is measured in std. deviation units, or equivalently, whether \( (R \beta - q) \) differs a "lot" from \( (R \beta - q) \) where "a lot" is measured in variance units

4. Which gives us the test statistic:
   \[ \frac{1}{\sigma^2} (R \beta - q)' [R [S'(X'X)^{-1} R']^T] (R \beta - q) = F_{r,k} \text{ is } F \text{ "big" in an } F_{r,k} \text{ distribution?} \]

   i.e., is there a good probability of an F this large or large if null were true?

5. Notes:
   a) This is the \( V(b) \) estimated; thus, if \( \hat{b} \) estimated with "nice" properties and \( V(b) \) estimated with "nice" properties (BLUE, or something)
   then \( F \) (Wald) test will also have "nice" properties...

   b) In cases (like examples a&b) where we are testing single hypotheses, the F-stat reduces to
   \[ \text{if } T \sim \chi^2_{n-k}, T^2 \sim F_{n-k} \]

   \[ \Rightarrow \text{test: } V T \sim \chi^2_{n-k}, T^2 \sim F_{n-k} \]
I.C. * Imposing the Hypotheses to Begin With & Checking the Degradation of Fit Caused by that Imposition

1) In the case where we hypothesize that a coefficient or a set of them is equal to zero, the imposition of that hypothesis as a restriction is equivalent to dropping that (these) variable(s) from the regression:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \]

\[ H_0: \beta_2 = 0 \quad \text{and} \quad \beta_0 = 0 \quad \Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \]

2) Suppose, instead:

a) \[ H_0: \beta_2 = 3 \quad \text{and} \quad \beta_0 = 1 \], then (\[ y - 3x_2 - x_3 \] = \[ \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \]) embodies the restriction.

b) \[ H_0: \beta_2 = 2 \beta_0 \], then \[ y = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + x_3) + \beta_3 x_2 + \ldots \] embodies the restriction.

c) In case (c), if you impose the restriction \[ y = 3x_2 \] from \( y \) before regressing on the rest, this forces \( b_2 = 3 \) and \( b_0 = 1 \) to be fixed while we estimate the rest of \( \beta_1 \). In case (b), the estimate of \( \beta_2 \), the coefficient on \( x_2 \), will produce a coefficient on \( x_2 \) that is twice that of \( x_2 \) by construction.

Practically, then, in case (a) you create a variable \( y^* \), which is \((y - 3x_2 - x_3)\) if use that as dependent variable. In (b) you create \( x_2^* = (x_2 + x_3) \) & use that in place of \( x_2 \).

3) Forcing OLS to labor under restrictions like this, you are certain that the restricted \( \hat{\beta} \) will be higher & (equivalently) that restricted \( R^2 \) will be lower. **Note** these disembowel \( R^2 \) from predicting the original \( y \) that are worse (not based on \( y^* \))

**Question is:** Is fit so much worse that it is unlikely it could have been this much worse just by chance even though restrictions (hypotheses) were true?

Answer: \[ \left( \frac{AR^2}{Ak} \right) = F \rightarrow EF \]

As before, less often there is to the right of \( F \) in a \( F_{ak,\infty} \) distribution, the less likely we have seen so much degradation of \( R^2 \) simply by chance.

---

I.D. Directly Restricted Least Squares: we could also rephrase the OLS problem as one of constrained optimization:

\[ \min (y - X_\beta) \] subject to \( \beta_0 = \beta_1 \)

logarithmic multiplier

\[ \Rightarrow \] (first order)

\[ \Rightarrow \] (second order)

\[ \Rightarrow \] \( \alpha X'X + R'\alpha = X'y \)

\[ \Rightarrow \] \( \alpha R^* + \alpha \lambda = X'y \)

writing \( \alpha X'X \) in matrix form:

\[ \begin{bmatrix} X'y & X' \beta_0 \\ \beta_0' & R^* \end{bmatrix} \begin{bmatrix} \alpha \\ \lambda \end{bmatrix} = \begin{bmatrix} X'y \\ \beta_0' \end{bmatrix} \]

This can be solved for \( \beta^* = \beta - \frac{1}{X'y + R^* X'y - R^* \beta^*} R' \beta_0 \) some positive semi-definite matrix.
I. D. Lagrange Multiplier Tests \\
Restricted: 

\[ \hat{b}^{\text{restricted}} = b + Q^{\text{est}} \cdot R' \left[ RQ^{\text{est}} \cdot R' \right]^{-1} (Rb - \hat{\theta}) \]

\[ V(\hat{b}) = V(b) - \text{some positive definite matrix} \]

1) I.e., imposing restrictions lowers \[ V(b^*) \]. Why? \( \Rightarrow \) we are "assuming" more information, there is less to estimate.

2) Lagrangian multiplier tests are based on the question: is \( \hat{\theta} = 0 \)?

From above, 

\[ \hat{\theta} = \left[ RQ^{\text{est}} \cdot R' \right]^{-1} (Rb - \hat{\theta}) \]

If \( \hat{\theta} = 0 \), the restrictions are not binding, i.e., they imposed no penalty on fit & (from 1 above) improved the precision of \( \hat{b} \) estimates.

We may return to these later, for now just know that a fourth option exists.

II. Test Options, Review

A. Wald F-tests: are coefficients estimated "for" hypothesis(es)

1. Simplest Special Case: Scalar Functions of Coefficients
   a) \[ \frac{g(b) - g(\beta)}{\sqrt{V(g(b))}} = T \sim t_{n-k} \]
   e.g., \( \beta = 1 \Rightarrow \alpha(1) \cdot 2e(b) = T \sim t_{n-k} \)

b) (Some thing) \[ (b_2 + b_3) \Rightarrow \frac{(b_2 + b_3)^2}{V(b_2) + V(b_3) + 2W(b_2, b_3)} = T \sim t_{n-k} \]

2. More generally, a set of such linear restrictions is tested by

\[ \left[ (Rb - \hat{\theta}) ', [RV(b)R']^{-1} (Rb - \hat{\theta}) / \hat{\theta} \right] = F \sim F_{j, n-k} \]

3. General Notes: a) The Wald approach (e.g. F-tests) is good when we have confidence in \( b \) & \( V(b) \) estimates
   e.g., if CLRM, then Wald tests asymptotically valid; if CMLRM, then Wald tests valid.

b) \[ J \cdot F = (Rb - \hat{\theta}) ', [RV(b)R']^{-1} (Rb - \hat{\theta}) / \hat{\theta} = X^2 \]

   where \( X^2 \sim \chi^2 \) in the limit

* which means, an asymptotically equivalent test to a Wald F-test is a Wald \( \chi^2 \) test based on some quantity, as F is only without the divided by J

B. Degradation of Fit Tests: \( \left( \Delta R^2 / \Delta k \right) = \frac{(1 - R^2_{(\text{est})})/(n-k)}{(1 - R^2_{(\text{est})})/(n-k)} \] \( \Rightarrow F \sim F_{k, n-k} \)

C. Similar Tests to \( R^2 \)-ratio are likelihood & Lagrange Multiplier tests.
II. C. Log-likelihood & Lagrange-Multiplier Tests

1. Likelihood-Ratio (LR) tests are basically same as F-ratio tests, only we worry about whether the degradation of likelihood is significant enough to have been likely to occur by chance alone.

\[ LR = -2 \ln(L_R) - \ln(L_U) \sim \chi^2 \]

- LR: likelihood of restricted model
- \( \chi^2 \): \# of restrictions
- L_U: likelihood of unrestricted model

Lagrange-Multiplier tests amount to the question of whether a (hypothetical) coefficient on the whole set of restrictions would be zero in a regression. Thus, it is really exactly the same as an F-ratio or \( \chi^2 \) test for an appropriately constructed auxiliary regression.

Usually, in that auxiliary regression, \( N \cdot R^2 \sim \chi^2 \)

II D. Tests of Structural Change: (an important example of testing restrictions)

1. Recall we have assumed (CLRM) that \( \beta \) is somehow, with mean \( E(\beta) \) equal to zero, \( \sigma^2 \) I notice that this implies \( \beta \) constant across observations. Suppose we thought the \( \beta \) of or that we wanted to test the CLRM assumption that \( \beta \) constant in our sample

2. Consider grouping your data into two sets of observations:

\[
\begin{bmatrix}
Y \\
X_1 \\
X_2
\end{bmatrix}
= \begin{bmatrix}
X_0 & 0 \\
0 & X_1
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1
\end{bmatrix}
+ \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2
\end{bmatrix}
\]

\( X_0 \) and \( X_1 \) are applied to the model for the first group \( \beta_0 \) and \( \beta_1 \) are for the second.

In our example, suppose we have

\[ y = (\text{donated to Democrats}) - (\text{donated to Republicans}) \]

\[ X = [X_0 \ X_1] \]

where \( X_0 \) is control, \\
and \( X_1 \) is marital status

if women, or alternatively, that it does not \& you'd like to test that implication of you just suspect there's a systematic difference (this test is less acceptable)

Socially, you don't have reasons for suspecting a difference

you could break up your sample:

men \[ \begin{bmatrix}
Y \\
X_0 \\
X_1
\end{bmatrix}
= \begin{bmatrix}
X_0 & 0 \\
0 & X_1
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1
\end{bmatrix}
+ \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2
\end{bmatrix}
\]

women \[ \begin{bmatrix}
Y \\
X_0 \\
X_1
\end{bmatrix}
= \begin{bmatrix}
X_0 & 0 \\
0 & X_1
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1
\end{bmatrix}
+ \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2
\end{bmatrix}
\]

That is, you take your original data & sort it into the groups. Then you take your \( X \)-matrix & double it in width equal to \( (\text{group 1}) \times (X) \& (\text{group 2}) \times \)

\[ \text{dummy} = 1 \text{ if } i \text{ is in group 1} \]

ditto for group 2

Page 181 of 364
More generally, you could have \( i \) groups (e.g. 50 states) and expect to allow different coefficients for each group:

\[
Y_i = \beta_{i\text{men}} + \beta_{i\text{women}} + \beta_{i\text{Income\text{men}}} + \beta_{i\text{Income\text{women}}} + \epsilon_i
\]

Written out for a regression program, it might be:

\[
Y = \beta_{\text{men}} + \beta_{\text{women}} + \beta_{\text{Income\text{men}}} + \beta_{\text{Income\text{women}}} + \epsilon
\]

Now, how would you test hypotheses that coefficients same across groups? 

Option 1: Estimate separate regression for each group and test a single regression for whole sample, not allowing the coefficients to differ by group. Clearly the latter is a restriction on the former so the former should fit better.

Base test: 

\[
\text{F} = \frac{(\sum_{i=1}^{k} \hat{e}_i')^2 / (2k - k)}{\sum_{i=1}^{k} \hat{e}_i^2 / (n - 2k)}
\]

Most generally:

\[
\text{F} = \frac{(\sum_{i=1}^{k} \hat{e}_i')^2 / (2k - k)}{\sum_{i=1}^{k} \hat{e}_i^2 / (n - 2k)}
\]

This is called a “Chow Test” & it tests for the equality of parameters across a set of subsamples.

3. Obviously, you can do the same thing allowing only some coefficients to vary across subsamples (restricting only some coefficients to be the same across subsamples).

a) Very common example is to allow for/ test for a different intercept across subsamples:

\[
\begin{bmatrix}
Y_i \\
Y_i
\end{bmatrix} = 
\begin{bmatrix}
1 & 0 & X_i \\
0 & 1 & X_i
\end{bmatrix} 
\begin{bmatrix}
\beta_m \\
\beta_w \\
\beta
\end{bmatrix} + 
\begin{bmatrix}
\epsilon_i \n\\
\epsilon_i
\end{bmatrix}
\]

Notice that this allows for a simple additive difference by the groups. Each individual of group \( i \) differs from a group \( j \) case with same \( X_i \) only by (stochastic term and) some constant.
II.D.3: Differences in Subsets of Coefficients

(a) difference constants (interecepts) : notice also that this amounts
to nothing more than including a dummy variable (c) for group membership
in the sample regression with the same intercept for all cases.

(b) Testing is exectly as before:

(i) \[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon \]

\[ H_0 : \beta_4 = 0 \]

\[ X_1, X_2, X_3, X_4 \]

\[ X_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \]
\[ X_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \]
\[ X_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \]
\[ X_4 = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \]

(c) Notice that you could also achieve the
same thing by creating just one dummy variable & the constant

(ii) \[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \]

\[ H_0 : \beta_4 = 0 \]

\[ X_1, X_2, X_3 \]

\[ X_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \]
\[ X_2 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \]
\[ X_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \]

(c) Similarly, we can allow some slope coefficients to vary across
subsamples & restrict others.

(d) Before we could estimate

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon \]

by directly creating 2 variables out of \( X_4 \) or by
using dummy variables:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \]

(ii) In either case, we could then test

\[ H_0 : \beta_4 = \beta_5 \]

(iii) Notice that, again, we could take one group as
the base case and consider the other as a devition
from them:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon \]

then \( \beta_4 \) in (c) is \( \beta_4 \) in (i) & \( \beta_5 \) in (c)
and the equivalent hypothesis is

\[ H_0 : \beta_4 = 0 \]

(d) Upshot of all this: We can test for "structural" i.e. parameter-change
in the model. Exactly as always this can be done as hypothesis tests on coefficients or
as tests on degradation of fit.

*There are usually a number of ways to specify a model & a number
of equivalent (asymptotically) best for each hypothesis. The specification
we choose to discuss here is but one way to look at a transparent base case & no base case
so that the coefficients tell you what they do & are not just residuals that are by itself.
III. D. 3.6. Cleaning Up Structural-Change Tests when \( V(e) \) unequal

- If \( V(e) \) unequal, then the pooled regression(s) may be estimated with all the subgroups in the same regression are not BLUE.

Why? because \( V(e) \neq 0 \) but further \( V(e) \neq 0 \) differ for each group.

- When we come to heteroskedasticity, we will see an optimal solution to this, but even when doing anything, we know that

\[
(\hat{\Theta}_1 - \hat{\Theta}_2) = (V(\hat{\Theta}_1) + V(\hat{\Theta}_2))(\hat{\Theta}_1 - \hat{\Theta}_2) = W \rightarrow X^2
\]

- If sample 1 & sample 2 are independent

In the limit, this test valid even if the sample have differ stochastic term variance, so long as the samples are independent.

IV. An intuitive introduction to Recursive Regression & Tests Based on Them & Related Procedures

A. Chow Tests: focus on parameter stability across sub-samples. More generally, we may be concerned about or interested in parameter stability throughout the sample or the maintenance of the CNLRM assumptions throughout the sample ⇒ concepts of recursive regression & cross-validation.

1. Use the first \((j-1)\) observations to predict the \(j\)th: \(\hat{y}' = \hat{x}_j' \hat{b} + \hat{e}_j\)

2. forecast \(y_j = \hat{y}_j = \hat{x}_j' \hat{b} + \hat{e}_j\)

3. So, if CNLRM, then

\[
\begin{align*}
\text{if } C(\text{ue}) \text{LRM, then } W_j \sim N(0, \sigma^2) \\
\text{if } C(\text{ue}) \text{LRM, then } W_j \sim N(0, \sigma^2)
\end{align*}
\]

⇒ a number of tests since, if CNLRM, then each \(W_j\) iid \(N(0, \sigma^2)\)

4. CUSUM test (cumulative sum of \(W_j\)) is based on \(W_j = \frac{W_j}{\sigma_j}\)

which has mean zero & a known variance & distribution. The variance is proportional to \(j\) so CUSUM plot looks like:

5. CUSUM: basically the same test only now we look at a function of \(W_j\)^2

6. Forecast Tests: using 1 period ahead forecasts like this, we can also check directly whether forecast errors in any given period are "significant" i.e. are large relative to std. errors.

Basic Options:
- Predict next observation only each time
- At each observation, predict all the rest
B. Recursive Coefficients: basically track the coefficients as you add observation-by-observation rather than residuals.

- Little or no direct tests from this (other than whether, given previous data, next single observation significantly Δ's the coefficient), but intuitively a nifty thing to have graphed: can observe how β estimates are updated (Bayesian updating in action) as info (obs.) is added.

- Detection of Jumps: Trends are signs of something not kosher with the CMLR™ as modeled by your regression. Host likely parameter variation in this base (and still could be something else unfortunately)

C. Moving Window Estimation: a related concept

⇒ (again esp. in time-series) take a block of observations of fixed size & "move that window through your sample" estimating a new regression at each stage.

E.g. Data from 1930-1990; annual, take 10-year blocks. Estimate 1930-1939, then 1931-1989, then... to the end

⇒ similar to recursive regression, but sample size is not increasing. Basically opposite intuition is at heart here; assumption is parameters (or something else) such as 𝜃 is changing through your sample & you want to plot its course (usually over time, but perhaps across the country, e.g., W or N to S or across the continent).

D. Cross-Validation: idea here is (again similar to estimate repeatedly, but dropping an observation or group of obs. at a time using that estimation to "predict" the dropped obs. graph:

- Useful for picking out "odd" observations, perhaps more so than the usual residuals (by the latter are based on regressions including those points). As such, the latter try to fit "odd" observations & if they're "odd" enough they'll have a lot of influence, & thus small (usual-type) residuals (much more on this later, for now a picture)

While reg. line & "odd" obs., leave other obs. out

⇒ huge "cross-validated" residual

⇒ Esp. useful & distinguishable from "studentized" residuals (more later) when groups to be omitted suggest themselves theoretically:

- E.g. Obs. on countries 1, 2, 3,...J over times 1, 2, 3,...J

⇒ May leave out country at a time to "predict" it & demand using the others; may provide evidence of places where the model is "not working"
V. Testing Non-Linear Restrictions

A. Suppose our theory leads to predictions which are not expressible as a linear function of the coefficients. Suppose we want to talk about predicted values which are non-linear functions—as always we would not want to present any results without also presenting some indication of the uncertainty of our results (std. err., cont. int., hypothesis test).

\[ \text{Variance of non-linear functions of } \beta. \]

B. Hypothesis Testing:
1) \( H_0: f(\beta) = \beta \) (first, the simple case where the restriction is a simple scalar (not \( \beta \)). Ex: \( b_0 + \beta = 1 \)

a) As always, we ask whether our estimate of \( f(\beta) \) or more generally of \( f(\beta) \) is "for" from its hypothesized level (1 or more generally, \( \gamma \)).

\[ Z = \frac{f(\beta) - \gamma}{\text{SE}(f(\beta))} \]

b) As you see, as always, "for" is measured in std. dev. units.

c) Calculating \( f(\beta) - \gamma \) is no problem; you just plug your coefficient estimates into the restriction. Our example: \( \sqrt{b_0 (b_0 + b_1) - 1} \)

2) Std. Error of \( f(\beta) - \gamma \), which is just std. err. of \( f(\beta) \) because \( \gamma \) is a constant: if Thus has no variance or covariance, is the issue.

a) Leads to Taylor-series approximation:

- For any function, \( f(x) \), a reasonable approximation for \( X \) near some value \( x_0 \) is \( f(x_0) + \frac{df}{dx}(x-x_0) \).

b) So, then consider approximating \( f(\beta) \) near the value \( \beta \):

\[ f(\beta) \approx f(\beta) + \frac{df}{d\beta}(\beta - \beta) \]

(c) The advantage of the approximation is that it is linear.

b) What then is the variance of \( f(\beta) \) using the approx.?

\[ \text{Var}(f(\beta)) \approx \text{Var}(f(\beta) + \frac{df}{d\beta}(\beta - \beta)) \]

(c) Under the null hypothesis, \( \beta \) is the truth, so it doesn't vary or covary, only \( \beta \) does.

b) \( \text{Var}(f(\beta)) \approx \text{Var}(f(\beta)) \) and the std. err. is the square root of that.

C. Our Example: \( f(\beta) = b_0 + (b_0 + b_1) \)

\[ \frac{df}{d\beta} = \begin{bmatrix} b_0 & b_1 \end{bmatrix} = \begin{bmatrix} \frac{df}{d\beta} \end{bmatrix} \]

and using our estimates,

\[ \text{Var}(f(\beta)) = \begin{bmatrix} \frac{df}{d\beta} \end{bmatrix} \begin{bmatrix} b_0 + b_1 \end{bmatrix} = \begin{bmatrix} \frac{df}{d\beta} \end{bmatrix} \begin{bmatrix} b_0 + b_1 \end{bmatrix} \]

\[ \Rightarrow \text{Test Stat: } Z = \frac{(b_0 (b_0 + b_1) - 1)}{\sqrt{V}} \]

D. How is \( Z \) distributed? Asymptotically normal (CLT)

C. When can we get away with such approximations? One thing we need is that if \( b_0 \) is an estimate (unbiased, efficient...) of \( \beta \), then \( f(\beta) \) is an equally good estimate of \( f(\beta) \). This is called invariance.
\[ y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \varepsilon_t \]
\[ x_t = 0 \quad d x_{t+i} = 1 \quad \forall i \geq 0 \]

\[ \Rightarrow y_{t+1} = \beta_2 x_{t+1} \]
\[ y_{t+2} = \beta_1 \beta_2 x_{t+1} + \beta_2 x_{t+2} \]
\[ y_{t+3} = \beta_1 (\beta_2 x_{t+1} + \beta_2 x_{t+2}) + \beta_2 x_{t+3} \]
\[ y_{t+4} = \beta_1^3 \beta_2 x_{t+1} + \beta_1 \beta_2^2 x_{t+2} + \beta_1 \beta_2 x_{t+3} + \beta_2 x_{t+4} \]

\[ y_{t+n} = \sum_{i=0}^{\infty} \beta_i \beta_2 x_{t+i} \]

\[ = \frac{1}{1-\beta_1} \cdot \beta^2 x_{t+n} \quad \text{for } 0 < \beta_1 < 1 \]

LR effect of \( x_t \):
\[ \frac{\beta_2}{1-\beta_1} = \text{LRE} \]

\[ V (\hat{\beta_2} \cdot (1-\hat{\beta_1})^{-1}) \approx \left[ \frac{d\text{LRE}}{d\beta} \right]' V(\hat{\beta}) \left[ \frac{d\text{LRE}}{d\beta} \right] \]

\[ \approx \left[ \frac{\hat{\beta_2}}{(1-\hat{\beta_1})^2} \right]' \left[ \frac{\hat{\beta_2}}{(1-\hat{\beta_1})^2} \right] \]
I. Concluding Topics in Hypothesis Testing

A. Testing Nonlinear Restrictions

1. Format (Wald Strategy):

\[ H_0: f(\beta) = q \Rightarrow z = \frac{f(b) - q}{\text{s.e.}(f(b))} \sim^* N(0,1) ; \text{n.b., usually use } t_{n-k} \]

2. Issues:

a) Calculating \( f(b) \) given estimates \( b \) is no problem

b) But how do we calculate \( V(f(b)) \) for \( f(b) \) nonlinear?

3. Taylor-Series Approximation (Linearization) \( \Rightarrow \) “Delta Method”

a) \( f(b) \approx f(\beta) + \nabla_{\beta} f(\beta) \cdot (b - \beta) \); recall \( \nabla_{\beta} f(\beta) = \frac{\partial f(\beta)}{\partial \beta} \) is vector

b) \( V\{\nabla_{\beta} f(\beta) \cdot (b - \beta)\} = [\nabla_{\beta} f(\beta)]^2 V(b) = [\nabla_{\beta} f(\beta)]' V(b) [\nabla_{\beta} f(\beta)] \)

4. So our test statistic:

\[
\begin{align*}
 t & = \frac{f(b) - q}{\sqrt{[\nabla_{\beta} f(\beta)]' V(b) [\nabla_{\beta} f(\beta)]}} \\
 & = \frac{f(b) - q}{\sqrt{[\frac{\partial f(\beta)}{\partial \beta_0} \quad \frac{\partial f(\beta)}{\partial \beta_1} \quad \cdots \quad \frac{\partial f(\beta)}{\partial \beta_k}]' V(b) [\frac{\partial f(\beta)}{\partial \beta_0} \quad \frac{\partial f(\beta)}{\partial \beta_1} \quad \cdots \quad \frac{\partial f(\beta)}{\partial \beta_k}]} \\
 & \sim^* N(0,1) ; \text{(often use } t_{n-k} \text{)}
\end{align*}
\]

5. Examples!
D. Whole key to this section: $\text{V}(f(b)) \equiv \left[ \frac{\partial f(b)}{\partial b} \right] \left[ \text{V}(b) \right] \left[ \frac{\partial f(b)}{\partial b} \right]^T$

$\Rightarrow V(f(b)) \approx \left[ \frac{\partial f(b)}{\partial \beta} \right] \left[ \text{V}(\beta) \right] \left[ \frac{\partial f(b)}{\partial \beta} \right]^T$

$C = \left( \frac{\partial f(b)}{\partial \beta} - g \right) \left( \frac{\partial f(b)}{\partial \beta} - g \right)^T$

$C$ is distributed $\chi^2$ asymptotically.

II. Choosing Between Non-Nested Models:

- All of our hypothesis tests so far have amounted to testing restrictions on a given model. Suppose, however, that we have two (or more) competing models, which do not share the same variables so that one cannot completely be expressed as a restriction on the other.

A. Option 1: Artificial nesting: just create a new matrix $W = X'UZ$ (no unique variables + $\mathbf{V}^*$ shared variables).
2. We could then interpret test of 
the Z model (H1) \( H_0: \theta = 0 \) vs. \( \theta \neq 0 \). Does it support or reject (or neither) the null hypothesis of \( H_0 \)?

b) What about the coefficient \( \alpha \)? Is it statistically significantly different from zero (or not)?

2) An alternative: Bonferroni method. Test all \( \alpha \) at the same time.

C. An alternative: Bonferroni method. Test all \( \alpha \) at the same time.

1. Consider: \( Y = (1-\alpha)(X\beta) + \alpha(Z\gamma) + \epsilon \), where \( \alpha + 1-\alpha = 1 \).

2. Test \( H_0: \alpha = 0 \) vs. \( \alpha \neq 0 \). If the test is significant, then \( \alpha \) is likely not zero.

3. If the test is significant, then \( \alpha \) is likely not zero.

4. If \( \alpha \) is not zero, then the model is likely not encompassing. If \( \alpha > 0 \), then the model is not encompassing.

5. If \( \alpha < 0 \), then the model is not encompassing.

6. If \( \alpha = 0 \), then the model is not encompassing.

Problem: This is not a problem because if all \( \alpha \) are zero, then the model is not encompassing.

Personal View: From a modeling perspective, this is exactly what should happen.

This is viewed as a problem because if all \( \alpha \) are zero, then the model is not encompassing.

The usual hypothesis testing is exactly what should happen.
C. Cox Tests: based on the likelihood-ratio type of test
- in linear model, asymptotically equivalent to D'M's J-test.
  - Greene talks through the steps; I won't repeat them here, just note that
    it's same Test: H0: X is correct set of regressors
    Ha: Z is """
  - except that: it relies more directly on normality (he it is
    based on likelihood-ratio which is valid for
    MLE estimators which OLS is only if we
    are in the CALRM world
  - on the other hand, based on LR & MLE as
    it is, it is more directly translatable to
    non-linear models (logit/probit, etc.) than
    M-D's J-test
  - has all the same epistemological issues as M-J's J-test
    (as always, logical issues are immutable, whether they're raised
    in a mathematical context or not; they're there if they're the same)
III. D. Lagrange-Multiplier Tests

1. Logics: If null true, then constraints are opt., not binding:
   a. If null true, then impose it as constraint on optimization, \( \nabla_{\theta} \ln L = 0 \) still
   b. ... & lagrangian multipliers, \( \lambda = 0 \)

2. Version 1a \( \Rightarrow \) \( b^* = b - \text{some positive-definite matrix} \)

   & test statistic: \( \left[ R'Q^{-1}R \right]^{-\frac{1}{2}} (Rb - \hat{\theta}) \sim \chi^2 \)

   Version 1b \( \Rightarrow \) same, but also, often, in auxiliary post-estimation regression, coefficients are \( \hat{\theta} \), so contrast them

E. Linear Combination & Joint hypotheses by linear-algebra expressions

1. \( x'b = \chi \Rightarrow \frac{(x'b - \chi)}{\sqrt{x'V(x)b}} = t \sim t_{n-k} \)

2. \( Rb = \chi \Rightarrow \frac{1}{2} (Rb - \chi)' V(k)'(Rb - \chi) \sim F_{1,n-k} \)

F. Non-linear functions of parameter estimates:

1. "Delta Method": asymptotic variance of linear-approx to \( f(\hat{\theta}) \):
   \( V(f(\hat{\theta})) \approx \left[ \frac{\partial f(\theta)}{\partial \theta} \right]' V(\hat{\theta}) \left[ \frac{\partial f(\theta)}{\partial \theta} \right] \)

2. Simulation (a.k.a., parametric bootstrap):

   \( \hat{\theta} \sim N(\theta, -\hat{H}^{-1}) \), so draw \( \hat{\theta}_i \) from MVN with that mean vector \( \theta \) and that variance-covariance matrix, & calculate
   \( f(\hat{\theta}_i), i = 1, \ldots, Z \) (with \( Z \) large) times.

   \( V(\hat{f}(\theta)) = \frac{1}{Z} \sum (f(\hat{\theta}_i) - \hat{f}(\theta))^2 \)

( Clarify: http://gking.harvard.edu/clarify/docs )
IV. Structural Change:
   A. Chow Tests -- degradation of fit strategy...
   B. Dummy Variables, Dummy-variable Interactions -- Wald strategy
   C. Brief Intro to recursive-estimation uses of test strategies.

V. Non-nested Hypothesis Testing
   A. In all our tests so far, the null hypothesis could be expressed as a restriction on the alternative (more general). If not this overlap, e.g.
   \[ Y = X\beta + \varepsilon_1 \quad H_0 \]
   \[ Y = Z\gamma + \varepsilon_2 \quad H_1 \]
   with \( X \cap Z \neq \emptyset \) and both non-empty.

   B. Option 1: artificially nest the models --
   \[ Y = \left[ X \cap Z \right] \beta^* + \left[ Z \setminus X \right] \gamma^* + \left[ X \cap Z \right] \omega + \varepsilon, \]
   then test \( \beta^* = \gamma^* \), but...what of \( \omega \)?
   ... neither model to nor \( H_0 \) properly reflected in these tests of \( \beta^* = \gamma^* \).

C. Davidson & McKinnon’s J-test:
   1. Consider:
   \[ Y = (1-\alpha)X\beta + \alpha Z\gamma + \varepsilon \]
   \( \alpha = 1 \) or \( \alpha = 0 \)...
   ... but \( \alpha \gamma = (1-\alpha)\beta \Rightarrow \) infinite set of \( \alpha, \gamma, \beta \) satisfy, and any perfect collinearity in \( X \cap Z \).

   2. Dars suggest:
   (1) Reg \( y \) on \( X \) & save \( \hat{\gamma} \)
   (2) Reg \( y \) on \( Z \) and \( \hat{\gamma} \)
   (3) Test coefficient on \( \hat{\gamma} = 0 \) tests \( X \) model against \( Z \) model. Reject implies \( X \) model
   (4) Reverse to test \( Z \) model v. \( X \) model.
   (5) “Problem”: all four possibilities exist
   (6) “Encompassingness” interpretation.

3. See Greene on Cox-V Zang tests (test more on ML). Clarke
   2001 PA & “distribution free” nonnested tests.
I. C(N) LRM
   A. \( \chi = X\beta + \varepsilon \)
   B. \( \varepsilon \sim (0, \sigma^2 I) \)
   C. \( E(\varepsilon | X) = 0 \)
   D. \( X \) of full column-rank
   E. \( X \) nonstochastic
   F. \( \varepsilon \sim N(0, \sigma^2 I) \)

II. Least-squares (if \( N \) also \( M \))
    estimates:
    \[ b = (X'X)^{-1}X'\chi = A\chi \]
    A. \( E(b) = E(A\chi) = E(AX\beta + A\varepsilon) = \beta \)
    B. \( V(b) = V(A\chi) = A V(\chi) A' = (X'X)^{-1}X'\sigma^2 I X (X'X)^{-1} \)
       \( = \sigma^2 (X'X)^{-1} \)
       \( = \text{Cramer-Rao lower bound} \)
    C. Estimate \( \hat{\beta} \) by
       \( \hat{V}(b) = \sigma^2 (X'X)^{-1} \)
       \( \hat{V}(b) = f(\frac{\hat{\sigma}^2}{\sigma^2}, V(\chi), R_{\chi X\chi}^2) \)
       \( \hat{V}(b) = f(\sigma^2, V(\chi), R_{\chi X\chi}^2) \)

III. Hypothesis Testing
     & Confidence Intervals

A. Confidence Intervals:
   \[ \Pr(b - t \cdot s.e.(b)) < \beta < b + t \cdot s.e.(b)) \]
   \[ \approx \frac{1}{\sqrt{s^2(X'X)^{-1}}} \text{critical stat from } T_{n-k} \]

B. Wald Tests (Logic):
   1. \( \frac{(\text{Estimate} - \text{Hypothesis}) / \text{s.e. (estimate)}}{\sqrt{\text{Var (Estimate)}}} \sim T_{n-k} \)
   2. \( \frac{(\text{Estimate} - \text{Hypothesis})^2 / \text{Var (Estimate)}}{F} \sim F_{\alpha/n-k} \)
      \( \frac{1}{\frac{1}{(Rb - q)'[V(b)](Rb - q)}} \)
   3. \[ \text{Confidence Regions: } t \sqrt{k - \beta} \leq [V(b)]^{-1/2} \leq \sqrt{k - \beta} \]

C. Degradation of Fit Tests (Logic):
   1. \( \frac{\Delta R^2 / k}{1 - R^2_{\text{adj model}}} \sim F_{\Delta k, n-k_{\text{adj model}}} \)
   3. \( 2 \times (\text{lnL}_0 - \text{lnL}) \sim \chi^2_{k} \)
   2. \( \frac{\Delta \varepsilon^2 / \Delta k}{[\text{Variance}_{\text{adj model}}]} \sim F_{\Delta k, n-k_{\text{adj model}}} \)
Topics in Linear Regression:

I. Indicator or Dummy Variables: Nonlinearity in Variables;

1. An indicator or dummy variable is a nominal variable that takes a value of 1 if the observation belongs to some defined group and zero otherwise. E.g.: gender (male=0, female=1), Incumbent (yes=1, no=0), LEFT (1=plurality, 0=majority), etc.

A. Return to a "typical" regression model:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon \]

1. Notice some things about the "intercept," \( \beta_0 \):
   a) it is the expected value (mean) of \( Y \) controlling for all the \( X \)'s:
   \[ E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \]
   or, loosely:
   \[ E(Y) = E(Y|X_1,\ldots,X_k) = \beta_0 \]
   b) it is the expected value (prediction) of \( Y \) when all the \( X \)'s are 0:
   \[ E(Y|X_1=0,\ldots,X_k=0) = \beta_0 \]
   c) it is also seen as the coefficient on a particular vector of observations, namely:
   \[ \beta_0 \]
   i.e., \( \beta_0 \) times a vector of constants (one by convention, but notice that if we use 2 as something only in estimating their relationship, we get a coefficient of \( \frac{1}{2} \)).
   d) it "takes out" the "levels" of \( X \) in the model:
      \[ Y = \beta_0 + \beta_1 X_1 + \epsilon \]
      \[ (\overline{Y} - \mu_Y) = \beta_0 (\overline{X_1} + \overline{X}) + \epsilon - \overline{\epsilon} \]
      if \( \overline{\epsilon} = 0 \), then \( \overline{Y} = \beta_0 + \mu_Y \)

B. Indicator variables (dummies), then, are similar to constants in a number of ways:

1) suppose we take \( X_0 = [1] \)
   d) break it up into \( X_0 = [1,0,0,\ldots,0] \)

Then, \( Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon \)

and the same model as above, except that we have allowed for a different intercept (or, equivalently):

2) so suppose we have:
   \[ Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \]
   then we'd have:
   \[ E(Y|X_2=1) = \beta_0 + \beta_2 \]
   \[ E(Y|X_2=0) = \beta_0 + \beta_1 \]

Notice

a) \( \beta_1 \) is the intercept (mean of \( Y \) given \( X_2 = 0 \)) if \( X_0 = 1, X_2 = 0 \)

b) \( \beta_2 \) is the intercept (mean of \( Y \) given \( X_2 = 0 \)) if \( X_0 = 0, X_2 = 1 \)

c) \( \beta_2 - \beta_0 \) is a fixed difference (additive) in mean by group \( X_0 \) and group \( X_2 \) -- fixed or whatever the level of \( X_1 \); group \( X_2 \) is expected to have \( Y = (\beta_2 - \beta_0) \) higher value than group \( X_0 \) with some \( X_1 \)

D. Example:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \]

\[ E(Y|X_1=1) = \beta_0 + \beta_1 \]

\[ E(Y|X_1=0) = \beta_0 + \beta_2 \]

Notice:

a) \( \beta_1 \) is the intercept (mean of \( Y \) given \( X_1 = 0 \)) if \( X_1 = 1, X_2 = 0 \)

b) \( \beta_2 - \beta_0 \) is a fixed difference (additive) in mean by group \( X_1 \) and group \( X_2 \) -- fixed or whatever the level of \( X_1 \); group \( X_2 \) is expected to have \( Y = (\beta_2 - \beta_0) \) higher value than group \( X_0 \) with some \( X_1 \)
C) This implies that dummy variables are an easy and flexible way to perform difference of means tests.

1. Suppose you have two groups: female & male
   a) let $X_{0i} = 1$ if female, 0 if male, then estimating
      \[ Y = \beta_0 + \beta_1 X_{0i} + \epsilon \]
      is an easy way to check if the mean of $Y$ is the same in the two groups: $H_0: \beta_1 = 0$
   b) identically, you could estimate:
      \[ Y = \beta_0 + \beta_1 X_{0i} + \epsilon \]
      The test for equal means in the two groups is now $H_0: \beta_1 = 0$

2. Suppose you want to know if some group of observations (from one observation to the whole set less one) is significantly different (by a simple additive factor) from the rest:
   e.g. if $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_{0i} + \epsilon$ where $X_{0i} = \{1$ if male, head of household \}
   0 otherwise
   if $\beta_2 \neq 0$, then
   \[ E(Y \mid \text{not male} \text{ h-o-b}) = \beta_0 + \beta_1 X_1 \]
   \[ \neq E(Y \mid \text{male} \text{ h-o-b}) = \beta_0 + \beta_1 + \beta_2 X_{0i} \]
   i.e. or suppose you have obs. from 1950-1990 annually:
   a quick way to check for a simple (additive) difference in 1974-1980 (orises years approx) would be to create a dummy = 1 for those years & 0 elsewhere.

3. This sort a: "is there a simple additive difference between nominal groups" question is only unambiguously informative in its answer if the other $X$'s are satisfactorily measured in the $X_i = 1$ & $X_0 = 0$ groups.

   E.g. from our problem set a while back: many of you wanted to create a var. "GER" (1 if country named Germany

   Now, $Y = \beta_0 + \beta_1 \cdot \text{GER} + \beta_2 \cdot X_1 + \epsilon$

   \[ \hat{\beta}_1 = \frac{Y_{\text{GER}} - \beta_0 - \beta_2 X_{\text{Germany}}}{\text{obs in Germany}} \]

   suppose for example I estimate this and get $\hat{\beta}_1 = .5$. Then
   I could say the German observation is .5 larger than what we'd predict, given its $X_2$ & the mean from other observations, but this is only meaningful if $X_2$ is coded/measured exactly as it was for the other observations so that $\hat{\beta}_2$ is the right coefficient in Germany also.

4. As we saw last week, dummy variables are also useful in considering the effects of some variable(s), $X$ (the $X$s), on $Y$ to differ across groups:

   \[ Y = \{A + \beta X_1 + \epsilon \} \text{ in group 1} = \{C + \beta X_1 + \epsilon \} \]
   \[ (A_0 + \epsilon) + \{D + \epsilon \} \text{ in group 2} = (C_0 + \epsilon) + \{D + \epsilon \} \]

   \[ Y = \{A + \beta X_1 + \epsilon \} \text{ in group 1} = \{C + \beta X_1 + \epsilon \} \]
   \[ (A_0 + \epsilon) + \{D + \epsilon \} \text{ in group 2} = (C_0 + \epsilon) + \{D + \epsilon \} \]
Big conclusion:

"Dummying out" a single observation is exactly the same as dropping it, except:

1. $R^2$ is bigger (in a probably meaningless way).
2. We get a coefficient on the dummy which is meaningful, if all the other X's are coded consistently.

Your estimate of $\hat{\beta}$ will not depend on the observation out.

$R = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$

Your estimate of the Std. Err. of the Regression Coefficient will be the same as if you had left it.

Suppose $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 = -2$. The effect of observations = (in this case) the actual value.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.

You will fit that observation perfectly whenever $X_{i1}$ is in that subset.
Thus, sometimes we create dummies for small groups of observations as a quick (and very crude) check on whether those observations belong in the sample.

Sets of Dummies (more than one group):

1. Disjoint Sets: E.g. German Parties
   - CDU/CSU
   - Social Democrats (SPD)
   - Free Democrats (FDP)
   - Greens & Other (GRO)

   \[ Y = \beta_0 + \beta_1 \cdot CDU + \beta_2 \cdot SPD + \beta_3 \cdot FDP + \beta_4 \cdot X_1 + \beta_5 \cdot X_2 + \ldots + \beta_k \cdot X_k + \epsilon \]

   \[ \beta_0 \text{ is } E(Y | CDU=0, SPD=0, FDP=0, X_1=0, \ldots, X_k=0) \]

   i.e. \( \beta_0 \) is the conditional (on the \( X \)'s) mean for the green & other group.

   *This group is called the “base-line” or “reference” group (or year, or party or city, whatever).*

   (b) The 0-group (“naught group”) is called the “reference” or “base-line group.”

   \[ \beta_0, \beta_2, \beta_3, \beta_4 \] now measure the conditional means of the CDU, SPD, & FDP as they deviate from the reference group mean.

   \[ E(Y | CDU=1, \text{rest}=0) = \beta_0 + \beta_1 \]

   \[ E(Y | SPD=1, \text{rest}=0) = \beta_0 + \beta_2 \]

   etc.

   (c) So, \( H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \) tests what? that conditional mean (CDU) doesn’t differ from Green & Others

2. Over-lapping Sets: E.g. Gender, Marital Status (partially overlapping)
   - Age (30+, 50+, 65+)

   \[ Y = \beta_0 + \beta_1 \cdot FEM + \beta_2 \cdot MSTAT + \epsilon \]

   \[ E(Y | \text{Male, single}) = \beta_0 + \beta_2 \]

   \[ E(Y | \text{Fem, single}) = \beta_0 + \beta_1 + \beta_2 \]

   etc.

   (b) \[ Y = \beta_0 + \beta_1 \cdot 30+ + \beta_2 \cdot 50+ + \beta_3 \cdot 65+ + \epsilon \]

   \[ E(Y | A \leq 30) = \beta_0 + \beta_2 \]

   etc.

   *Just as before – define your vars to facilitate interpretation (e.g. you could always redefine groups to get*
Comments on Dummies

1. Too often dummies are theoretical crutches.

- Survey in 50 US states: "each US state has different "culture" so they'll respond differently" amounts to a dummy variable for each state (or very little more).

  a) Dummies are frequently best left to capture variation in which you are not primarily interested. Usually an effect about which you would theorize would warrant greater attention to the connection of theory to measurement & empirical evaluation.

  b) Sometimes dummies are useful in exactly this way: you may consider it a good test for your theory that when applied across all 50 states using survey data there are no longer differences in mean across these states:

    E.g. suppose you thought GDP growth = f(Education in State) in State

    If you had GDP growth for 50 states over 200 years. Then if

    \[ \text{GDP} = \beta_0 + \sum_{i=1}^{50} \beta_i \text{STATE}_i + \beta_{EDU} + \epsilon \]

    was estimated & an F-test of \( \beta = 0 \) at 1\% fails to reject, you might feel good about how well your theoretical model is doing in that you don't need dumb (dummy) things.

2. Spline Regression: Involves using dummies to fit lines like this

    That is, it's a technique of using dummies not only to allow slopes / intercepts to vary across groups, but also to make the fitted lines "link up".

    (See Greene on this). Generally, though, it will again be preferred to derive the expected shape of the line from your theory & then estimate a function (non-linear in the X's) which allows such a form.

    E.g. y

    True relationship could be estimated by splining like so, but if you expected a hump like this, why not est y = 8 + 2x + 2x^2 + e?

    Few things in socio-pol-ec really have "kinks" like spline functions. Many may be more smoothly non-linear.

    Use a smooth non-linear function of x, esp. since these will usually estimate where "turning points are. Spline typically require you to assume them in the specification.
II. Non-Linearity in Variables

A. We've already talked about some "non-linear in variables" possibilities. Here are some commonly useful examples:

1. (Natural) logs: \( y = \beta \ln x + \varepsilon \Rightarrow \frac{dy}{dx} = \frac{\beta}{x} \) (graphs assume \( \beta > 0 \))

   a) lots of effect at small \( x \)s, increasing less as \( x \) gets bigger.

   \[ \frac{dy}{dx} = e^{\beta x} \varepsilon \beta = e^{\beta x} \]

   (again, assuming \( \beta > 0 \))

   \[ \frac{dy}{dx} = \frac{\beta}{x} \]

   (assuming \( \beta > 0 \))

   \[ \lim_{x \to 0} e^{\beta x} = 0 \]

   \[ \lim_{x \to +\infty} e^{\beta x} = +\infty \]

   \[ e^0 = 1 \]

   \[ e^\infty = \varnothing \]

2. Squares:
\( y = \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon \)

   \( \beta_1 > 0 \) \& \( \beta_2 > 0 \)

   \[ \beta_1 > 0 \) \& \( \beta_2 < 0 \)

   Max occurs at:
   \[ x = -\frac{\beta_1}{2\beta_2} \]

   Min occurs at:
   \[ x = -\frac{\beta_1}{2\beta_2} \]

   \[ x = -\frac{\beta_1}{2\beta_2} \]

   \( \beta_1 < 0 \) \& \( \beta_2 > 0 \)

   \[ \beta_1 < 0 \) \& \( \beta_2 < 0 \)

   \[ x = -\frac{\beta_1}{2\beta_2} \]

   \[ x = -\frac{\beta_1}{2\beta_2} \]

Example: I once had a hypothesis that \( E(y|x) \) was highest when \( x \approx 0.5 \), should taper toward zero at \( x = 1 \) if \( x = 0 \) \( \left( \text{permissible range of } x \right) \). Estimated \( y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \varepsilon \)

3. Polynomials: Recall Taylor-series approximations; \( \Rightarrow \) estimates such that \( \beta_0 \), \( \beta_1 \), \( \beta_2 \), \( \varepsilon \leftarrow 0 \) what have the approximations to arbitrary function by

\[ \Rightarrow \text{can approximate arbitrary function by a polynomial}\]

\( y = f(x) \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots \)
I. Interactive Hypotheses: Specifying, Estimating, Evaluating, & Interpreting
(see Franzese & Kam (UK Press, forthcoming) for Brambor et al. PA '05)

A. Effect of X on Y depends on Z. Examples

1. Examples
   - Z = 1 or Z ≥ Z is "necessary condition" for effect of X on Y;
     e.g., ifideal point princ. & agent equal, then no princ-agent effects
   - Institutions: e.g., NoP = f(NGS, DM, NGSE, DME)
     \( \frac{d\text{NoP}}{d\text{NGS}} = f(DM) \) & \( \frac{d\text{NoP}}{d\text{DM}} = f(NGS) \)

2. Institutions, Institutions: DM affects whether optimal pub-pul. mnp.
   more distr. or redist. (pub good)
   \( \uparrow \text{PD} \Rightarrow \text{this effect:} \)
   e.g., \( \uparrow \text{PD} \Rightarrow \text{better able respond to these demands} \)
   \( \uparrow \text{PD} \Rightarrow \text{partisan as opposed distinct viewpoints} \)

3. "Context Matters": \( \frac{\partial Y}{\partial X} \) depends on characteristics of "context"
   \( \uparrow \text{PD} \Rightarrow \) culturally more distr.
   \( \uparrow \text{PD} \Rightarrow \) culture

4. Delegation/Princ-Agent/MultiHands on Wheel
   \( Y = h(z) \cdot f(x) + (1 - h(z)) \cdot g(z) \)
   \( \frac{\partial Y}{\partial z} = \frac{\partial h(z) \cdot f(x)}{\partial z} + \frac{\partial (1 - h(z)) \cdot g(z)}{\partial z} = b(z) \)

B. One Simple & Useful Way to Model "Effect of X depends on Z":
   "Linear-Multiplicative Interaction Term"

1. \( Y = b_0 + b_X X + b_Z Z + b_{XZ} X Z + \ldots + e \)

2. \( \Rightarrow \frac{\partial Y}{\partial X} = b_X + b_{XZ} Z \)
   \( \Rightarrow \frac{\partial Y}{\partial Z} = b_Z + b_{XZ} X \)
   \( \Rightarrow \frac{\partial Y}{\partial X} \equiv \frac{\partial Y}{\partial Z} \equiv \frac{\partial Y}{\partial X \partial Z} = b_{XZ} \)

3. \( b_X (b_{XZ}) \) are "effect of X (of Z)" when Z (X) = 0.
   a) "Main Effect:"
   b) Generally good idea to include \( b_X \):
      i) Occam's Razor: gives simpler linear-additive a shot
      ii) Allow Intercept to Effect Line: prudent caution -- guards
   against nonlinear interaction or
   (see below)
   iii) However, neither logically nor statistically necessary -- \( b_X = 0 \)
      if both theory & evidence suggests not needed, then omit;
      may be good idea -- efficiency.

4. Nothing new regarding estimation -- sidebar on "Hierarchical/Structural" Models
   See Franzese & PA '05)

C. Evaluating Hypotheses:
1. \( \frac{\partial Y}{\partial X} \) or \( \frac{\partial Y}{\partial Z} \) ≥ 0:
   \( \frac{\partial Y}{\partial X} = b_X + b_{XZ} Z \Rightarrow b_X = 0 \equiv b_Y = b_{XZ} = 0 \)
   \( \Rightarrow \frac{\partial Y}{\partial Z} \geq 0 \Rightarrow b_Z + b_{XZ} Z \geq 0 \)
\[ \begin{align*}
0 \quad & E(Y) = f(x, B) \\
2 \quad & E(\varepsilon) = 0 \\
3 \quad & Var(\varepsilon) = \sigma^2 I \\
4 \quad & Cov(\varepsilon, X) = 0 \quad E(\varepsilon | f(x, \beta)) = 0 \\
5 \quad & \nabla_B f(x, B) \text{ of full col. rank} \\
\hline
X \text{ non-stoch} \quad \varepsilon \sim N(0, \sigma^2 I) \\
\end{align*} \]

\[ \begin{align*}
\text{Min} \quad & \left[ y - f(x, B) \right] \cdot \left[ y - f(x, B) \right] \\
- \frac{1}{2} \left[ \nabla_B f(x, B) \right] \cdot \left[ y - f(x, B) \right] \quad \text{wrt} \quad B & = 0 \\
- \frac{1}{2} \left[ \nabla_B f(x, B) \right] \cdot \dot{y} + \frac{1}{2} \left[ \nabla_B f(x, B) \right] \cdot \dot{f}(x, B) & = 0 \\
\left[ \nabla_B f(x, B) \right] \cdot \dot{y} & = \left[ \nabla_B f(x, B) \right] \cdot \dot{f}(x, B)
\end{align*} \]
A. Over-Fitting: "...your inferences about how the systematic vs. stochastic parts of the world actually work will be wrong.

Again, this is an epistemological point as much as anything.

Unless you actually think the world is 100% systematic and you think your model is 100% correct, explaining everything is a huge mistake -- it actually explains nothing.

True however you go about "explaining" and "inferring." (opinion)

⇒ Many Quantitative & Most Qualitative analyses probably over-fit their data (what historically happened)

⇒ No theory should fit every event in its domain perfectly (pretty sure this is not opinion)

B. "Included Variable Bias": don't forget that you want to include causal priors not causal posteriors

i.e. If your theory is:

\[ X \rightarrow Z \rightarrow Y \]

If you want to estimate the effect of X (e.g., full stop) then the fact that Z is correlated with Y(X is irrelevant.

\[ \text{the effect of X controlling for Z is not what you want to estimate, so leave Z out.} \]

i.e. "bias" always means relative to the true model you care about. Suppose:

\[ X \rightarrow Z \rightarrow Y \]

but also: (i.e., not all of Z caused by X)

⇒ create: \[ Z^* = Z - \hat{Z} \hat{X} \] (i.e., residuals from reg. of Z on X)

⇒ Estimate: \[ Y = \beta + \beta X + \beta Z^* + \epsilon \]

⇒ Efficiency gain

⇒ Causally Prior does not necessarily mean temporally prior (in the simplest sense).

E.g., Forward-looking Expectations (whether Rational Expectations or otherwise) in Lucasian sense.

⇒ People's expectations of what will happen change their actions, then (expected) outcomes can cause prior events

⇒ n.b. Actually this is still temporally prior:

people form expectations prior to actions.

In this sense, (I think) prior causes are always temporally prior as well, just need to think about it a bit.
III. Interaction Terms A Interactive Hypotheses

A. Quite frequently in social science (especially as our theoretical progress has continued), our hypotheses come in the form: "the effect of [some variable] X [other variable] on [some dependent variable] depends on [some other variable]." Examples:

1. The effect of education [some individual on higher income] depends on the level of education in the economy as a whole.
2. The effect of constituent interests (say, pro-anti-tariff) on policy depends on the nature of democratic representation (majority vs. plurality).
3. The effect of economic conditions (UE, Infl, Growth) on presidential approval depends on economic conditions abroad (as a measure of how citizens might evaluate the domestic economy's performance relative to what they might think would be in current global circumstances).

B. Consider these statements that "something depends on" something else the same way we always should, i.e., something = f (something else), i.e., something is a function of something else:

1.) Then above becomes, letting $I_i = \text{indiv. } i$'s income, $E_{dj} = \text{avg. ed. in economy } j$,

$$I_i = f(E_{di}, E_{dj}) = E_{di}$$

Consider: $I_i = \beta + \alpha E_{di} + \epsilon$

but: "effect of $E_{di}$ on $I_i$ depends on $E_{dj}$." i.e., $\beta = \text{effect of } E_{di} \text{ on } I_i = \alpha_0 + \alpha_1 E_{dj}$

$\Rightarrow I_i = \beta + (\alpha_0 + \alpha_1 E_{dj}) E_{di} + \epsilon$

$\Rightarrow I_i = \beta + \alpha_0 E_{di} + \alpha_1 E_{di} E_{dj} + \epsilon$

$\Rightarrow \frac{\partial I_i}{\partial E_{di}} = \alpha_0 + \alpha_1 E_{dj} = \beta.$

n.b. $\frac{\partial I_i}{\partial E_{di}} = \alpha_0 E_{dj}$ thus, we have implicitly assumed that the effect of $E_{dj}$ on $I_i = \frac{\partial I_i}{\partial E_{dj}}$ is zero when $E_{dj}$ is zero.

*More generally, we wouldn't want to assume such a thing. Similar to including a constant, we'd want to estimate the "constant" in the effect line: $\frac{\partial I_i}{\partial E_{di}} = \alpha_0 + \alpha_1 E_{dj}$ which means we should rewrite the model:

$$I_i = \beta + \alpha_0 E_{di} + \alpha_1 E_{di} E_{dj} + \alpha_2 E_{dj} + \epsilon$$

$\Rightarrow \frac{\partial I_i}{\partial E_{di}} = \alpha_0 + \alpha_1 E_{dj}$  \(\text{(the effect of } E_{di} \text{ on } I_i \text{ depends on } E_{dj})\)

$\Rightarrow \frac{\partial I_i}{\partial E_{dj}} = \alpha_2 + \alpha_1 E_{di}$  \(\text{(the effect of } E_{dj} \text{ on } I_i \text{ depends on } E_{di})\)

2.) How now do we evaluate some hypotheses we might have?

a) The effect of individual education level [depends on the national Ed level]: on individual income

$$\frac{\partial I_i}{\partial E_{di}} = f(E_{di}) = \alpha_0 + \alpha_1 E_{di} ; \text{ does this depend on } E_{dj} ?$$

\(\alpha_0 \neq 0 \rightarrow \text{linear model. \(\alpha_0 \neq 0 \), if}\)

b) The effect of national ed. level [on individual income]

depends on the individual ed. level

$$\frac{\partial I_i}{\partial E_{dj}} = f(E_{dj}) = \alpha_0 + \alpha_1 E_{di} ; \text{ does this depend on } E_{dj} ?$$

\(\alpha_0 \neq 0 \rightarrow \text{standard } t\text{-test if}\) \(\alpha_0 \neq 0 \), if does

$$\Rightarrow H_0 : \alpha_0 = 0 \Rightarrow \text{some test}$$

These are the same hypotheses, may sound like different
III. Interactive Terms & Hypotheses

B. 2: Some hypotheses:

b. Effect of $E_i$ on $I_j$ is a function of $E_j$ \( \Rightarrow \) Effect of $E_i$ on $I_j$ is a function of $E_j$.

\[
2 \left( \frac{\partial^2 I_j}{\partial E_i \partial E_j} \right) = \alpha_i = \left[ \frac{\partial \left( \frac{\partial^2 I_j}{\partial E_i} \right)}{\partial E_i} \right]
\]

l) How does the effect of $E_i$ on $I_j$ change as $E_j$ changes?

\[
\frac{\partial I_j}{\partial E_i} = \alpha_i + \alpha_j E_j, \quad \frac{\partial (\alpha_i + \alpha_j E_j)}{\partial E_j} = \alpha_i,
\]

\[
\frac{\partial I_j}{\partial E_j} = \alpha_j + \alpha_j E_j, \quad \frac{\partial (\alpha_j + \alpha_j E_j)}{\partial E_j} = \alpha_j.
\]

- So how about a confidence interval for the effect of $E_i$ on $I_j$?

\[
\text{Effect} = \frac{\partial I_j}{\partial E_i} = \alpha_i + \alpha_j E_j.
\]

\[
V(\text{effect}) = V(\alpha_i + \alpha_j E_j) = V(\alpha_i) + E_j^2 V(\alpha_j) + 2 \text{Cov}(\alpha_i, \alpha_j) E_j
\]

\[\Rightarrow \text{Conf. Interv.} (\alpha_i + \alpha_j E_j) \pm t \cdot \sqrt{V(\text{effect})}
\]

- So, coeff. on var. by itself is just effect when the other term is zero.
- Notes: Effect depends on level of other variable ($\alpha_i \neq 0$)
- So does std. error of the effect: $V(\text{effect}) = f(E_j, E_j^2)$

\[\Rightarrow \text{Why does it tend to bow in?}
\]

\[\Rightarrow \text{V(} \text{effect) greater with } E_j^2\]

\[\Rightarrow \text{V(} \text{effect) less with } E_j^2\]

\[\Rightarrow \text{W(} \text{effect) tends to be negative}.
\]

\[\Rightarrow \text{Why?} \quad \Rightarrow E_i \& E_j \text{ tend to be positively correlated}
\]

\[\Rightarrow \alpha_i \text{ coeff. } \& \alpha_j \text{ coeff. tend to be negative.}
\]

\[\Rightarrow \text{Fact that both } \frac{\partial I_j}{\partial E_i} \text{ and } \frac{\partial (\partial I_j}{\partial E_i} \text{ depend on } E_j
\]

\[\Rightarrow \text{Hypoth. tests that "effect of } E_i \text{ on } I_j = 0"\]

\[\Rightarrow \text{well depend on the level of } E_j \text{ at which you evaluate them (i.e. the conf. int. above)}
\]

\[\Rightarrow \text{Test that } \alpha_i = 0 \text{ is just test that effect of } E_i \text{ is zero when } E_j = 0. \text{ This may be (in fact, frequently is) not a substantively meaningful question.}
\]

\[\Rightarrow \text{Symmetric for } \alpha_j \text{ and } \alpha_k \text{ of } \alpha_i \text{ for the intercepts of the respective effect lines - just as with other "intercept-type-things" we often don't care a whole lot about them.}
\]

\[\Rightarrow \text{slope of the two effect lines are the same (} \alpha_i \text{).}
C. Effect of $E_d$ on $Y_i$ is not zero \( \Rightarrow \) we just saw that, given $\alpha_i \neq 0$ we evaluate this at same level of $E_{d_i}$. The results will vary.

1. **However, hypothesis that individual $i$'s income depends on his education level is more general.**

\[
\frac{\partial Y_i}{\partial E_{d_i}} = \alpha_0 + \alpha_1 E_{d_i} \neq 0
\]

2. So, one more general test would be $H_0: \{ \alpha_1 = 0 \}

\Rightarrow \text{ Wald or Fratio or LR-type test of } \alpha_0 \neq \alpha_1

\text{ Symmetric for individual } i \text{'s income depends on } E_{d_i}.

\text{individual } i \text{'s income depends on neither her/his education nor the } \text{ ini. ed. level.}

3. Of much use most of the time: are these "effect line" plots

\[
y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x \cdot z + \epsilon
\]

\(\text{cont. interval (or simply } \pm \text{ s.e.)} \Rightarrow \text{which is like a } 95\% \text{ c.i. or something}

\]

C. So, many of the non-linearities we saw before & also many of the uses of dummy variables can also be seen as interactive terms.

1. Diff. slope & intercepts in two groups:

\[
x \{ \begin{cases} \beta_1 & \text{group } 1, \\ \beta_0 & \text{group } 2 \end{cases} \}
\]

\[
y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x \cdot z
\]

\text{slope in group } 1 = \beta_2 + \beta_3 \text{ & in group } 2 = \beta_3 \text{ if group } 1

\text{intercept in group } 1: \text{E}(Y|X=1, z=0) = \beta_0 + \beta_1

\text{in group } 2: \text{E}(Y|X=1, z=0) = \beta_0

2. Squared Terms:

\[
y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon
\]

\[
\frac{\partial Y}{\partial x} = \beta_1 + 2 \beta_2 x
\]

\[
V(\frac{\partial Y}{\partial x}) = V(\beta_1) + 4x^2 V(\beta_2) + 2x \cdot 2 \text{ Cov}(\beta_1, \beta_2)
\]

3. You can mix & match all this stuff too:

\[
y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \beta_4 z \cdot x + \beta_5 z \cdot x^2 + \epsilon
\]

\(\text{again, the danger is:}

\(\text{we will have a lot of correlation here, over-fitting again on this later)}

\[
\frac{\partial^2 Y}{\partial x^2} = \beta_2 + 2 \cdot 2 \beta_5 x + \beta_5 z + 2 \beta_5 x \cdot z
\]

\[
V(\frac{\partial^2 Y}{\partial x^2}) = V(\beta_2) + \text{X}^2 V(\beta_5) + \text{X} \cdot \text{X} V(\beta_5) + 2 \cdot \text{X} \cdot \text{X} \cdot 2 \text{ Cov}(\beta_2, \beta_5) + 2 \cdot \text{X} \cdot X^2 \cdot \text{ Cov}(\beta_5, \beta_5)
\]
A. In most of our preceding discussions, we have been assuming that
we know the correct model. So really, there remains the somewhat
prior question of how we
determine the correct model for the "true" model or, still more realistically,
the appropriate candidate model.

B. How do we choose between formally competing models? (See Figure 3)

1. Nested models: If one model is just an addition or deletion of variables in

   the other, we can use the likelihood ratio test or the Chow test

2. More general: Approaches to model selection

   a) Stepwise: forward selection, stepwise regression, etc.
   b) Information theory
   c) Cross-validation
   d) Other methods

3. Less mechanistic approaches to model selection

   a) Goodness of fit tests: R^2, adjusted R^2, etc.
   b) Diagnostic plots and residuals
   c) Subjective criteria

4. In practice, we don't know the correct model. So instead of asking the question of
   how we determine the correct model for the "true" model or, still more realistically,
   how we determine the appropriate candidate model, we might ask ourselves,
   how do we choose among the candidate models that remain after we have
   eliminated the incorrect models?

5. If we have a large number of candidate models, we need a computer to help us
   with the calculations.
IV. C.1. Omitted Variable Bias

- \( b^* \) biased if \( AX_2 \neq 0 \) \( \Rightarrow \beta_2 \neq 0 \)

- How much is \( b^* \) biased? by \( +A \hat{X}_2 \hat{\beta}_2 \)

- coefficient on \( X_2 \) in fitting \( y \)

- \( \text{matrix of coefficients on } X_1 \text{ in fitting } X_2 \)

  - in simple trivariate case: \( y = X_1 \beta_1 + X_2 \beta_2 + e \)

  - but we estimate \( \hat{y} = X_1 \hat{b}_1 + e \)

  \[ \Rightarrow \hat{b}_1^* = \hat{b}_1 + b_{1x_2} \cdot b_{2x_2} \]

  - what it ought to be

- with more than one included or omitted, this depends on all the coefficients in a way that could be figured out, but is more difficult intuitively, still...

- You can figure out sign of bias if can say something about \( b_{1x_2} \) and \( b_{2x_2} \)

- It can guess the magnitude of bias if can say "about size \( \hat{b}_1^* \)"

- What about \( \text{s.e.}(\hat{b}_1^*) \)?

  - if \( X_2 \) relevant, \( \text{s.e.}(\hat{b}_1^*) \geq \text{s.e.}(\hat{b}_1) \)

  \[ \Rightarrow \text{estimate larger/smaller} \]

  \[ \Rightarrow \text{biased} \]

  \[ \Rightarrow \text{bias estimate of true} \ (\text{s.e.}) \quad \text{-- the true stochastic variation} \]

- Possibilities

  1. \( \beta_2 = 0 \) \( \Rightarrow \) leaving \( X_2 \) out was right thing to do, \( b^* \) is BLUE

  2. \( \beta_2 \neq 0 \), but \( AX_2 = 0 \) \( \Rightarrow \hat{b}_1^* \) unbiased (i.e. \( b_{1x_2} \) too high) \( \Rightarrow \) T-student too low likely

  \[ \Rightarrow \text{make your mistakes/tests tougher on yourself (inefficiency) -- generally considered a lesser sin (wasteful)} \]

- Recall \( s_{xx} = \sum (x_i - \bar{x})^2 \)

- \( \text{s.e.}(\hat{b}_1^*) \) biased by \( V(\hat{b}_1^*) = \frac{s_{xx}}{\hat{\sigma}_e^2} \)

- \( \Rightarrow \) Tough to say much about whether T-type tests too hard or too easy

- "Upshot: easy to say (I almost always do) that "you left something out" by itself, this doesn't say anything yet, any such criticism will need to consider all these known properties of such omissions"

- Again, this a fact of inference -- the presence or absence of the problem has nothing to do with whether we evaluate our evidence statistically -- the myth merely clarifies what is a logical fact

b) omitted, irrelevant variables:

- Good: You want to omit irrelevant stuff:

  1. since irrelevant, no bias for includeds (b's or s.e.'s)

  2. if you include it, then s.e. unchanged

\[ \text{s.e.}(\hat{b}_1) = \frac{s_{xx}}{n} \cdot V(\hat{b}_1) \cdot \text{Cov}(X_1, X_2) \]

\[ \Rightarrow \text{or unchanged} \]
c) included relevant variables: Good. You want to include relevant stuff.
   (generally speaking)
   benefit: b's unbiased
   cost: typically s.e. (b's) higher — correctly higher though

   \[ \text{Mean Squared Error} = (\text{bias})^2 + \text{Variance} \quad (\text{saw this before}) \]

   \[ \Rightarrow \text{If we want min. MSE, then some bias may be} \]
   \[ \text{acceptable if \text{Variance goes down enough to more} } \]
   \[ \text{than compensate.} \]
   \[ \Rightarrow \text{Greene has more on this, but biased estimators have} \]
   \[ \text{received little attention \& use (except perhaps for Bayesian/s) } \]
   \[ \text{n.b. if don't know amount or even sign of bias,} \]
   \[ \text{hypothesis testing becomes impossible} \]

d) included irrelevant variables:
   True Model is \( y = X_\beta + \epsilon \) (i.e. C.W. LRN)
   we estimate \( y = X_\beta + X_\epsilon + \epsilon \)

   \[ \text{No bias: } E(b_\beta) = \beta, \quad E(b_\epsilon) = 0 \quad (\text{by OLS}) \]

   \[ \text{Std. Errs: } V(b_\beta) \text{ too high!} \]

   \[ \text{Why? } V(b_\beta) = f(S_\epsilon^2, V(X), \sigma^2) \]

   \[ \Rightarrow \text{V(b_\beta) too high} \]

   \[ \text{too high} \]

   \[ \text{prob. slightly too high} \]

   \[ \text{almost certainly increased} \]

   \[ \text{more costly the smaller is n} \]

   \[ \text{cannot have decreased} \]

V. Cleaning Up:
A. "Over-fitting":
   Recall world has systematic (X) \& stochastic (\epsilon) components. In limited samples (i.e. not infinite) if you keep adding X's then just by chance in your sample, you'll start to fit \epsilon also.

   How? Intuition: suppose you start creating dummy variables for individual observations. Then when you have n=k, you have fit every observation perfectly.

   Alternative Intuition: for any function \( f(x, \epsilon) \), we can approximate it arbitrarily closely by Taylor Expansion (linear approximation). If you keep adding variables, you'll fit the function \( f(X, \epsilon) \) too well in your sample — fitting some of the peculiarities in your data as well as its systematic components

   So what? Isn't fitting as much as possible good?

   No. If your estimated model is fitting peculiarities in your particular sample, then next sample you apply the model to will give you wrong answers. (worse than had you fit only systematic component.) Need to understand that this happens "bc your model fit" stuff beyond the true systematic component -- as such, even if you never get another sample, your answer is still wrong about how the systematic \& stochastic...
FEMALE, MARRIED
FEMALE, UNMARRIED
MALE, MARRIED
MALE, UNMARRIED

\[ \beta_0, \beta_1, \beta_2, \beta_3, \gamma_0, \gamma_1, \gamma_2, \gamma_3 \]

FEMMAR, FEMUNM, MALMAR, MALUNM

FEMMAR: \[ \alpha_0 = \beta_0 + \beta_1 + \beta_2 + \beta_3 \]
FEMUNM: \[ \alpha_1 = \beta_0 + \beta_1 \]
MALMAR: \[ \alpha_2 = \beta_0 + \beta_2 \]
MALUNMAR: \[ \alpha_3 = \beta_0 \]

FEM, MAL, MAR, UNMAR

Partial Collinearity

"Effect of Female":
\[ \frac{\Delta Y}{\Delta GEN} = \beta_1 + \beta_3 \text{MAR} = (\alpha_1 - \alpha_3) \times \text{MAR} \]

Need some: \( \omega_0, \omega_1, \omega_2, \omega_3 \)

FEM, MAL, MAR, FEMMAR

\[ \frac{\Delta Y}{\Delta GEN} = (\omega_0 - \omega_1) + \omega_3 \times \text{MAR} \]
I. Multicollinearity

A. Perfect Collinearity: \( \text{Var}(\hat{\beta}_k) = \sigma^2 \frac{1}{n-k} \) for an exact linear function of the other \( X \)s, then \( R_k^2 = 1 \) \& \( V(\hat{\beta}_k) = 0 \) (or undefined)

- This is because as we've shown before, if \( X_k \) is a perfect linear function of some other \( X \) (or \( X_k \)), then the \( \hat{\beta} \) which solves \( \min \beta (X'X)^{-1} (X'Y) \), the OLS problem is not unique.

- Equivalently, \( X_k \) perfectly collinear with some combo of other \( X \)s means \( X \) not of full column rank (this by def, full column rank which then means that \( (X'X) \) is not of full column rank, which means \( (X'X) \) has determinant zero \( \rightarrow \) therefore \( (X'X)^{-1} \) doesn't exist.

Example: (from Greene)
\[ C = \beta_0 + \beta_1 \text{non-\text{arabic}} + \beta_2 \text{salary} + \beta_3 \text{income} + E \]

- but, \( \text{income} = \text{salary} + \text{non-arabic} \), \( \Rightarrow \) for any \( \beta_0, \beta_1, \beta_2, \beta_3 \) which work for this model,
- \( \Rightarrow \) for any \( \beta_0, \beta_1, \beta_2, \beta_3 \), \( \beta_0 + \beta_1, \beta_2 + \beta_3 \) \& their sum \( \beta_0 + \beta_2 + \beta_3 + \beta_4 \) also solve the problem
- \( \Rightarrow \) an infinite set of \((\beta_0, \beta_1, \beta_2, \beta_3)\) could solve this \( \neq \) not any \( \beta_0, \beta_1, \beta_2, \beta_3 \) but any in a relationship as given here.
- Thus, we say the model are unidentified parameters of this model.

Perfect Collinearity \( \Rightarrow \) Model is unidentified \& therefore unanswerable.

B. (Near) Multicollinearity:

1. Perfect Collinearity is neither a dumb mistake or a failure of the researcher to think clearly about the question she wants the data to address.
   When it's not, there's still not much that can be done about it absent more info (assumptions). In other words, that question can't be addressed by statistical analysis & addressing it otherwise does not solve the problem. Only more info/assumptions for changing the question will "solve" the problem.

2. Correlation among \( X \)s, but not perfect correlation:
   a) Fact: Nowhere have we assumed to date that \( X \)s were uncorrelated; only that not perfectly correlated -- 
   \( \Rightarrow \) OLS is best among linear, unbiased estimators.

   b) It's just that best isn't going to be particularly good when the \( X \)s highly correlated. Symptoms usually ascribed to multicollinearity:
      i) Unstable Parameter Estimates: small changes in the data can have large impact on parameters
      ii) High Std. errors for individual coefficients even if groups are jointly significant & \( R^2 \) high
      iii) Coefficients have "abnormal" or "implausible" magnitude in the sense that, naively interpreted, forgetting that you're controlling for highly correlated over factors, they appear
I. B. Multicollinearity

Reminder: these coefficients are still BLUE, so, they are

1) "correctly" unstable in the sense that these are
   the minimum variance, unbiased, linear b's (MVUE)
2) "correctly" high variance in that these
   s.e.'s, though large, are as small as possible
   among all linear unbiased estimates
3) the coefficients are not "wrong" sign or
   "implausible" magnitude --- given these data
   that is the unbiased \( b \) --- it is either the theory
   that was "implausible" or the theorist who failed
   to keep in mind that the b's are the effect
   of \( x \), keeping all other \( x \)'s fixed.

So, problem w/ multicollinearity is that \( b \)'s have large (though correctly
large) standard errors. If that parameter estimates are potentially "over-
"sensitive to slight differences in the data. Why?

1) Coefficients in multiple regression are effects holding all other indep.
variables fixed. When \( x \) correlated with other \( x \)’s, there isn’t much
info. to be had on this question. Recall Ballentine:

\[
b_0 \text{ is based on } \frac{\text{Cov}(B, A)}{\text{Var}(B)} \text{ after we net out the variation of } C. \text{ So, it's based on } \frac{\text{B}^T \text{B}}{\text{B}^T \text{X} \text{X}^T \text{B}}.
\]
When \( C \) & \( B \) highly correlated, it looks more like:

\[
\begin{array}{ccc}
A & B & C \\
\odot & \odot & \odot \\
\end{array}
\]

2) Very little info on which to base \( b_0 \) (A, b)

(ii) The "problem" of multicollinearity is, thus, exactly the same
as the "problem" of few observations: too little information
in the available evidence about the question being asked.

Solutions:

1) Source of "problem" estimated \( \sqrt{\text{Var}(b)} = \frac{\text{SE}(b)}{\sqrt{n}} \)
   first point & it cannot be stressed enough: if you have no more information
   to bring to bear & if you cannot or will not change the question being
   asked somehow, then there is nothing that can be done because,
   multicollinearity or not, OLS is the BEST LINEAR UNBIASED ESTIMATOR.

Thus, all the so-called "solutions" amount to bringing more info to bear for
changing the question being asked of the data.

a) Drop (some) \( x \)’s: this amounts to changing the question, no longer "effect of
   \( x \)" controlling for \( x \), but "effect of \( x \)" not controlling
for \( x \)" --- if \( x \) belongs in the model (\( \text{cov}(x, y) \neq 0 \)),
then biased (\( x \) correlated w/ \( x \), or there was no "problem"

b) Test \( b_k = 0 \) & \( b_0 = 0 \) rather than \( b_k = 0 \). This is obviously changing the
   question, but perhaps it’s all you need/want/you can say.

c) Impose Restrictions: e.g. setting \( b_k = 0 \) (which is same as dropping \( x \))
is bringing outside info (or assumptions) to bear.
   \( \rightarrow \) any such restrictions (e.g. \( b_k = b_0 \)) could help --- 1) the restriction
   is true that is, if it is not, this is bias.

d) Relax the unbiasedness criterion for estimation (Ridge Estimators). This
   also is changing the question. More over, with \( b \) biased, we do not know
how to make inferences unless we know how it is biased & we don't know any
8.3 Multicollinearity: "solutions"

c. Get more data
   i) ideally, if you can get new data where R_s^2 is lower, then
      can get much better estimates (use all the data)
   ii) even if you can't, more data (more info) is still better (just
      not as much better)

   Multicollinearity is just a version of a small N problem -- get more
   data if you can.

d. Collapse some of your X's into composite indicators:

   Suppose theory says "material wealth of a country contributes to its
   degree of democracy" if you use G=CD/Pop, C=DM/Pop, & D=Income/Pop, as measures of
   "material wealth"

   Options:
   (1) \( Y = \beta_0 + \beta_1 G + \beta_2 C + \beta_3 D + \epsilon \)
   (2) \( Y = \beta_0 + \beta_1 (C+D) + \epsilon \)

   \( b_1 \) estimate of \( \beta_1 \) may be quite imprecisely measured in (1)
   \( \beta_1 \) may be quite imprecisely measured in (1)
   \( R^2 \) (mult. corr. of C with G&D) is likely high.
   \( R^2 \) would probably be more precisely estimated

   or you could use F-test of \( b_1, b_2, b_3 \) since
   that's what you asked about anyway

   Suppose you actually wanted to distinguish \( \beta_1, \beta_2, \beta_3 \) though
   estimate (1) that's best you can do (get more
   data if you can).

   There's a possibly more efficient way to go about option (2).
   Suppose we think that G, C, D are all (imperfect) indicators of some
   underlying characteristic -- say "material wealth"

   Then consider problem of finding some X variable
   which could capture as much as possible variation
   in (G, C, D). This is "principle components" analysis --
   or more generally, factor analysis.

   We're not going to go very deeply into it, but some thoughts

   i) Principle components is kinda like Backward Regression:
      rather than finding \( b \) which gives \( Y = Xb \) which best fits
      \( Y \), we take \( Y_1, Y_2, Y_3 \) and try to find the single X
      which would predict all \( \{ \text{set of } X \}'s \)
      three as best as possible.

   ii) Also called "orthogonal decomposition" because you take
      some N-dimensional array of vectors \( \{ G, C, D \} \) break
      into orthogonal vectors

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]

   \[ \begin{pmatrix} \end{pmatrix} \]
iii) This sort of hints at the main weakness of principle components: The orthogonal components are just some mathematical construct from the original data – just some recombination of the info (if you use all N factors) or some part of the info (if you use a subset). As such, there is no necessary link between a component and any theoretical variable whatever.

   * E.g., we collapse G, C, & D to \( Y^* \) which is the principal component of G, C, & D. We are sure \( M^* \) is the single vector which captures the most possible of any single vector of the total variation in G, C, & D. We are not sure that that empirical entity, \( M^* \), is in fact a measure of "material wealth." We can argue it is. We can argue it is a good indicator or "proxy" of material wealth – and perhaps we're on good ground in this example – but there's not much by way of theory behind this.

   * **Upshot:** Sometimes collapsing like this is a good idea – perhaps \( K \) is already large (many independent vars) or you'd really prefer not to expand it a whole lot more. Perhaps you have strong reason to believe some set of \( X \)'s are indicators of a single (or smaller set) of underlying factors, but don't have strong theory on just how they combine into these underlying factors. However, if you only care about this underlying factor's effect, often you'd be better served by optimization which uses all available info, not just the orthogonal parts.

   * Always, getting more (relevant) data is a good idea if possible.

   * Multicollinearity is not a problem with OLS or the CNLRM but rather an (unfortunate) fact about your data. Incidentally, this means that the notion of a "test for multicollinearity" is nonsense. Just measure it.\(^{\text{Fk}}\).

   * Once you've exhausted all info you could possibly bring to bear, if once you've considered all these different questions you might ask or ways you might ask the same question, there is nothing you can really do about it any further other than consider the sensitivity of your results to slight changes in the data. (more on this in a bit)

II. Missing Data:

A. There is no relationship between which observations are missing and the object of the study (the dependent variable).

   * OLS unbiased & consistent
   * Inefficient in the sense that the missing observations could have helped

   1. Some X's missing: basically nothing you can do about this

      * why? \( \hat{y} = X\beta + \varepsilon \) if the rest of the CNLRM

      * for any missing observation, you may have X's, but

      * by the CNLRM model, you have no info about \( \varepsilon \) in those cases, so, you have no usable statistical variation.
II. Missing Data

A. Random Missing Data $\Rightarrow$ unbiased, consistent but perhaps inefficient

1) $Y$'s missing = can't do anything about it, just to drop the obs.

(but read on!)

2) Some $X$'s missing:
   
   a) if random $\Rightarrow$ unbiased, consistent but perhaps inefficient to drop
   
   b) if systematic: (e.g., $X$ income, high incomes don't answer)
       
       In the $\Rightarrow$ c) if relationship $y \Rightarrow x$ is non-function of $X$,
       then unbiased $\Rightarrow$ consistent to drop. (Inefficiency still)
       
       $\Rightarrow$ basically, this is same as any other
       
       extrapolation outside the dataset.
       
       $\Rightarrow$ you must assume the model (with
       the ANOVA assumptions) continues
       
       to hold beyond the sample
       
   $\Rightarrow$ if relationship $y \Rightarrow x$ not same, then cannot
   extend results beyond "type" of $X$'s in your
   sample (a form of selection bias if you do)

   c) in Multiple Regression: $Y = \beta X + \epsilon$;
      but some $Z$'s for $X$'s missing,
      
      i) Could use $X$ to predict missing $Z$'s if $Z$'s to predict
         missing $X$'s = (ignoring some minor but complicated
         issues) some efficiency gain

      ii) Better still, use as much outside info as possible
          to model the missing $X$'s for $Z$'s.

   Update $\Rightarrow$ ignoring some issues which arise here which, to be
   more correct would have to be dealt with. Much of the
   time, unless scattering of missing $X$'s in your data
   matrix leads to many incomplete observations on ($Y$, $X$)
   but relatively fewer missing $Y$'s, my intuition/reading
   of the literature is that consensus is, if you're better
   off just dropping or there not a lot to be gained.

   c) if systematic in the $X$'s: i.e., if the reason data are
      missing is related to the phenomenon being studied.

B. Selection Bias (A massively abbreviated introduction)

1) Suppose $Truth = \Rightarrow y \Rightarrow x$ but you can only
   observe $y$ at $X$'s

   $\Rightarrow$ then you'd estimate $\hat{b}$, which is obviously biased.
   This is selection bias $\Rightarrow$ non-random
   missing data

2) Notice, however, that there might be some way to use
   the fact that $X$'s are cut-off $\Rightarrow$ too (probabilistic)

3) Also, if you can think about selection process, i.e., why
   you only see $y \leq X$ observations, then you can model
   that relationship too.

4) Basically, selection bias exists when there is a systematic
   reason you are missing some observations on $y$. N.B., if
   that's so, then you've left out some info about $y$ if
   you just drop the observations. So, it's a version of omitted-
   variable (or omitted-equation or omitted-variable) bias.
III. Grouped Data: if relationship you wish to estimate exists at some level of aggregation (e.g. individual, second, province), but you observe the data at a greater level of aggregation (e.g. county, year, country), then

1) Parameter Estimates are Less Efficient
2) "Fit" of regression is "better", sometimes dramatically so

\[ Y_i = X_i \beta + \varepsilon_i \]

but you observe only \( \bar{Y}_j \) \& \( \bar{X}_j \), where \( \bar{Y}_j \) \& \( \bar{X}_j \) are averages of some groups of \( i \)'s.

i) Simple addition shows that \( \bar{Y}_j = \bar{X}_j \beta + \bar{e}_j \) for some \( \beta \)

- So \( \bar{Y}_j \) \& \( \bar{X}_j \) is still unbiased (consistent as \( j \to \infty \), \# of groups, not \# of \( i \)'s necessary (in OLS))
- In fact, it's BLUE given that this grouping is unavoidable
- If the size of the groups, \( j \), is the same.

ii) If size of groups varies, then

\[ \text{var}(\bar{Y}_j) = \frac{1}{n_j} \text{var}(\bar{X}_j) \]

\[ \Rightarrow \text{OLS not BLUE} \]

- It is unbiased (in most cases) \& consistent (\( \text{in } j \))
- But inefficient even given that nothing can be done about the aggregation.

What to do? \[ \text{var}(\bar{Y}_j) = \frac{1}{n_j} \text{var}(\varepsilon) \]

\[ \Rightarrow \text{var}(\bar{Y}_j) = (\text{var}(\varepsilon))^2 \cdot \frac{1}{n_j} \Rightarrow \text{var}(\varepsilon) = \text{var}(\varepsilon) \cdot n_j \]

So, multiply everything (\( \bar{Y}_j, \bar{X}_j \)) by \( n_j \) \& you get: BLUE

\[ \Rightarrow \text{This is the "Weighted Least Squares" correction for heteroskedasticity} \]

Much more on this later.

3) In either event, the regression based on averages is definitely inefficient relative to the less aggregated regression (assuming the model, i.e. the theory, applies to the less aggregated grouping).

Why? Averaging some group of \( X \)'s \& the corresponding group of \( Y \)'s gets rid of all variation of \( X \) \& of all \( Y \) around their \( \bar{Y} \).

- Wasting Info is always inefficient.

\[ \Rightarrow \text{unless theory does not apply to lesser aggregation, always use the lesser aggregation} \]

\[ \Rightarrow \text{don't mean to imply at all that this "unless" is} \]

\[ \Rightarrow \text{is necessary} \]

4) \( R^2 \) is typically higher (S.E.R. \( (\varepsilon) \) is typically lower).

\[ \text{why? } \sum_{i=1}^{n} e_i^2 > \sum_{j=1}^{J} j \sum_{i=1}^{n_i} e_i^2 \]

\[ = \sum_{j=1}^{J} n_j (\text{var}(e))^2 \]

\[ = \sum_{j=1}^{J} (\text{var}(e))^2 \]

\[ \Rightarrow \sum_{i=1}^{n} e_i^2 \]

\[ \Rightarrow \text{S.E.R. } \frac{\sum\text{var}(e)^2}{\sum\text{var}(e)^2} < \frac{\sum\text{var}(e)^2}{\sum\text{var}(e)^2} \]

\[ \Rightarrow \frac{\text{S.E.R.}}{\text{R}^2} > \frac{\text{S.E.R.}}{\text{R}^2} \]
IV. Measurement Error & Proxy Variables

A. What if we have poor measurement of X as it relates to the theory? I.e., suppose we are arguing that \( y = f(x^*, \varepsilon) \) but unfortunately we don't have \( x^* \) exactly but some (flawed) measure of it.

\[
\text{Theory: } \quad y = x^* \beta + \varepsilon \\
\text{Estimable: } \quad y = Xb + \varepsilon \quad \text{where} \quad X = x^* + u
\]

- \( b = \frac{\text{Cov}(X, y)}{\text{Var}(X)} = \frac{\text{Cov}(X^* + u, y)}{\text{Var}(X^* + u)} = \frac{\text{Cov}(X^*, y) + \text{Cov}(u, y)}{\text{Var}(X^*) + \text{Var}(u)} = \beta \frac{\text{Var}(x^*)}{\text{Var}(x^*) + \text{Var}(\varepsilon)} \)

\[\text{assuming } \varepsilon = \text{measurement error uncorrelated with } y \text{d uncorrelated with } x^* \text{ (true } x) \]

\[\text{(dependent variable)}\]

- but, we wanted \( b^* = \frac{\text{Cov}(x^*, y)}{\text{Var}(x^*)} \) which is higher unless \( \text{Var}(x^*) \) is \( \sigma_u^2 \) is 0

B. \( \text{Var}(b) \) when \( x^* \) measured with random error

- \( \text{estimated } \text{Var}(b) = \frac{\sigma_u^2}{n} (X'X)^{-1} = \text{Cov}(x^*, \text{Var}(X)) \)

\[\text{estimated std. error } \approx \begin{cases} \text{or } & \text{based on } x > \sigma_u^2 \text{ based on } x^* \\
& \text{bc } x \text{ worse predictor of } y \\
& \text{than } x^* \implies \text{(i.e., estimated)} \\
& \text{but } \text{Var}(x) \uparrow (\text{bc } \text{Var}(x^*) + \text{Var}(\varepsilon)) \rightarrow \text{up std. err.} \end{cases} \]

C. But \( T = \frac{b}{\text{se}(b)} \) definitely goes down

- This is tremendously complicated to prove, but intuitively we know \( b \) is too low. I.e., we know both numerator and denominator of \( \text{se}(b) \) are too high. It can be shown that \( b \) is "more too low" than \( \text{se}(b) \) is too low if \( \text{se}(b) \) is too low at all.

\[\text{In bivariate case, generally measurement error in } x \text{ makes our hypothesis tests too hard (a lesser evil).} \]

D. How wrong (biased) is our \( \hat{b} \)?

- Asymptotically: \( \hat{b} \sim n \text{ Cov}(x^*, y) / \text{Var}(x^*) + \text{Var}(\varepsilon) \) whereas \( \beta^* \) (based on \( x^* \)) is asymptotically \( \beta \) (by Gauss-Markov OLS is BLUE)

1) So: \[ \lim_{n \to \infty} \hat{b} = \beta \left( \frac{\sigma_u^2}{\sigma_x^2 + \sigma_u^2} \right) \]

2) \( \hat{b} \) is biased toward zero (attenuated) by an amount proportional to good variation in \( x \) relative to garbage variation in \( x^* \)

\[\text{Estimates of } \beta; \text{i.e., estimates of relationships are only good in proportion to how good the data is} \]

(\text{and of course, this is one of those logical issues having nothing to do with machine learning; the math only clarifies what we should have known already})

3) If we have some info about \( \sigma_u^2 \) (measurement error) we can figure out what effect it's having on our results (i.e., invalidate case)

Even an intuition about \( \sigma_u^2 \) relative to \( \sigma_x^2 \) should help (with eq. 1)
E. Error in \( Y \) (bivariate or multivariate)

Truth: \( \mathbf{Y}^* = \mathbf{X}\beta + \mathbf{\varepsilon} \)

Observed: \( \mathbf{Y} \) which is \( \mathbf{Y}^* + \mathbf{U} \) (again, \( \mathbf{Y} \) is assumed random & uncorrelated with \( \mathbf{X} \) or \( \mathbf{Y}^* \)).

Then estimating \( \mathbf{Y} = \mathbf{X}\hat{\beta} + \mathbf{e} \)

Gives \( \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \)

\( = A \mathbf{Y} = A (\mathbf{Y}^* + \mathbf{U}) \)

\( = A (\mathbf{X}\beta + \mathbf{\varepsilon} + \mathbf{U}) \)

\( = \beta + A \varepsilon + A \mathbf{U} \)

⇒ Measurement Error in \( \mathbf{Y} \) just adds the measurement error into the residual. No big deal beyond this.

F. Measurement Error with Multiple Regression:

\( \mathbf{Y} = \mathbf{X}\beta + \mathbf{\varepsilon} \) but we observe \( \mathbf{Y} = \mathbf{X}\mathbf{U} + \mathbf{\varepsilon} \) (i.e. all \( \mathbf{X} \)'s or any one \( \mathbf{X} \)'s measured with error).

* Even assuming each \( \mathbf{U} \) i.i.d., uncorrelated with anything,

\[ \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \] is biased -- all of the \( \hat{\beta} \)'s are biased.

Why? \( \hat{\beta} \) is \( (\mathbf{X}'\mathbf{X})^{-1} \) mixes it all around -- each element of \( (\mathbf{X}'\mathbf{X})^{-1} \) depends on all the products & cross-products of the \( \mathbf{X} \)'s. (Remember that from matrix algebra week?)

* Generally we are unable to say anything much more precisely since each element of \( (\mathbf{X}'\mathbf{X})^{-1} \) (which is needed to estimate both \( \hat{\beta} \) & \( \mathbf{V}(\hat{\beta}) \)) (remember) depends on a whole bunch of unknowns (all the \( \mathbf{U} \)'s). [Depressing, huh?]

* One thing we can still say is that certainly the bigger is \( \mathbf{U}^2 \) relative to \( \mathbf{\varepsilon}^2 \) (or more precisely in Matrix notation, the more positive definite is \( \mathbf{V}(\mathbf{X}\mathbf{U}) - \mathbf{V}(\mathbf{\varepsilon}) \) the less severe the problem.

This is little more than intuition at this point though -- if same as before at that -- the degree of trust in estimates is a function of the "signal-to-noise ratio".

G. Proxy Variables -- Sometimes the variables suggested by theory are simply unmeasurable or even unmeasurable even in principle. One frequently cited example is intelligence. Others might be capital, human, physical, social.

If, however, we can suggest some observable correlate of the unobservable factor, e.g., IQ Test score, education years, fixed machinery stock, density of associational networks, we can use them as {	extit{proxy}} variables. All the measurement error features apply...
moreover it is now not clear what the coefficients’ magnitude means beyond the clarity with which we might grasp how the proxy relates to the underlying factor. Nonetheless, Monte Carlo experimentation broadly suggests that using proxies is better than nothing (usually) if that’s all you can do.

V. Regression Diagnostics, Influential Points, & Sensitivity Analysis

(See Handout & Greene p6)

A. Outliers & Influential Points: recall that OLS, since it’s based on minimizing squared residuals can be quite sensitive to "extreme" observations

1. Extreme Y’s (i.e. Yi far from y)
   a) if Yi relatively close to regression relation given by rest of data, not inordinate influence:
   b) if, however, Y is far from the "normal" X-Y relationship, a Y-outlier can have massive effect on the estimated relationship:

   Notice: an influential outlier is not necessarily associated with a large residual.

2. Extreme values of X:
   a) Yi in same relation to extreme X, as other (X, Y), then X-outliers actually are good: this produce smaller std. errors of b (remember?)
   b) X which are abnormally related to Y can have massive influence

Notice: these examples reveal that it is most typically observations extreme in both X & Y that have influence (1b & 2b) but this need not be the case:

B. Leverage: You may have noticed from the above that "abnormal" Y’s alone do not produce influence (1b). It is only observations far from X which have the potential (no guarantee) of being inordinately influential.

Distance of X_i from X is this potential we refer to as leverage.

One useful measure of this is from the diagonal of the "fitted-value-maker" matrix, \( N = X(X'X)^{-1}X' \)

\[ p_{ii} = X_i(X'X)^{-1}X_i \]

is the i-th element of the diagonal of N and measures leverage (potential)

(From here, move on to Jim Demer's notes on Regression Diagnostics -- do not cite, please, these are forthcoming)
TOPICS IN DIAGNOSTICS
Importance of Regression Diagnostics
Leverages
Studentized Residuals
Outlier tests
Influence Diagnostics
Useful Plots
Partial Regression Plots
Normality Diagnostics-q-q plots
Non-linearity tests–non-linear smoothers
Collinearity Diagnostics
Heteroscedasticity Tests
Autocorrelation Tests

CASE DIAGNOSTICS: Residuals, Leverages, Influence.

Anomalous observations, even singletons, can completely destroy a regression, given the sensitivity of the least squares loss function to atypical points. It is therefore important to identify observations that may be dominating the estimates, and to understand how much effect they have on the fit.

Kinds of outliers.

There are several kinds of outliers that can arise in a regression problem, all of which are interesting:

1.) Outliers in the x-space. Points that lie far from the centroid of the data matrix X have tremendous leverage on the least squares fit. Such cases can be highly influential if they deviate from the general pattern in the rest of the data.

2). Outliers in the Y direction. Such points are most dangerous when they have leverage.

3). Outliers in the residuals (large deviations from the fit).
Regression Analysis
J. DeNardo

Outliers of one kind need not show up in the other guises. In particular, neither type 1, nor type 2, nor both in combination, necessarily result in type 3. High leverage observations are not necessarily influential, they are only potentially so.

Leverages

The leverage is a basic measure of how far each case lies from the center of the X’s (usually taken as the joint mean or centroid). Leverages are the diagonal elements, $p_{ii}$, of the projection matrix $P$.

$$P = X(X'X)^{-1}X'$$
$$p_{ii} = x_i'(X'X)^{-1}x_i$$
$$p_{ij} = x_i'(X'X)^{-1}x_j$$

where $x_i$ is the $i$th row of the data matrix $X$, and $x_j$ is the $j$th row. Cases with high leverage have the greatest potential for influencing a fit, though that potential will not be realized if the leverage point conforms with the general pattern of relationship in the rest of the data.

It can be shown that

$$1/n \leq p_{ii} \leq 1/c$$

where $c$ is the number of rows identical to $i$ in the data matrix (including $i$ itself). If the data have no replicate observations (typical in non-experimental fields), the upper bound will be 1. The leverage will increase toward the upper bound as the observation gets further away from the joint mean. The lower bound occurs when $x_i$ equals the joint mean of the X’s exactly.

Studentized Residuals

High leverage points attract the least squares fit toward themselves. As a result, they tend to generate small residuals, with little variation around the fit. Recall that
Regression Analysis
J. DeNardo

\[
E(e'e') = E(M\mu \mu'M') = \\
= \sigma^2M \\
= \sigma^2(I - P)
\]

with

\[
\text{Var} \ (e_i) = \sigma^2(1 - p_{ii})
\]

and

\[
\text{Cov} \ (e_i,e_j) = -\sigma^2(1 - p_{ij})
\]

As you can see, the residual variance approaches zero as the leverage, \( p_{ii} \), approaches 1. The fitted residual goes to zero for high leverage cases, \textit{whatever} the value of \( Y_i \).

\textit{Internally Studentized Residuals}

To remove the heteroscedasticity induced by the leverages, we can standardize the least squares residuals by dividing each one by its estimated standard deviation:

\[
r_i = \frac{e_i}{s \sqrt{1 - p_{ii}}}
\]

where \( s^2 = e'e/n - k \).

The standardized residuals behave much like a Student's t random variable, with \( E(r_i) = 0 \), and \( \text{Var}(r_i) = 1 \). However, their distribution is not precisely Student's t, because the numerator and denominator are not independent.

\textit{Externally Studentized Residuals}

Another approach to standardizing the OLS residuals uses an estimated standard deviation that is independent of the residual itself. This is accomplished by deleting the \( i \)th observation, and computing the estimated disturbance variance \( s_{ii}^2 \) on the remaining cases, for all \( i \). The result is the \textit{externally studentized residual}:
$r_i^* = \frac{e_i}{s(i) \sqrt{1 - p_{ii}}}$

These residuals are distributed as Student's t, with $n - k$ d.f. They can be referred to a Student’s t table, to make a formal test for outliers.

**A Dummy Variable Test for Outliers**

Suppose you suspect that a particular case is an outlier from the fit (type 3). Append a dummy variable to the data matrix which equals 1 for the $i$th case and zero everywhere else. The estimated t-ratio associated with this variable is exactly $r_i^*$! When the t-ratio is sufficiently large, one can reject the null hypothesis that the $i$th case belongs to the model. Warning: if one has a hundred cases, with normally distributed disturbances, $\mu$, then there will typically be 5 observations for which $r^*$ will be “significant” at the .05 level. Throwing these cases away, simply because they are “outliers,” will produce biased estimates of the disturbance variance, and all its derivative quantities. Note: with the dummy variable included in the model, the estimates of the other coefficients will be unaffected by the $i$th case.

Neither the $r$ nor $r^*$ residuals are independent. In practise, the internally and externally studentized residuals tend to be very similar, and generally preferable to the raw OLS residuals. Most packages compute them routinely now, including DataDesk and X-Lisp-Stat, SST, SAS, etc.

**Influence Diagnostics**

A *stable* fit should not change very much when the data change only slightly. A number of diagnostic statistics investigate the stability of the fitted estimates when the data are slightly perturbed. Typically, these analyses proceed by deleting one observation at a time, and investigating what happens to the fit.
Remark: Diagnostic measures based on casewise deletion are commonly used to study statistical stability. Examples include a. the Jackknife, b. cross-validation, and c. influence diagnosis. The approach should be compared to the classical measurement of stability (via confidence intervals, etc.), based on a replication (or sampling) model. When the model is believed, no replication is actually done.

*The Sherman-Morrison-Woodbury Theorem.*

On the face of things, casewise deletion appears dauntingly tedious, especially for large datasets. We don’t really want to recompute 1500 regressions to get the studentized residuals in a survey-based study. As it turns out, one can compute everything much more efficiently. The basis for casewise diagnostics is the Sherman-Morrison-Woodbury Theorem, which shows how the inverse of matrix changes when we modify it by deleting one row.

Let \( A \) be a square non-singular matrix, which is \( k \times k \). (In our application, \( A \) will be \( X'X \).) \( z \) is a \( k \times 1 \) column vector. (For us, \( z' \) will be the \( i \)th row of the matrix \( X \).) You can verify that \((A - zz')\) becomes the \( X'X \) matrix with the \( i \)th row not involved (a rank one modification). Then

\[
[A - zz']^{-1} = A^{-1} + \frac{A^{-1}zz'A^{-1}}{1 - z'A^{-1}z}
\]

Notice that no new matrix inversion is required to update the original inverse, making the computations and calculations much simpler.

The point of departure for all the case statistics described below is:

\[
[X_{-i}'X_{-i}]^{-1} = (XX)^{-1} + \frac{(XX)^{-1}x_{i}x_{i}'(XX)^{-1}}{1 - x_{i}'(XX)^{-1}x_{i}}
\]

\[
= (XX)^{-1} + \frac{(XX)^{-1}x_{i}x_{i}'(XX)^{-1}}{1 - p_{ii}}
\]
Regression Analysis
J. DeNardo

Regression Diagnostics

This equation relates \( [X_i'X_i]^{-1} \) to \( (XX)^{-1} \). After multiplying both sides by \( X'y - x_i'y_i \) and simplifying, we get the useful statistic, \( DFBETA_i \):

\[
 b - b._i = \frac{(XX)^{-1}x_ie_i}{1 - p_{ii}}
\]

This is the change in the estimated coefficient vector produced by deleting the \( i \)th case. If we multiply both sides by \( x_i' \), we have:

\[
 \hat{y}_i - \hat{y}_{i,-i} = \frac{p_{ii}e_i}{1 - p_{ii}}
\]

This is the impact on the predicted value for the \( i \)th case, produced by deleting the \( i \)th observation. Standardizing by the standard deviation of the fitted values \( [\text{Var}(\hat{y}) = \sigma^2P] \) gives the diagnostic statistic called DFFITS:

\[
 \text{DFFITS} = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{p_{ii}}} = \frac{p_{ii}e_i}{1 - p_{ii}} \frac{1}{s_{-i}\sqrt{p_{ii}}} = \left( \frac{p_{ii}}{1 - p_{ii}} \right)^{1/2} \frac{e_i}{s_{-i}\sqrt{1 - p_{ii}}} = \left( \frac{p_{ii}}{1 - p_{ii}} \right)^{1/2} r_i^*
\]

DFFITS is the number of estimated standard errors that the fitted value for case \( i \) changes, if the \( i \)th observation is deleted from the data. The effect is the product of two terms; one increases with the leverage, \( p_{ii} \) (type 1 outlier), and the other with the externally studentized residual (type 3 outlier).

Another popular measure of influence is \textit{Cook's Distance}. \( D \) measures distance between the estimated coefficient vectors, with and without the \( i \)th case. The raw distances, \( d_i = b - b._i \), are normalized into standard error units, and combined into a composite scalar summary:
Cook's $D_i = \frac{(b - b_{-i})' (X'X) (b - b_{-i})}{k s^2} = \left( \frac{p_{ii}}{1 - p_{ii}} \right) \frac{r_{i}^2}{k}$

As you can see, the mathematics of diagnostics is rich and interesting, with lots of surprising connections and computational shortcuts. Amazingly, all of these case deletion diagnostics can be computed on one pass, as functions of the leverages and residuals.

**DIAGNOSTIC PLOTS**

This rich array of diagnostic measures, when displayed in a modern graphical computing environment (exploiting color, selection highlighting, brushing, linking, slicing, lassoing, and rotation), provides detailed information about the quality of a regression fit and its statistical stability. These capabilities often produce an embarrassment of riches, since we can easily compute dozens or hundreds of regressions, each of which generates dozens of plots and diagnostics. A good package makes this work move quickly by exploiting a "hot-button" interface or "cognistics" (basically situation sensitive macros). Port-overs from the mainframe era typically have painfully slow interfaces (command-line-like), making the user spell out every step one by one.

Whatever you do, never assume that numerical statistical summaries tell the whole story. Graphical displays are indispensable for interpreting what your regressions mean.

**Inventory of useful plots.**

Here is a list of plots that often prove useful. In the early stage of an analysis, problems show up right away and a complete cycle of plots usually isn't necessary. As one moves toward a better fit, the investigation proceeds in greater depth.
Regression Analysis
J. DeNardo

1. Scatterplots
   a. Of pairs of predictors $x_k$, $x_j$.
   b. Of pairs $y, x_j$.
   c. The scatterplot matrix combines a & b, giving a generalization of the variance-covariance (or correlation) matrix.

2. Residual plots
   a. Studentized residuals vs. predicted.
   b. Studentized residuals vs. predictors, $x_j$.
   c. In time series: residuals vs. lagged residuals.
   d. Residuals vs candidate predictors, or influence diagnostics.

3. Univariate plots (for $y$, $x$, and $r$, leverages, DFFITS or COOK).
   a. Histograms, stem-plots, box-plots, kernel density plots.
   b. Probability plots (q-q plots are linear for normally distributed variables).

4. Partial regression plots.
   a. $y$ vs $x_j$, controlling for all other $x_i$.
   b. A plot matrix of these for all $x$.

5. Rotating Plots
   a. $y$ vs 2 $x$'s.
   b. $y$ vs $y$-hat + an influence statistic. Etc.
   c. $e$ vs $y$-hat + an influence statistic, or vs. $x$'s Etc.
   c. For more variables, 2 rotators with 4-6 variables, using color or highlighting to show connection.

6. Linked plots
   a. All of the above can be enhanced by linking plots together, by coloring cases, or labelling points so connections between the displays become evident.

Experience and imagination are both helpful.
Regression Analysis
J. DeNardo

Regression Diagnostics
- 9

What are we looking for in these plots?

1. Have the Gauss-Markov conditions been violated? (linearity, homoscedasticity, non-autocorrelation, independence of the X’s and μ’s).

   Unfortunately, the last condition is critically important but the hardest to check. Remember that least squares residuals are forced to be orthogonal to the regressors. The first three can be checked more easily.

2. Is normality of the disturbances an unreasonable assumption?

3. Is collinearity a problem?

4. Are particular cases or clusters distorting or dominating the fit? Which ones, and how much?

5. Do the cases fall into distinct groups, calling for separate treatment?

6. Do some cases not fit the model? Why?

If the answer to any of these questions is YES, then the OLS summary statistics become dubious, unreliable, unstable, or hard to interpret. Corrective measures are then called for (like GLS, re-specifying the list of regressors, transformations, robust regression, non-linear regression, partitioning of the dataset, case deletion, collection of new information, etc).

Murder Data Example

We still have some theory to cover, but let’s go ahead with an example, using the Leamer Murder Data (see E. Leamer, “Let’s Take the Con out of Econometrics”). Just for reference, here are the variables:
In "Open Only," the data in the DataDesk file "Death Penalty Data" contains measurements from 44 states, collected during 1950. The measured variables are as follows:

Dependent Variable
- Murder Rate = Murder rate per 100,000, FBI estimate.

Independent Deterrence Variables
- Pr(Conv|M) = Probability of conviction for murder, given commission. Defined by Pr(Conv|M) = C/Q-hat, where C = convictions for murder and Q-hat = Murder Rate * POP, POP = the state population in hundreds of thousands. Note that Murder Rate is an estimate based on a sample from each state.
- Pr(Ex|Conv) = Probability of execution, given conviction. This is the average number of executions 1946-50, divided by C, the number of convictions.
- Time for M = Median time served in months for murder by prisoners released in 1951.
- Exec = A dummy equal to 1 if Pr(Ex|Conv) > 0. Measures presence or absence of a "working" death penalty.

Independent Economic Variables
- INC = Median income of families in 1949.
- Poverty = Fraction of families in 1949 with less than one-half INC.
- Labor Force = Labor force participation rate. Fraction of adult population (16 years or older) in the labor force (employed, unemployed, or in the Armed Forces.)
- UNEMP = Unemployment rate.

Independent Sociological and Environmental Variables
- NON-W = Percentage non-white.
- YOUTH = Percentage 15-24 years old.
- URB = Percentage urban.
- MALE = Percentage male.
- HOME = Percentage of intact families with both husband and wife present.
- SOUTH = A dummy equalling 1 for Southern states.
- POP = State populations in hundreds of thousands.

Weighting Variable
- SQRTNF = Square root of the population of the FBI reporting region. This variable can be ignored for our purposes. Used to construct weighted least squares estimates.

Here's a regression model that theory or prejudice or curiosity might have lead us to consider:
Regression Analysis
J. DeNardo

Dependent variable is: Murder Rate
No Selector
R squared = 64.6%  R squared (adjusted) = 61.0%
s = 2.788 with 44 - 5 = 39 degrees of freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>553.753</td>
<td>4</td>
<td>138.438</td>
<td>17.8</td>
</tr>
<tr>
<td>Residual</td>
<td>303.184</td>
<td>39</td>
<td>7.77394</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.58429</td>
<td>5.630</td>
<td>0.459</td>
<td>0.6488</td>
</tr>
<tr>
<td>Pr(Ex</td>
<td>Conv)</td>
<td>4.12951</td>
<td>6.469</td>
<td>0.638</td>
</tr>
<tr>
<td>Time for M</td>
<td>-0.021329</td>
<td>0.0077</td>
<td>-2.77</td>
<td>0.0085</td>
</tr>
<tr>
<td>Poverty</td>
<td>9.97774</td>
<td>18.02</td>
<td>0.554</td>
<td>0.5829</td>
</tr>
<tr>
<td>Non-W</td>
<td>24.5693</td>
<td>4.116</td>
<td>5.97</td>
<td>≤ 0.0001</td>
</tr>
</tbody>
</table>

In many articles, we would find a table like this, perhaps with some F-tests pitting deterrence theory vs. sociological determinism. Cook book data analysts would be happy enough to let things stand here. The R² is “good” afterall. The only glitch is that the estimate for the deterrent effect of the death penalty has the “wrong” sign. The coefficient is not significant, however. Let’s see how much further the diagnostics and plots can take us into the problem.

First some scatter plots:
These residuals are grossly heteroscedastic. The wedge is obvious.
The residual distribution looks fairly normal, but with fatter tails than expected. I’ve selected the tails in the histogram, and cued up the state names for the highlighted cases. The poorest fitting cases are mostly Southern.

Here I’ve added a non-linear (Lowess) smoother to the residual plot. There is no obvious non-linearity, but perhaps a faint tilt (nothing to write a paper about.)

Just for curiosity, what’s the relationship between the ordinary and the studentized residuals? Very linear.

Now let’s check out some of the influence diagnostics.
In this rotating plot, one case stands way out from all the rest—Nevada. In this fully linkable DataDesk Table, I've selected the most extreme cases on the Cook's D measure. The table will give you a concrete feeling for how the computed influence statistics look, but tables are hard to digest.
## Regression Analysis

### J. DeNardo

<table>
<thead>
<tr>
<th>State</th>
<th>Murder Rate</th>
<th>EStudRes</th>
<th>leverages</th>
<th>DFFITS</th>
<th>Cook</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>19.2523</td>
<td>2.6852324</td>
<td>0.15225479</td>
<td>1.1379802</td>
<td>0.22342149</td>
</tr>
<tr>
<td>AR</td>
<td>7.5286</td>
<td>-1.0616314</td>
<td>0.08525005</td>
<td>-0.32409316</td>
<td>0.02093906</td>
</tr>
<tr>
<td>AZ</td>
<td>5.6567</td>
<td>-0.4952436</td>
<td>0.05144756</td>
<td>-0.1153752</td>
<td>0.00027135</td>
</tr>
<tr>
<td>CA</td>
<td>3.2094</td>
<td>-0.74887349</td>
<td>0.04349177</td>
<td>-0.15968629</td>
<td>0.00515803</td>
</tr>
<tr>
<td>CO</td>
<td>2.8048</td>
<td>0.43231789</td>
<td>0.09403111</td>
<td>0.13927795</td>
<td>0.00396228</td>
</tr>
<tr>
<td>CT</td>
<td>1.4085</td>
<td>-0.46211295</td>
<td>0.09982054</td>
<td>-0.15887888</td>
<td>0.00483377</td>
</tr>
<tr>
<td>DE</td>
<td>6.1778</td>
<td>0.25719558</td>
<td>0.04534163</td>
<td>0.05650611</td>
<td>0.00064377</td>
</tr>
<tr>
<td>FL</td>
<td>12.1511</td>
<td>0.99483773</td>
<td>0.08572047</td>
<td>0.28162065</td>
<td>0.01586623</td>
</tr>
<tr>
<td>IA</td>
<td>1.3423</td>
<td>-0.02653618</td>
<td>0.10460490</td>
<td>-0.00812533</td>
<td>0.00001355</td>
</tr>
<tr>
<td>ID</td>
<td>3.7062</td>
<td>-0.23986400</td>
<td>0.07794819</td>
<td>-0.08198488</td>
<td>0.00137759</td>
</tr>
<tr>
<td>IL</td>
<td>5.3496</td>
<td>0.97066680</td>
<td>0.15094384</td>
<td>0.28222503</td>
<td>0.01595384</td>
</tr>
<tr>
<td>IN</td>
<td>4.7234</td>
<td>-1.7486842</td>
<td>0.08765345</td>
<td>0.38810510</td>
<td>0.00101325</td>
</tr>
<tr>
<td>KS</td>
<td>3.809</td>
<td>0.38918002</td>
<td>0.03592081</td>
<td>-0.26213903</td>
<td>0.00110353</td>
</tr>
<tr>
<td>KY</td>
<td>10.4411</td>
<td>2.0165321</td>
<td>0.14935223</td>
<td>-0.00427937</td>
<td>0.00000376</td>
</tr>
<tr>
<td>LA</td>
<td>9.5804</td>
<td>-0.62560572</td>
<td>0.11497362</td>
<td>-0.07363227</td>
<td>0.01378592</td>
</tr>
<tr>
<td>MA</td>
<td>1.0156</td>
<td>-0.20354093</td>
<td>0.05753585</td>
<td>-0.22557845</td>
<td>0.00388624</td>
</tr>
<tr>
<td>MD</td>
<td>7.522</td>
<td>-0.01733197</td>
<td>0.13502051</td>
<td>0.24098216</td>
<td>0.01036132</td>
</tr>
<tr>
<td>ME</td>
<td>1.3094</td>
<td>0.34893118</td>
<td>0.07350372</td>
<td>0.3810510</td>
<td>0.00146709</td>
</tr>
<tr>
<td>MN</td>
<td>1.6721</td>
<td>-0.80442544</td>
<td>0.03721475</td>
<td>-0.67834490</td>
<td>0.00155957</td>
</tr>
<tr>
<td>MO</td>
<td>7.0718</td>
<td>1.2252607</td>
<td>0.27708551</td>
<td>-0.31917975</td>
<td>0.02042257</td>
</tr>
<tr>
<td>MS</td>
<td>11.7936</td>
<td>-1.0956886</td>
<td>0.10073621</td>
<td>0.26628806</td>
<td>0.01422052</td>
</tr>
<tr>
<td>MT</td>
<td>2.7072</td>
<td>-0.95364354</td>
<td>0.07348898</td>
<td>0.14862782</td>
<td>0.00448786</td>
</tr>
<tr>
<td>NC</td>
<td>13.2054</td>
<td>0.94550943</td>
<td>0.05319433</td>
<td>0.17450356</td>
<td>0.00623355</td>
</tr>
<tr>
<td>NE</td>
<td>3.4836</td>
<td>0.62704379</td>
<td>0.22687264</td>
<td>0.76734490</td>
<td>0.00793140</td>
</tr>
<tr>
<td>NH</td>
<td>0.807</td>
<td>0.32213561</td>
<td>0.13304787</td>
<td>0.21324764</td>
<td>0.00914553</td>
</tr>
<tr>
<td>NJ</td>
<td>2.3161</td>
<td>-0.50472662</td>
<td>0.05481278</td>
<td>2.4593624</td>
<td>1.1500463</td>
</tr>
<tr>
<td>NM</td>
<td>3.4709</td>
<td>-0.88552791</td>
<td>0.66677853</td>
<td>0.36299356</td>
<td>0.02680061</td>
</tr>
<tr>
<td>NV</td>
<td>8.3135</td>
<td>1.7385942</td>
<td>0.27436647</td>
<td>-0.02223690</td>
<td>0.00010147</td>
</tr>
<tr>
<td>NY</td>
<td>1.5683</td>
<td>-0.59029657</td>
<td>0.04978694</td>
<td>-0.08696035</td>
<td>0.00161713</td>
</tr>
<tr>
<td>OH</td>
<td>4.1335</td>
<td>-0.09714371</td>
<td>0.29817113</td>
<td>-0.33943501</td>
<td>0.02306626</td>
</tr>
<tr>
<td>OK</td>
<td>3.8393</td>
<td>0.50754356</td>
<td>0.1075126</td>
<td>-0.09011792</td>
<td>0.00166269</td>
</tr>
<tr>
<td>OR</td>
<td>1.8328</td>
<td>-0.98033390</td>
<td>0.07624993</td>
<td>-0.17581739</td>
<td>0.00630883</td>
</tr>
<tr>
<td>PA</td>
<td>3.536</td>
<td>-0.31366689</td>
<td>0.12411637</td>
<td>-1.2895695</td>
<td>0.30666752</td>
</tr>
<tr>
<td>RI</td>
<td>1.1084</td>
<td>-0.46705767</td>
<td>0.27912380</td>
<td>0.07015524</td>
<td>0.00131779</td>
</tr>
<tr>
<td>SC</td>
<td>8.8992</td>
<td>-2.0724147</td>
<td>0.18202185</td>
<td>0.64006699</td>
<td>0.06871519</td>
</tr>
<tr>
<td>SD</td>
<td>1.2739</td>
<td>0.16991863</td>
<td>0.04596011</td>
<td>0.35649508</td>
<td>0.02523819</td>
</tr>
<tr>
<td>TN</td>
<td>15.2657</td>
<td>2.9182056</td>
<td>0.09046328</td>
<td>-0.26333740</td>
<td>0.01397544</td>
</tr>
<tr>
<td>TX</td>
<td>11.1535</td>
<td>1.1302506</td>
<td>0.08968931</td>
<td>0.31520287</td>
<td>0.01975169</td>
</tr>
<tr>
<td>UT</td>
<td>1.7368</td>
<td>-0.83895180</td>
<td>0.07447309</td>
<td>-0.28956913</td>
<td>0.00167031</td>
</tr>
<tr>
<td>VA</td>
<td>11.9782</td>
<td>1.11111815</td>
<td>0.10177520</td>
<td>0.07200980</td>
<td>0.00106309</td>
</tr>
<tr>
<td>WA</td>
<td>3.0412</td>
<td>0.21392579</td>
<td>0.06761989</td>
<td>-0.1521234</td>
<td>0.00468495</td>
</tr>
<tr>
<td>WI</td>
<td>0.8477</td>
<td>-1.0752551</td>
<td>0.04196462</td>
<td>-0.07650287</td>
<td>0.00119857</td>
</tr>
<tr>
<td>WV</td>
<td>2.8339</td>
<td>-0.72684397</td>
<td>0.06253145</td>
<td>0.06253145</td>
<td>0.00119857</td>
</tr>
<tr>
<td>WY</td>
<td>2.8915</td>
<td>-0.29621486</td>
<td>0.06253145</td>
<td>0.06253145</td>
<td>0.00119857</td>
</tr>
</tbody>
</table>

To see how the influential points may have affected our regression coefficients, we can study the partial regression plots. Remember that the bivariate regression slopes in these pictures are the same as the multiple regression coefficients. We can therefore see
Already we must wonder if the standard model describes these data. These plots look weird. One problem is that the murder variable is highly skewed. Did you realize that from looking at the scatterplots? Many novices wouldn’t notice.

Moving along to the regression fit, here are the externally studentized residuals versus the predicted murder rates.
how each coefficient has been determined by particular observations. (Evidence of non-linearity, non-constant variance, and separation into groups also show up here.)

We see that the critical estimate for the deterrent effect of the death penalty is dominated by a single observation—Nevada. Notice that the coefficient is essentially determined by two observations rather than 44. (One observation has many replicates). We can check the effect Nevada has on the coefficients by deleting the case and refitting.
Yikes. Even the sign of the estimate depends on this lone case. Here’s what happens to the whole regression when we delete Nevada. Compare these estimates to the originals.

Dependent variable is: Murder Rate
No Selector
44 total cases, of which 1 are missing
R squared = 66.9%  R squared (adjusted) = 63.4%
s = 2.719 with 43 - 5 = 38 degrees of freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>567.430</td>
<td>4</td>
<td>141.858</td>
<td>19.2</td>
</tr>
<tr>
<td>Residual</td>
<td>280.844</td>
<td>38</td>
<td>7.39053</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.80010</td>
<td>5.521</td>
<td>0.652</td>
<td>0.5183</td>
</tr>
<tr>
<td>Pr(Ex</td>
<td>Conv)</td>
<td>-11.0777</td>
<td>10.78</td>
<td>-1.03</td>
</tr>
<tr>
<td>Time for M</td>
<td>-0.018774</td>
<td>0.0076</td>
<td>-2.46</td>
<td>0.0187</td>
</tr>
<tr>
<td>Poverty</td>
<td>6.61106</td>
<td>17.67</td>
<td>0.374</td>
<td>0.7104</td>
</tr>
<tr>
<td>Non-W</td>
<td>27.7815</td>
<td>4.418</td>
<td>6.29</td>
<td>≤ 0.0001</td>
</tr>
</tbody>
</table>

Are you a believer in regression diagnostics now? In many different ways, this regression is a mess. Nobody should believe the estimates in the first table. Much work remains to be done before we have a satisfactory description of these data.
Partial Regression Plots

The algorithm for the partial regression plots is simple:

1. Regress $y$ on all the $X$'s except the candidate variable $x_j$. Save the $e_y \mid x-j$ residuals.
2. Regress $x_j$ on the remaining regressors. Save the $e_{x_j} \mid x-j$ residuals.
3. Plot the $e_y \mid x-j$ residuals versus the $e_{x_j} \mid x-j$ residuals.
4. The bivariate regression line in this plot yields the same coefficient for $x_j$ as does the multiple regression. The plot therefore allows us to see how the multiple regression coefficient is determined by the data, even when the $X$'s are confounded.
5. The residuals in the partial regression plot are the same as the residuals in the full regression.

In the presence of collinearity among the regressors, bivariate scatterplots of $y$ vs $x_j$ can be highly misleading. For the same reasons that bivariate regressions become biased when relevant variables are excluded from the model, so too can bivariate scatterplots give distorted impressions of the relationship between $y$ and $x_j$. (When and how we'll discuss shortly).

Plots of regression residuals against additional candidate regressors (in their raw form) can be similarly misleading. These plots suffer from the same difficulties as the step-wise regressions discussed earlier. The trouble with sequential regressions is that the first regressors to go in the fit pick up the confounded effects of excluded regressors they covary with.

Partial regression plots control for the confounding among the regressors (at least the linear confounding). They are the better way to go.

Why does the double-residual regression give the multiple regression coefficient? Here is a geometric demonstration.
We can write the multiple regression model as:

\[ y = X\beta + \gamma z + \mu \]

where \( z \) is the "added variable" of interest. The OLS method requires that we minimize
\[ |y - Xb - gz|^2 \]

over \(b\) and \(g\). One can first minimize the function over \(b\), for fixed \(g\), and then minimize over \(g\). The first minimum is attained at

\[ b = (X'X)^{-1}X'(y - gz) \]

and is equal to

\[
|y - X(X'X)^{-1}X'y - gz - gz|^2 \\
= |(I - X(X'X)^{-1}X')y - g(I - X(X'X)^{-1}X')z|^2 \\
= |My - gMz|^2
\]

Recall that \(M\) is the antiprojector corresponding with \(X\) (\(M\) projects vectors onto the subspace orthogonal to \(X\)). Thus \(My\) is the residual vector gotten by regressing \(y\) on \(X\) (the part of \(y\) orthogonal to \(X\)) and \(Mz\) are the residuals gotten from regressing \(z\) on \(X\). The remaining minimization is clearly equivalent to a bivariate regression of \(My\) on \(Mz\).

Recalling earlier results on orthogonal projection for vectors:

![Diagram](image)

we found
Regression Analysis
J. DeNardo

\[ b = \frac{y'x}{x'x} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\text{Cov}(x,y)}{\text{Var}(x)} \]

The corresponding minimization over g therefore gives (by orthogonal projection)

\[ g = \frac{(My)'Mz}{(Mz)'Mz} \]
\[ = \frac{y'M'Mz}{z'M'Mz} \]
\[ = \frac{y'Mz}{z'Mz} = \frac{\text{Cov}(My,Mz)}{\text{Var}(Mz)} \]

This shows algebraically why the partial regression plots give the multiple regression coefficients.


Q - Q Plots

All of the standard hypothesis tests and confidence intervals assume that the disturbances, \( \mu \), are normally distributed. It is therefore useful to investigate whether the assumption is valid, at least approximately. The normal probability plot (an example of the
quantile-quantile or Q - Q plot, with a Gaussian reference) provides a simple graphical test. For normally distributed variables, this plot yields a straight line.

The algorithm for making the normal probability plot is as follows:

1. Put the data \( x_{(i)} \) in rank order, letting \( i \) be an index for the position of each case in the ranked list; \( i = 1 \) for the smallest to \( n \) for the largest.

2. The quantile associated with the order statistic \( x_{(i)} \) is the cumulative proportion of cases equal or smaller to \( x_{(i)} \) in the distribution. To avoid values of zero and one (which makes comparison with continuous theoretical distributions difficult), these proportions can be defined by:

\[
f_i = \frac{(i - \frac{1}{2})}{(n + \frac{1}{2})}
\]

Thus if we have 101 cases, the median, \( x_{(51)} \), is the

\[
f_i = 50.6666667/101.3333333 = .5000000049342
\]

quantile of the data.

3. Quantile plots of \( f_i \) vs \( x_{(i)} \) are effective for locating the median, quartiles, and other quantiles of a distribution. Steep sections of the plot are regions of high density in the distribution of \( x \). The plot is essentially the empirical cumulative distribution function of the data.

4. Suppose \( z \) is distributed Standard Normally, and \( F(z) \) is the cumulative distribution function (cdf). \( F(z) \) is the comparison distribution to which we refer \( x \).
5. By definition, \( z_i = F^{-1}(f_i) \) is the \( f_i \) quantile of the standard normal distribution. To find the median of the standard normal, for example, we compute \( F^{-1}(0.5000) = 0 \). The inverse of \( F \) gives the z-score in a standard normal distribution whose cumulative probability = 1/2. \( F^{-1}(0.975) = 1.96 \).

6. Graphing each \( x_{(i)} \) against the corresponding \( z_i = F^{-1}(f_i) \) produces the quantile comparison or q-q plot. If the \( x_i \) are in fact drawn from the reference distribution, then—within sampling error—the quantiles of the two distributions are the same, and the plot is linear.

7. Of course, we can choose any theoretical cdf for comparison that we like. Modern regression packages routinely produce the q-q plot with a normal reference distribution.

Here are some examples:

Northridge Aftershock Magnitudes (> 2.0)
Regression Analysis
J. DeNardo

Energy-GNP Dataset:

A Mixture of Normal Distributions

A Uniform Distribution
A Normal (5,10) Distribution

With practice you will learn to recognize the various departures from normality captured by these plots.

NON-LINEAR SMOOTHERS

A basic assumption in least squares regression is that a linear relationship exists between the X’s and y. To check this assumption, we can fit a non-linear smoother to the scatter plots, residual plots, or to the partial regression plots. There are many choices described in the literature, including moving averages (also called digital filters), running medians (made popular by Tukey’s 1977 EDA book), splines (piecewise, smooth polynomials), and many more. Each possibility appears in many variants.

Lowess

Lowess is useful for bivariate and time-series data. The basic idea is simple: for each x value, fit a regression line to nearby (x,y) observations and then take as the smooth, the predicted value of y at x. Using this technique, the smooth can be computed for all values of x, whether they appear in the sample or not.

A basic parameter of the lowess technique (as for other smoothers) is the fraction of the data, f, contained in the smoothing window. The wider the window, the more local irregularities get
smoothed out. (A common default for $f$ is $2/3$, but many different values can be illuminating.)

To make the fitted regressions robust, the technique computes an iteratively reweighted least squares fit. The (bi-square) weights are a declining function of the fitted residuals:

$$w_i = \begin{cases} 
(1 - (r_i/6M)^2)^2 & |r_i| < 6M \\
0 & |r_i| > 6M 
\end{cases}$$

where $M = \text{MAD}(r_i) = \text{median}_i(|r_i - \text{median}_j(r_j)|)$

The fit is recomputed over several iterations until the weights converge. Here are some examples:

Here's an example where the non-linearity is subtle (log energy consumption vs. log gnp, 1970, DCs and LDCs):
COLLINEARITY DIAGNOSTICS

Statistical consequences of collinearity:

The problem caused by collinearity is not bias, but instability in the regression estimates—the same consequence that follows from having a small sample. The problem arises from a lack of information in the design matrix—i.e. a lack of independent variation in the X’s. Some common symptoms include:

1. Small changes in the dataset produce wild swings in the estimated coefficients.
2. Coefficients may have large standard errors and low significance levels, even though they are jointly highly significant and the R² is large. If one is not careful, relevant variables can be dropped from the model.
3. Coefficients may have the wrong sign or implausible magnitudes.
4. Coefficient estimates may be dominated by noise, when the noise becomes the only part of the data that isn’t collinear.

Here are two pictures that help to appreciate the problem:
Notice that the fit is well determined in one direction. We can make good forecasts, but determining the separate impacts of Gallup and Harris will be difficult.

Suppose

\[(1) \quad y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \mu \]

but that in our sample we observe

\[X_2 = \frac{2}{3} X_1.\]

Then

\[(2) \quad y = (\beta_1 + \frac{2}{3} \beta_2) X_1 + \beta_3 X_3 + \mu = \gamma X_1 + \beta_3 X_3 + \mu\]

In this case, $\gamma$ can be estimated quite accurately:

\[
\text{Var} (b_1 + \frac{2}{3} b_2) = \text{Var} (b_1) + \frac{4}{9} \text{Var}(b_2) + \frac{4}{3} \text{Cov} (b_1, b_2)
\]

where $\text{Cov} (b_1, b_2) < 0$.

Of course, $\gamma$-hat estimates the composite effect of $X_1$ and $X_2$—we can’t identify their separate effects. But if our purpose is forecasting, and the relationship between $X_1$ and $X_2$ persists out of sample, then we’ll be o.k.

Another way to appreciate the problem is by considering the joint confidence ellipsoids for the estimates:
95% Joint Confidence Ellipse for $\beta_1, \beta_2$
with Strong Collinearity

\[
\begin{align*}
\beta_2 \\
(0,0) \\
(\beta_1)
\end{align*}
\]

\[
\text{Cov}(x_1, x_2) \gg 0
\]

95% t-interval

In the separate t-tests, we can not reject $H_0: \beta = 0$ for either coefficient (nor many other hypotheses). The joint hypothesis $H_0: \beta_1 = 0 \cap \beta_2 = 0$ will be decisively rejected, however. Notice that if by chance we overestimate one coefficient, we will tend to underestimate the other, perhaps even with the wrong sign.

**Detecting Collinearity**

1. The most commonly used and most inadequate diagnostic is the correlation table. The pairwise correlations can not reveal multivariate collinearities i.e. linear combinations that are highly correlated in the data matrix.

2. Much better are the auxiliary regressions of each $x_j$ against the remaining independents. Such regressions not only reveal multiple collinearities, but yield useful diagnostic information about the instability created by them:
Denote by $R_j^2$ the auxiliary $R^2$ from regressing $x_j$ on the remaining $x_{i\neq j}$. With centered $x$'s we get:

$$\text{Var}(\hat{\beta}_j) = \frac{s^2}{x_j'x_j} \frac{1}{(1 - R_j^2)}$$

As we add variables collinear with $x_j$ to the model, the standard error for $x_j$ blows up.

The *Variance Inflation Factor* (VIF) is

$$\text{VIF}_j = \frac{1}{(1 - R_j^2)}$$

These are useful diagnostics that provide a concrete sense of how much instability the collinearities produce. The VIFs do not detect multiple near singularities. If more than a single linear dependence exists, the auxiliary regressions may themselves suffer from multicollinearity, making it difficult to diagnose the sources of the problem. The VIFs by themselves do not identify the origin of the collinearity.

One can go much further by computing eigenvalue-eigenvector decompositions of the matrix $X'X$. Doing so provides information about how the variance of each particular coefficient has been affected by collinearities in the data. In general, the impact is spread over several coefficients, but not uniformly. For details, see Belsey, Kuh, and Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (Wiley).

**HETEROSCEDASTICITY DIAGNOSTICS**

Heteroscedasticity can also degrade inferences in OLS regressions, again by inflating the variances of the estimates. Many
formal tests are described in the literature, including the White test, the Goldfeld-Quandt test, and Glesjer's test. In everyday practice, it is convenient to have a simple diagnostic readily at hand.

The most common graphical diagnostic is (studentized) residual vs. fitted. The familiar wedge pattern in residuals is especially common in counted and measured data, where the dependent is highly skewed:

Here's a clear example from the Murder dataset.

![Residual vs Fitted Plot](image-url)

Sometimes the pattern isn't so obvious, however, and a more sensitive procedure is helpful. Here are some widely recommended plots that are easy to construct:

1. residual $^2$ vs y-hat.
2. $|\text{residual}|$ vs y-hat.
3. log $|\text{residual}|$ vs log (y-hat). If the standard deviation of the disturbances is proportional to a power $\theta$ of the mean response (an important model of heteroscedasticity), then this plot will be approximately linear and its slope will provide information about the value of $\theta$. 
4. Each of these functions of the residuals vs. $x$, or $z$. Particularly when $x$ or $z$ measure the "size" of the cases (e.g. population, GNP, or area.)

Here are the absolute murder residuals vs. predicted, with a lowess smoother fitted to the plot. The dependency of spread on the level $y$-hat is clear, even when we set aside the largest residuals that may be dominating the visual impression:

Here are the same residuals vs. fraction non-white in each state:
Now, Spherical Disturbances (i.e. \( V(\epsilon) \neq 0 - I \))

I. Background

Discussion

\[
\begin{align*}
Y &= X\beta + \epsilon \\
E(\epsilon) &= 0 \\
\text{Cov}(\epsilon) &= \Omega \\
\text{NARMA} &= \text{null column rank} \\
\epsilon_i &\sim N(0, \sigma^2) \text{ iid}
\end{align*}
\]

Have been dealing with \( \sqrt{v} \) in particular: incorrect set of \( X \)'s, non-linear in \( X \).

Now we move on to consider variations of \( \sqrt{v} \).

\[ V(\epsilon) = \sigma^2 \Omega \] since each \( \epsilon \) can have diff. variance (\( \sigma^2_i \)). Think of the \( \sigma^2 \) here as pulling out some base-line obs.' variance to which others are scaled. It's in that sense arbitrary usually.

In that sense it is arbitrary, usually.

A. What did we use \( V(\epsilon) = \sigma^2 I \) for?

A. \( \hat{\sigma}^2 \)

B. Distribution of \( \hat{\sigma}^2 \) (as in: hypothesis tests, confidence intervals, etc.)

It's taken to be the variance.

Unbiased

Consistent (i.e., that \( \hat{\sigma}^2 \) asymptotically normal)

1) Did use \( \sqrt{v} \) to prove OLS: efficient and that

B. What does \( V(\epsilon) = \sigma^2 \Omega \) mean?

\[
V(\epsilon) = E\left[\begin{array}{cccc}
\epsilon_1^2 & \epsilon_1 \epsilon_2 & \cdots & \epsilon_1 \epsilon_n \\
\epsilon_2 \epsilon_1 & \epsilon_2^2 & \cdots & \epsilon_2 \epsilon_n \\
\vdots & \vdots & \ddots & \vdots \\
\epsilon_n \epsilon_1 & \epsilon_n \epsilon_2 & \cdots & \epsilon_n^2
\end{array}\right]
\]

\[ V(\epsilon) = \sigma^2 \Omega = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}
\end{bmatrix}
\]

Why do we care if \( \sigma_{ij} \neq 0 \) for some \( i \neq j \)?

\( \sigma_{ij} \neq 0 \) implies that observations \( i \) and \( j \) are not independent. This means that, given obs. 1, obs. 2 is not completely new info.

\( \sigma_{ij} \neq 0 \) implies that observations \( i \) and \( j \) are not independent. This means that, given obs. 1, obs. 2 is not completely new info.

V.\text{arsa} 

\( \sigma_{ij} \neq 0 \) means that observations \( i \) and \( j \) have differing amounts of stochastic relative to systematic info. \( (y = f(X\beta)) \)

\( \sigma_{ij} \neq 0 \) means that observations \( i \) and \( j \) have differing amounts of stochastic relative to systematic info. \( (y = f(X\beta)) \)

\( \sigma_{ij} \neq 0 \) means that observations \( i \) and \( j \) have differing amounts of stochastic relative to systematic info. \( (y = f(X\beta)) \)

\[ \hat{\sigma}_{ij} \neq 0 \] might well imply that we've left something out that correlates across obs. For explaining different amounts across obs. might indicate omitted variables or a sort.
C. A More Concrete Example: Suppose we have observations on some $y$, say votes for incumbent, of some $x$ explanatory variable, say the unemployment rate. We observe $(y, x)$ in 20 countries over 40 years. Then

$$y_{ct} = x_c \beta + \varepsilon_{ct}.$$  

This is the model for country $c \in \{1, \ldots, 20\}$ and $t \in \{1950, 1951, \ldots, 1990\}$ or $t = \{1, \ldots, 40\}$.

$$V(\varepsilon_{ct}) = \sigma^2_e I_{20}.$$

$$\Rightarrow V(\varepsilon) = \begin{pmatrix} \sigma^2_e & 0 & \cdots & 0 \\ 0 & \sigma^2_e & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_e \end{pmatrix}_{20 \times 20}$$

$$\text{Cov}(\varepsilon_{ct}, \varepsilon_{ct}) = \sigma^2_e I_{20}.$$

$$\Rightarrow \text{just same other ctry year.}$$

1) The OLS LRM assumes $V(\varepsilon) = \sigma^2 I$.

$$\Rightarrow \text{No obs. in any ctry-year is correlated with any other obs, within or across both ctry & year.}$$

$$\Rightarrow \text{The variance of every obs. (within across ctry & year) is the same.}$$

2) Now this might well be overly restrictive. We might well suppose:

   a) Obs. corr. over time within ctry (usually called auto- or serial correlation).

   b) Obs. corr. across ctry's at same times (usually called contemporaneous correlation).

   c) Obs. diff. variance by ctry (ctry-wise heteroskedasticity).

3) Now suppose we are unwilling to make any restrictions on the parameters in $V(\varepsilon)$. We wish to leave it open that $\text{Cov}(1950, 1951)$ differs from $\text{Cov}(1970, 1972)$ that $\text{Cov}(\text{US, UK})$ differs from $\text{Cov}(\text{US, UK})$ that $V(\text{US, 1950})$ differs from $V(\text{GER, 1972})$ and so on. How many unique $\sigma^2$ are there to estimate? Then $\sigma^2$? Square: $NT \times NT$ elements. Only Diagonal & one of triangles are unique ($\sigma^2_{ii} = \sigma^2_{jj}$). So:

$$\frac{1}{2}(NT)^2 + \frac{1}{2}NT = \frac{1}{2}NT(NT+1)$$

parameters! 

Aside: notice that we can distinguish spatial & temporal correlation. These technically can be treated much the same.
4. Some Possible Assumptions limiting the number of parameters in \( V(\varepsilon) \) to estimate:

- **(a)** C(\(N\)) LRM & OLS: \( V(\varepsilon) = \sigma^2 I \)
  - 1 parameter

- **(b)** Time-Series:
  \( off(\varepsilon_t, \varepsilon_{t-1}) \) = some simple declining function of temporal distance between obs. One very common one is \( \varepsilon_t = \rho \varepsilon_{t-1} + \sigma \)
  - \( \sigma^2 \) & \( \rho \) parameters

- **(c)** Cross-Sections: \( V(\sigma^2) \) different for each group (\( N \) groups theoretically defined in your scenario)
  - \( 2N + N(N-1) \) parameters

- **(d)** Panel Data (some as TSCE, but many more C's than T's)
  - a single \( \rho \) for all individuals
  - perhaps a \( \sigma^2 \) for each \( T \)
  - \( T + 1 \) parameters

In general, there will be a trade-off between how realistic the maintained restrictions on \( \Omega \) are and how well the necessary parameters of \( \Omega \) may be estimated (i.e., how valid results based on estimating \( \Omega \) are likely to be across repeated samples).
II. Generalized Linear Regression Model (GLRM)

1. \( y = X\beta + \epsilon \)
2. \( E(\epsilon) = 0 \)
3. \( V(\epsilon) = \sigma^2 \Omega \)
4. \( \text{Cov}(X, \epsilon) = 0 \)
5. \( X \) of full rank
6. \( \epsilon \sim N(0, \sigma^2 \Omega) \)

A. Suppose, This is Model, but we estimate by OLS

\( \Rightarrow \) Finite Sample Properties of OLS

1. Still Unbiased:

\[
\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \epsilon)
\]

\[
= (X'X)^{-1}X'X \beta + (X'X)^{-1}X'\epsilon
\]

\( \Rightarrow \) exactly as before

\[
b_{ls} = \beta + (X'X)^{-1}X'\epsilon
\]

\( \Rightarrow \) \( E(b_{ls}) = \beta \) if \( E(\epsilon) = 0 \)

\( \Rightarrow \) so, long as \( \text{Cov}(X, \epsilon) = 0 \), \( b_{ls} \) unbiased

2. Still Consistent (for "well behaved" \( \Omega \))

\[
V(b_{ls}) = V(\beta + \epsilon) = \sigma^2 (X'X)^{-1}X'X(\sigma^2 \Omega)^{-1}X'X(\sigma^2 \Omega)^{-1}
\]

\[
\Rightarrow \text{usually rearranged to: } \sigma^2 (X'X)^{-1}X'X(\sigma^2 \Omega)^{-1}X'X(\sigma^2 \Omega)^{-1}
\]

Now, unbiased + \( \lim_{n \to \infty} V(b) = 0 \) is sufficient (not necessary) for consistency

\( \Rightarrow \) loose demonstration of consistency:

- \( \text{n.b. } (X'X)^{-1} \) twice, \( (X'X) \) once (with \( \sigma^2 \Omega \) in middle of \( \frac{1}{2} \))
- \( \sigma^2 \Omega \) in "denominator"
- \( \sigma^2 \Omega \) if \( (X'X) \to 0 \) faster than \( \sigma^2 \Omega \) does as \( n \to \infty \)

- In general, this means that off-diagonals of \( \sigma^2 \Omega \) must decline sooner or later as \( n \) grows (hence)
- And that \( X \)'s continue to exhibit variation as \( n \) increase sample (this usually taken as self-evident for any reasonably defined \( X \))

3. Still Normal or Asymptotically Normal:

\( a) \) If \( \epsilon \sim N(0, \sigma^2 \Omega) \), then \( b_{ls} \) is just a linear combo of

\( \beta \)s and \( \epsilon \)s, so \( b_{ls} \) is also normal; \( b_{ls} \sim N(\beta, \sigma^2 \Omega A') \)

\( b) \) If \( \epsilon \sim N(0, \sigma^2 \Omega) \), then \( b_{ls} \) is asymptotically normal,

\( \Rightarrow \) basically the same conditions as in \( a \) above hold.

4. However, OLS is \( \sigma \) Inefficient \( \Rightarrow \) have some info, namely

\( \Rightarrow \) \( V(\epsilon) = \sigma^2 \Omega \) not \( \sigma^2 \Omega \)

\( \Rightarrow \) \( V(b_{ls}) \neq \sigma^2 (X'X)^{-1} \)

\( \Rightarrow \)\( V(b_{ls}) = \sigma^2 (X'X)^{-1}X'\Omega X (X'X)^{-1} \)
Robust Estimation of the GLRM

Since OLS is unbiased and consistent but inefficient, producing the wrong standard errors, one possible fix would be to estimate by OLS, do nothing about the inefficiency, but address the standard error problem.

$\Rightarrow$ need to estimate $V(b) = (X'X)^{-1}X'(\Omega^{-2}XX'X)^{-1}$

A. As stated before, if we try to estimate $\Omega$ directly, without any restrictions on its form, there are too many parameters.

B. However, $[X'0^{-2}XX']$ which is the problem here is not $n \times n$ (like $\Omega$) but $k \times k$ (like $X'$). As we've seen above, what we need for consistency involves convergence of this whole thing. Similarly, to obtain a consistent estimate of standard errors, what we need is not element-by-element estimation of $\Omega$, but rather a consistent estimate of $(X'0^{-2}XX')$. For consistent standard errors, then, we need only consistent estimates of those $K(K+1)/2$ elements of this matrix.

1) Pure Heteroskedasticity: $V(\varepsilon) = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \Rightarrow N$ obs. & $N$ parameters

but $\{X'0^{-2}XX\} = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \sum_{i=1}^{n} \omega_i \varepsilon_i X_i'

\text{rather than try to estimate } \omega_i^2 \text{ directly, just use } \sum_{i=1}^{n} \omega_i^2 \varepsilon_i^2 \text{ as a column vector}

\text{and residual variance}

This produces a consistent estimate of the sum (but not of } \omega_i^2 \text{). This is all we need } \varepsilon_i.

What produces inconsistency of standard errors is how $\varepsilon_i \varepsilon_j$ tends to vary with $X_i X_j$.

$\Rightarrow$ White's heteroskedasticity-consistent standard errors: (State calls them)

Huber/White/Sandwich estimates.

So, use OLS to estimate $\hat{\beta}$ and then using White's robust s.e.'s:

$V(\hat{\beta}) = (X'X)^{-1}(\sum_i \varepsilon_i^2 X_i X_i')(X'X)^{-1}$

$\Rightarrow$ "Robust" fixes your standard errors in that they are now consistent for any heteroskedasticity which is a function of (the $X$'s and their squares, 3rd their cross-products).

It turns out, as we saw at the beginning, these are the only forms that produce inconsistency of the standard errors. It should be noted, perhaps, that White's s.e.'s and related forms are not unbiased (nor efficient, of course)--however, so far, Monte Carlo simulations have shown them to work surprisingly well.
Extensions to White's when $\Omega$ not simple heteroskedastic:

\[ X'Q^{-2}QX = \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i \xi_j \quad \text{which could be estimated by} \]

\[ Q_k = \sum_{i=1}^{n} \sum_{j=1}^{n} e_i e_j \xi_i \xi_j \quad \text{we would think} \]

Two Problems:
1. If $\sum e_i e_j \xi_i \xi_j$ not "well behaved" in the same sense as before, this sum may not converge to a constant as $n \to \infty$, but may instead grow without bound.
2. Whereas with true $\Omega$ we are sure its positive definite so that $X'QX$ is p.d., so that $V(b) > 0$, with $Q^*$ substituted for $(X'QX)$ we cannot be sure of p.d. anymore.

\[ \Rightarrow \text{Newey-West: Autocorrelation-Consistent Standard Errors} \]

\[ Q^* = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\tau_{i,j}} + \sum_{i=1}^{n} \sum_{j=1}^{n} e_i e_j \xi_i \xi_j \]

White's, which takes care of the $e_i \cdot e_j$ as it correlates to $X_k' X_k$, & cross-products.

\[ \text{Newey-West addition which takes care of} \]

setting $\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\tau_{i,j}}$ weights $e_i e_j$ by (inversely) the temporal distance btw observations.

\[ \Rightarrow \text{assumes/forces relation of } e_i \cdot e_j \text{ to fade as time from } i \text{ to } j \text{ increases} \]

so that $L$ is unfortunately fairly arbitrary maximum "lag-length" time separation at which you wish to consider $V(b)$ > 0, $E(e_i e_j)$ non-negligible.

\[ \Rightarrow \text{There are rules of thumb out there, but no good theory (some simulations)} \]

\[ \Rightarrow \text{I don't know of any tests or Monte Carlo simulations showing N-W works as well as White's seems to.} \]

*My reading is that jury is out on N-W as opposed to more direct fixes for serial correlation.

* Beck & Katz (APSR '95) Panel-Consistent Standard Errors

In exactly analogous spirit, B&K consider that if we are concerned with contemporaneous correlation across cross-sectional units in TSCS data, then we can estimate $\sum e_i e_j X_k X_k'$ appropriately to reflect this concern:

1. Re-configure errors into $E = T \times N$ shape:

2. Then $A' (E' E \otimes I_T) A = V(b)$

(See B&K for more, see my diss. appendix on PCSE's for brief intro to extension)
So, in the presence of OLS-Robust, we have: $b_{GLS}$ unbiased & s.e.($b_{GLS}$) consistent but may be unsatisfied with merely consistent standard errors.

$\Rightarrow$ Generalized Least Squares (GLS)

1. Suppose we happen to know $\Omega$ (we nearly will, but this is just to get the "in principle" facts out on the table)

$\Rightarrow$ $\Omega = C\Lambda C'$ where $C$ is the matrix of characteristic vectors of $\Omega$.

2. Define $\Lambda^\frac{1}{2} \equiv \{ \lambda^\frac{1}{2} \}$ (square root of $\Lambda$ element by element)

3. Then $\Omega = T \Lambda T'$

$\Rightarrow$ $\Omega^{-\frac{1}{2}} = T \Lambda^{-\frac{1}{2}}$ so $\Omega^{-1} = P P'$

$= C \Lambda^{-\frac{1}{2}} \Lambda^{-\frac{1}{2}} C' = C \Lambda C^{-\frac{1}{2}} C = C^{-1}$

Loosely speaking, then, $P$ is the square root of the inverse of the variance-covariance matrix of $\Omega$. We'll see in a moment why it is absolutely key.

B. Take the GLRM:

1.  $Y^* = X^* B + E^*$

Let $Y^* = \rho Y, X^* = \rho X, E^* = \rho E$

2. $V(E^*) = \rho^2 V(E) = \rho^2 \Omega$

Thus, $V(E^*) = \rho^2 I$.

$\Rightarrow$ The (Gauss-Markov) OLS is BLUE for the transformed model. (Applied to GLRM by transformation, this is called 'Aitken's Theorem'. GLS is BLUE.

4. So, if we know $\Omega$, we can find $P$ transform the variables and then everything is as before except $R^2$ may perhaps be S.E.E. (MSE, S.E.R)

Problem with these things is that GLS doesn't minimize the sum of squared errors, but a weighted sum of squared errors.

Furthermore, $R^2$ is purely descriptive. We usually prefer $R^2$ based on $b_{GLS}$ but original data (This no longer $\Omega$ but $\rho^2$).

Similarly, S.E.E is estimate of $\sigma^2$ in $\sigma^2 \Omega$ and so it is no longer exactly an estimate of $V(E)$ for all $i$. 
5. Otherwise, though, once we have the appropriate weighing matrix, we transform $\hat{\beta}$ then estimate $\hat{\sigma}^2$ and carry on tests or whatever exactly as before.

**Easy Example:** $V(\varepsilon) = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}$ (pure hetero)

- $Y = X\beta + \varepsilon$
  - Use $P = P = [(V(\varepsilon)^{-1})]^{1/2}$ (this loose notation)
  - $PY = PX\beta + PE$

$$\Rightarrow V(P\varepsilon) = P^2 V(\varepsilon) = \begin{bmatrix} n \sigma^2 \varepsilon_i \end{bmatrix} [\begin{bmatrix} 2 \sigma^2 \\ 5 \sigma^2 \end{bmatrix}] = \begin{bmatrix} 2n \sigma^2 \\ 5n \sigma^2 \end{bmatrix}$$

**Basically, we have**

$V(\varepsilon) = \sigma^2_i,$ then weight obs. by $\frac{1}{\sigma_i^2}$

$$\sqrt{\left[ \frac{1}{\sigma_i^2}\varepsilon_i \right]} = \frac{1}{\sigma_i} \sqrt{V(\varepsilon_i)} = \frac{1}{\sigma_i} \sigma_i^2 = 1$$

**N.b.** One more thing not to notice: if we know $\sigma_i^2$ V(\varepsilon) then we can weight appropriately so that $V(\varepsilon^*) = \frac{1}{\sigma_i^2}$ (all variances are 1)

But for OLS = BLUE all we need is $V(\varepsilon^*) = \sigma^2 I$, $\sigma^2$ need not be 1. This implies we don't need to know $\sigma^2$ exactly but rather proportionately. i.e. if we know $V(\varepsilon_i) = k \cdot \sigma^2_i$ we don't need to know $k$: weighting by $\frac{1}{\sigma_i}$ $\Rightarrow V(\varepsilon^*_i) = k$.

C. **Maximum Likelihood Estimation of the GLRM** (loose intro)

1. MLE is actually a remarkable intuitive approach to statistical estimation (it's merely that it can look technical on paper)

   Basic Idea: Each $Y_i$ is independent given the (controlling for the) $X_i$, draws from some probability distribution $f(X,\varepsilon)$, stochastic part.

   a) Indep. means that the joint probability (Actually likelihood) of all the data are the product of all the individual ones

   $$\Pr(Y_i) \propto \prod_i f(X_i, \varepsilon_i)$$

   b) So maximize this likelihood with respect to the parameter $\theta$ (in GLRM, parameters are $\beta$ and $\sigma^2$)

   estimates

   2. The MLE (which as with CNLREG) turns out to be the GLS estimator

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\sigma}^2 = \frac{S_{ML}^2}{n-k}$$

$$S_{ML}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Rightarrow S_{ML}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Rightarrow \hat{\sigma}^2 = \frac{S_{ML}^2}{n-k}$$

Page 261 of 364
More usually, we do not know $\Omega = \text{estimate its}
\text{(feasible GLS or FGLS)}$

A) Asymptotically, if $\hat{\Omega}$ is a consistent estimate of $\Omega$, (converges to
it as $n \to \infty$) then $\hat{\beta}_{GLS}$ is asymptotically efficient as is
$s.e. (\hat{\beta}_{GLS})$.

B) Some take this as an amazing & powerful result: we don't need that $\hat{\Omega}$
is an efficient estimate of $\Omega$, only that it is consistent, to get
asymptotically efficient results for $\beta$ & $s.e. (\beta)$.

C) The worry, of course, is that in limited samples there's no guarantee
that $\hat{\beta}$ performs better than $\beta$. The crux of the problem, of course,
is that we introduce some instability by estimating $\Omega$. My reading of
the literature is that the general prescription is not to try to estimate
too many parameters in $\Omega$ relative to the number of observations.
Some guidelines/examples:

1) Always try to use your head about what is likely to be true
about $\text{V}(\epsilon)$. E.g., if each obs. is an average of $n_i$
random variables, then $\text{V}(\epsilon_i) \propto \frac{1}{n_i}$
- is proportional to
Theory should not be put in the closet just because we're
now talking about "residual" properties.

2) Don't try to capture every conceivable nuance in the structure of
$\Omega$ in estimating it. Monte Carlo usually reveals that using fewer
parameter forms of $\Omega$ (where fewer is relative to # obs) does
better across repeated samples in practice.

3) Examples:  
- a) Simple AR(1): $\epsilon_t = \rho \epsilon_{t-1} + \gamma_i$
  related to this period's
  last period's shock
  $\gamma$ iid, $\text{V}(\gamma) = \sigma^2$
  $\Rightarrow \text{V}(\epsilon) = \left( \frac{1 - \rho^2}{1 - \rho^2} \right)$
  \begin{pmatrix}
  \rho & \rho^2 & \rho^3 & \cdots \\
  \rho^2 & \rho^4 & \rho^5 & \cdots \\
  \rho^3 & \rho^5 & \rho^7 & \cdots \\
  \vdots & \vdots & \vdots & \ddots
  \end{pmatrix}
  \Rightarrow$ only one
  more parameter:
  \( \rho \)
  n.b. \( |\rho| \) must
  be less than
  one for $\text{V}(\epsilon)$
  to be well behaved (more later)

- b) $\text{V}(\epsilon_i) = \sigma^2 f(z_i)$
  \( \Rightarrow \) i.e., variance can be modeled just
  like mean (Xb) was as a function
  of some variable or set of variables.
  e.g., $\text{V}(\gamma_i) = f(\text{population diversity of economy, etc.})$

- c) $\text{V}(\epsilon_j) = \sigma_j^2$ where \( j \)'s are groups of observations.

- d) Sometimes you know something for sure about variance:
  e.g., Average of $N_i$ obs
  "Quality of data" indexed by data-recorder or somehow else
  \( \Rightarrow \) Always use info, that you know "for sure" or reasonably surely
  like this, 

Page 262 of 364
"Robust" solutions & FGLS are not necessarily exclusive.

- It's usually permissible to use some info, you know "for sure" & also use robust if you think V(\epsilon) matrix to be related, but X, X', X' X in some undefined way.

**Example (my diss, chpts 3 & 4) Chpt 3: \text{Var}(\epsilon) different by county \Rightarrow FGLS

+ Possible contemporaneous correlation \Rightarrow Beck-Katz

Chpt 4: Obs. are averages over differing # of years \Rightarrow FGLS

+ Possible+Likely other heterosd contemp. corr

\Rightarrow White's \text{Var} Beck-Katz

---

**VI. Testing When V(\epsilon) = \sigma^2 \Omega: Final Note**

A. If FGLS only, then exactly as always -- be sure to use transformed data (for F-tests especially).

B. If Robust or FGLS+Robust, then you have done nothing to "fix" the fit. Accordingly, "degradation of fit" tests are not valid \Rightarrow use Wald-types (including T-tests).

n.b. Asymptotic tests w/ s.e.'s only consistent

---

**III. Estimating in Stages:**

A. Stage 1: Estimate by OLS

B. Stage 2: Use estimated residuals to estimate \hat{\Omega} (i.e., to estimate the reduced set of parameters for your structural \hat{\Omega})

C. Stage 3: FGLS \quad b_{FGLS} = (X' \hat{\Omega}^{-1} X)^{-1} X' \hat{\Omega}^{-1} y

\quad V(b_{FGLS}) = \sigma^2 (X' \hat{\Omega}^{-1} X)^{-1}

D. Stage 4: Optional Iteration \Rightarrow some times we repeat this until \text{bias doesn't change much} from iteration to iteration. I.e., use residuals from \text{(y-b}_{FGLS}) to re-estimate \hat{\Omega}. Calculate \text{bias using this new \hat{\Omega}}. Repeat until bias doesn't change noticeably (can be set arbitrarily small) from iteration to next.

---

Examples of this Procedure you May have Heard of Before

- Mostly for Time Series:
  - Cochrane-Orcutt (C-O)
  - Hildreth-Lu (H-Lu)
  - Iterated H-Lu
  - Prais-Winston (no nickname, aww...)
  - Iterated Prais-Winston
I. Within the GLRM:

- \( Y = X \beta + \varepsilon \)
- \( E(\varepsilon) = 0 \)
- \( V(\varepsilon) = \Omega \)
- \( \text{Cov}(X, \varepsilon) = 0 \)
- \( \text{rank}(X) = k \)

We now focus in a particular form of violation of CLRM assumption.

3. I.e., we focus in on a particular form of \( \Omega \), a particular set of properties of the stochastic part of the world.

(Pure) Heteroskedasticity

A. Definition: Refers to situation where \( V(\varepsilon_i) = \sigma^2_i \), but there is no covariance across observations.

\[ V(\varepsilon_i) = \sigma^2_i \]

We have independent observations, but it is possible each observation comes from a distribution with different variance.

B. This means \( V(\varepsilon_i) \) is diagonal:

\[ V(\varepsilon_i) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix} \]

\[ \Rightarrow \text{We could also write:} \]

\[ V(\varepsilon_i) = \sigma^2 \Omega \]

1. That is, we could pull a common factor \( \sigma^2 \) out of the variance of each; \( \sigma^2_i \), and write it:

\[ \sigma^2_i = \sigma^2 \Omega \]

2. But now it would be somewhat arbitrary what \( \sigma^2 \) to "pull out" like this. The common practice is to normalize it like so:

\[ \text{trace}(\Omega) = \sum_{i=1}^{p} \sigma_i^2 \]

we normalize: \( \sigma_i^2 \Omega \rightarrow \eta \)

3. \( \sigma^2 \Omega \) can be thought of as:

a) \( \Omega \) captured the variety of the variances of \( \varepsilon_i \), i.e., the \( \sigma_i^2 \) reflected relatively.

b) \( \sigma^2 \) provides the overall scaling of the variance.

C. By the way: if you have a heteroskedastic model and you estimate it as such, what M.S.E. (s.e.r. or s.e. = i.e., the variance) do you get?

- depends: some software provide simply \( \sum \sigma_i^2 \)
- others give estimates of \( \sigma^2 \)
- \( \sigma^2 \) is \( V(\varepsilon_i) = \Omega \) with \( \Omega \) normalized this way. \( \Rightarrow \) always check what a software package is actually reporting before you use it --- don't repeat anything you don't know what it is (obviously).

C. Common Precautions

- (Possibly Misleading)

1. Heteroskedasticity is commonly thought of as a cross-sectional phenomenon, and it's true it often arises there -- but there is no necessary link one way or the other.

2. It is also commonly thought of as "a problem" that must be "corrected." I would argue, however, we might better think of it as a "feature" of the social world which we want to "estimate," in some cases it may well be a "feature" we can take advantage of.
D. Examples: 1) Cross-section of city’s, states, etc. etc. -- basically cross-sections of observations on anything that involve input from multiple decision-makers
   ⇒ often be quite reasonable to expect variance to be inversely proportional to the number of such decision-makers
   ⇒ Most certainly, this is so, the dependent variable is directly an average of a group lower-level aggregation outputs.
   ⇒ In many settings, especially economic ones but not exclusively, "diversification" of multiple units aggregated into a single observation will also affect variance. E.g., Variance (City Output) or V(Simprofit) are typically decreasing in diversity of the "portfolio".

⇒ Illustrates a point: Existence or Not / Degree of Heteroskedasticity is itself no less a theoretical proposition to which data can speak than are statements about means

⇒ (ECY = XB is no less or more a theoretical proposition than is V(ε) = (Cβ))

2) For certain survey questions, it will often be reasonably to expect more or less variant answers the more educated is the respondent.

⇒ It is important to note that these are proportions about Variances (2nd Moments) not Means (1st moments). That is, the above example is not a statement about what more educated people will say (that's in the E(Y) = XB part +) but rather that the distribution of educated people's answers will have greater or lesser variance (V(ε) = (Cβ) +) -- spread of educated answers around XB will be larger or smaller than spread of less educated around the same regression line: E(Y) = XB.

2) Variances certainly can change in Time-Series Data as well:
   a) Structural Shift: --- Prior to Voting Rights Act & After --- Populations becoming wealthier, more educated, more, less.
      ⇒ Capital Mobility moves over time. Legal/ regime changes: Pre-Post Bretton Woods
   b) ARCH (GARCH, ARCH-M, GARCH-M): New long-form seemingly esoteric stuff -- basic notion is simple: variance produces more (or less) variance.
      ⇒ In other words, fluctuations (unpredictable movements) trigger other unpredictable movements (or force their own corrections).
      ⇒ These -- the save in finance, these days
      Basic Model: Y = XB + E
      V(ε) = (Cβ + E)

1) WRAP-UP: No reason our theorizing ought to be limited to 1st moment.
   ⇒ i.e., we can make predictions about variances as well as expected values.
   ⇒ That said, it's a good idea to make really sure your prediction is about the spread of outputs not about the outputs themselves are.
* Estimation by OLS & White's "Robust" V.Cov($\beta$):

A. Coefficient estimates, $\hat{\beta} = Ay = (X'X)^{-1}X'y$ are...
1. Unbiased across repeated samples they would be right on avg.
2. Consistent as sample size $\to \infty$, $\hat{\beta}$ converges to exactly $\beta$
3. Inefficient - we have more information we could have used so $V(\hat{\beta})$ is higher than it could have been

B. OLS estimate of $V(\hat{\beta})$; the usual $\sigma^2(X'X)^{-1}$ would be biased, inconsistent & garbage, $\sigma^2$

$$V(\hat{\beta}) = AV(\epsilon)A' = AV(\epsilon)A'$$
$$= AV(\epsilon)A'$$

both $\epsilon$ is fixed, no variance, so $V(\epsilon) = 0$

$$= (X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1} = 0 - (X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}$$

C. Defining $Q = (X'X)^{-1}$ & $Q^* = \frac{1}{n}\sum_{i=1}^{n} W_iX_iX_i'$ (the White's estimate of the middle term $X'\epsilon\epsilon'X$)

then, the following:

$$\frac{\sigma^2}{n} Q^{-1} Q^* Q^{-1}$$

is the "asymptotic variance" of $\hat{\beta}$

D. OLS estimation then has

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{n} Q^{-1}(Q^*Q^{-1}))$$

to write it out long-hand at least once:

$$\hat{\beta}_{res} \sim N(\beta, \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{1}{W_i} X_iX_i' \left(\frac{X'X}{n}\right)^{-1})$$

is asymptotically distributed normal with mean $\beta$ and variance [that whole mess]

E. Upshot: Asymptotically (i.e., for "large" samples), OLS estimation with White's standard errors produces: Unbiased & Consistent estimates if consistent $V(\hat{\beta})$ estimates, so is a decent way to proceed, just "inefficient".

However, how "large" is "large"? Cannot be said unambiguously but my read of KC is "not very" by around 150 or so

This tends not to do perfectly badly - inefficient (so if you could get more info from data could have smaller $V(\hat{\beta})$ but tests & c. i. s. and such tend not to be very optimistic.

F. How inefficient? Of course, it again depends, but for a simple white type of heteroskedasticity: $V(\epsilon) = \sigma^2 X_i^2$, Greene 12.2.1 shows that degree of inefficiency is a function of the kurtosis of $X$ - i.e., of how tall the tails of the empirical distribution of $X$ is relative to a Normal distribution. (Formally, kurtosis can be measured by $k = \frac{E(X_i^4)}{E(X_i^2)^2}$.) Typical for typical economic data $k = something between 2.5$ which implies confidence intervals based on OLS are 70% or more wider than the proper GLS estimate would produce.

I read the general conclusion as follows: (if it's simple & abstract and maybe obvious) the more variance your variances (i.e., farther from $\sigma^2$ if you are, proportionately) the less efficient you are. For $V(\hat{\beta}_{res}) = 2 to 4 times \sigma^2$, tough
III. OK, now suppose you use OLS without heteroskedasticity. How bad are things? That is, how far is \( \sigma^2 (X'X)^{-1} \) from \( \sigma^2 (X' \Omega X) (X'X)^{-1} \)? (loosely)

A. The \( \sigma^2 \) in \( \sigma^2 (X'X)^{-1} \) is estimated by \( \frac{1}{n-k} \sum_{i=1}^{n} e_i^2 \) as we know. If \( \text{V}(e_i) = \sigma^2 \) though, what does this estimate produce?

B. \( \sigma^2 = \frac{1}{n-k} \sum_{i=1}^{n} e_i^2 \) is like estimating \( \overline{\sigma^2} \), the average (or mean) of all the different variances. In fact, it's an unbiased estimate of that "mean of the variances" under fairly general circumstances. 

\[ \Rightarrow \text{consistent} \]

C. So now the question how far is \( \sigma^2 (X'X)^{-1} \) from \( \sigma^2 (X' \Omega X) (X'X)^{-1} \)? reduces to how far is \( X' \Omega X (X'X)^{-1} \) from \( I \). Why? Because if that last thing is \( I \), then \( \sigma^2 (X'X)^{-1} \) differs from \( \sigma^2 (X'X)^{-1} \) only in their estimates of \( \sigma^2 \) which we've just shown the OLS estimate is OK; not great.

D. So, how far is \( X' \Omega X (X'X)^{-1} \) from \( I \)? Note: this is like saying how far is \( X' \Omega X \) from \( X'X \) because if \( X' \Omega X = X'X \), then we have some matrix times its inverse which is \( I \).

E. So, key question is how much does \( X' \Omega X \) differ from \( X'X \)? Remember what \( \Omega \) is. It's that matrix of the relative variances:

\[ \Omega = \begin{bmatrix} \sigma_{1}^2 & \sigma_{2}^2 & \cdots & \sigma_{n}^2 \end{bmatrix} \]

Now, if the \( \sigma_i^2 \) are not a function of the \( x_i \), that is, if the heteroskedasticity is not related to the \( x_i \)'s so that (loosely) \( \text{Cov}(x_i, \sigma_i^2) = 0 \), then on average \( X' \Omega X \) is not different from \( X'X \).

\[ \Rightarrow \text{If heteroskedasticity is unrelated to the } x_i \text{'s (i.e., the } x_i \text{'s, the } x_i^2 \text{'s & the crossproducts } x_i x_j \text{), then the OLS standard errors are "right on average" -- i.e. they are "unbiased".} \]

Which, as Greene concludes (page 577), means loosely speaking: if the heteroskedasticity is not related to \( (X'X) \) then the OLS computations are not "wrong on average" though they are, of course, inefficient.

\[ \Rightarrow \text{Much of the motivation for using White's} \]
A. We already explored some graphical methods for rooting out heteroskedasticity. Recall that they amounted to looking for patterns in the "spread" of residuals. The trick is going from these informal "tests" to more formal ones is two-fold
1) Realizing that $\sigma^2$ is a reasonable estimate of $\text{E}(e^2) = \text{V}(e)$
2) Figuring out the distribution that applies to the test statistic, which we, as always, will leave to the "experts"

B. White's General Test
1) Recall: OLS std. err.'s inconsistent iff $\text{V}(e_i)$ is some $f(X'X)$
2) So, White's suggestion:
Regress $e_i^2$ on each $X$, each $X^2$, & each $X_1 \times X_2$ (i.e., the unique elements of this
3) If $\text{V}(e_i)$ is not a function of the $X$'s, $X^2$'s, & cross-product of the $X$'s, then this auxiliary regression should be insignificant. $X = X^2$ = enter just once, so also $X_1 \times X_2$
4) Two Stats: a) $n \cdot \bar{R}^2 \sim \chi^2_{k-1}$
where $n$ is # obs. in aux. reg.
$k$ is # of regressors in aux. reg. minus 1
$\bar{R}^2$ is $R^2$ in aux. reg.
$\chi^2$ is "asymptotically distributed"
$b)$ Could also calculate the F-stat of the aux. reg. Exact distribution of this is not known though
$\Rightarrow$ if $n \cdot \bar{R}^2$ bigger than some critical level from $\chi^2_{k-1}$, then you likely have heteroskedasticity, which is related to $X'X$.
$\Rightarrow$ Use "White's Robust" or WLS (here)

C. Goldfeld-Quandt Test:
1) Intuition: under homoskedasticity, however you group your data, $(\Sigma e^2_i)/n$ as an estimate of $\text{V}(e)$ in that group should be the same.
2) Group your data somehow. E.g., you think $\text{V}(e_i) = f(X_i) >$
$\Rightarrow$ sort your data by $X_i$
3) Separate a high $X_i$ group and a low $X_i$ group (some suggest leaving a small (less than vs the obs) middle group out.
4) Run your regression model separately in each sample. Then...
5) $\frac{(\Sigma e_1^2)/(n-k)}{(\Sigma e_2^2)/(n-k)} = F \sim F_{n-k,n-k}$
$\Rightarrow$ if $E \sim$ Normally, then this is exactly $F$ distribution
$\Rightarrow$ if not, distribution is unknown
D. Breusch-Pagan/Godfrey Test:

1) Intuition \( V(\varepsilon_i) = f(\bar{Z}) + \delta \)

2) Normalize your squared residuals \( q_i = \frac{e_i^2}{(\bar{e}_i^2)^{1/2}} \)
   (i.e., all \( e_i^2 \) measured relative to the average \( \bar{e}_i^2 \) )

3) Then \( LM = \frac{1}{2}[q_i'Z(Z'Z)^{-1}Z(q_i - n)] \)
   \( LM \sim \chi^2_{p - \# \text{ of variables in } Z} \)
   (basically this is a Wald test on the auxiliary regression of \( q_i \)'s on \( Z \) )

4) Again, if \( LM \) large, then likely have heteroskedasticity,
   which is a function of the \( Z \)'s.

(n.b. if \( Z \) is same as variables in White's test, these two tests are identical)

5) Again, there's some concern about what the distribution of
   \( LM \) actually is if \( \varepsilon \) not normal. Koeker & Bassett have suggested:
   \( \varepsilon \sim (e_1^2, e_2^2, \ldots, e_n^2) \) and \( T \sim (1, 1, \ldots, 1) \)

Then \( \bar{u} = e_1^2 \) and \( V = \frac{1}{n} \sum_{i=1}^{n} \bar{e}_i^2 - \frac{e_1^2}{n} \)

Then \( LM = \frac{1}{n} (\bar{u} - \bar{u})'Z(Z'Z)^{-1}Z'(\bar{u} - \bar{u}) \)

\( \Rightarrow \) asymptotically same as \( LM \) above, but absent normality it seems to work better in small samples.

(E. Likelihood Ratio Tests: 1) Maximum likelihood will estimate \( \hat{\beta} \) & \( \hat{\Omega} \) simultaneously, directly. Accordingly we can estimate \( \Omega \) in one go, then

2) Then, as usual, \( \chi^2 \)

3) An Example: Groupwise Heteroskedasticity

   Restricted: \( V(\varepsilon_i) = \sigma^2 \) \( \forall i \)

   Unrestricted: \( V(\varepsilon_i) = \sigma_i^2 \) \( \forall i \) in group \( g \)

\[ \text{let } S^2 = \left( \frac{\varepsilon_j^2}{n} \right) \text{ from OLS regression on whole sample } \]
\[ \text{let } S_g^2 = \left( \frac{\varepsilon_j^2}{n} \right) \text{ from OLS regression } \]
\[ \text{one at a time, } \]

Then \( n(S^2 - \sum_{g=1}^{G} n_g S_g^2) = LR \sim \chi^2_{G-1} \)

4) As always, if this too big, then likely have groupwise hetero.
F. Glesjer's Test: The obvious regression tests are basically valid.  

\[ \text{for any } \quad V(\varepsilon_i) = f(z_i^2) + \gamma \]

Just write the appropriate regression model and conduct the appropriate (Wald) test.

E.g.  \[ V(\varepsilon_i) = f(z_i^2) + \gamma \]
\[ \Rightarrow \varepsilon_i^2 = \beta_0 + \beta_1 z_i + \gamma \]

- Test on \( \beta_1 \) (using White's V-Cov matrix is recommended)

E.g. White's test is an example. So is Breach-Pagan/Godfrey.

E.g.  \[ V(\varepsilon_i) = \sigma^2 \exp[\alpha z_i^2] + \gamma \]  
(Variances must be positive; this model assures that all predicted values are positive)
\[ \Rightarrow \ln V(\varepsilon_i) = \ln \sigma^2 + \alpha z_i^2 + \gamma \]
\[ \Rightarrow \ln \varepsilon_i^2 = \beta + \beta_1 z_i + \beta_2 z_i + \ldots + \beta_k z_k + \gamma \]

- Wald test all of these

Thus, \( \ln \varepsilon_i^2 = \beta_0 \) would be homoskedasticity

I find this "Glesjer" approach by far the most intuitive.  
It is also the most powerful way to test the specific hypothesis embodied in the variance model. Some of the others (e.g. White's version of the "Glesjer") have an advantage of generality.

V. Efficient Estimation: Weighted Least Squares (WLS)

A. The Pure Heteroskedasticity Model:
1. \( Y = X\beta + \varepsilon \)
2. \( E(\varepsilon) = 0 \)
3. \( V(\varepsilon) = \sigma^2 \Omega = \sigma^2 \begin{bmatrix} \omega_1 & 0 \\ 0 & \omega_2 \end{bmatrix} \)
4. \( \text{Cov}(X, \varepsilon) = 0 \)
5. \( X \) of full rank
6. \( \varepsilon \) (thus \( \varepsilon \) normal)

B. \( \Omega \) is known (Base-line case, which never obtains, of course)
\[ \Omega^{-1} = \begin{bmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ 0 & 0 & \omega_3 \end{bmatrix} \]
\((\text{loose notation}) \quad \sqrt{\Omega} = \begin{bmatrix} \sqrt{\omega_1} & 0 & 0 \\ 0 & \sqrt{\omega_2} & 0 \\ 0 & 0 & \sqrt{\omega_3} \end{bmatrix} \)
\( \text{call this } P \)

C. Then \( (PY) = (PX)\beta + (PE) \)
\[ \Rightarrow V(PE) = P^2 V(\varepsilon) \]
\( \text{just by premultiplying everything by some } P \text{ matrix} \)
\[ \Rightarrow \begin{bmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ 0 & 0 & \omega_3 \end{bmatrix} \]
\[ \begin{bmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ 0 & 0 & \omega_3 \end{bmatrix} = \sigma^2 \]
So, if we know what \( V(\epsilon_i) \) is proportional to; i.e., if we know the relative variances of the observations, i.e., the \( \omega_i^2 \)’s, then multiply \( Y_i \) by \( \frac{1}{\omega_i} \) and all \( X_i \) by \( \frac{1}{\omega_i} \), call those \( Y^* \) and \( X^* \)

- OLS of \( Y^* \) on \( X^* \) is BLUE because \( Y^* \) is \( \epsilon^* \)

E. Intuition: Why do we want to “weight” some observations more than others?

\[
y_i = f(X_i, \epsilon_i)
\]

We are trying to estimate the systematic relationship between \( X \) and \( Y \)

- With heteroskedasticity, the proportion of information in obs. \( i \) that is systematic & proportion that is stochastic (random) varies.

\[\Rightarrow\] We want to weight more systematic observations more, or equivalently, we want to weight more stochastic (less systematic) obs. less.

- Thought experiment: You ask 100 people the temperature. You happen to know how accurate each person is in terms of their expected squared error. What do you do? You take a weighted average of their statements, weighing more accurate people more. Same thing here.

- Visual Picture

If you knew this, wouldn’t you want to discount obs 2? and somewhat obs 4? Obs 1 and 3 seem more reliable about where mean for that obs is. I.e., where \( E(y_i) \) is, which is what we’re trying to figure out:

\[ E(y_i) = x_i \beta \]
II. Efficient Estimation when \( \Omega \) known:

1. WLS:
   \( \begin{align*}
   & a) \quad Y^* = PY = \Omega^{1/2} Y \\
   & b) \quad X^* = PX = \Omega^{1/2} X
   \end{align*} \)

   Transformed variables \( \Rightarrow \) OLS of \( \beta \) on \( X^* \)

   \[ (X^*\ ')(X^*')^{-1}X^*\ 'Y = \beta \]

   is BLUE

   \[ \text{wls} \]

   \[ \text{wls} \]

   \[ V(\beta)^* = \sigma^2 \ (X^*\ 'X^*)^{-1} \]

2. WLS in one step:

   \[ \begin{align*}
   \beta &= (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} Y \\
   \text{wls} \\
   V(\beta)^* &= \sigma^2 \ (X' \Omega^{-1} X)^{-1} \\
   \text{wls}
   \end{align*} \]

VI. Estimation of \( \Omega \): Feasible GLS (or WLS), FEGLS (or FWLS)

A. Two-Step FWLS:

1. Estimate \( Y = X\beta + \epsilon \) by OLS

2. Use \( \hat{\epsilon} \) estimates from this to model \( V(\epsilon) \) as we did in testing

3. Use estimated \( V(\epsilon^2) \) to construct \( \hat{\Omega}^{-1/2} \)

4. Do WLS using \( \hat{\Omega}^{-1/2} \)

B. Example:

1. \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \)

   GDP growth in city \( i \)

   same vector of stuff that matters

   residuals

2. Hypoth. \( V(\epsilon_i) = \alpha_0 + \alpha_1 Z_i + \gamma \)

   \( Z \) = sectoral diversity of city’s output

   \( \alpha_1 < 0 \) (by guess)

- models: \( (\text{new})^2 = \beta_0 + \beta_1 Z + \gamma \)

   \( \Rightarrow \) this is done so that all \( \epsilon_i^2 \) from the

   model will be positive

   estimate by OLS, fitted values are \( (\hat{\epsilon}_i^2) \)

   \( \Rightarrow \) so, expected values are \( \hat{\epsilon}_i^2 \)

   \( \Rightarrow \) so, weight by \( \frac{1}{\hat{\epsilon}_i^2} \)

   i.e. \( \hat{\sigma}^2 \), the weighting matrix, is

\[ \hat{\sigma}^2 = \begin{bmatrix}
\frac{1}{\hat{\epsilon}_1^2} & 0 \\
0 & \frac{1}{\hat{\epsilon}_2^2} & \cdots & 0 \\
0 & 0 & \cdots & \frac{1}{\hat{\epsilon}_n^2}
\end{bmatrix} \]

3. Estimate WLS, using \( \hat{\beta}^* = \hat{\sigma} Y \) and \( \hat{X}^* = \hat{\sigma} X \)

C. Properties of \( \hat{\beta} \):

0. Unbiased

3. Consistent

0. Asymptotically Efficient

Properties of \( \text{var} \ (\hat{\beta}) \):

0. Unbiased

3. Consistent

0. Asymptotically Efficient
VII. Upshot:

A. White's Robust S.E.s are almost always a good idea.
   1. They're consistent precisely in the case where OLS S.E.s are not, when $V(\hat{\beta}) = f(XX)$
   2. They're not much less efficient than OLS S.E.s except in very small samples perhaps
   3. Properties in very small samples are hard to the draw --
      but then so are properties of nearly anything in subsamples

B. White's General Test & Glejser's Specific Tests are the most direct ways to go about considering heteroskedastic possibilities. I like the following process:
   1. Think about $V(\hat{\beta}) = f(Z)$. model and test it
   2. Consider White's General Test in comparison to your specific one as revealing whether there's more heteroskedasticity than your model speaks to,
      or the dangerous type.
   3. Consider OLS + White's or White's or OLS depending on outcome of (2)
      But: be wary of overfitting. My rule is always to have as many obs as parameters
      in $\theta$ you're going to estimate. I like 10 as a bare minimum.

C. No less theorizing should go into the probability distributions from which you consider your data to have been drawn than goes into your theories about the mean ($\mu_s$) of those distributions. (Unfortunately, usually a lot less goes in.)

D. However, also in principle, it is generally a good idea to check first whether some $Z$ variable belongs in the $Y = X\beta + E$ equation (the theory about $E(Y)$) then proceed to consider

$$V(\hat{\beta}) = V(\hat{\beta}) = f(Z)$$

Reason is that if you have $Z$ out of mean model but it belongs in (social) then $E_i$ related to $Z$ which assures that $E_i^2$ is too, but this has nothing to do with heteroskedasticity really. It's rather indicative of mis-specification.
I. Preliminaries: Serial (and Spatial) Correlation

A. To date, we have been assuming that:

1. $y = \beta x + \varepsilon$
2. $E(\varepsilon) = 0$
3. $V(\varepsilon) = \sigma^2 \Omega$ (in the CMLRM, $\Omega = I$
4. $\text{Cov}(x, \varepsilon) = 0$
5. $\text{Rank}(x) = k$

Well, nothing much

B. At one level, it is quite possible that nothing needs to change at all here when we move to the situation are not "randomly sampled" in the usual sense, but rather gathered over time, or gathered in some other way in which we expect them to be related rather than independent in their raw form. We simply want to "model the temporal process" same as we model any other process.

C. An Example to Focus Discussion

$A_t = f(\beta, \gamma, \varepsilon)$ - stochastic part of the world

1) The modelling, or theory-building, exercise is, of course, deciding what variables or factors go in the matrix $x$
2) So in this example, what might affect Presidential Approval?
   a) Economic Conditions: i) UE
      ii) Infl
      iii) GDP Growth
   b) Events: i) Vietnam -- maybe War Deaths, TV Coverage,...
      ii) Watergate -- maybe Dummy, or Dummy for pre/post
      iii) Iran Hostages -- maybe linear or non-linear
      iv) Events Variable: "really found the flag" effects
   c) What about the Past? History?
      i) The idea here is that it's quite possible, in fact likely, that this year's approval rate may be affected by last year's (even controlling for UE & Infl & Events).
      ii) Interpreting: i) Surely, we always have left something out of the model. Past "things" persist over time, so next year may be indicative of those "things" (controlling for other $x$s)
   ii) Maybe approval adjusts slowly to changing conditions? Perhaps we should be more careful when conditioning on those conditions that has filtered down to approval survey respondents? Perhaps because it takes a while for folk to come to their "long-run" view on how much to credit or blame the president?
   iii) In some scenarios, perhaps not this one, it's costly to change $y$, or $y$ can only change sluggishely, then past levels are sticky
D. The Point is that our first line of action should probably be to model the way the past affects the future. More than likely this is not a “nuisance” or a “problem” in the data that we ought to “correct”, but rather this is information, this is something systematic about the social world that we would, presumably, like to understand better.

\[ \text{Just like heteroskedasticity was better viewed as something about which we theorize, to model, to estimate, … perhaps even to discover, autocorrelation, that is, the temporal process, is likewise only more so if anything.} \]

E. The Shadow of the Past: There are, in general, at least three ways we can think about past values of \( Y \) affecting the future:

1. \( Y_t = f(X_t, Y_{t-1}, \varepsilon_t) \rightarrow \text{the past level of} Y \text{ affects the future, all of} Y \text{ is “sticky”} \)
2. \( Y_t = f(X_t, \varepsilon_t, Y_{t-1}) \rightarrow \text{the past expected or equilibrium level of} Y \text{ or the systematic part of} Y \text{ is sticky} \)
3. \( Y_t = f(X_t, \varepsilon_t, \varepsilon_{t-1}) \rightarrow \text{the past unexpected or disequalibrating or stochastic part of} Y \text{ affects the future (is sticky)} \)

4. Since \( Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + \varepsilon_t \), any two of these exactly determines the third, i.e. we cannot distinguish \( \beta_1 \) from \( \beta_2 \) and so on, (in a linear model)

5. Some examples of the idea underlying each of the above:

- a) Past Level Affects the Future: perhaps you prescribe to an incrementalist view of public budgets. That is, you think budgets adjust very slowly, or do not respond immediately, entirely, to changes in underlying causes. You think, in particular, that all of the current budget—the parts which have fully adjusted to underlying causes—the parts which have not—are resistant to change, i.e. will not respond immediately to further changes in the “underlying causes”.

  \[ Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + \varepsilon_t \]

- b) Past Expected Level Affects the Future: perhaps instead you think that only spending that is “in line” with underlying causes of the moment build constituencies that fight change. Unpredictable errors in previous budgets are easily removed (if in fact are removed) next budget.

  \[ Y_t = \beta_0 + \beta_1 X_t + \beta_2 (\beta_1 X_{t-1}) + \varepsilon_t \]

- c) Past Stochastic Realizations (Shocks) Affect the Future: now to the contrary, believe it’s precisely the inexplicable spending that builds such constituencies.

  \[ Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \]

...
II. E. 5. c. This last conceptualization, that past shocks persist, yields the most common model in time-series econometrics:

\[ Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \]

\[ \varepsilon_t = \rho \varepsilon_{t-1} + \gamma_t \]

\( \gamma_t \) iid mean 0, variance \( \sigma^2 \)

My guess as to its dominance in econometrics is the seeming concurrence with the notion of constant freely equilibrating markets; if \( X_t \) captures the objective economic conditions of the situation, then \( Y_t \) is the equilibrium level. Any change in \( X_t \) produces the corresponding change in \( Y_t \). Any tendency for this not to be so in a systematic way - i.e., for mistakes in \( E(Y_t | X_t) \) to correlate in a predictable manner - must come from the non-systematic part of the world. Thus, "stickness" cannot be in the equilibrium relationships but must be in stochastic elements.

This is just a guess, but the point is that these are each defensible models on statistical grounds. In their own terms, it is a matter of theory & empirical testing which applies in various cases. (Recall, though, the ambiguity involved in any two determining the third - we cannot distinguish any one from any one other, or we cannot distinguish all one from its corresponding two others in combination.)

II. F. Summary: 
1) The past usually affects the future (geographic proximity often matters, etc.). Therefore, we should model this relationship.
2) There are three inter-related ways to model this temporal process, each having different substantive implications. (However, there is some ambiguity in that if more than one type is present, we cannot say which is operating)
3) Modelling this relationship can in fact be quite simple (haven't told you how yet). (It isn't always.)
4) Statistical inference, likewise, can be completely unaltered (it isn't always however haven't told you about this yet); likewise,
5) Substantive interpretation can also be completely unaltered (it usually isn't & haven't told you about this yet either).
II. (Some of) The Key Statistical-Inference Issues

A. First, most broadly: Absent adequate modelling of the temporal process...

1) ... You Don't have \( T \) independent observations (independent sources of information) even though you have \( T \) observations

\[ \Rightarrow \text{Sid, Eric's will probably be over-optimistic} \]

2) ... You've left something out (the temporal process part) of the model

\[ \Rightarrow \text{Omitted Variable. Across model samples there is no reason to believe that what you left out is correlated with what you included which is why OLS is nonetheless unbiased inconsistent.} \]

However, within a given sample of time-series data (especially "trended" data) it's probably virtually certainly the opposite case. So, within a sample it's basically like omitting relevant fixed bias term -- it's just that across samples these biases will go in different directions & net out; the inefficiency does not.

3) ... \( V(\varepsilon) = \Omega \neq 0, \varepsilon \), which means there remains information in the residuals (different variances for covariances) that you could & should exploit

B. Absent adequate modelling of the temporal process, OLS is...

1) Unbiased -- i.e., across repeated samples it is, right on average, there's generally no way to know which way & how much it's "biased" within your own sample by being left out the "shadow of the past"

2) Consistent -- as \( T \to \infty \) this "within-sample" issue disappears. So, if we had infinitely long time-series of data, OLS would produce the right coefficients on \( X \) with or without adequate modelling of temp. process.

\[ \text{But: } T \text{ may have to be quite large indeed to get much solace from consistency in this setting.} \]

because each new obs. \( \beta \) is not wholly new info.

So, suppose \( p \neq 0 \), then each new obs. is really only new info, \( \Rightarrow \) if we would have been happy with 100 independent obs., we'd want 1000 obs.

3) Inefficient -- in the sense of there's information not being used (\( \Omega \)), so the true \( V(\beta_{OLS}) \) is greater than it truly would be using all the information.

\[ \text{But: In this case, the estimated } V(\beta_{OLS}) \text{ is very likely to be too low because we're pretending to have } T \text{ independent observations when we don't, so it's probably better to focus on...} \]

4) \( V(\beta_{OLS}) \) is Biased, Inconsistent, & Just plain wrong.

(In typical time-series I've looked at -- it's usually WAY wrong/ the amount of lying about the # of independent observations for the importance of the omitted time factor is usually LARGE.)
III. What to do about it

A. First note that from a purely mathematical standpoint, heteroskedasticity and correlation of observations are the same type of problem (there's one technical difference we'll discuss momentarily). Namely OLS is unbiased if consistent (in one sense), but inefficient and produces the wrong standard errors (the wrong estimate of \( V(\hat{b}) \)).

B. Option 1: So we could just accept the inefficiency & fix the standard errors \( \Rightarrow \) Newey-West, Heteroskedasticity & Auto-correlation consistent standard errors

1. Estimate \( y = \beta_0 + \beta_1 x \) by OLS

2. Estimate

\[ V(\hat{\beta}) = \sigma^2 (X'X)^{-1} X' \Omega X (X'X)^{-1} \]

using a White-type substitution for the \( X' \Omega X \) part

\[ \Rightarrow \text{est. } V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \left[ \sum \text{variance of } e_i, e_i = \sum \sum W_{ij} (x_i,x_j) + \sum \text{covariances between } e_i, e_j \right] \]

\[ \Rightarrow \text{est. } V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \left[ \sum \text{variance of } e_i, e_i + \sum \text{covariances between } e_i, e_j \right] \]

where:

- \( L \) is a maximum lag-length to be determined
- \( W_{ij} = \frac{1}{1 + L} \)

These complications are necessary to ensure that \( X' \Omega X \) is positive definite, which means that for any matrix \( A \), all the diagonal elements of \( A (X'\Omega X)^{-1} A' > 0 \), for \( A (X'X)^{-1} \). \( A (X'\Omega X)^{-1} \) is the estimated \( V(\text{cov}(\hat{b})) \), which must have positive diagonal elements since those are the variances.

B. The lag-length \( L \) is the time-space between observations beyond which you are comfortable the correlation is negligible. Newey-West themselves suggest using the following:

\[ L = \lfloor 4 (T/100)^{1/4} \rfloor \]

I suggest instead use obvious cut-offs -- e.g. if in our budget example, budgets are rewritten from scratch every 4 years, there's an argument for using \( L = 4 \).

\[ \Rightarrow \text{and } L \text{ should be proportional to what you think } \rho \text{ is.} \]

E.g. if you think \( \text{cor}(e_0, e_0) = .9 \) (and this is constant), then \( \text{cor}(e_0, e_0-x) = .9^x \). So if \( L \) of about 10 \( \Rightarrow \text{cor}(e_0, e_{10-x}) \leq .35 \), which checking Greene Table 13.3, should be OK. L of 20 or 30 will leave \( \rho = .10 \) at the end, which would be OK by many standards.
C. Model the Temporal Process Directly (Option 2)

1. Option 2a: \( y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \varepsilon_t \)
   - If \( \varepsilon_{t-1} \) in this model are independent, constant variance, then OLS is BLUE.
   - How do you implement this?
     - Just create (or variable(s) equal to \( y_{t-1}, (y_{t-1}, \ldots) \)
     - In a good time-series package (i.e., not STATA) you can usually just use \( y_{t-1} \) and the like exactly as any other variable, not having to create it first.
     - In Stata: "generate ylag1 = y_{t-1}\"
     - Then estimate by OLS (or whatever other process).

2. Option 2b: \( y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \varepsilon_t \)
   - If \( \varepsilon_{t-1} \) in this model are \( \text{id} \), then \( \Rightarrow \text{OLS is BLUE} \) — actually just need a constant variance, not necessarily like so.
   - How do you implement this?
     1. OLS without the \( \varepsilon_{t-1} \)
     2. Save fitted's, \( \hat{y}_t \), as a variable
     3. (generate logs if you have to as you do in Stata)
     4. OLS with the lagged fitted values
     (you can keep doing a cycle of 2 & 3 until the estimated \( \beta \)'s do not change from round to round)
     - Asymptotically, whether you iterate or not theoretically makes no difference. There's not much Monte Carlo evidence, either way, but still I'd say standard practice is to iterate.

3. Option 2c: \( y_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \beta_2 x_t + \varepsilon_t \)
   - If \( \varepsilon_t \) in this model (most author's replace \& with some other letter here, to "avoid confusion" -- I'm not sure it helps from where we're at right now) anyway
   - If \( \varepsilon_t \) in this model is \( \text{id} \), then OLS is BLUE.
   - How do you implement this?
     1. OLS without the lagged \( \varepsilon \)'s
     2. Save residuals, \( \hat{\varepsilon}_t \), & create the lagged \( \varepsilon \)'s variables
     3. OLS with the lagged \( \varepsilon \)'s
     Iterate 2 & 3 if you like
   - (To be continued.)
I. SUMMARIZING WHAT HAS GONE BEFORE

A. We will always want to begin by considering the process that generated our data, considering critically, for example, the \( \text{C}(N) \text{-LRM} \) assumption that observations are independent.

- Certainly, in virtually every time-series context the assumption that observations in temporal proximity are independent will generally be untenable (this may be generalized; perhaps, that proximate obs. are often likely to be correlated -- whether proximity is temporal, geographic, or whatever).

B. We then noted that absent any appropriate or sufficient measures in our modeling, the application of the \( \text{C}(N) \text{-LRM} \) to dependent obs. could be thought of as introducing several familiar problems:

1. It's like trying about how many independent pieces of information we have: that \( N \) independent sources of info. but \( N \) partially dependent a partially independent sources of info. The appropriate partialling remains to be done.

2. It's also like leaving out a variable: namely, the "past" or "history" or the "temporal process." Now, granted, across repeated samples there's no reason to suppose this omitted factor (omitted variable) is correlated with your included variables, so omitting it does not produce bias across samples; you'll be right on average. It does, however, make you inefficient, since the "past" is usually by far the single most important factor, you are "inefficient" by a mile usually. So much, in fact, that in any given sample with fairly correlated data, there's little reason to believe anything simple OLS might tell you (absent adequate modeling of the temporal process).

3. It's also like failing to capitalize on information in the \( V \text{-Cov} \) of your residuals. Like any failure to use info., it's inefficient. (Just happens to typically very inefficient in common time-series contexts.)

C. Then we turned to the specific statistical implications of estimating simple OLS (without adequate controls) anyway.

1. \( \beta_{OLS} \) is unbiased: \( \mathbb{E}(\beta_{OLS}) = \mathbb{E}(\beta) = \mathbb{E}(A(X'\beta + \epsilon)) = AX'\beta + A\mathbb{E}(\epsilon) = I\beta + \beta = 0 = \beta \)

2. \( \beta_{OLS} \) is consistent: \( V(\beta_{OLS}) = V(\beta) = A'V(y)A = (X'X)^{-1}X'V(y)X(X'X)^{-1} \)

   so long as \( (X'\epsilon) \) converges to zero, this whole thing converges to zero. Unbiased + variance goes to zero \( \Rightarrow \) consistent

   why? Basically, if \( \lim_{N \to \infty} X'\epsilon = 0 \), then \( X'V(\epsilon)X \) is converging to \( (\text{some constant})X'X \) which is multiplied by two \( (X'X)^{-1} \) terms.

   \( \Rightarrow \) really loose talk here. So we have a constant-like thing divided by \( (X'X) \), which together goes to zero as \( N \to \infty \). \( \Rightarrow \) variance goes to zero.\n
3. \( \beta_{OLS} \) is inefficient: As we saw in heteroskedasticity case, we want to weight by \( \frac{1}{\sigma^2} \). I.e., we want to count systematically abs. more; here the analogy is we only want to count the "independent parts" of each piece of info.
I. Correlated Obs: We do OLS anyway $$\Rightarrow$$ b_{ols} unbiased, consistent, but inefficient.

4. $$V(b_{ols}) = V(Ay) = AV(y)A' = (X'X)^{-1}X'(X'X)^{-1}$$
   * but OLS prints out $$\hat{V}(b_{ols}) = \sigma^2(X'X)^{-1}$$
   $$\Rightarrow$$ The reported V-Cov Mat of OLS b's are just plain wrong.

D. We then turned to what to do about it:

1. Forget the inefficiency in b estimates, & "fix" the V-Cov($b_k$) matrix to a consistent estimate $$\Rightarrow$$ Newey-West "Robust" V-Cov matrix

   my advice: inefficiency can be huge -- see if there's something you can do about it first. It's often easy to do something, if often the case that that simple something does much to all the way home.

2. Model the Temporal Process Directly
   a) Lagged 1 (or multiple lags)
   b) Lagged Y (or multiple lags)
   c) Lagged X (or multiple lags)

   $$\Rightarrow$$ if remaining residuals are such that $$V(\varepsilon) = \sigma^2 I$$ & independent (constant variance), then this returns you to the C(N)LRM world. (If OLS is one, I call BLUE)

3. (F)GLS (actually, this is G, i.e. in disguise more or less, and vice versa).
   a) As noted last time we met GLS, the idea here is to find
      $$P$$ such that $$V(P\varepsilon) = \sigma^2 I$$. Equipped with $$P$$, we can transform the data into $$Y^* = PY, \quad X^* = PX$$ and regress $$Y$$ on $$X$$ by OLS
      since $$E^* = PE$$, will follow the C-NLRM.
   
   b) Since $$V(P\varepsilon) = PV(\varepsilon)P' = P\sigma^2 \Omega P'$$, it's relatively easy to see that we want $$P = \sigma^{-1/2} \Omega^{1/2}$$, so that this becomes
      $$\sigma^{-1/2} \Omega^{-1/2} \Omega^{1/2} = \sigma^{-1/2} \Omega^{1/2} = \sigma^{-1/2} I$$
      which, moving things around in a way we generally couldn't but here can, because of special properties of $$\Omega$$, becomes
      $$\sigma^{-1/2} \Omega^{1/2} = \sigma^{-1/2} \Omega^{1/2} = \sigma^{-1/2} I$$
   
   c) So what is this "magic" transforming matrix? In the AR(1) case, i.e., where
      $$E_t = \rho E_{t-1} + \nu_t$$, with $$\nu_t$$ i.i.d., we have
      i) $$Cov(E_t, E_{t-1}) = Cov(\rho E_{t-1}, \nu_{t-1}) = \rho Cov(E_{t-1}, E_{t-1}) + Cov(\nu_t, \nu_{t-1})$$
      ii) $$Cov(E_t, E_{t-1}) = \cdots = \rho^t Cov(\varepsilon_{t-1}, \varepsilon_{t-1}) = \rho^t V(\varepsilon_{t-1}) = \sigma^2 \rho^t$$ (b/c $$\varepsilon_t$$ iid)
      iii) For constant variance, then,
      $$V(\varepsilon) = V(\varepsilon)| _ {\rho^t} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{t-1} \\ \rho & \rho^2 & \rho^3 & \cdots & \rho^{t-1} \\ \rho^2 & \rho^3 & \rho^4 & \cdots & \rho^{t-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \rho^{t-3} & \cdots & 1 \end{bmatrix}$$

   (V) Now what's $$V(\varepsilon)$$?
   $$V(\varepsilon) = V(\rho E_{t-1} + \nu_t) = \rho^2 V(E_{t-1}) + V(\nu_t) + 2 \rho Cov(E_{t-1}, \nu_t)$$
   $$V(\varepsilon) = \rho^2 V(E_{t-1}) + \sigma^2 + 0$$ (b/c $$\varepsilon_t$$ iid)
   $$\Rightarrow$$ now, constant variance $$\Rightarrow V(\varepsilon) = V(\varepsilon) \rightarrow \sigma^2 = \rho^2 \sigma^2 + \sigma^2$$
   $$V(\varepsilon) = \sigma^2 = \frac{\sigma^2}{\rho^2} \sigma^2$$
c. \[ V(\theta) = V(\xi) \]

for AR(1) model

\[
\begin{bmatrix}
\rho & \rho^2 & \ldots & \rho^{T-1} \\
1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\
\rho^2 & 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\
\quad & \ddots & \ddots & \ddots & \ddots & \ddots \\
\rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \\
0 & \rho^{T-1} & \rho^{T-2} & \ldots & 1
\end{bmatrix}
\begin{bmatrix}
\xi_1 \\
\xi_2 \\
\xi_3 \\
\vdots \\
\xi_T
\end{bmatrix}
= \begin{bmatrix}
\sigma^2 \\
\rho \sigma^2 \\
\rho^2 \sigma^2 \\
\vdots \\
\rho^{T-1} \sigma^2
\end{bmatrix}
\]

\[ \Omega^{-\frac{1}{2}} \]

\[ \Omega^{-\frac{1}{2}} = P = 
\begin{bmatrix}
1 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ldots & 0 \\
0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 1
\end{bmatrix}
\]

(d. Now, what's the square root of the inverse of this? Or, more precisely

"\( \Omega^{-\frac{1}{2}} \)"

\[ \equiv P = 
\begin{bmatrix}
1 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ldots & 0 \\
0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 1
\end{bmatrix}
\]

(take my word for it)

"now, this is what it is"

\[ \Omega^{-\frac{1}{2}} = P = 
\begin{bmatrix}
1 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ldots & 0 \\
0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 1
\end{bmatrix}
\]

(c. So, the transformed data are:

\[
Y_t - \rho Y_{t-1} = \rho Y_{t-1} - \rho Y_{t-2}
\]

\[
Y_t^{*} = Y_t - \rho Y_{t-1}
\]

\[
Y_t + \rho Y_{t-1}
\]

\[ X_{t-1} - \rho X_{t-1}
\]

\[ X_{t-1} + \rho X_{t-1}
\]

1) That is, the transformation is to take "partial differences." The intuition here is, as before: we want to "count" only the new info in obs t, i.e., loosely: \( \rho(\text{obs. } t-1) \) of it is not new — that is, just the part related to the previous obs.

2) Also, notice that this "partial differencing" amounts to some thing as estimating a coefficient on a lagged estimated residual:

\[ Y_t = X_t \beta + \xi_t \]

Partial differenting:

\[ Y_t^{*} = X_t \beta + \xi_t^{*} \]

\[ Y_t^{*} - \rho Y_{t-1}^{*} = (X_t - \rho X_{t-1}) \beta + \xi_t - \rho \xi_{t-1} \]

\[ Y_t = X_t \beta + \rho Y_{t-1} - \rho X_{t-1} \beta + (\xi_t - \rho \xi_{t-1}) \]

\[ Y_t = X_t \beta + \rho (Y_{t-1} - X_{t-1} \beta) + \xi_t \]

\[ Y_t = X_t \beta + \rho \xi_{t-1} + \xi_t \]

\[ \text{Estimating a Coefficient on lagged residual} \]

[Comment: Only thing is that latter method has a harder time dealing with the first observation -- not a big deal in large to medium samples.]
I. D. 4. The fact that "partial differencing" amounts to estimating a coefficient on the lagged residual is a very convenient thing because

a) $\Omega^{-1}$ quickly gets extremely messy as we start to consider more lags & ARIMA(p,q) structures

b) But, the general principle that we can simplify it to "controlling for lagged residuals" basically remains & remains relatively simple.

II. Practical Estimation of Time-Series Models in Stata

A. No such thing. If you're going to be doing a lot of TS, consider learning another package (E-Views, RATS, LinDep, etc.)

B. Stata "Features"

1. `varname[-n-]` can be used in an appropriately sorted data set to create lagged variables. All such variables have to be created first using "gen varl = varname[-n-]"

   Differences must be created likewise: "put name you want here
gen difvar = var1-var[-n-]"

2. AR(p) Models can be estimated by typing "corc" (Gchrime-Orcutt), "prais" (Prais-Winston), "blu" (Hildreth-Lu) instead of "reg" or "fit".

   All three are roughly equivalent ways of estimating models with AR(p) error processes ($\epsilon_t = \rho \epsilon_{t-1} + \eta_t$)

3. Newey-West V-Cov Matrices can be calculated using "newey" in place of "fit" or "reg"

4. Near as I can tell, that's it. More complicated AR(X) & MA(X) processes are unavailable without your programming them. Not is any "newey" "corc" combination or whatsoever. Reportedly, stata's producers hope to bring some "time-series features" online next generation.

   C. Thankfully, for the occasional doubter, these limited abilities then suffice.

III. Estimation of Lagged-Dependent Variable (LDV) models where residuals remain correlated.

A. One of the first things proven in most econometrics texts is that OLS is biased & inconsistent (as well as inefficient) when one uses LDV's in the presence of autocorrelated residuals. This is of course correct, but that last underlined clause is key. It's "wrong" only if the true residuals are correlated controlling for the lag.

   First, let's demonstrate the problem; then we'll consider testing for it. I'll hint at a "solution" (my comment: "solution" is that the problem is over-blown & the "solution" ain't all that bloody great).

B. The "Problem": if $Y_t = X_\beta + Y_{t-1} + \epsilon_t$

   $\epsilon_t = \rho \epsilon_{t-1} + \eta_t$ iid

2) then $\text{Cov}(Y_{t-1}, \epsilon_t) = \text{Cov}(Y_{t-1}, \rho \epsilon_{t-1} + \eta_t) = \text{Cov}(Y_{t-1}, \rho \epsilon_{t-1}) + \text{Cov}(Y_{t-1}, \eta_t) = \rho \text{Cov}(Y_{t-1}, \epsilon_{t-1}) + \eta_t$

   $\text{Cov}(Y_{t-1}, \eta_t) = \rho \text{Cov}(Y_{t-1}, \epsilon_{t-1}) + \rho \text{Cov}(\epsilon_{t-1}, \eta_t)$

   $\Rightarrow \text{Cov}(Y_{t-1}, \epsilon_t) = \rho \text{Cov}(Y_{t-1}, \epsilon_{t-1}) + \rho \text{Cov}(\epsilon_{t-1}, \eta_t) \neq 0$
Simplified Exposition Bias if LDV w/Gen Errs:

\[ Y_t = X_t \beta + \delta Y_{t-1} + \varepsilon_t \]

\[ \varepsilon_t = \rho \varepsilon_{t-1} + \omega_t \quad \text{with} \quad \omega_t \sim \text{iid} \]

\[
\text{Cov}(Y_{t-1}, \varepsilon_t) = \text{Cov}(Y_{t-1}, \rho \varepsilon_{t-1} + \omega_t) \\
= \text{Cov}(X_t \beta + \delta Y_{t-2} + \varepsilon_{t-1}, \rho \varepsilon_{t-1} + \omega_t) \\
= \rho \text{Cov}(\varepsilon_{t-1}, \varepsilon_{t-1}) + \rho \times \text{Cov}(Y_{t-2}, \varepsilon_{t-1}) \\
= \rho \frac{\sigma^2}{1 - \rho^2} + \rho \times \text{Cov}(Y_{t-1}, \varepsilon_t) \\
\]

\[ \Rightarrow (1 - \rho^2) \text{Cov}(Y_{t-1}, \varepsilon_t) = \rho \frac{\sigma^2}{1 - \rho^2} \]

\[ \Rightarrow \text{Cov}(Y_{t-1}, \varepsilon_t) = \frac{\rho \sigma^2}{(1 - \rho^2)(1 - \rho^2)} \neq 0 \]

\[ \text{... UNLESS } \rho = 0 \]

\[ \boxdot \]
III: B. Since Cov($Y_{i+1}$, $e_t$) = $\rho \sigma^2 \frac{p}{1 - \rho^2}$, it should be equal to 0.

The fourth C.I.R.M assumption (that Cov($X^*, e_t$) = 0) is violated. That assumption was necessary to OLS unbiasedness & consistency.

3) OK, so if $Y_{i+1} = X_t \beta + \phi Y_{i} + e_t$

and $E_t = \rho e_{t-1} + \omega_t$, with $\omega_t$ iid,

then OLS is biased & inconsistent. By how much?

\[
\lim_{n \to \infty} b_{\text{OLS}} = \beta + \frac{\rho(1-\rho^2)}{1+\rho^2} \sigma^2
\]

(i.e., the inconsistency (or the "large sample bias") The $\beta$ are wrong too, by a similar consideration.

a) Is this a lot? Depends, but in my experience, $\sigma^2$ is usually large (between .50 & 1) and $\rho$ is usually about .5 as large at most the few times I've estimated both.

\[
\Rightarrow \text{ This is usually not such a big deal (see Table)} \text{ as is made up of it in most tests (at least for examples I've seen)}
\]

b) It can, however, be serious, & we'd definitely want to check if we're in danger if we can (of bias).

C. The Instrumental Variables "cure" for the Cov($X^*, e_t$) "disease"

1. The Problem is that if $X$ includes $Y_{i}$ & $e_t$ is auto-correlated, then Cov($X, e_t$) $\neq 0$ & we have a problem: $\beta$ biased & inconsistent.

2. A "solution" (we'll see it again in more detail):

a) Find some $Z$ such that Cov($Z, X$) $\neq 0$

$\Rightarrow$ Cov($Z, e_t$) $\neq 0$

That is, find some variable(s) that are uncorrelated with the residual (no way to test this, it's a theoretical proposition) but nonetheless correlated with $X$ (i.e., in this case, in particular with $Y_{i-1}$).

b) Then $b_{\text{IV}} = (Z'X)^{-1} Z'Y$ is unbiased consistent (asymptotically efficient) in practice, it's inefficient to a degree Cov($Z, X$) is or less meaningless not as high as it could be it still have Cov($Z, e_t$) $= 0$

IV. Testing for Autocorrelation at Residuals

A. Graphical Methods: Plot $e_t$ over time (i.e., in order from $t_0$ to $T$ along $X$ axis & $e_t$ on $Y$ axis).

Look for tendencies to have residuals stay on one side or other of zero for non-random amounts of time or to oscillate regularly.

\[
\begin{array}{c}
\text{This is strong positive correlation,} \\
\text{Usually much harder to see than this} \Rightarrow \text{statistical methods}
\end{array}
\]

\[
\text{This is strong negative correlation (rare in most types of data).}
\]
IV. Testing for Autocorrelated Residuals

B. The Durbin-Watson Statistic: we're looking for indication that residuals "go together" with some temporal linkage.

1. Consider

\[ DW = d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \]

a) Note: \( E(e_t) = 0 \) for all \( t \)

b) If \( e_t \) & \( e_{t-1} \) do not "go together" there should be no tendency for \( e_t - e_{t-1} \) to be larger or smaller than \( e_t - 0 \).

c) It will be twice as variant though

\[ \Rightarrow d = 0 \)

d) If \( e_t \) & \( e_{t-1} \) are uncorrelated, \( E(d) = 2 \)

e) If \( e_t \) & \( e_{t-1} \) positively correlated, then \( (e_t - e_{t-1})^2 \) is generally smaller than \( e_t^2 \) \( \Rightarrow \) \( d < 2 \). The opposite for negative correlation.

2. So, \( DW \approx 2 \) \( \Rightarrow \) no 1st-order correlation (1st-order means one lag, 2nd-order is two, and so on)

a) \( \Rightarrow \) Problem 1: DW distribution is a bit indeterminate. We don't get one clear value below or above which we reject \( H_0 \): no 1st-order autocorrelation. Get two values \( d^{\text{lower}} \) and \( d^{\text{upper}} \) for each \( p \)-level (actually, we get a range for which these are the bounds). If \( DW < d^{\text{lower}} \) then reject; if \( DW = d^{\text{upper}} \) then don't reject; if in between: shrug. Most folk balk at \( DW < 1.7 \) or \( DW > 2.3 \)

b) \( \Rightarrow \) Problem 2: whatever we decide on the statistic's range, it can be large for small samples, DW tells us nothing about any higher-order correlations which may or may not be present.

3. Problem 3: In presence of a lagged dependent DW is incorrect.

\[ Durbin's h: \quad h = \frac{1}{n} \sum_{t=1}^{n-1} e_t e_{t-1} \quad \text{where} \quad S_e^2 \text{ is the estimated variance of the coeff. on the lag } \]

\[ h \sim N(0, 1) \]

High values of \( h \) (those with low \( p \)'s associated with them) imply rejection of \( H_0 \): no first-order autocorr.

4. Despite all these problems & squabbies, DW is still the most commonly cited & reported serial correlation stat. As I said, most folk are uncomfortable with \( DW < 1.7 \) or \( DW > 2.3 \) (approx)

- State will give you a DW only if you explicitly ask for it by using "regdw" instead of "reg", or "fit," for "cor," "pars," 4 "nl" it reports DW for untransformed & transformed regressor.

- Becker should indicate transformation necessary & whether it was sufficient if all is well. If not, need to do more (or less).

5. There is some claim that DW are powerful tests for \( AR(1) \). This could only be so if we always take \( d^{\text{upper}} \) reject for any \( DW < d^{\text{upper}} \), but for many samples, this would amount to always rejecting. Personally, I think DW is just a quick, crude 1st look. It is somewhat better, but much less well known.
1. Breusch-Godfrey: Just Regress $e_t$ on $X_t \& e_{t-1} \ldots e_p$.
   a) If no $AR(p)$ or $MA(q)$ process up to the $p^{th}$ order, then this regression should have no explanatory power.
   b) $T \cdot R^2 \sim \chi^2_p$ if $T \cdot R^2$ too high, then $e_{t-1} \ldots e_p$ explained too much of $e_t$ to be mere chance.
   c) This is very intuitive.
   d) works for AR or MA equally well.
   e) works with or without lagged $y$'s (just be sure they're in $X$ where they belong).
   f) quite reasonably powerful for test of $p^{th}$ order.
   g) this is about as good as it gets.
   h) Supposed problem: you have to choose $p$. This is same for any test, though, so not really relevant.

$\Rightarrow$ These are easy and should be used if reported much more than they are, especially anywhere $AR(p), MA(q), \forall p, q > 1$ is expected for where lagged $y$'s are included.

2. Box-Pierce: $Q \&$ Ljung-Box $Q$
   a) Box-Pierce: $Q = T \sum \frac{E_i^2}{T}$ where $E_i = e_{i-1} - \frac{1}{T} \sum_{i=1}^{T} e_t^2$

   $\Rightarrow$ This is kind of a mix of a $DW$-like thing calculated for all lags up to $L$, squared (to get $e^2$ things) and multiplied by $T$ like a Breusch-Godfrey.

   $\Rightarrow$ it's a summary statistic of $DW$ or $e^2$ for lags up to $L$.

   b) Ljung-Box refined this a bit to:

   $Q' = T(T+2) \sum \frac{E_i^2}{T}$

   *Same notion though, same indication it has better small-sample properties.

$\Rightarrow$ c) either way, $Q \sim \chi^2_L$ if too high or start implies some auto-correlation in $e_t$ at $1^{st}$ to $L^{th}$ order.

3. Autocorrelation & Partial Autocorrelation: Graphs/Tests/Table
   a) You could also simply estimate $Corr (e_t, e_{t+k})$ for any and all $k$ you like along with standard errors. AC plots.

   b) Similarly, you could also estimate $Corr (e_t, e_{t-1} | e_{t-1}, e_{t-2}, \ldots e_{t-1})$ along with standard errors.

   d) These can be intuitively revealing especially graphically once you begin to be able to distinguish AR & MA processes in them.

   b) You can download a set of procedures to do this in stata from the stata web site.

4. Probably all of these are preferable to DW, especially in time-series domain. Usual practice is to report $DW$ and leave it at that if $2.3 > DW > 1.7$. Better, perhaps, would be to do that & add footnote referring to these other tests (whatever relevant ones you conducted), especially with lagged $y$.
V. Summary Wrap-Up of What to Do

A. Consider process generating your data: do you expect independent observations controlling for X? 

B. Try to Model the Dependence Process 
   - Consider a lag of Y. Make sense? How many lags? 
   - How about lagged X's? Delayed response of Y to X? Longing response of Y to X? (often lagged Y can approximate latter adequately).

C. Does This Take Care of the Matter (i.e., Test for remaining correlation in residuals) (My experience is that usually it is no)
   - Breusch-Godfrey LM test, Lagrange multiplier test
   - Box-Pierce/Q-Q, and PAC, ACF suggest remaining autocorrelation

D. Try Modelling the Error Process (also or instead, depending on above)
   - i.e., FG-LL by Cor, prais, or hlu. Not usually much difference. 
     Cor is standard, but some indication prais & hlu more efficient, especially in smaller samples. Go with care.
     If you want to use standard, avoid getting into it; prais or hlu otherwise.
     Try Newey-West if indications (e.g., Breusch-Godfrey LM test) are that autocorrelation remaining is not terribly large (say \( \hat{z}_p < .33 \))
   - especially if this

V. Cleaning Up

A. What's all that "unit root" stuff I've heard about: 

1. It's complicated. 
2. Recall that we want \( \frac{1}{n} X'X \) to converge to something (the variance-covariance matrix of regressors) 
   - that we want \( V(\varepsilon) \) to be defined (less than 00). 
   - Usually the latter is the focus of our worries in this context. 
3. If \( \varepsilon_t = \rho \varepsilon_{t-1} + \varepsilon_t \) with \( \varepsilon_t \) iid, then 
   \[
   V(\varepsilon) = \rho^2 V(\varepsilon_{t-1}) + V(\varepsilon_t)
   \]
   with constant \( V(\varepsilon) \) which we have on which we get 
   by transforming appropriately
   \[
   V(\varepsilon) = \rho^2 V(\varepsilon) + \sigma^2 
   \]
   \[
   \Rightarrow V(\varepsilon) = \frac{\sigma^2}{1-\rho^2} 
   \]
   - Only if \( |\rho| < 1 \) is this appropriately defined. 
   - \( |\rho| = 1 \) \Rightarrow \( \sigma^2 = \sigma^2 \) yikes
   - \( |\rho| > 1 \) \Rightarrow \( \sigma^2 - \rho^2 < 0 \) yikes

4. So, we want \( |\rho| < 1 \). Check to be sure this is so. I.e., hypoth. test on \( \rho = 1 \) (or \( \rho = -1 \), whichever is appropriate), you want to reject

5. Under null of \( \rho = 1 \), \( t = \frac{\varepsilon_t}{\sqrt{\hat{\rho}}} \) is not distributed as a \( t_{n-p} \) but something else. Basically, you need \( r \) this bigger than usual. At least \( d \) is good.
V. Cleaning Up

A. Unit Roots

5. So, check to be sure $|\beta| < 1$ by at least 3 s.e. $\equiv 4$ would be comfortable.

6. What if $|\beta|$ too close to 1 (or over)?

"Usually this means simply that you "have't controlled for enough" Why? note,

$$E_t = \rho E_{t-1} + \varepsilon_t$$

$$= \rho (\rho E_{t-2} + \varepsilon_{t-1}) + \varepsilon_t$$

$$= \rho (\rho (E_{t-3} + \varepsilon_{t-2}) + \varepsilon_{t-1}) + \varepsilon_t$$

and so on

$$= \sum_{s=0}^{\infty} \rho^s \varepsilon_{t-s}$$

Current residual is a weighted sum of all past "shocks." If $\rho < 1$, this weight diminishes as shocks recede into the past... shocks 20 years ago have $\rho^{20}$ lingering effects, whereas last year's shock has $\rho$ lingering.

With $|\rho| > 1$, this means that events/shocks further back into the past have more influence than more recent ones. Usually, this does not make sense. For $|\rho| = 1$, this means shocks/events way back in the past have same magnitude as more recent ones. Also, often doesn't make sense.

You're probably over-estimating correlation in stochastic component because you left too much in the stochastic component. Consider specifying the model to "control for more stuff," especially if more of the kind of stuff that's relatively invariant over time.

B. Interpreting Dynamic Specifications:

$$y_t = \omega y_{t-1} + x_t \beta + \varepsilon_t$$

with $\varepsilon_t$ iid or autocorrelated doesn't matter here.

$$\frac{\partial y_t}{\partial x_t} = \beta$$

same as always

but next period, $y_{t+1}$ is larger (smaller) by $\beta \Delta x_t$ units which means $y_{t+1}$ is $\omega \Delta y_t$ larger (smaller) in addition to the $X + \beta$ effect.

example

$\beta > 0$

$\omega > 0$

\begin{align*}
  \beta & > 0 \\
  \omega & > 0 \\
  X & \text{ goes up 1 unit} \\
  \Delta x_t & \text{ stays there} \\
  & \text{ so-called "permanent shock"} \\
  \end{align*}

\begin{align*}
\text{at time } t_0, \\
X & \text{ goes up 1 unit} \\
\text{ and stays there} \\
\text{ same except } \\
\text{ in this case} \\
\omega & \text{ is a temporary shock} \\
\text{ and } & \text{ the effect lingers into} \\
\text{ the future, decaying at rate } & \omega
\end{align*}
V. B. Interpreting Dynamic Specifications: \( y_t = \omega y_{t-1} + \beta x_t + \varepsilon_t \)

1. Effects of \( \Delta x \) fade away at rate \( \omega \)

2. Long-Run effect:
   a) for temporary shock, \( \forall t < 1 \), the effect eventually fades away completely \( (\lim_{t \to \infty} \omega^t = 0) \)
   b) for a permanent shock, we have cumulative effects:

   \[ \text{long-run effect is} \quad \Delta y = \Delta x \sum_{t=0}^{\infty} \omega^t \]

   \[ \frac{1}{1-\omega} \times \text{the "immediate" effect, } \beta \Delta x \]

   \[ \Delta x \cdot \left( \frac{1}{1-\omega} \right) \beta \]

3. Duration of Effects:

   • for the AR(1) type model, i.e. with lagged \( y \), all effects linger forever, but with \( |\omega| < 1 \) they do fade away to nothing.

   • To talk about "duration" of effects, we can think in "half-lives" and the like, i.e. how long until 50% of the effect has faded away (a half-life)

\[ \Rightarrow 0.5 = \omega^T, \text{ solve for } T \]

   • easy way:

\[ \ln(0.5) = x \ln(\omega) \Rightarrow x = \frac{\ln(0.5)}{\ln(\omega)} \]

   • more usually we ask how long for a large % to fade away, such as 90%, or 99% or something. Can solve for any % you like though:

\[ 90\%-\text{life:} \quad 0.1 = \omega^T \Rightarrow x = \frac{\ln(0.1)}{\ln(\omega)} \]

   • n.b. for 2%, want \((1-2)\) here.

4. This simple LDV model is also called a "partial adjustment" model (PA) because it models \( y \) as adjusting slowly (partially each period) to changes in \( x \), eventually reaching its new (long-run) equilibrium level. PA's make sense in a lot of contexts: a) where it's costly for agents to adjust \( y \) rapidly.

   b) as a short-hand for effects of \( x \) that linger in a decaying manner.

5. What about correlation in residuals? i.e. what about \( \rho \) for \( \varepsilon_t \) and \( \varepsilon_{t-1} \)? Changes in \( x \) produce changes in \( E(y) \) not the like? in \( \varepsilon \), so by itself correlated residuals produces no dynamics in the systematic component. Only in the stochastic component. There, a shock, \( \varepsilon_t = 1 \) in period \( t \) produces (assume all other \( \sigma = 0 \))

\[ \varepsilon_0 = \gamma_0 \]

\[ \varepsilon_1 = \rho \varepsilon_0 \]

\[ \varepsilon_2 = \rho^2 \varepsilon_0 \]  

\[ \Rightarrow \varepsilon_t = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-1} \]

\[ \Rightarrow \text{Same sort of thing, but only in the stochastic component} \]

6. Dynamic Specifications get much more complicated than this, but a general rubric for interpreting them is to put them in a form:

\[ y_t = \sum \omega_i y_{t-i} + \sum \beta_i x_{t-i} + \varepsilon_t \]

\[ \Rightarrow \text{which might also be expandable} \]
Interpreting Dynamic Specifications:

6. \( Y_t = \sum \omega_i Y_{t-i} + \sum X_{t-i} \beta_i + \epsilon_t \) & Use a spreadsheet

\[ \Delta Y_t = \omega_1 \Delta Y_{t-1} + \omega_2 \Delta Y_{t-2} + \beta_1 \Delta X_t + \beta_2 X_{t-1} + \epsilon_t \]  
(an error-correction model)

\[ Y_t - Y_{t-1} = \omega_1 (Y_{t-1} - Y_{t-2}) + \omega_2 Y_{t-2} + \beta_1 (X_t - X_{t-1}) + \beta_2 X_{t-1} + \epsilon_t \]

\[ Y_t = (1 + \omega_1) Y_{t-1} + (\omega_2 - \omega_1) Y_{t-2} + \beta_1 X_t + (\beta_2 - \beta_1) X_{t-1} + \epsilon_t \]

\[ \Rightarrow \text{an E.C. model is just a type of PA model.} \]

e.g. unit-root worries? Is \((1 + \omega_1, + \omega_2 - \omega_1) \approx (1 + \omega_2)\) sufficiently below \(1\)? IF, is \(\omega_2\) sufficiently below \(0\).

One of Beauties of E.C. Model is the interpretability of its coefficients:

- \(\omega_1\) is the short-term persistence in \(Y\) (how do changes persist into future changes)
- \(\omega_2\) is the decay rate (i.e., \((1 + \omega_2)\) is the PA rate)
- \(\beta_1\) is the long-term or level persistence
- \(\beta_2\) is the short-term (or changes or immediate) relation between \(X\) & \(Y\)

Used\(\text{A Spreadsheet}\) to present such results is often helpful.

<table>
<thead>
<tr>
<th>(t)</th>
<th>A1</th>
<th>B1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ T_{large} \]

\[ Y_t = \hat{\omega}_1 \ast (A3 - A2) + \hat{\omega}_2 \ast A2 + \hat{\beta} \ast (B4 - B3) + \hat{\beta} \ast B3 \]

\[ \Rightarrow \text{Response of} \ Y \text{ to one unit permanent increase in} \ X \text{ (ceteris paribus; ceteris zero actually)} \]
<table>
<thead>
<tr>
<th>Expected Value of Lagged Y Coefficient as Multiple of True Coefficient</th>
<th>Gamma = the coefficient on lagged Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>0.1</td>
<td>NA</td>
</tr>
<tr>
<td>Rho = the coefficient on lagged epsilon</td>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
<td>NA</td>
</tr>
<tr>
<td>0.4</td>
<td>NA</td>
</tr>
<tr>
<td>0.5</td>
<td>NA</td>
</tr>
<tr>
<td>0.6</td>
<td>NA</td>
</tr>
<tr>
<td>0.7</td>
<td>NA</td>
</tr>
<tr>
<td>0.8</td>
<td>NA</td>
</tr>
<tr>
<td>0.9</td>
<td>NA</td>
</tr>
<tr>
<td>1</td>
<td>NA</td>
</tr>
</tbody>
</table>
Models for Time-Series-Cross-Section Data

I. Data formats and terminology

A. TSCS: repeated obs across space (cntrys, regions, dyads, grps, individ’s, etc.) & time (decades, years, months, etc.)

B. Nested Data Types (subsets, starting w/ most general):

1. Multilevel/Hierarchical Data:
   a) multiple sub-unit observations per unit: \( y_{ij} \)
   b) Examples: students w/in classrooms; survey respondents w/in states or countries
   c) More than two levels also possible: students w/in classrooms w/in schools; survey respondents w/in states w/in countries.

2. Time-Series-Cross-Section:
   a) Sub-units are time periods: \( y_{it} \)
   b) Variants: units are same each time period (e.g., countries over time); units new sample each period (e.g., repeated surveys)

3. Panel (Longitudinal) Data:
   a) Sub-units are time periods; units are same each time period.

4. Notes; Further Complications:
   a) Terminology not consistently applied:
      (1) Some distinguish more by N×T dimensions, calling large-N, small-T panel data and larger-T (smaller-N) time-series cross-section
      (2) Formally, no difference whether N>T, T>N, by how much, or else, but dimensions crucial to estimation strategy practical efficacy since some work well as \( T \rightarrow \infty \), some as \( N \rightarrow \infty \) or \( NT \rightarrow \infty \)

   b) Irregular periodization

   c) Non-rectangular datasets and missing data
II. Notation, most-general (linear) model:
\[ y_{it} = \alpha_{it} + x'_{it} \beta + \varepsilon_{it}; \; \varepsilon \sim (0, \Sigma_{it}); \; i = 1..N, \; t = 1..T, \; n = NT \]

A. Nothing necessarily changes (all data are TSCS data):
1. If willing assume Gauss-Markov:
\[ y_{it} = \alpha + x'_{it} \beta + \varepsilon_{it}; \; \varepsilon \sim (0, \sigma^2 I) \]
\[ y = XB + \varepsilon; \; \varepsilon \sim (0, \sigma^2 I) \]
\[ \text{Cov}(X, \varepsilon) = 0 \]
\[ \{ \varepsilon \sim N(0, \sigma^2 I) \} \]

a) where last line not necessary OLS, if absent lean on CLT for distribution of estimates; necessary for ML=OLS.
b) Nothing new; this C(N)LRM => OLS=BLUE.

(1) BLUE: Best Linear Unbiased Estimator =>
(2) Coefficient-Estimate Properties: unbiased, consistent, efficient
(3) V-Cov(b)-Estimate Properties: unbiased, consistent, efficient

2. Similarly, if want to relax to
\[ y = XB + \varepsilon; \; \varepsilon \sim (0, \sigma^2 \Omega) \]
\[ \text{Cov}(X, \varepsilon) = 0 \]
\[ \{ \varepsilon \sim N(0, \sigma^2 \Omega) \} \]

a) w/ last line as above, nothing new: G(N)LRM => GLS=BLUE, FGLS=asymptotically BLUE, where asymptotically BLUE means:
b) FGLS Properties: Consistent, Asymptotically Efficient
3. Likewise, TSCS=collection of time-series, so all discussed regarding time-series models applies (with appropriate care to respect breaks b/w units in the series; e.g., $y_{2,1(t-1)} \neq y_{1,T}$.

B. Why separate chapters/sections on TSCS & panel then?

1. Substantively, interesting opportunities for testing
   a) Both cross-sectional & cross-temporal variation for leverage
   b) Examples:


   (2) *Individual Behavior*: CS: Impossible distinguish impact various aggregate-level factors (e.g., turnout & voting/electoral institutions & conditions); TS: impossible distinguish impact various unit-level factors (e.g., turnout & SES); TSCS=>both become possible.

   (3) *Institutions*: TS: vary little or 0 almost by def.; CS: covary w/ all other (relatively) fixed aspects (e.g., “culture”); TSCS: leverage

c) General Principle: (CS: little leverage)(TS: little leverage) = TSCS: little $\times$ little = more (*important that multiplicative*)

2. Statistically, interesting challenges for estimation
   a) Heteroskedasticity + Serial Correlation both likely now.
   b) New Beast: Contemporaneous (spatial) correlation.
   c) Quasi-philosophical issues of repeated sampling; n.b., key question is always population to which you’re trying to infer & relation of sample thereto. E.g., I disagree w/ Beck on “fixed $N$”.
   d) Questions of Parameter (Model) Stability: *really interesting*, like (b) above, because they’re so substantively central/“neat”:

      (1) Is effect of some institution (or drug or policy or...) same in all countries (or persons or firms or...)? If not, why & how does it vary?
(2) Does money-growth have the same effect on output everywhere and "everywhen"? Or does it vary? Latter question more interesting if we ask "and how?" e.g. more nominal contracting $\Rightarrow$ more effect? Pre-Robert Lucas more effect than post?

(3) Does electoral-institution or election-context or party-systemic feature have same effect on voting behavior of all individuals? How do these effects vary with individual characteristics like education?

C. In Summary:

1. TSCS usually offers best leverage these kinds of questions

2. Departure from before lies in plausibility of key assumptions:
   a) Parameter Stability:

   $$y_{it} = \alpha + x_{it}'\beta + \varepsilon_{it} \Rightarrow (\alpha_{it}, \beta_{it}) = (\alpha, \beta) \ \forall i, t$$

   b) Spherical Errors (homoskedasticity+uncorrelated):

   $$\varepsilon \sim (0, \sigma^2 \Omega) \text{ more plausible than } \varepsilon \sim (0, \sigma^2 I)$$


III. From the Most-General (& Inestimable) Form Down:

A. Most-General Form:

$$y_{it} = f_{it}(x_{it}, \beta_{it}, \varepsilon_{it}); \ \varepsilon \sim (0, \Sigma_{it}); \ i = 1..N, \ t = 1..T, \ n = NT$$

Notes: $x^1 = 1; \ \beta^1 = \alpha; \ x$ may contain time-space lags $x$ or $y$.

1. Parameters=$K+\frac{1}{2}(NT)^2+\frac{1}{2}NT \ per \ function, \ per \ observation!$

2. MASSIVELY under-identified$\Rightarrow$impose structure to reduce parameterization; from where? Theory & Substance (& Assume)
B. Virtually always assume:

1. $f_{it}(\cdot) = f(\cdot) \, \forall i, t$ : same fnctn relates $X_{it}, \beta_{it}, \epsilon_{it}$ to $y$ in all obs; may be stronger than needed; could parameterize changes $f_{it}(\cdot)$ or allow it to vary across but not within groups of obs $\{it\}$.

2. $\Sigma_{it} = \Sigma \, \forall i, t$ : each obs draw from distribution with same variance-covariance across obs; may be stronger than needed…

3. Parameters: $K(NT)+\frac{1}{2}(NT)^2+\frac{1}{2}NT$ per NT obs.$\Rightarrow K+\frac{1}{2}(NT+1)$ per obs. Still **way, way** too many.

C. Next, can assume constant coefficient-vector: $\beta_{it} = \beta \, \forall i, t$

1. Parameters: $K/(NT)+\frac{1}{2}(NT+1)$ per obs. Still **way** too many.

2. May be stronger than needed, can allow: $\beta_{it} = g(z, \gamma, \eta_{it})$, with parameters$<NT-K$-parameters($\Sigma$).

D. Still must reduce parameterization $\Sigma$.

1. Fully general var-covar structure not estimable; nothing to learn from history &/or comparison if insist all unique.

2. Plausible/practically-realistic variance-covariance structures:

   a) **Sphericity**: from $\sigma^2\Omega$ to $\sigma^2I \Rightarrow \text{from } \frac{1}{2}(NT)^2+\frac{1}{2}NT$ to $1$.

   b) **Panel Heteroskedasticity**: from $V(\epsilon_{it}) = \sigma^2_i \Rightarrow N$ parameters.

   c) **Serial Correlation**: $\epsilon_u = \rho \epsilon_{i,t-1} + u$ (2 params), or $\epsilon_u = \rho_t \epsilon_{i,t-1} \ (N+1 \text{ params})$

   d) **Parks-Kmenta**: panel het&ser+unique $\sigma_{ij} = \sigma_{ji} \, \forall ij$, but not $t \Rightarrow 2N+\frac{1}{2}N(N-1)\ldots$needs lots of $T$.

   e) **Many other plausible parameterizations…**
\[ x, \varphi = \mathcal{U}, \varphi = \mathcal{X} = (X| \delta) \Lambda = (\delta) \Lambda \]
IV. From simplest model upward (parsimony principle):

\[ y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}, \quad V(\varepsilon_{it}) = \sigma^2 \]

A. Pool all data and estimate by OLS

1. Advantages:
   a) Gives maximal leverage estimating parameters
   b) Consistent w/ general theories
   c) BLUE, iff this right model...

2. Disadvantages:
   a) Might not be right model.
   b) What might go wrong (specification error; omitted-variables):
      (1) Nonsphericity: \( V(\varepsilon) = \sigma^2 \Omega \neq \sigma^2 I \)
      (2) Unobserved (unmodeled) Unit (e.g., country) Effects: \( \alpha_i \neq \alpha \)
      (3) Unobserved (unmodeled) Time (sub-unit) Effects: \( \alpha_t \neq \alpha \)
      (4) Unobserved (unmodeled) Coefficient Variability: \( \beta_i \&/or \beta_t \neq \beta \)
   c) What might go wrong: an example in a picture
B. 1st Defense: Model It! (besides, substance, not nuisance)

1. If, for example, expect some pattern non-sphericity, this likely because you expect some systematic…

   a) …variation in effect of \( x_{it} \) across \( i,t \)

      (1) \( \Rightarrow \) what looks like heteroskedasticity if model \( \beta \) as a constant

      (2) \( \Rightarrow \) model the interactive (or group-wise varying) effect:

         \[ \beta_{it} = \gamma_0 + \gamma_1 z_{it} (+\phi_i) \] \( (\Rightarrow \) linear-interaction model \)

         \[ \beta_{it} = \gamma_i (+\phi_i), \text{ or } \beta_{it} = \gamma_t (+\phi_t), \text{ or } \beta_{it} = \gamma_s (+\phi_s) \] \( (\text{same}) \)

   b) …dependence of \( y_{it} \) on \( y_{i,t-1} \), &/or \( y_{it} \) on \( y_{jt} \)

      (1) \( \Rightarrow \) what looks like serial &/or spatial error-correlation if fail model the temporal &/or spatial dynamics in outcome, \( y \)

      (2) \( \Rightarrow \) model the temporal &/or spatial dynamics:

         \[ y_{it} = \alpha_{(i,t)} + \beta_{(i,t)} x_{i,t} + \rho_{(i,t)} y_{i,t-1} + \varepsilon_{it} \]

         \[ y_{it} = \alpha_{(i,t)} + \beta_{(i,t)} x_{i,t} (+\rho_{(i,t)} y_{i,t-1}) + \theta_{(i,t)} \sum_{j \neq i} w_{ij} y_{ji} + \varepsilon_{it} \]

2. Implications of “Model It!” Strategy; first, call all RHS: \( X\beta \)

   \[ \hat{\beta}_{OLS} = (XX)^{-1} X'y = (XX)^{-1} X'X\beta + \varepsilon \]

   a) \( = (XX)^{-1} X'X\beta + (XX)^{-1} X'\varepsilon \)

   \( = \beta + (XX)^{-1} X'\varepsilon \Rightarrow E(\hat{\beta}) = \beta \text{ if } E(\varepsilon) = 0 \) (as usual)

   \[ V(\hat{\beta}_{OLS}) = V[(XX)^{-1} X'y] = (XX)^{-1} V(\varepsilon) X(X'X)^{-1} \]

   \( = (XX)^{-1} X'\sigma^2 I X (XX)^{-1} \)

   b) \( = \sigma^2 (XX)^{-1} X'X (XX)^{-1} \)

   \( = \sigma^2 (XX)^{-1} \) (as usual)

   c) OLS=BLUE if model right \((^\text{spatial})\); what if imperfect/incomplete?
3. For example, suppose use just linear-interaction, when linear-interaction w/ error (=random coefficients)

\[
\text{Truth: } y_{it} = x_{it} \left( \gamma_0 + \gamma_1 z_{it} + \phi_{it} \right) + \varepsilon_{it}
\]

\[
\text{Model: } y_{it} = x_{it} \left( \gamma_0^* + \gamma_1^* z_{it} \right) + \varepsilon_{it}^*
\]

\[
\Rightarrow \hat{\beta}_{OLS} = \left( \left[ x \quad xz \right]' \left[ x \quad xz \right] \right)^{-1} \left[ x \quad xz \right]' y
\]

\[
= \left( \left[ x \quad xz \right]' \left[ x \quad xz \right] \right)^{-1} \left[ x \quad xz \right]' \left[ \gamma_0 x + \gamma_1 x \cdot z + \phi \cdot x + \varepsilon \right]
\]

a)

\[
= \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} + \left( \left[ x \quad xz \right]' \left[ x \quad xz \right] \right)^{-1} \left[ x \quad xz \right]' \left[ \phi \cdot x + \varepsilon \right]
\]

\[
\Rightarrow E(\hat{\beta}) = \beta \text{ if } E(\phi \cdot x) = 0, \ E(X'\varepsilon) = 0
\]

\[
\text{V}(\hat{\beta}_{OLS}) = \text{V}\left[ (X'X)^{-1} X'y \right] = (X'X)^{-1} X' \text{V}(x \cdot \phi + \varepsilon) X (X'X)^{-1}
\]

b)

\[
= (X'X)^{-1} X' \sigma^2 \Omega X (X'X)^{-1}
\]

c) OLS=unbiased, but inefficient coefficients; wrong s.e.'s

d) However, some easy fixes (explained below…)

4. E.g., unit-specific effects or coefficients, but model only part of that parameter heterogeneity (mis-specification: OVB)

\[
\text{Truth: } y = X\beta + Z\gamma + \varepsilon
\]

\[
\text{Model: } y = X\beta^* + \varepsilon^*
\]

\[
\hat{\beta}_{OLS} = (X'X)^{-1} X'y = (X'X)^{-1} X' (X\beta + Z\gamma + \varepsilon)
\]

a)

\[
= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'Z\gamma + (X'X)^{-1} X'\varepsilon
\]

\[
= \beta + F_{ZX}\gamma + (X'X)^{-1} X'\varepsilon \text{ where } F_{ZX} \text{ is OLS } Z \text{ on } X
\]
\[
\mathbf{V}(\hat{\beta}_{OLS}) = \mathbf{V}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\right)
\]
\[
\text{but } \mathbf{V}(\hat{\beta}_{OLS}) = \mathbf{V}\left(\hat{\mathbf{F}}_{2X}\hat{\gamma} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\right)
\]
b) \[
\text{and } \mathbf{V}(\hat{\beta}_{OLS})_{OLS} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}
\]
c) I.e., completely-standard omitted-variable bias (OVB).

5. (Time-)Serial Dependence:

a) If temporal dynamics specified in systematic component suffice (no residual/stochastic-component correlation remaining, which is testable), OLS=BLUE.

b) If insufficient, OLS inconsistent, but still:

\[
E\left(\hat{\rho}_y\right) = \rho_y + \frac{\rho_\varepsilon (1-\rho_y^2)}{1+\rho_\varepsilon \rho_y}
\]

(1) Magnitude of the problem:

(2) And can (partially) address s.e. part of problem as below…

![Absolute Bias in OLS-Estimates rho(Y)=F(rho(e),rho(Y))](image-url)
Percentage Bias in OLS-Estimates \( \rho(Y) = F(\rho(e), \rho(Y)) \)

\[\begin{array}{c}
\text{Bias} \\
150.0\% \\
140.0\% \\
130.0\% \\
120.0\% \\
110.0\% \\
100.0\% \\
90.0\% \\
80.0\% \\
70.0\% \\
60.0\% \\
50.0\% \\
40.0\% \\
30.0\% \\
20.0\% \\
10.0\% \\
0.0\% \\
\end{array}\]

\[\begin{array}{c}
\text{rho(e)} \\
0.15 \\
0.3 \\
0.45 \\
0.6 \\
0.75 \\
0.9 \\
0.7 \\
0.35 \\
\end{array}\]

\(\begin{array}{c}
140.0\% - 150.0\% \\
130.0\% - 140.0\% \\
120.0\% - 130.0\% \\
110.0\% - 120.0\% \\
100.0\% - 110.0\% \\
90.0\% - 100.0\% \\
80.0\% - 90.0\% \\
70.0\% - 80.0\% \\
60.0\% - 70.0\% \\
50.0\% - 60.0\% \\
40.0\% - 50.0\% \\
30.0\% - 40.0\% \\
20.0\% - 30.0\% \\
10.0\% - 20.0\% \\
0.0\% - 10.0\%
\end{array}\)

c) Possible to model temporal dependence in both \( y \) & \( \varepsilon \) by NLS:
\[ y_t = X_t \beta + \rho_y y_{t-1} + \varepsilon_t; \quad \varepsilon_t = \rho \varepsilon_{t-1} + v_t; \quad v_t \sim (0, \sigma_v^2 I_N) \]
\[ \Rightarrow y_t = X_t \beta + \rho_y y_{t-1} + \rho \varepsilon_{t-1} + v_t \]
\[ = X \beta + \rho_y y_{t-1} + \rho \varepsilon_{t-1} \left( y_{t-1} - X_{t-1} \beta - \rho_y y_{t-2} \right) + v \]
\[ = (X_t - \rho \varepsilon_t X_{t-1}) \beta + \rho_y y_{t-1} + \rho \varepsilon_{t-1} \left( y_{t-1} - \rho_y y_{t-2} \right) + v \]

d) Note, however, indeterminacy in total = systematic + stochastic, so any two of possible lag \( y \), lag \( x \), lag \( \varepsilon \) => third, etc.

6. Spatial Dependence:

a) Situation more complicated. OLS inconsistent even if model spatial-dependence fully.

b) However, still generally better to model than to omit it, & we'll talk about redressing the simultaneity in this case later if time.
7. Summary:

a) If can model theoretical/substantive reason for deviation from C(N)LRM, in TSCS data or elsewhere, do so, and, if & insofar as successful, strategy is optimal in all regards.

b) Insofar as possible, “Model It!” in model of the first-moment, $E(y)$, i.e., the systematic component; for two reasons:

   1. Usually, theoretical/substantive info re: **systematic** component
   2. Observationally, only info on stochastic component (i.e., second moment) conditional on info in first moment (systematic component)
   3. May not be possible or theoretically/substantively correct; could have theory/substance info about second moment (variance). E.g.:
      (a) DepVar=average varying # lower-level outcomes $\Rightarrow V(\epsilon)$~1/#.
      (b) Thry/Subst e.g.: edu (info) not affect response; rather reliability or accuracy or theoretical-explicability of response $\Rightarrow V(\epsilon)=f(\text{edu})$
      (c) In such cases, still “Model It!” (in second moment, this means to model reduced parameterization of $\Omega$).

c) As seen, insofar as fail model fully deviations CLRM, problems arise essentially as omitted-variable bias in worst cases, but as “just” inefficiency and wrong s.e.’s in many other cases, so...

C. Redresses, mostly partial &/or imperfect, of deficiencies in implementation of the Model It!™ strategy

1. “Robust” or “Sandwich” Variance-Covariance Estimators

a) Key Insight: $\Omega$ in general has $\frac{1}{2}n(n+1) > n$ parameters; however, for consistent v-cov estimation, need consistent est only of $X'\Omega X$, which many fewer ($<\frac{1}{2}k(k+1)$, depending on parameterization).

b) Start Here: $V\left(\hat{\beta}_{LS}\right) = \sigma^2 (X'X)^{-1} X'\Omega X (X'X)^{-1}$. This differs from OLS formula, $V\left(\hat{\beta}_{OLS}\right) = \sigma^2 (X'X)^{-1}$, only insofar as $X'\Omega X$ differs from $X'X$, which is only insofar as elements of $\Omega$ covary with the elements of $X'X$, i.e. the $\omega_{ij}$ w/ $x_i^2$, $x_j^2$, &/or cross-products of $x$'s.
c) I.e., after matrix multiplication, see v-cov estimates using LS formula off by factor of: \( \sum_{i,j,s,t} e_{it}e_{js} (x_{it}x'_{js}) - \sum e_{it}^2 I_k \).

d) . . , we can “fix” our v-cov estimates, i.e. render them “robust”, i.e., consistent, to presence of certain pattern of non-sphericity by replacing \( X'\Omega X \) in correct formula w/ some \( \sum_{i,j,s,t} e_{it}e_{js} (x_{it}x'_{js}) \) configured to reflect that pattern of non-sphericity.

\[
\hat{V}_s \left( \hat{\beta} \right) = (X'X)^{-1} X'\Omega X (X'X)^{-1} \equiv (X'X)^{-1} \hat{Q}(X'X)^{-1}
\]

2. Cases:

a) Pure Heteroskedasticity (White’s Robust Standard Errors):

\[
\hat{Q} = \sum_{i} e_{i}^2 (x_{i}x'_{i})
\]

b) Panel Heteroskedasticity (Rob, yesterday morning):

\[
\hat{Q} = \sum_{i} \left( \sum_{t} e_{it}^2 (x_{it}x'_{it}) \right)
\]

c) Het. & (Time) Auto-Correlation (Newey-West HAC, Rob):

\[
\hat{Q} = \sum_{i} \left( \sum_{t} e_{it}^2 (x_{it}x'_{it}) \right) + \sum_{i} \left( \sum_{s=1}^{L} w_{i}e_{it}e_{i,t-s} (x_{it}x'_{i,t-s} + x_{i,t-s}x'_{i,t}) \right)
\]

where \( L = \text{max lag-length considered appreciable} \) & \( w_{i} = 1 - \frac{t}{L-1} \)


\[
\hat{Q} = X' \left( \frac{E' E}{T} \otimes I_T \right) X
\]

where \( E = \text{the } T \times N \text{ matrix estimated residuals} \)
e) Many others possible. Several “cluster” types, e.g., designed for various multilevel/hierarchical data structures (sub-units not time)

\[
\hat{\beta} = \frac{1}{N-k}(X'X)^{-1} \left( \sum_{j=1}^{n_c} \left( \sum_{i=1}^{n_j} e_i x_i \right) \left( \sum_{i=1}^{n_j} e_i x_i \right)' \right) (X'X)^{-1}
\]

where \( n_j \) = # obs. \( i \) in macro-level (cluster) \( j \), & \( n_c \) = # clusters

f) Small-sample adjustments have been suggested for each.

3. FGLS: Feasible Generalized-Least-Squares

a) Consistent V-Cov Estimates only address “inconsistency” of s.e.’s, do not address bias or efficiency of coefficient estimates (although require consistent coefficient-estimates for formal properties) or “unbiasedness” and “efficiency” of s.e.’s.

b) To improve efficiency coefficient (\& s.e.) estimates—still not directly or formally redress any bias concerns arising from other problems, OVB e.g., and still reliant on “first-stage” consistency—we can parameterize and estimate \( \hat{\Omega} \), use it to transform data to such that C(N)LRM applies.

c) Example: Parks-Kmenta FGLS for TSCS:

1. Panel-specific AR(1) in residuals => \( N \) parameters

2. Panel-specific \( \sigma_i^2 \) => \( N \) parameters

3. Dyad-specific \( \sigma_{ij} \) => \( N(N-1) \) parameters (n.b., symmetric)

4. =>N(N+1) pars=>unless T>>2N, inadvisable (Beck-Katz ‘95)

5. NOTE: Could offer more theoretically structured (\& thereby parametrically reduced) structure non-sphericity pattern => greater efficiency & better small-sample properties. E.g., just contemp corr. => N(N-1) parameters needs T>>N.

d) FGLS properties: consistent & asymptotically efficient.
given $\hat{\Omega}$, let $P \equiv \hat{\Omega}^{-\frac{1}{2}}$, then:

$$Py = PX\beta + Pe \implies \hat{\beta}_{FGLS} = \left[ (PX)' (PX) \right]^{-1} (PX)'Py$$

$$\hat{\beta}_{FGLS} = \left[ X'P'PX \right]^{-1} X'P'y = \left[ X'\hat{\Omega}^{-1}X \right]^{-1} X'\hat{\Omega}^{-1}y$$

$\implies$ unbiased, consistent, asympt'ly efficient if $C(X, \epsilon) = 0$

$$V(\hat{\beta}_{FGLS})_{FGLS} = \left[ X'\hat{\Omega}^{-1}X \right]^{-1} X'\hat{\Omega}^{-1}V(y)\hat{\Omega}^{-1}X \left[ X'\hat{\Omega}^{-1}X \right]^{-1}$$

$$V(\hat{\beta}_{FGLS})_{FGLS} = \sigma^2 \left[ X'\hat{\Omega}^{-1}X \right]^{-1} X'\hat{\Omega}^{-1}\hat{\Omega}^{-1}X \left[ X'\hat{\Omega}^{-1}X \right]^{-1}$$

$$V(\hat{\beta}_{FGLS})_{FGLS} = \sigma^2 \left[ X'\hat{\Omega}^{-1}X \right]^{-1} X'\hat{\Omega}^{-1}X \left[ X'\hat{\Omega}^{-1}X \right]^{-1}$$

$$V(\hat{\beta}_{FGLS})_{FGLS} = \sigma^2 \left[ X'\hat{\Omega}^{-1}X \right]^{-1}$$

$\implies$ "unbiased, consistent, asympt'ly efficient" (as above)

D. Unit & Period Fixed & Random Effects, w/ Example (gratitude to Greg Wawro):
1.1 Unobserved Country Effects and LSDV

- In the model:

\[
\text{Govt Spending}_{it} = \beta_0 + \beta_1 \text{Openness}_{it} + \beta_2 Z_{it} + \varepsilon_{it}
\]

It might be argued that the level of government spending as a percentage of GDP differs for reasons that are specific to each country (e.g., solidaristic values in Sweden). This is also known as cross-sectional heterogeneity.

- If these unit-specific factors are correlated with other variables in the model, we will have an instance of omitted variable bias. Even if not, we will get larger standard errors because we are not incorporating sources of cross-country variation into the model.

- We could try to explicitly incorporate all the systematic factors that might lead to different levels of government spending across countries, but places high demands in terms of data gathering.

- Another way to do this, which may not be as demanding data-wise, is to introduce a set of country dummies into the model.

\[
\text{Govt Spending}_{it} = \alpha_i + \beta_1 \text{Openness}_{it} + \beta_2 Z_{it} + \varepsilon_{it}
\]

This is equivalent to introducing a country-specific intercept into the model. Either include a dummy for all the countries but one, and keep the intercept term, or estimate the model with a full set of country dummies and no intercept.
1.1.1 Time Effects

- There might also be time-specific effects (e.g., government spending went up everywhere in 1973–74 in OECD economies because the first oil shock led to unemployment and increased government unemployment payments). Once again, if the time-specific factors are not accounted for, we could face the problem of bias.

- To account for this, introduce a set of dummies for each time period.

\[
G_{out\text{ }Spending}_{it} = \alpha_i + \delta_t + \beta_1 \text{Openness}_{it} + \beta_2 Z_{it} + \varepsilon_{it}
\]

- The degrees of freedom for the model are now \( NT - k - N - T \). The significance, or not, of the country-specific and time-specific effects can be tested by using and \( F \)-test to see if the country (time) dummies are jointly significant.

- The general approach of including unit-specific dummies is known as Least Squares Dummy Variables model, or LSDV.

- Can also include \( (T-1) \) year dummies for time effects. These give the difference between the predicted causal effect from \( x_{it}\beta \) and what you would expect for that year. There has to be one year that provides the baseline prediction.

(Recall pictures of pooled regression in presence of omitted unit-specific intercepts \&/or coefficients.)

1.2 Testing for unit or time effects

- For LSDV (including an intercept), we want to test the hypothesis that

\[
\alpha_1 = \alpha_2 = \ldots = \alpha_{N-1} = 0
\]
Can use an $F$-test:

$$F(N - 1, NT - N - K) = \frac{(R^2_{UR} - R^2_R)/(N - 1)}{(1 - R^2_{UR})/(NT - N - K)}$$

In this case, the unrestricted model is the one with the country dummies (and hence different intercepts); the restricted model is the one with just a single intercept. A similar test could also be performed on the year dummies.

1.2.1 How to do this test in Stata?

* After the `regress` command you type:

1. If there are (N-1) country dummies and an intercept
   ```
   test dummy1=dummy2=dummy3=dummy4=\ldots=dummyN-1=0
   ```
2. If there are N country dummies and no intercept
   ```
   test dummy1=dummy2=dummy3=dummy4=\ldots=dummyN
   ```

1.3 LSDV as Fixed Effects

* Least squares dummy variable estimation is also known as *Fixed Effects*, because it assumes that the variation in the dependent variable, $y_{it}$, for given countries or years can be estimated as a given, fixed effect.

* Before we go into the justification for this, let us examine which part of the variation in $y_{it}$ is used to calculate the remaining $\beta$ coefficients under fixed effects.
A fixed effects model can be estimated by transforming the data. To do this, calculate the country mean of $y_{it}$ for all the different countries. Let the group mean of a given country, $i$, be represented as $\bar{y}_i$.

Let the original model be

$$y_{it} = \alpha_i + x_{it} \beta + \varepsilon_{it} \tag{1.1}$$

Then:

$$\bar{y}_i = \alpha_i + \bar{x}_i \beta + \bar{\varepsilon}_i.$$

If we run OLS on this regression it will produce what is known as the “Between Effects” estimator, or $\beta_{BE}$, which shows how the mean level of the dependent variable for each country varies with the mean level of the independent variables.

Subtracting this from eq. 1.1 gives

$$(y_{it} - \bar{y}_i) = (\alpha_i - \alpha_i) + (x_{it} - \bar{x}_i) \beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

or

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i) \beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

If we run OLS on this regression it will produce what is known as the “Fixed Effects” estimator, or $\beta_{FE}$.

It is identical to LSDV and is sometimes called the *within-group estimator*, because it uses only the variation in $y_{it}$ and $x_{it}$ within each group (or country) to estimate the $\beta$ coefficients. Any variation between countries is assumed to spring from the unobserved fixed effects.
Note that if time-invariant regressors are included in the model, the standard FE estimator will not produce estimates for the effects of these variables. Similar issue w/ LSDV.

- IV approach to produce estimates, but requires some exogeneity assumptions that may not be met in practice.

- The effects of slow-moving variables can be estimated very imprecisely due to collinearity.

1.4 What Types of Variation do Different Estimators Use?

- Let us now determine the sum of squares ($X'X$) and cross-products ($X'y$) for the OLS estimator and within-group estimator in order to clarify which estimator uses what variation to calculate the $\beta$ coefficients.

- Let $S_{xx}$ be the sum of squares and let $S_{xy}$ be the cross-products. Let the overall means of the data be represented as $\bar{y}$ and $\bar{x}$.

- Then the total sum of squares and cross-products (which define the variation that we use to estimate $\hat{\beta}_{OLS}$) is:

$$S_{xx}^T = \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x})(x_{it} - \bar{x})'$$

$$S_{xy}^T = \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x})(y_{it} - \bar{y})$$
• The within-group sum of squares and cross-products (used to estimate $\hat{\beta}_{FE}$) is:

$$S^W_{xx} = \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'$$

$$S^W_{xy} = \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)$$

• The between-group sum of squares and cross-products (used to estimate $\hat{\beta}_{BE}$) is:

$$S^B_{xx} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$

$$S^B_{xy} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})$$

• It is easy to verify that:

$$S^T_{xx} = S^W_{xx} + S^B_{xx}$$

and:

$$S^T_{xy} = S^W_{xy} + S^B_{xy}$$

• We also have that:

$$\hat{\beta}_{OLS} = [S^T_{xx}]^{-1}S^T_{xy} = [S^W_{xx} + S^B_{xx}]^{-1}[S^W_{xy} + S^B_{xy}]$$

and

$$\hat{\beta}_{FE} = [S^W_{xx}]^{-1}[S^W_{xy}]$$

while,

$$\hat{\beta}_{BE} = [S^B_{xx}]^{-1}[S^B_{xy}]$$
The standard $\hat{\beta}_{OLS}$ uses all the variation in $y_{it}$ and $x_{it}$ to calculate the slope coefficients while $\hat{\beta}_{FE}$ just uses the variation across time and $\hat{\beta}_{BE}$ just uses the variation across countries.

- We can show that $\hat{\beta}_{OLS}$ is a weighted average of $\hat{\beta}_{FE}$ and $\hat{\beta}_{BE}$. In fact:
  \[
  \hat{\beta}_{OLS} = F^W \hat{\beta}_{FE} + F^B \hat{\beta}_{BE}
  \]
  where $F^W = [S_{xx}^W + S_{xx}^B]^{-1} S_{xx}^W$ and $F^B = [I - F^W]$

1.5 Random Effects Estimation

- Fixed effects is completely appropriate if we believe that the country-specific effects are indeed fixed, estimable amounts that we can calculate for each country.

- Thus, we believe that Sweden will always have an intercept of 1.2 units (for instance). If we were able to take another sample, we would once again estimate the same intercept for Sweden. There are cases, however, where we may not believe that we can estimate some fixed amount for each country.

- In particular, assume that we have a panel data model run on 20 countries, but which should be generalizable to 100 different countries. We cannot estimate the given intercept for each country or each type of country because we don't have all of them in the sample for which we estimate the model.

- In this case, we might want to estimate the $\beta$s on the explanatory variables taking into account that there could be country-specific effects that would enter as a random shock from a known distribution.
• The appropriate model that accounts for cross-national variation is random effects:

\[ y_{it} = \alpha + x_{it}' \beta + u_i + \varepsilon_{it} \]

In this model, \( \alpha \) is a general intercept and \( u_i \) is a time-invariant, random disturbance characterizing the \( i \)th country. Thus, country-effects are treated as country-specific shocks. We also assume in this model that:

\[
E[\varepsilon_{it}] = E[u_i] = 0 \\
E[\varepsilon_{it}^2] = \sigma^2_{\varepsilon}, E[u_i^2] = \sigma^2_u \]
\[
E[\varepsilon_{it} u_j] = 0 \quad \forall \ i, t, j; \ E[\varepsilon_{it} \varepsilon_{js}] = 0 \quad \forall \ t \neq s, i \neq j; \ E[u_i u_j] = 0 \quad \text{for} \ i \neq j.
\]

• For each country, we have a separate error term, equal to \( w_{it} \), where:

\[ w_{it} = \varepsilon_{it} + u_i \]

and

\[ E[w_{it}^2] = \sigma^2_{\varepsilon} + \sigma^2_u, \quad \text{and} \quad E[w_{it} w_{is}] = \sigma^2_u \quad \text{for} \ t \neq s. \]

• It is because the random effects model decomposes the disturbance term into different components that it is also known as an error components model.

• For each panel (or country), the variance-covariance matrix of the \( T \) disturbance terms will take the following form:

\[
\Sigma = \begin{bmatrix}
(\sigma^2_{\varepsilon} + \sigma^2_u) & \sigma^2_u & \sigma^2_u & \cdots & \sigma^2_u \\
\sigma^2_u & (\sigma^2_{\varepsilon} + \sigma^2_u) & \sigma^2_u & \cdots & \sigma^2_u \\
\vdots & \vdots & \ddots & \cdots & \vdots \\
\sigma^2_u & \sigma^2_u & \sigma^2_u & \cdots & (\sigma^2_{\varepsilon} + \sigma^2_u)
\end{bmatrix} = \sigma^2_{\varepsilon} I_T + \sigma^2_u i_T i_T'
\]

where \( i_T \) is a \( (T \times 1) \) vector of ones.
The full variance-covariance matrix for all the $NT$ observations is:

$$
\Omega = \begin{bmatrix}
\Sigma & 0 & 0 & \cdots & 0 \\
0 & \Sigma & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \Sigma
\end{bmatrix} = I_N \otimes \Sigma
$$

The way in which the random effects model differs from the original OLS estimation (with no fixed effects) is only in the specification of the disturbance term. When the model differs from the standard Gauss-Markov assumptions only in the specification of the errors, the regression coefficients can be consistently and efficiently estimated by Generalized Least Squares (GLS) or (when we don’t exactly know $\Omega$) by Feasible Generalized Least Squares (FGLS).

Thus, we can do a transformation of the original data that will give the FGLS estimates and creates a new disturbance matrix that conforms to Gauss-Markov.

1.6 FGLS Estimation of Random Effects

* The FGLS estimator is

$$
\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}y)
$$

To estimate this we will need to know $\Omega^{-1} = [I_N \otimes \Sigma]^{-1}$, which means that we need to estimate $\Sigma^{-1/2}$:

$$
\Sigma^{-1/2} = \frac{1}{\sigma_e} \left[ I - \frac{\theta}{T} i_i i_i' \right]
$$

where

$$
\theta = 1 - \frac{\sigma_u}{\sqrt{T \sigma_u^2 + \sigma_e^2}}
$$

* Then the transformation of $y_i$ and $X_i$ for FGLS is

$$
\Sigma^{-1/2}y_{it} = \frac{1}{\sigma_e} \begin{bmatrix}
y_{i1} - \theta \bar{y}_i \\
y_{i2} - \theta \bar{y}_i \\
\vdots \\
y_{iT} - \theta \bar{y}_i
\end{bmatrix}
$$
with a similar looking expression for the rows of $X_i$.

- It can be shown that the GLS estimator, $\hat{\beta}_{RE}$, like the OLS estimator, is a weighted average of the within ($FE$) and between ($BE$) estimators:

$$\hat{\beta}_{RE} = \hat{F}^W \hat{\beta}_{FE} + (I - \hat{F}^W) \hat{\beta}_{BE}$$

where:

$$\hat{F}^W = [S^W_{xx} + \lambda S^B_{xx}]^{-1} S^W_{xx}$$

and

$$\lambda = \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + T \sigma^2_u} = (1 - \theta)^2$$

- If $\lambda = 1$, then the random effects model reduces to OLS. There is no country-specific disturbance term, so the regression coefficients are most efficiently estimated using the OLS method.

- Where $\lambda = 0$, we have that the country-specific shocks far dominate the other parts of the disturbance. Then the random effects model reduces to the fixed effects model. We attribute all cross-country variation to the country-specific shock, with none attributed to the random disturbance $\varepsilon_{it}$ and just use the cross-time variation to estimate the slope coefficients.

- To the extent that $\lambda$ differs from one, we can see that the OLS estimation involves an inefficient weighting of the two least squares estimators (within and between) and GLS will produce more efficient results.

- To estimate the random effects model when $\Omega$ is unknown, we use the original OLS results, which are consistent, to get estimates of $\sigma^2_\varepsilon$ and $\sigma^2_u$. 
1.7 Testing between Fixed and Random Effects

- If the random, country-specific disturbance term, $u_i$, is correlated with any of the other explanatory variables, $x_{it}$, then we will get biased estimates in the OLS stage because the regressors and the disturbance will be contemporaneously correlated.

- The coefficient on $x_{it}$ will be biased and inconsistent, which means the OLS estimates of the residuals will be biased and the $\hat{\beta}_{RE}$ will be biased. This sets us up for a Hausman test:

  $H_0: \ E[u_i x_i] = 0$; Random effects appropriate $\Rightarrow \hat{\beta}_{RE}$ is approximately equal to $\beta_{FE}$ but is more efficient (has smaller standard errors).

  $H_1: \ E[u_i x_i] \neq 0$; Random effects is not appropriate $\Rightarrow \hat{\beta}_{RE}$ will be different from $\hat{\beta}_{FE}$ (and inconsistent).

- In this setting, the Hausman test statistic is calculated as:

  $$W = \chi^2_K = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' \hat{\Sigma}^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE})$$

  where

  $$\hat{\Sigma} = \text{var}[\hat{\beta}_{FE}] - \text{var}[\hat{\beta}_{RE}]$$

  If the Hausman test statistic is larger than its appropriate critical value, then we reject random effects as the appropriate specification.

- Greene, p. 298, also shows how to perform a Breusch-Pagan test for random effects based on the residuals from the original OLS regression. This tests for the appropriateness of OLS versus the alternative of random effects. It does not test random effects against fixed effects.
1.8 How to Do this in Stata

- `xtreg depvar [varlist], re` for random effects
- `xtreg depvar [varlist], fe` for fixed effects
- `xtreg depvar [varlist], be` for between effects

To perform the hausman test, type `xthausman`

After `xtreg depvar [varlist], re`

To run the Breusch-Pagan test for random effects versus OLS, type `xttest0`

After `xtreg depvar [varlist], re`

V. A Fuller List of Useful Stata™ Commands for TSCS (gratitude to Suzie De Boef)
1 Basics

All the Stata commands for pooled time-series cross-sections is listed under the “xt” commands. Before you use these, however, you need to tell Stata that you have this two dimensional data structure. To do so we specify the variables associated with time and with unit in each do-file that you write to run xt commands:

. iis county
. tis year

TSCS data must be in what Stata calls “long” format. Long format looks like this:

<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>dep</th>
<th>indep1</th>
<th>indep2</th>
<th>indep3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1900</td>
<td>13.6</td>
<td>75</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>1901</td>
<td>13.9</td>
<td>82</td>
<td>1</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1900</td>
<td>12.2</td>
<td>75</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>1901</td>
<td>12.1</td>
<td>82</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Wide” format looks like this:

<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>dep</th>
<th>indep1</th>
<th>indep2</th>
<th>indep3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1900</td>
<td>13.6</td>
<td>75</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>1900</td>
<td>12.2</td>
<td>75</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1901</td>
<td>13.9</td>
<td>82</td>
<td>1</td>
<td>110</td>
</tr>
<tr>
<td>2</td>
<td>1902</td>
<td>12.1</td>
<td>82</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To go from wide to long (or long to wide) you use the “reshape” command. See the Stata help file for details on this command.

2 Key -xt- commands

- `xtsum`: gives you the same information as the summarize command — gives means and standard deviations and does so both within units and between them.
• **xttab**: similar to tabulate, but again breaks down both within units and between them.

• **xtreg**: To fit fixed or random effects models, options include
  
  - `fe` estimates a fixed effects model,
  - `re` estimates a gls random effects models (default),

• **xtgls**: options include
  
  - `panel()` indicates the type of panel (unit) variability; options include:
    * `iid`: independent, common-variance panels, yielding a single estimate of $\sigma^2$,
    * `heteroscedastic`: uncorrelated units, each with its own variance estimate, yielding $N$ distinct estimates of $\sigma_i^2$,
    * `correlated`: heteroscedastic, spatially correlated panels, yielding $\frac{N(N-1)}{2} + N$ parameters ($\frac{N(N-1)}{2}$ covariances $\sigma_{it}, \sigma_{jt}$ and $N$ variances estimates $\sigma_i^2$),
  
  - `corr()` indicates the within-unit temporal correlation:
    * `indep` gives estimates no temporal correlation,
    * `ar1` gives estimates with common AR(1) errors, estimates a single value of $\hat{\rho}$.
    * `psar1` gives panel-specific AR(1) estimates, yielding $N$ separate estimates of $\hat{\rho}_t$.

  - A summary of the various xtgls options and the number of parameters each estimates like:

    |                        | No AR(1) | Common $\hat{\rho}$ | Separate $\hat{\rho}_t$ |
    |------------------------|----------|----------------------|-------------------------|
    | $\sigma_i^2 = \sigma^2$, $\text{cov}(\sigma_{it}, \sigma_{jt}) = 0$ | $k + 1$  | $k + 2$              | $N + k + 1$             |
    | $\sigma_i^2 \neq \sigma^2$, $\text{cov}(\sigma_{it}, \sigma_{jt}) = 0$ | $N + k$  | $N + k + 1$          | $2N + k$                |
    | $\sigma_i^2 \neq \sigma^2$, $\text{cov}(\sigma_{it}, \sigma_{jt}) \neq 0$ | $\frac{N(N-1)}{2} + N + k$ | $\frac{N(N-1)}{2} + N + k + 1$ | $\frac{N(N-1)}{2} + 2N + k$ |

• **xtpce**: estimates models with Beck and Katz’s (1995) “panel corrected standard errors”:
  
  - `corr` is the same as for xtgls. xtpce will automatically do a Prais Winston regression to deal with temporal correlation if either of the two AR options are specified,
  
  - hetonly corresponds to the panel(heteroscedastic) option in xtgls,
  
  - independent corresponds to the panel(independent) option in xtgls,
  
  - the default is heteroscedastic, spatially correlated panels.

• **xtregar**: estimates fixed and random effects models with AR(1) errors. This is similar to the xtreg command, in that one specifies `fe` or `re` for fixed or random effects models, respectively. The model also gives estimates of fixed or random effects and common $\hat{\rho}$.

• **Useful tests**:
  
  - `xttest2`: used after xtgls or xtregar; implements a Breusch-Pagan test for cross-unit correlation. Uses estimated residuals from the most recent model to test whether the cross-unit correlations are identically zero. The test is distributed $\chi^2$ with $\frac{N(N-1)}{2}$ degrees of freedom.
  
  - `xttest3`: modified Wald statistics that tests the hypothesis that $\sigma_i^2 = \sigma^2$ for all $i$ (homooscedasticity). The statistic is distributed $\chi^2$ with $N$ degrees of freedom.
I. TSCS & Endogeneity

Data which are observed across time and space are referred to as "panel data" or "time-series-cross-section data" or "longitudinal data.

1. Examples: Surveys of the same individuals over some time period (panels).
2. Formally, it makes no difference whether you observe more cross-sections than time-periods (\(n \geq T\) or \(T \leq n\)), as a practical matter, however, it may not be feasible or advisable to estimate the same model in these different circumstances.

C. Still, since formally the same, & this is a bit intro, here's the basic idea: we model

\[ y_{it} = X_{it}' \beta + \epsilon_{it}, \]

where the subscript \(i\) refers to individual \(i\) at time \(t\).

D. So, first thing: If we are willing to assume:

1. \( E(\epsilon_{i}) = 0 \)
2. \( V(\epsilon_{i}) = \sigma^2 I \)
3. \( Cov(X, \epsilon_{i}) = 0 \)
4. \( Rank(X) = k \)

Then: nothing new; it's (W)LRM \( \Rightarrow \) OLS is BLUE.

E. Similarly, if we wish to relax *3 to \( V(\epsilon_{i}) = \sigma^2 \Omega \), nothing changes: it's the (W)LRM \( \Rightarrow \) G-LS is BLUE. But GLS is impossible \( \Rightarrow \) FGLS or "Robust" techniques.

\[ i \Rightarrow F. \] So the only reason there are separate chapters (14 and parts of 15 am 16) on this is because substantively this sort of data tends to raise interesting possibilities for testing & statistically presents interesting challenges for estimation.

1. Substantively interesting: in a nutshell: both cross-sectional and cross-temporal variation can be utilized for comparative leverage. Eq.

a) Effects of Bicameralism: US time-series no help, US states cross-section little help (1 uni-cameral). US TSCS (still little help). Democracies & Democracies over time \( \Rightarrow \) new we have leverage.

b) Unemployment Incidence (or Illness Incidence, Treatment response, whatever): need helps a lot to get leverage across time & space.

c) Institutions in Political Science & Econ.

d) In 1970... or... they don't vary much over time, but some vary across countries/regions, whatever, but so do lots of other things in a covered space.

e) Same help but limited from elsewhere, in relatively fixed manner.

\[ i \Rightarrow iii) \] TSCS is (not much help) (limited help) (reasonable help) (important that this is multiple active
I. F. 2. TSCS Statistically Interesting
   a) Heteroskedasticity + Autocorrelation likely
   b) New: Best: Contemporaneous Correlation also seems likely now
   c) Quasi-Philosophical Issues of "Repeated Sampling"

   - Do we view \( T \to \infty \) as relevant asymptotic dimension?
   - or \( n \to \infty \)? Both? (What exactly does last mean?)
   - Does it matter statistically? (Usually)
   - Really interesting questions of "parameter stability." (These really interesting like b)
   - above b, c. They're so substantively critical.

   Eg. Is effect of some institution(same in all countries)
   - it not, why and how does it change?
   - Does money-growth have the same effect on output everywhere and "everywhere"? Or does it vary?
   - Latter question more interesting if we ask "and how?" e.g. more nominal vs. more effect? Pre-Robert Lucas more effect than post?
   - Does drug have same effect regardless of behavior patterns of individuals?

II. TSCS "Methods Intro."

A. First thing is to look at:

\[
V(\varepsilon) = \begin{bmatrix}
  a_1^2 & a_1 a_2 & a_1 a_3 & \cdots & a_1 a_{n-1} \\
  a_2 a_1 & a_2^2 & a_2 a_3 & \cdots & a_2 a_{n-1} \\
  a_3 a_1 & a_3 a_2 & a_3^2 & \cdots & a_3 a_{n-1} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  a_{n-1} a_1 & a_{n-1} a_2 & a_{n-1} a_3 & \cdots & a_{n-1}^2
\end{bmatrix}
\]

- B1. Reasonable, perhaps to suggest 1) different variance for each \( c \)
- B2) different \( AR(2) \) for each \( c \) (1st order covariance)
- \( \ell_c \)

\[
\begin{align*}
\sigma^2_{\varepsilon} &= \frac{1}{2} n(n-1) \times \text{the } (0, \sigma^2) \text{ parameters,} \\
\end{align*}
\]

and \( n \) of the \( \sigma^2 \) parameters \( \Rightarrow \text{might be reasonable if } T \gg N \)

\[
\text{so then } \# \text{ obs } \frac{N}{2} \to 1 \times \frac{1}{2} n(n-1) + 2n
\]
It's the substance of the process by which you believe your data were generated that has to guide you. The model of the past, not the description of the data at hand, underlies the whole endeavor. Finally, you could think to allow a different $\rho$ for different times and different places, different variances too, maybe covariance too. At one level, perhaps this seems reasonable, but after all, both times and places differ, but this puts us all the way back to $(nT-1) + nT + \frac{1}{2}nT(nT-1) + 3nT$ parameters, or covariance parameters $\Rightarrow$ Can't infer anything $\Rightarrow$ If you want to infer anything anyhow from history, you must be making some assumptions (explicit or implicit) about the process generating your data.

5. Last Words: Remember that our first line of defense is to model the process directly.

Unemployment, $y_i = \sum (y_{i,t} - \bar{y}) \Rightarrow$ why not include GDP abroad in stuff? (i.e., contemporaneous may be reflecting that economic conditions in $i$ depend in part on conditions in $j$. So perhaps:

$$\sum_{j\neq i} \left(\frac{\text{Bilateral Trade } i,j}{\text{Total Trade } i}\right) \cdot (\text{GDP growth } j)$$

belongs in your equation? (There's a slight endogeneity problem here, but for reasonable $n/(>10?\text{ maybe?)}, I\text{ think it's a big deal})$

Otherwise, nothing really new is introduced by TSCS as regards assumption $3$.

C. Constant Parameters in TSCS:

1. Do you think the intercept ($\beta_0$) is same for all $i$? What about for all $t$? E.g., might their be some fixed (or random) difference across your $i$'s that you haven't accounted for in your $X$'s?
2. If so, some options
   a) Best Option: figure out what you’re missing a include it
   b) Maybe allow for a fixed difference across units by putting a dummy in for each (avoiding perfect collinearity though).

   
   ⇒ This is depressingly theoretical though. Try for (A) as much as possible.

   c) There are times when (b) might be excessive (esp. when \( n > T \) under which condition, (b) calls for many new variables, putting serious strain on deg. free (\( nT - k \)).)

   ⇒ thing called “random effects”. Basically, rather than treat each cross sect. unit as having some fixed difference, (ie. fixed across repeated samples) to be estimated, assume the differences you observe in your sample came from some distribution. Rather than estimate n fixed effects, try to estimate the parameters of the distribution from which the observed effects may have come. E.g., if they came from a normal, only one parameter: \( \mu, \sigma^2 \).

3. Ditto to the concerns regarding constant \( \beta \) for the other coefficients:
   a) Best Option: Model why the coefficients change (e.g., estimating Unemployment = \( \beta_0 + \beta_1 \text{Money} + \beta_2 \text{Staff} \))

   we considered that \( \beta_i \) might vary by degree of nominal contracting in the economy ⇒ measure that and estimate the appropriate interactive model

   \[ \text{UE} = \beta_0 + \beta_1 \text{Money} + \beta_2 \text{NomCont} + \beta_3 \text{NomCont} \times \text{Staff} \]

   b) Let the coefficients vary by groups
   c) in a “fixed” way as above
   d) in a “random coefficient” way as above

Conclude with a graph on why different patterns are important to consider.
I. Cleaning Up on TSCS

A. At one level, there's nothing new here:
   1. We still model \( Y = X\beta + \epsilon \)
      - \( \text{E}(\epsilon) = 0 \)
      - \( \text{V}(\epsilon) = \sigma^2 I \) or possibly \( \text{V}(\epsilon) = \sigma^2 \Omega \)
      - \( \text{Cor}(X, \epsilon) = 0 \)
      - \( \text{Rank}(X) = k \) (Normality)

   ⇒ Nothing intrinsically new in allowing this model to apply to observations
   across space (i.e., individuals, countries, states, etc.) and time (years months).

2. What's new perhaps is the plausibility of some these assumptions
   - \( \beta \) constant across all \( i \) \& \( t \)? That's implied by \( Y_{it} = X_{it} \beta + \epsilon_{it} \)
   - More plausible sets of restrictions on the structure of \( \Omega \).
     In particular: the plausibility of group-wise heterogeneous group-specific autocorrelation & contemporaneous correlation across groups.

   ⇒ In both cases, our first line of action is to model these interesting possibilities:
     \( Y_{it} = \alpha_i + X_{it} \beta + \epsilon_{it} \) ⇒ different intercept for each cross-section = a fixed diff.

     - \( \alpha_i \) can be treated as fixed ("fixed effects") in the level of \( Y \) by group controlling for \( X \)

   ⇒ What is called the "within estimator" or "Least-Squares Dummy Variable" (LSDV) model) across repeated samples if you drew another sample individual \( i \) would still have the same \( \alpha_i \)

   ⇒ To estimate \( \beta \) across group variation in \( \alpha_i \) is netted out by the dummies.

   1. \( \alpha_i \) can be estimated as "random", not fixed across repeated samples in which case you estimate the parameters of the distribution from which the \( \alpha_i \) are drawn rather than as fixed effects (this is the random effects estimator).

   ⇒ Again \( \alpha_i \) can be taken as fixed or as random.

   2. \( Y_{it} = \alpha_i + \alpha_t + X_{it} \beta + \epsilon_{it} \) ⇒ different intercept for each time-period

      - \( \alpha_i \) is a fixed difference in level of \( Y \)
      - \( \alpha_t \) controlling for \( X \) for each time-period

   ⇒ Usually this will be too general \((N + T + N_i T_i \text{ parameters})\) to estimate. Terribly painful to recognize it as one very general starting point perhaps.

   ⇒ Again differences in \( \beta \) can be modeled as fixed or random. Which?

   ⇒ Always: we are usually best advised to model differences in \( \beta \) by interactions.

   e.g. \( \text{App}_{it} = \beta_0 + \beta_1 \cdot \text{UE}_{it} + \beta_2 \cdot \text{Inf}_{it} + \epsilon_{it} \) → May argue \( \beta_1 \& \beta_2 \) not same for every president? Why? Maybe Democrats suffer more approval loss for \( UE \& \text{Rep's} \) for \( \text{Inf} \)? Or v.v.? ⇒ Dummy for party & dummy \( UE \& Dummy \cdot \text{Inf} \).
B. On the importance of examining the constant $\beta$ assumption:

1. Is it safe to "pool" observations across your $n$ units, $i$, over $T$ time-periods, $t$?

   a) What can happen if $Y_{it} = \alpha_{i} + \beta_{i}X_{it} + \epsilon_{it}$ does not hold for all units $i$, but rather the correct model is $Y_{it} = \alpha_{i} + \beta_{i}X_{it} + \epsilon_{it}$?

   b) Simple OLS produces estimates of $b_{i}$, which are a weighted average of $b_{\text{within}}$ and $b_{\text{between}}$, i.e.:

   (i) $b_{\text{within}}$ is: $Y_{it} - \bar{Y}_{i}$ regressed on $X_{it} - \bar{X}_{i}$; i.e. the group means are dropped out before regressing (including a dummy for each group achieves the same thing).

   (ii) $b_{\text{between}}$ is: $Y_{i}$ regressed on $X_{i}$; i.e. the $n$ group means of $Y$ are regressed on the corresponding $n$ group means of $X$.

   (iii) You could use only variation within groups over time (i.e., $Y_{it}$ regressed on $X_{it}$); or only variation across groups, averaging over time. (This is two different estimates: one based on a time series of $NT$ observations, one based on a cross-section of $N$ observations. Simple OLS takes a weighted average of these. Is this OK? If the observations truly follow the same model, it's not only OK but it's efficient.

2. How could (should) you check?

   a) Run $T$ cross-section regressions of your model $Y_{i} = \alpha_{i} + \beta_{i}X_{i} + \epsilon_{i}$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

   b) Does $\beta$ bounce around a lot (relative to its s.e.'s)? If so, your model is not specified well enough to hold across your whole TSCS sample.

D. (a) May not always be possible: how about checking your model with $N$ dummies by group (fixed effects by group) and $T$ dummies by time-period (fixed effects by time-period). (That's 2 regressions. Are $\beta$ estimates stable? (compared to s.e. (b)))

   b) May also be impossible sometimes—e.g., have some variables which do not vary over time (e.g., perf. colin. with group dummies) or do not vary across groups (e.g., perf. colin. with time dummies).

   $\Rightarrow$ Cross-Validation: (one-out sensitivity analysis) re-estimate $N$ times, each time omitting a group. Is $\beta$ very sensitive?

   (This is like a DFBeta)
There is tons more to be said about TSCS data, but I’ll close with an example of some of the options outlined. Suppose you think public debt increases as unemployment rises, but that, controlling for this, there may well be differences in the level of debt across countries. In fact, in the way debt responds to unemployment, i.e., you think:

$$D_{it} = \alpha_i + \beta_i \cdot UE_{it} + \epsilon_{it}$$

Now, further suppose that what you have in mind is that bicameral polities have different debt levels, controlling for UE, than unicameral polities, and that debt responds differently to UE in bicameral than unicameral polities.

1. Option 1: a) Estimate $$D_{it} = \alpha_i + b \cdot UE_{it} + \epsilon_{it}$$

that is, each country is allowed to have a different intercept and slope. Easiest way is just to estimate each country separately.

$$N$$ estimates each of $$\alpha$$ and $$b$$

b) Regress the estimated $$\alpha_i$$ on a variable measuring bicameralism

(Alternatively you can regress the estimated $$\beta_i$$ on the same variable)

N.b. These should use WLS, why? Because each $$\alpha_i$$ & each $$\beta_i$$ have a different (estimated) variance. They are coefficient estimates having a s.e. ($$\hat{\alpha}_i$$) and s.e. ($$\hat{\beta}_i$$) associated with them, should weight by [s.e.].

2. Option 2: Let $$B_i = \{1 \text{ if } i \text{ is bicameral}\}$$

Regress $$D_{it} = \beta_0 + \beta_1 \cdot B_i + \beta_2 \cdot UE_{it} + \beta_3 \cdot B_i \cdot UE_{it} + \epsilon_{it}$$

⇒ The heterogeneous coefficients ($$\alpha_i$$ & $$\beta_i$$) are modelled directly here as an interactive effect in the pooled model.

N.b. Cannot put group dummies in this model bc they will be perfectly collinear with $$B_i$$. 
II. Endogeneity (Simultaneity, Reverse Causality, Bi-Directional Causality)

A. In social science, the proposition that Y causes X, X causes Y is absolutely ubiquitous. At some level it may be said, perhaps, that everything in some degree causes everything else.

E.g. 1) Econ: Quantity Supplied $\leftrightarrow$ Quantity Demanded (in fact, in equilibrium, $Q_S = Q_D$).
    - via prices
    - Real Wage $\leftrightarrow$ Inflation
    - via expectations & directly
    - Consumption $\leftrightarrow$ Income
      - via macro-economic identities $Y = C + I + G + X - M$
      - etc., etc., etc.

2) Poly Sci: Candidate Spending $\leftrightarrow$ Election Outcome
   - via expectations (candidates spend more if expect close election)
   - Constituency Interest $\leftrightarrow$ Policy
     (more broadly, social change structure)
   - Military Spending $\leftrightarrow$ Threat of War
     (more generally, any spending & thing it's targeted at)
   - etc., etc., etc.

3) Other Fields & Combinations of Fields:
   - Exercise $\leftrightarrow$ Health
   - Income Inequality $\leftrightarrow$ Income Growth
   - Immigration $\leftrightarrow$ (Econ & Soc. Conditions at Home)
     & (Emigration)
   - etc., etc., etc.

B. Most Simply & Fairly Generally, it is quite common & reasonable in a great many contexts to suppose:

\[ Y = f(X, \ast) \]  
\[ X = g(Y, \ast) \]

\( Y \) is a function of \( X \) & possibly other stuff
\( X \) is a function of \( Y \) & possibly other stuff

1. First point is that without more information / assumptions / restrictions, this is indeterminate. Nothing systematic can be said. E.g., just to illustrate, suppose f\( (\cdot) \) g\( (\cdot) \) are linear in \( X \& Y \) only:

\[ Y = a \cdot X \]
\[ X = b \cdot Y \]

\[ \Rightarrow Y = a \cdot b \cdot Y \Rightarrow a \cdot b = 1 \Rightarrow a = \frac{1}{b} \]

\[ \Rightarrow \text{any } a \& b \text{ such that } a = \frac{1}{b} \text{ solves this system} \]

\[ \Rightarrow \text{any } a \& b \text{ such that } a = \frac{1}{b} \text{ are equally indistinguishable} \]

\[ \text{effects of } X \text{ on } Y \text{ and } Y \text{ on } X: \]

\[ \text{e.g. } (a = 1, b = 1), \ (a = 5 \times 10^3, b = \frac{1}{5000}) \]

\[ \text{etc., etc., etc.} \]

\[ \Rightarrow \text{If we want to make any arguments regarding systematic relationship between } X \& Y \text{ we must bring more information to bear} \]
II.B.2 A graphical illustration of the endogeneity problem:

Suppose this represents the market for apples.

\( Q^*(p, \alpha) \) is the supply of apples which is, of course, a function of the price sellers could get for an apple, and maybe some other stuff too.

\( Q^d(p, \alpha) \) is the demand, a function of what buyers must pay and maybe other stuff.

\[ \Rightarrow \text{Problem is we only observe } (p^*, q^*) \text{ the amount that actually sold at a price at which it sold. An infinite # of } Q^d \text{ and } Q^s \text{ curves might have intersected at that point. So, too, if we see } (p^*, q^*) \text{ moving over time, an infinite # of } \]

\( Q^d \text{ or } Q^s \text{ shifts could have accounted for that move. How then can we learn anything about } Q^d \text{ or } Q^s \text{?} \]

Another Example: Cops & Crime

\[ \text{Cops (Crime, } \alpha \text{)} \]

\[ \text{Cops (Crime, } \alpha \text{)} \]

\[ \text{Crime (Cops, } \alpha \text{)} \]

\[ \text{Crime (Cops, } \alpha \text{)} \]

\[ \Rightarrow \text{We only observe } (C^*, C^*) \text{; how can we find out the effectiveness of cops as crime reducers? How about the response rate of cop-hiring to crime?} \]

3. The statistical manifestation of the endogeneity problem:

a) \( Y = X_\beta + e + \varepsilon \)

\( \Rightarrow \text{Cov}(X, \varepsilon) = \text{Cov}(Y_\beta + Z\gamma + E_X, E_Y) \)

b) \( X = \gamma_\alpha + Z\omega + E_X \)

\[ \Rightarrow \text{Y contains } E_Y \Rightarrow \text{this not O}\]

\( \Rightarrow \text{Cov}(X, \varepsilon) = 0 \text{ violated} \)

\[ \text{Cov}(Y, E_X) = \text{Cov}(X_\beta + Z_\gamma + E_Y, E_X) \neq 0 \]

\[ \Rightarrow \text{same problem} \]

c) So, endogeneity (simultaneity) \( \Rightarrow \text{Cov}(X, e) \neq 0 \Rightarrow \text{OLS is...} \)

i) biased

ii) inconsistent (and inefficient)

Std. Errs also biased and inconsistent.

d) How bad is it? Loosely, it depends on \( \alpha \) relative to \( \beta \) (i.e. true \( \alpha \) relative to true \( \beta \)).
II. B. 3.c. Estimate \( Y = \beta X + \epsilon \) by OLS, how bad is it?
\[ X = \alpha Y + \omega + \epsilon \]

Think as (i) Regress \( Y \) on \( X \neq Z \): bias & inconsistency, loosely speaking, is proportional to \( 1/\|\alpha\|/\|\beta\| \) (for \( X \& Y \) in same scale)

(ii) Regress \( X \) on \( Y \& Z \): bias & inconsistency roughly proportional to \( 1/\|\beta\|/1/\|\alpha\| \) (for \( X \& Y \) in same scale)

\( \Rightarrow \) Very loosely: the more important the omitted causal direction relative to the included one, the worse the bias.

II. C. What can be done about it?

1. The Graphical Intuition: Let’s go back to supply & demand for apples.

\[ Q^s(p, \cdot) \]

\[ Q^d(p, \cdot) \]

\[ Q^s(p, \text{weather}, \cdot) \]

\[ Q^d(p, \text{weather}, \cdot) \]

\( \Rightarrow \) With enough variation in \( Q^s \) explained by variations in weather, can trace out \( Q^d \) curve well.

\( \Rightarrow \) Similar, we need something in demand equation which is not in supply equation to estimate the supply curve. E.g., reports on the medical benefits of apples.

\[ Q^s(p, \cdot) \]

\[ Q^d(p, \text{report}, \cdot) \]

\( \Rightarrow \) What we want to do is find stuff that affects \( Y \) which affects \( X \) only in so far as it affects \( Y \& \) stuff that affects \( X \) only in so far as it affects \( X \).

N.b. Actually, this is somewhat more than we need (it’s sufficient, not necessary). Strictly what we need is some way to tie down how much of movements in \( Y \) associated with some factor \( & \) how much of movements in \( X \) associated with that factor. These all or nothing restraints are the easiest to implement & by far the most common in the literature.
II. C; 8. What can we do about it, the mathematical exposition:

\[ Y = f(X, Z, \varepsilon) \]

what we need is something in \( Z \)
\[ X = g(Y, Z, \varepsilon) \]

which is not also in \( Z \xrightarrow{\text{Crime Example}} \)

\[ \# \text{Cops on Street} = f(\text{Crime}, \text{Mayor's Party}, \text{Election Year}, \text{Economic Conditions}, \ldots) \]
\[ \text{Crime} = g(\text{Cops}, \text{Punishment}, \text{Economic Conditions}, \text{Age Structure of Population}, \ldots) \]

Arguments:
1. Cops are hired to keep crime down: \( \frac{\partial \text{Cops}}{\partial \text{Crime}} > 0 \)
2. Republicans hire more cops: \( \frac{\partial \text{Cops}}{\partial \text{Rep}} > 0 \)
3. Mayor and Governors hire more cops around election years: \( \frac{\partial \text{Cops}}{\partial \text{Election Year}} > 0 \)
4. Nog's hire more cops when economic conditions worse (as a jobs program): \( \frac{\partial \text{Cops}}{\partial \text{Economy}} < 0 \)
5. Criminals do less crime when more cops around: \( \frac{\partial \text{Crime}}{\partial \text{Cops}} < 0 \)
6. Criminals do less crime when punishment more severe: \( \frac{\partial \text{Crime}}{\partial \text{Punishment}} < 0 \)
7. Criminals do more crime when e.cond. bad: \( \frac{\partial \text{Crime}}{\partial \text{Economy}} < 0 \)
8. Youths do more crime: \( \frac{\partial \text{Crime}}{\partial \text{Youth}} > 0 \)

A system of equations is identified if there is at least 1 variable per equation which is not in the other equations (there are other ways to identify a system). Here, \( Z \) = \{Mayor's Party, Gov's Party, Election Year, Ec. Conditions\}.

These factors help identify problems. This does not help.

\( Z \) = \{Punishment, Age Structure, Ec. Conditions\}.

A system is just identified if there are \((q-1)\) unique variables in \( Z \).

It is "overidentified" if \( q \) or more, where \( q \) is the number of equations.

3. What to do about it, the statistical exposition:

Problem is \( \text{Cov}(X, \varepsilon) \neq 0 \)

\[ \Rightarrow \text{find some } Z \text{ (or set of } Z \text{'s) such that } \lim_{n \to \infty} \text{Cov}(Z, \varepsilon) = 0 \text{ and yet } \lim_{n \to \infty} \text{Cov}(Z, X) \neq 0 \]

Such a \( Z \) is called an "instrumental variable" on IV.

The statistical presentation is probably the least intuitive at first, however, it is also the most general in the sense that any "problem" which amounts to a \( \text{Cov}(X, \varepsilon) \neq 0 \) problem has the same IV "solution."

E.g. Measurement Error: \( X = X^* + U \Rightarrow \text{Cov}(X, \varepsilon) \neq 0 \)

\[ \Rightarrow \text{find } Z \text{ such that } \lim_{n \to \infty} ZU = 0 \text{ but } \lim_{n \to \infty} Z^* = 0 \]

* Lagged \( Y \) with AR in \( \varepsilon \) \( \Rightarrow \text{Cov}(X, \varepsilon) \neq 0 \Rightarrow \text{find } Z \text{ such that...} \]

* Random Effects with \( X \) variables that don't vary over time

\[ \Rightarrow \text{find } Z \text{ such that } \lim_{n \to \infty} Z^* = 0 \text{ but } \lim_{n \to \infty} ZX \neq 0 \]
Tying Down the Endogenous Systems: A Loose Fit

**Equation One**

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
\vdots \\
Y_M
\end{bmatrix} =
\begin{bmatrix}
\beta_1 & \beta_2 & \beta_3 & \cdots & \beta_M \\
\beta_1 & \beta_2 & \beta_3 & \cdots & \beta_M \\
\beta_1 & \beta_2 & \beta_3 & \cdots & \beta_M \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\beta_1 & \beta_2 & \beta_3 & \cdots & \beta_M
\end{bmatrix}
\begin{bmatrix}
X_1 \\
X_2 \\
X_3 \\
\vdots \\
X_M
\end{bmatrix}
\]

**Equation Two**

\[
\begin{bmatrix}
E_1 \\
E_2 \\
E_3 \\
\vdots \\
E_M
\end{bmatrix}
\]

**Equation M**

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
\vdots \\
Y_M
\end{bmatrix} =
\begin{bmatrix}
\delta_1 & \delta_2 & \delta_3 & \cdots & \delta_M \\
\delta_1 & \delta_2 & \delta_3 & \cdots & \delta_M \\
\delta_1 & \delta_2 & \delta_3 & \cdots & \delta_M \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\delta_1 & \delta_2 & \delta_3 & \cdots & \delta_M
\end{bmatrix}
\begin{bmatrix}
X_1 \ (n) \\
X_2 \ (n) \\
X_3 \ (n) \\
\vdots \\
X_M \ (n)
\end{bmatrix}
\]

- Coefficients on all the Y's in all the equations
- Coefficients on all k of the X's in each of the M equations
- A 6th observation on all the Y's
- 6th obs on the k endog. vars X_1, X_2

**A:** In Matrix Algebra, this looks like:

\[
Y_i = \Gamma + X_i' B = E
\]

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_M
\end{bmatrix} =
\begin{bmatrix}
\Gamma \\
\chi \\
\vdots \\
\chi
\end{bmatrix} +
\begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_M
\end{bmatrix} +
\begin{bmatrix}
E_1 \\
E_2 \\
\vdots \\
E_M
\end{bmatrix}
\]

Define \( V(E) = \sum_{i=1}^{M} (X_i' B)^2 \)

\( V(E) \) is a symmetric, positive definite matrix.

**B:** If we simply regress each of the endogenous variables \( Y \) on each of the independent variables \( X \), we can find what's called the "reduced form":

\[
Y = X \Pi + \Theta
\]

where \( \Pi \) is a \((k \times M)\) reduced-form coefficient matrix \( \Pi \) (these are a "blob" of all the ways the \( \Pi_i \) of \( B \) interact through)

\( \Theta \) is a \((M \times M)\) reduced-form covariance matrix of residuals \( \Theta \) (a "blob" of all the ways the errors vary and relate across equations)

**C:** In the full model, \( A \), there are:

- \( \Pi \) is \((k \times M)\) reduced-form coefficient matrix
- \( B \) is \((k \times M)\) reduced-form parameter estimates
- \( \Sigma \) is \((M \times M)\) symmetric
- \( \Sigma \) is \((M \times M)\) symmetric \( \Sigma \) parameters in \( \Sigma \) to estimate
- \( \Pi \) is \((k \times M)\) reduced-form coefficient matrix
- \( \Pi \) is \((k \times M)\) reduced-form coefficient matrix
- \( \Omega \) is \((M \times M)\) symmetric
- \( \Omega \) is \((M \times M)\) symmetric

\[
\begin{bmatrix}
M^2 + KM + \frac{1}{2} M(M+1) \\
M^2 + KM + \frac{1}{2} M(M+1)
\end{bmatrix} = \begin{bmatrix}
M^2 \\
M^2
\end{bmatrix}
\]

Where do we get the extra leverage to tie down the endogenous system?
where does the extra info come from to tie down the remaining

1) Notice in \( I^2 \) that the diagonal elements are the coefficients
on \( y_t \) in the \( i \)-th equation. This amounts to an arbitrary scaling
coefficient really. I.e. usually we just put 1 on the left hand
side and only other things on the right. This is same as
setting all \( y_t=1 \), perfectly legitimate -- in fact any constant
is legit, the 1 is just easiest for interpretation.

\[ \Rightarrow M \] parameters tied down, \( M^2-M \) or \( M(M-1) \) left

2) Identities: If you happen to be so lucky as to know some
things are true for certain \( y_t \) as accounting or other identities,
then you can go right ahead and impose these facts rather
then attempt to estimate them. General rule: 1 fact = 1 piece
of info = 1 parameter identified.

3) Exclusions: This is what we were talking about before --
find something you can leave out of some equations (i.e.
exclude or equivalently, force to have coefficient zero)
General Rule: every parameter of \( B_t \) or \( F_t \) you force to be
zero (by excluding that variable from that equation) ties
down another parameter.

4) Linear Restrictions: This is just the more general version
of (3): Any coefficient you can fix to anything (zero
or otherwise) is one less thing to estimate \( \Rightarrow \) one
more parameter identified.

5) Non-linearities: This is very complicated, but knowing how
exactly the variables \( y_t \) or parameters enter the model I is obviously
more info & accordingly can be used if you know how to
help identify the system. (Simple example: for one \( y_t \), \( y_t=f(x_t, x_{t-1}) \)
but for all others, only \( x_t \) not \( x_{t-1} \) matters \( \Rightarrow \) just like any other
exclusion then)

6) Restrictions on the Residual Covariance Matrix: Again this is
complicated, but broadly any further info is useful in identifying.
A common use of these are Vector Auto-Regression (VAR)
techniques which force a sort of triangular \( \Sigma \) thereby reducing the
unique parameters in \( \Sigma \) to considerably less than \( E(M+1) \).

Overwhelmingly, the usual source of additional info (assumptions) are
exclusions. (variables in some \( y_t \) equations are left out of all \( y_t \) equations).
Second most common are VAR & related techniques.

III.E. A Word on VAR's

- These amount to Regressing every \( y_t \) on it's own lags & lags
  of the other variables.

- Finding exogenous \( x_t \)'s, or enough exogenous \( x_t \)'s, is
given up upon a rather some restriction of the error-covariance
matrix is assumed.

\[ \Rightarrow \] The coefficients in VAR are not very substantively meaningful, but
the "response" of each \( y_t \) over time to "unexplainable shocks" to
other variables (unexplainable by all the lags that is) can be traced in what’s called an “impulse-response” function.

1. My view: these are useful descriptive devices, suggestive often of interesting empirical relationships; however, the neat “magic” of identifying by obscure assumptions regarding the error-variance structure is usually just that: “magic” in the sense of “illusion.”

2. You should probably know what Granger Causality is (related to VAR techniques).

Basically: \[ Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_L Y_{t-L} + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_L X_{t-L} + \epsilon_t \]

* If \( \alpha_1 = \alpha_2 = \ldots = \alpha_L = 0 \), then \( X \) does not granger-cause \( Y \).

Thus, Granger-Cause means simply that controlling for previous levels of \( Y \), previous levels of \( X \) help predict \( Y \).

\[ \Rightarrow \text{This is not “causality” then at all but just another kind of correlation.} \]

* One Problem: this assumes that nothing can cause anything that predates it. This is wrong if any of the outcomes being considered involve (presumably human) actors who are forward looking.

* Another Problem: it remains the case that this is merely correlation of \( X \)'s with \( Y \)'s, it still says nothing about causality.

\[ \Rightarrow \text{General Rule: Causality cannot be “proven” or “tested for” -- closest we can come is to assume something is causal, then see if other things are too.} \]

Corollary: Endogeneity & Exogeneity likewise cannot be proven or tested per se. So-called “tests” are based on assuming some identifying restrictions then testing more of them.

\[ \Rightarrow \text{Problem is rejection does not actually leave you able to say it’s the extra restrictions that are problematic.} \]

Likewise, Acceptance could be because initial identifying assumptions are wrong.

\[ \Rightarrow \text{Identification / Endogeneity / Exogeneity / Causality: these are things that stand or fall on the basis of the arguments supporting them, which evidence is then consistent with or not consistent with. It is important, then, first to know what assumptions/arguments/restrictions you are making, second what these imply substantively, and third what evidence would then be consistent or inconsistent with them.} \]

OK, enough preamble, on to the usual practice of identifying and estimating systems of equations: Two-Stage Least-Squares (TSLS or 2SLS).
A. First write down the equation(s) you want to estimate:

\[ \text{Crime} = f(\text{#Cops, Economic Conditions, Punishment, Age, Structure, } E) \]

B. Decide which are endogenous and which exogenous:

Endogenous:
\[ \text{Crime} \rightarrow \# \text{Cops, Economic Conditions, Severity of Punishments} \]

Exogenous:
\[ \text{Crime} \rightarrow \text{Age, Structure} \]

C. Think about some things that would cause the endo. factors but

would not cause \( y \) (crime) except through these factors. You

need at least one unique such factor per endogenous right-hand

side variable.

\( \# \text{Cops} \): Mayor's Partisanship, Gov's Partisanship, Election-Year

(last year's #cops?)

\( \text{Econ. Conditions} \): 1st be more specific -- say if what

causes crime is specifically Unemployment, inflation, & income distribution.

\( \text{Severity of Punishment} \): Partisan Elections, (last year's

severity? same caveat?)

D. Now, gather your exogenous factors & call them instruments, \( Z \)

\[ \text{Crime}_t = \beta_0 + \beta_1 \text{#Cops}_t + \beta_2 \text{Econ. Conditions}_t + \beta_3 \text{Severity}_t \\
   + \beta_4 \text{Age}_t + \epsilon_t \]

Exogenous Factors (potential instruments):

\( Z = \{ \text{Mayor's Party}_t, \text{Gov's Party}_t, \text{Election-year}_t, \text{Cops, Severity}_t, \text{Econ. Conditions}_t \} \)

E. Make sure \( # IV's \geq # \text{endog. factors} \)

\( \Rightarrow \) "Order" condition for identification -- usually all we check.

F. Estimate Instrumental Variables by 2SLS (over)
6. Two-Stage Least-Squares Estimation

- Let \( X \) be the matrix of right-hand-side variables, endog. exog.
- Let \( Z \) be the matrix of instruments plus any exog. variables from \( X \)
- Then regress \( X \) on \( Z \) and use predicted values to estimate \( y \).
  
  i.e. \( N_X X \Rightarrow \hat{X} \)  
  \( (\hat{X}'\hat{X})^{-1}\hat{X}'y \Rightarrow b_{2SLS} \)
  
  use \( y - Xb_{2SLS} = e_{2SLS} \)

More slowly now:

1. Regress each endogenous variable in \( X \) on all the exogenous variables (those in \( Z \) and in \( X \)). Call the fitted values \( \hat{X} \).

2. Regress \( y \) on \( \hat{X} \) (all the exogenous variables in \( \hat{X} \)).

3. Use \( y - Xb_{2SLS} = e_{2SLS} \) for your residual calculations (such as \( V(b_{2SLS}) \)).

(under the identifying assumptions being true), \( b_{2SLS} \) is unbiased, consistent, efficient (amongst those using this \( Z \) matrix of info).

In our example:

1. Regress \( \hat{X} \) on \( \hat{X} \) (note that \( \hat{X} \) includes the variable age as a column of \( \hat{X} \)).

2. Regress each exogenous conditions, severity of punishment on \( \{ \text{Mayor's Party, Gov's Party, Elect Year, UE national, Infl national, Infl Dist ntl, Cops t-1, Severity t-1, Age} \} \)

3. Calculate: \( \hat{\sigma}^2 (\hat{X}'\hat{X})^{-1} \) \( V(e)=\hat{\sigma}^2I \)

Just like the measurement error case, the story becomes depressing if the identifying assumptions are false. Results are trustworthy here to degree exogeneity assumptions are true just as there it was to degree data measurement was "true".
I. Introduction & the Linear Model

A. Herebefore, we have been considering dependent variables that are...

1. Continuous: all values in a range (preferably -∞ to +∞ but we had some transformations for >0 only; for example)

2. Interval: the spaces between the possible values of \( x \) are meaningful: e.g., \( x = 2 \) really is twice \( x = 1 \) and \( x = 3 \) is as far from \( x = 4 \) as \( x = 2 \) is from \( x = 1 \), etc.

3. Limited Discrete Variable: \( x \) is 0, 1, or 2 and \( x = 2 \) is not twice \( x = 1 \).

B. But there are plenty of other modes in which the social world provides us data! More importantly, there are plenty of other types of variables about which we are interested in constructing and empirically evaluating theory.

Tools: 1. Nominal: the values the variable takes on only signify membership in some qualitative category (party, sex, race, employment, unemployed, single, married, etc.)

2. Ordinal: the values the variable takes on signify the ranking of the possible outcomes, but the distances between adjacent categories is not meaningful (rankings)

3. Count: the values are discrete (not continuous), ordered, and interval since they are counts of event occurrences
I.B. Limited & Discrete Variables

1. Limited Variables:
   a) Truncated: Values above or below some point do not occur or are not observed.
   b) Censored: Like truncated, except that values above or below some point all appear as a single outcome (usually that boundary point).
   c) Duration: Observations on variable recording elapsed time during which some condition is present (absent).

I.C. Dichotomous, Nominal Variables - The Linear Model

1. We've already seen those of course - on the right hand side, question is, what can and should be done if the dependent variable takes on the values {0 or 1} only?
   Option 1: "What the heck, OLS works pretty well in a bunch of situations as an approximation, why not just use it here?"

2. Graphical Look: 
   \[ Y = \begin{cases} 1 & \text{if } x > x_0, \text{ or } X \geq x_0 \\ 0 & \text{if } x \leq x_0 \end{cases} \]
   C. We could fit a line to this.
   d) If the data really were generated available as more differentiated than a binary indicator, then it would be wasteful to dichotomize to this. Notice there just ain't a whole lot being revealed by these sets of \((x, y)\) as opposed to what would have been in a more typical scatter plot (i.e., the left there). Did you notice that it would have been a little harder to eyeball where the line goes in the graph above relative to one at left? Kinda indicates you need more (0, 1) obs. to get as much info as continuous (typical graph).

3. So we could fit a line. Suggests OLS:
   a) Problem 1: especially if phenomenon truly is nominal (i.e., 0-1).
      \[ \text{OLS makes lots of nonsense predictions: } \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7, \gamma_8 \]
      Interpret the line as hitting the underlying probability that \(Y = 1 \Rightarrow \gamma = P(Y = 1) = p \)
      \[ Y \text{ is a Bernoulli random variable } \]
      \[ Y = \begin{cases} 1 \text{ with prob. } p \\ 0 \text{ with prob. } 1-p \end{cases} \]
      \[ \Rightarrow Y \text{ has mean } p, \text{ variance } p(1-p) \]
   b) Problem 2: \( Y \) is a Bernoulli random variable
      \[ Y = \begin{cases} 1 \text{ with prob. } p \\ 0 \text{ with prob. } 1-p \end{cases} \]
      \[ \Rightarrow \text{ heteroskedasticity} \]
   c) Problem 3: Is it reasonable to assume that the relationship between \(Y\) and \(X\) is linear? Don't you think that far \(p\) very near 0 or very near 1 further changes in \(X\) should have little effect? (This is related to prob. 1).
Problem 4 (with OLS applied to binary data): Residuals are not normal. $E = Y - X\beta = \{1 - X\beta\}$ \Rightarrow bounded, specifically related to $X\beta$, etc \Rightarrow not normal.

On the other hand: $E* = Y* - X\beta - P - X\beta \Rightarrow$ could be normal, perhaps except still bounded.

That's back to the original hand:

$I.D.$ Linear Regression is easy if we're familiar with it, so what if we use it anyway?

$\Rightarrow$ Linear Probability Model (LPM)

$Y = X\beta + E$, $\hat{Y}$ is presumed to estimate $\hat{y}^* = \hat{p} = \text{prob}(Y = 1)$

Given that we know $V(Y) = p(1-p)$, we should use WLS here.

- estimate $\hat{Y} = X\hat{\beta} + \hat{E}$ by OLS, save residual
- generate weighting vectors: $W = \frac{1}{\hat{p}(1 - \hat{p})}$
- estimate $\hat{Y} = X\hat{\beta} + \hat{E}$ by WLS, using $W$ as the weight

$\implies$ LPM: unbiased estimate of the parameters of this (linear probability) model

- consistent
- efficient among linear unbiased estimates

2. I.e., the weighted LPM is BLUE, so what's the big deal?

a) Still get non-sense $\hat{\beta}$ predictions (e.g. by the way, $\hat{\beta} \geq 1$ and $\hat{\beta} \leq 0$ are gonna cause problems in step 6 above too).

- but so long as you're not too concerned about $\hat{\beta}$ per se, but rather care about $\hat{p}$ this ain't that big a deal.

b) Its only BLUE among linear probability models if it's just not terribly defensible that probability is linear.

$\Rightarrow$ Real story is that WLS is still B.L.U.E. Approx. Truth (BLUAT), but in this case it's likely that "best" is pretty crummy especially out away from $p = 0.5$.

Visual Aid:

- WLS of the LPM is most reasonable for $X$ near $X$ (center) and $Y$'s pretty evenly 0-1 mixed.
- can start to get really poor as an approx. pretty quick as we get away from either of these.

If you're familiar with Taylor series approx., then our LPM is first-order Taylor approx. to the non-linear model around $\approx$.
3. So, simply put, the problem with the LPM for the underlying probability of a binary variable is that it’s simply the wrong model; probability is non-linear. (Obviously, this is a theoretical conviction, but it can be shown that, strictly speaking, LPM is a poor model for some pretty ridiculous things. If we impose the MLE constraints, OLS or WLS typically does not, which is why it’s an OK approx for some limited purposes.) Use a non-linear model.

II. Non-linear Models of Probability:

A. The Log-Odds or Logistic Regression

1. \( \theta = \text{Odds} = \frac{p}{1-p} \)
   - i.e., Odds that \( Y=1 \) are \((1-p):p\); e.g., if odds are \(3:1\), then \( Y=1 \) is 3 times as likely as \( Y=0 \)
   - if \( 0<p<1 \), then odds are \( 0<\theta<\infty \) which gets rid of one bound.

2. \( \ln(\theta) = \ln(\text{Odds}) = \log-odds = \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p) \)
   - if \( 0<p<1 \), then \( \ln(\text{Odds}) \) is unbounded: \(-\infty < \ln(\theta) < \infty \)

3. So, one possibility is to replace \( \ln(p) = X\beta \) with \( E(\ln\left(\frac{p}{1-p}\right)) = X\beta \)
   - that is, we just change the model of \( p \) to a non-linear one with the properties we want.

   a) dropping the \( E(\cdot) \) to keep notation simple: (*commonly seen versions)
      \[
      \star \ln\left(\frac{p}{1-p}\right) = X\beta \\
      \Rightarrow \frac{p}{1-p} = e^{X\beta} \\
      \Rightarrow \hat{p} = \left(1+e^{X\beta}\right)^{-1} e^{X\beta} = \left[1 + e^{X\beta}\right]^{-1} e^{X\beta} = \left[1 + e^{X\beta}\right]^{-1} \\
      \Rightarrow \hat{p} = \left[1 + e^{-X\beta}\right]^{-1} \\
      \Rightarrow \hat{\theta} = \frac{\hat{p}}{1-\hat{p}} = \frac{1}{1+e^{-X\beta}} \\
      \Rightarrow \hat{\theta} = \left[1 + e^{-X\beta}\right]^{-1} \\
      \Rightarrow \hat{\theta} = \left[1 + e^{-X\beta}\right]^{-1}
      \]

4. This is the prediction in terms of \( \hat{\gamma}^* \), but what about the prediction of \( Y \)? We predict \( Y=1 \) if \( \hat{\gamma}^* > 0 \) if \( \hat{\theta} > 0 \).

   These translate to: \( \hat{Y}=1 \) if \( \hat{\theta} > 0 \) if \( \hat{\theta} > 0 \).

B. The other Possibility is more direct and more interpretable. Let \( \phi(\gamma) \) be the standard normal p.d.f.; let \( \Phi(\gamma) \) be the associated c.d.f. (cumulative dist.)

Then, let \( Y^* \sim \Phi(\gamma) \), \( Y=1 \) if \( Y^* \geq 0 \) if \( \Phi(\gamma) > 0.5 \).

This can be arbitrarily chosen, so set it to 0 for simplicity: \( p(Y=1 | Y^*) = \Phi(\gamma) \) if \( Y^* > 0.5 \).
"Thy" Prob(event j occurs) = Prob(Y = 1) = F(xβ, params)

let \( F() \) be lin-add, then

\[ \lim_{x \beta \to +\infty} P(Y = 1 | x) = 1 \]

\[ \lim_{x \beta \to -\infty} P(Y = 1 | x) = 0 \]

[1] probably add more "thy" that rates of this convergence \( \to 0 \) smoothly \( \Rightarrow \) 'sigmoid' (S-) shape

\[ \Rightarrow \text{Many Options:} \]

\[ \text{Probit: } \quad \text{Prob}(Y = 1 | x) = \int_0^x + \Phi(\theta) \, d\theta = \Phi(x\beta) \]

\[ \text{Logit: } \quad \text{Prob}(Y = 1 | x) = \frac{e^{x\beta}}{1 + e^{x\beta}} = \Delta(x\beta) \]

\[ \text{Weibull: } \quad P(Y = 1) = e^{-e^{x\beta}} \quad \text{relaxes symmetry} \]

\[ \text{comp. log-log: } \quad P = \frac{1 - (e^{-e^{x\beta}})}{e^{x\beta}} \quad \text{also relases symm.} \]
III.C.3. The effect of $\Delta X$, the method without derivatives

a) Evaluate $\hat{\beta}$ at some meaningful set of values of $X$.

b) Hold other $X$'s fixed & change your focused on $X$ some meaningful # of units, re-evaluate $\beta$

c) $\hat{\beta}_b - \hat{\beta}_a$ is then the estimated effect of the change in $X$ on the probability $Y=1$, holding all else constant at these levels.

d) example:

$$y_i = \begin{cases} 1 & \text{if individual } i \text{ votes} \\ 0 & \text{if individual } i \text{ doesn't vote} \end{cases}$$

$$X = \begin{cases} \text{age} \\ \text{income} \\ \text{sex} \\ \text{race} \\ \text{whatnot} \end{cases}$$

Hold age, income, race & whatnot fixed...

- to sample means
- to some substantively important level

Evaluate

$$\hat{\beta}_a = \text{prob}(Y=1 \mid X, \text{ sex= male})$$

$$\hat{\beta}_b = \text{prob}(Y=1 \mid X, \text{ sex= female})$$

$$\hat{\beta}_b - \hat{\beta}_a$$ is then the amount by which a female $i$ is estimated to be more likely to vote than a male with these same $X$ characteristics.

→ Use Tables

→ Use Graphs:

\[ i \]

holding all else to their means

or whatever
That is, we model: 
\[ P(Y = 1) = P(Y^* < 0) = \int_{-\infty}^{0} f(y^*) \, dy^* = \int_{-\infty}^{0} \Phi(y^*) \, dy^* = \Phi(0|X\beta) \]

In one line: 
\[ P(Y_i = 0) = \int_{-\infty}^{0} \Phi(y_i^* | X_i\beta) \, dy_i^* = \Phi(0|X_i\beta) \]

It looks complicated, but the logic is straightforward: The underlying probability of \( Y = 1 \) depends on some unobservable variable, \( Y^* \), being greater than some threshold. Since scale of \( Y \) is arbitrary, then, set threshold to anything you like (could instead set \( Y^* \) to what you like a dummy threshold for each \( Y_i \)). Easiest is zero. So question becomes: is \( Y^* > 0 \)? Answer: \( Y^* \) is distributed normal, mean \( X\beta \), variance 1, what is probability that \( Y^* > 0 \) given \( X\beta \) can be relaxed to nonstochastic, but usually not enough info to sort it out.

III. C. Back to Logit (it's easier to work with, even though Probit has the better theoretical underpinnings, so reasons we don't explore it here; it has to do with choice theory).

So, we conceptualize: 
\[ \hat{\beta} = \left[ 1 + e^{-X\beta} \right]^{-1} \]

1. What does it look like? Sigmoid curve
   (probit looks the same, but has thinner tails—usually no noticeable difference)

2. What is the effect of an increase in \( X^* \)? Depends on where you are in the curve. (ditto for probit)

\[ \text{logit: } \frac{\partial}{\partial X} \left[ \log \left( \frac{1}{\Phi(\Phi(X^*))} \right) \right] = \frac{e^{-X^*}}{1 + e^{-X^*}} \cdot \frac{e^{-X^*}}{1 + e^{-X^*}} \cdot \beta = \hat{\beta} \cdot (1 - \hat{\beta}) \beta \]

\[ \text{probit: } \frac{\partial}{\partial X} \Phi(\Phi(X^*)) = \Phi(X^*) \cdot \beta \]

\[ \text{p.d.f. of std. normal evaluated at } X \beta \text{ (again max at } X \beta = 0 \text{, goes to zero as } X \beta \to \infty \text{ or } X \beta \to -\infty) \]
5. So, OLS is the MLE of $\beta$ for the CLRM. Unbounded least squares is the maximum likelihood estimator of $\beta$. It is well known that the estimation of $n-2$ parameters of a $\mathcal{N}(X\beta, \Omega)$ regression model has $n-2$ degrees of freedom.

$$\sum_{i=1}^{n} (y_i - x_i' \beta) \cdot x_i = 0$$

$$\sum_{i=1}^{n} (y_i - x_i' \beta)^2 = \frac{\sum_{i=1}^{n} (y_i - x_i' \beta)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$n \cdot (\sum_{i=1}^{n} (y_i - x_i' \beta) \cdot x_i \cdot (y_i - x_i' \beta)) = 0 = \frac{\sum_{i=1}^{n} (y_i - x_i' \beta) \cdot x_i \cdot (y_i - x_i' \beta)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (y_i - x_i' \beta) \cdot x_i \cdot (y_i - x_i' \beta)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (y_i - x_i' \beta) \cdot x_i \cdot (y_i - x_i' \beta)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 2.3$$

4. Maximum Likelihood refers to a model that maximizes the likelihood of observing the data you actually observed given the model.

$$LL = \frac{1}{2} \ln(\frac{1}{\sigma^2}) + \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i' \beta)^2$$

3. Since each $y_i$ is independent, the probability of all the $y_i$ is the product of their individual marginal probability density functions.

$$f(y_i | x_i, \beta) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - x_i' \beta)^2}{2\sigma^2}}$$

2. The normal probability density function $f(y | \mu, \sigma^2)$ is of full rank $p$. It is based on an independent normal distribution with mean $\mu$, and variance $\sigma^2$. 

1. Maximum Likelihood Methods - a quick review of the classical linear models. Recall that one way to specify a linear probability model is $y_i = X_i \beta + \epsilon_i$. This is a draw from an independent normal distribution with mean $X_i \beta$ and variance $\sigma^2$. An alternative is logarithmic least squares where the model is $\ln(y_i) = X_i \beta + \epsilon_i$.
B. Maximum Likelihood Estimation of Binary Dependent Variable Models.

1. Start with first principles (assumptions) just like in linear models
   a. \( Y_i \) independently Bernoulli with probability \( P_i \)
      i.e.: \( Y_i = 1 \) with probability \( P_i \)
      \( Y_i = 0 \) with probability \( 1 - P_i \)

   b. \( P_i = g(X_i; \beta) \)
      (i.e., the probability that \( Y_i \) equals 1 is some function of \( X_i \) and the parameters \( \beta \))

   i) for logit:
      \[ P_i = \frac{1}{1 + e^{-x_i \beta}} = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \]

   ii) for probit:
      \[ P_i = \int_{-\infty}^{x_i \beta} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \Phi(x_i \beta) \]
      (the cumulative standard normal, evaluated at \( x_i \beta \))

   c. \( X_i \) of rank \( k \) (no perfect linear dependence among \( X_i \)'s)

2. Then write down the probability function for each obs. i.e:
   \[ f(Y_i; X_i; \beta) = P_i Y_i (1 - P_i)^{1 - Y_i} \Rightarrow \text{probability } X_i = 1 \]

3. Since each \( Y_i \) is independent (given the \( X_i \)'s) the joint pdf is:
   \[ L = \text{likelihood} = \prod_{i=1}^{n} P_i^{Y_i} (1 - P_i)^{1 - Y_i} \]

4. Maximum Likelihood Proceeds by maximizing this likelihood over the parameters \( \beta \) given the data \( (X, Y) \)--using logarithms since this is easier:
   \[ \ln(L) = \sum_{i=1}^{n} \{ Y_i \ln P_i + (1 - Y_i) \ln (1 - P_i) \} \]
   \[ \ln(L) = \sum_{i=1}^{n} \{ Y_i \ln P_i + (1 - Y_i) \ln (1 - P_i) \} \]
   \[ \max_{\beta} LL \Rightarrow \frac{\partial LL}{\partial \beta} = 0 \Rightarrow \sum_{i=1}^{n} \{ Y_i \cdot \frac{\partial P_i}{\partial \beta} - (1 - Y_i) \cdot \frac{1}{P_i} \cdot \frac{\partial P_i}{\partial \beta} \} = 0 \]

   \[ \Rightarrow 0 = \sum_{i=1}^{n} \{ Y_i \cdot (1 + e^{-x_i \beta}) \cdot \frac{\partial e^{x_i \beta}}{\partial \beta} - (1 - Y_i) \cdot \frac{1}{e^{x_i \beta}} \cdot \frac{\partial e^{x_i \beta}}{\partial \beta} \} \]

   \[ \Rightarrow 0 = \sum_{i=1}^{n} \{ Y_i \cdot (1 + e^{-x_i \beta}) \cdot \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} - (1 - Y_i) \cdot \frac{1}{e^{x_i \beta}} \cdot \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \} \]

   \[ \Rightarrow 0 = \sum_{i=1}^{n} \{ Y_i \cdot (1 + e^{-x_i \beta}) - (1 - Y_i) \cdot \frac{1}{e^{x_i \beta}} \cdot (1 + e^{-x_i \beta}) \} \]

   \[ \Rightarrow 0 = \sum_{i=1}^{n} \{ Y_i - (1 + e^{-x_i \beta}) \} X_i \]

   \[ \Rightarrow \sum_{i=1}^{n} \{ Y_i - (1 + e^{-x_i \beta}) \} X_i = 0 \]

   for probit they are:
   \[ \sum_{i=1}^{n} \{ Y_i - \Phi(x_i \beta) \} \cdot \frac{\phi(x_i \beta)}{\Phi(x_i \beta)} (1 - \Phi(x_i \beta)) \cdot X_i = 0 \]
IV.B.5. Maximum Likelihood Estimation of Logit & Probit:

The Normal Equations:

a) \[ \sum_{i=1}^{n} \left( y_i - \left(1 + e^{-x_i \beta}\right)^{-1} \right) \cdot x_i = 0 \] for logit

and

b) \[ \sum_{i=1}^{n} \left( y_i - \Phi(x_i \beta)\right)^2 \cdot \frac{\phi(x_i \beta)}{\Phi(x_i \beta)(1-\Phi(x_i \beta))} \cdot x_i = 0 \]

The normal equations implicitly give \( \beta \) as a function of \( X \& Y \), which maximizes the likelihood of having observed the data-generating model just as CMLRM normal equations implicitly gave \( \beta \) as a function of \( X \& Y \), which maximized the likelihood of observing the data given the model.

i) The solution of these equations is the answer to the question: "If \( y_i = 1 \), is given by this function (the logit or probit function, e.g. \( \frac{1}{1+e^{-\beta}} \)), what values for \( \beta \) maximize the likelihood of having observed this set of \((y_i, x_i)\) observations -- i.e., this configuration of \( X \)'s & \( Y \)'s?"

ii) The difference between the logit/probit normal equations and the CMLRM normal equations are that the former could be solved explicitly for \( \beta \): \( \beta = g^{-1}(\gamma) \), whereas the latter cannot.

\[ \Rightarrow \] The computer repeatedly searches for the values of \( \beta \) which maximize the likelihood function. It stops when likelihood gets only negligibly higher or no higher) for any change in any \( \beta \) in any direction.

(it can be shown that the likelihood function for both probit & logit is "strictly concave" this means it has only one peak - so the maximum found is the global maximum)

6. Estimating the Variance:

Notice that there is a likelihood function changing more for movements of \( \beta \) from the optimum than the "flatter" previous likelihood function. In this sense, comparing any \( \beta \) with some other
estimate of $\hat{\beta}$ leads to greater relative likelihood of $\hat{\beta}$ to $\tilde{\beta}$ in the former case than the latter.

\begin{align*}
\text{Case a) } & \quad \text{Case b) } \\
\hat{\beta} & \quad \tilde{\beta} \\
\beta & \quad \beta
\end{align*}

- For $\hat{\beta}$ & $\tilde{\beta}$ equally distant in the two cases,

\[
\frac{\text{likelihood}(\hat{\beta})}{\text{likelihood}(\tilde{\beta})} \quad \frac{\text{likelihood}(\hat{\beta})}{\text{likelihood}(\tilde{\beta})}
\]

$\Rightarrow$ more certain of $\hat{\beta}$ relative to alternatives $\tilde{\beta}$ in case (a) than in case (b).

So, the more the likelihood function "curves away" from the optimum $\hat{\beta}$, the more certain is our estimate of $\hat{\beta}$.

$\Rightarrow$ Variance estimates will be based on curvature of likelihood function at $\hat{\beta}$. To be specific:

\[
V(\hat{\beta}) = -\frac{1}{\left[ \frac{\partial^2 \ln(L)}{\partial \beta^2} \right]_{\hat{\beta}}} \quad \text{i.e., negative the inverse of the second derivative of the log-likelihood function evaluated at $\hat{\beta}$.}
\]

- Why "negative"? Second derivative is negative because $\ln(L)$ curves down at $\hat{\beta}$ (see above figures). We want positive signs here—
  we're concerned about magnitude of curvature, we know its direction is down but that's unimportant here.

- Why "inverse"? The absolute value of second derivative is higher in case (a), revealing more curvature; more certainty which means smaller standard errors (certainty & standard errors are inversely related).

- Why "second derivative at optimum"? Because that's the mathematical definition of "curvature away from $\hat{\beta}$".

The variance-covariance matrix of $\hat{\beta}$ is a standard output of all MLE programs. It can either be calculated explicitly from (c) using the analytic solution to $\frac{\partial \ln(L)}{\partial \beta}$ if the analytic solution is possible, or it can be computed empirically by computer simulation.
G. An Example of a simple logit model that does have an explicit MLE analytical solution (i.e., that wouldn't need to be estimated by quantitative, computer-search methods). (It's pretty much the only example.)

1. The model with just a constant
   a) \( Y \sim \text{Bernoulli} \) with probability \( p \), independent
   b) \( p_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} = \frac{e^{x_i \beta}}{e^{x_i \beta} + 1} \)
   c) \( \text{Rank}(X) = k = 1 \) (\( X \) is just a vector of ones.)

2. Likelihood of \( y_i \)
   \[
   p_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} = \frac{e^{x_i \beta}}{e^{x_i \beta} + 1} \]

3. Joint likelihood of the \( n \) independent \( y_i \)'s
   \[
   L = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)} \Rightarrow \ln(L) = \sum_{i=1}^{n} y_i \ln p_i + (1-y_i) \ln (1-p_i)
   \]

4. The MLE Estimate:
   
   \[
   \frac{\partial \ln(L)}{\partial \beta} = \sum_{i=1}^{n} \left[ y_i \left( \frac{1 + e^{\beta}}{e^{\beta}} \right) \left( \frac{e^{\beta}}{1 + e^{\beta}} \right) + (y_i - 1) \left( \frac{1 + e^{\beta}}{e^{\beta}} \right) \left( \frac{e^{\beta}}{1 + e^{\beta}} \right) \right] = 0
   \]

   \[
   \Rightarrow \sum_{i=1}^{n} \left[ y_i \left( \frac{1 + e^{\beta}}{e^{\beta}} \right) - \frac{e^{\beta}}{1 + e^{\beta}} \right] = 0
   \]

   \[
   \Rightarrow \sum_{i=1}^{n} y_i = \frac{e^{\beta}}{1 + e^{\beta}} \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{n}
   \]

   \[
   \text{Duh! your best guess for } \beta \text{ if all } y_i \text{ the same is just the proportion of ones, i.e., } \frac{1}{n} \]
V. Interpreting Results from Logit/Probit Estimations

A. Substantive Meaning of Coefficients

1. In all models, we are interested in the effect of \( X \) on \( Y \). In this case, the effect of \( X \) on the probability \( Y=1 \) is directly at issue.

\[
\text{Logit: } \frac{\partial \Pr(Y=1)}{\partial X_i} = \frac{\partial \phi}{\partial X_i} = \phi \left( 1 + e^{-X} \right) \frac{\partial}{\partial X_i}
\]

\[
= (-1) \left( 1 + e^{-\beta} \right)^{-2} e^{-\beta} \cdot (-\beta)
\]

\[
= \frac{e^{-\beta}}{1 + e^{-\beta}} \cdot \beta
\]

\[
= \frac{1}{1 + e^{\beta}} \cdot \frac{e^{\beta}}{1 + e^{\beta}} \cdot \beta = \frac{\beta}{1 + e^{\beta}} \cdot \beta = \beta (1 + \beta) \beta
\]

- The effect of a unit change in \( X \) on the probability \( Y=1 \) is not constant. It depends on where in the "S-curve" you are evaluating this effect.

a) When \( X \beta \) is near \( 0 \), so \( \beta \) is near 1, the effect is (approximately 1) \( \beta \approx 0 \)

b) When \( X \beta \) is near \( 0 \), so \( \beta \) is near \( \frac{1}{2} \), the effect is (approximately \( \frac{1}{2} \)) \( \beta \approx (\frac{1}{2}) \beta \)

c) Things to Note:

(i) Sign of \( \beta \) is the sign of the effect of \( X \) on \( \Pr(Y=1) \)

(ii) Magnitude of \( \beta \) relative to some other \( \beta_i \) is relevant for comparing the magnitude of the effect of \( X \) relative to \( X_i \)

(iii) The effect of a unit change in some \( x \) on \( \Pr(Y=1) \) depends on \( \beta \), which in turn depends on all \( X \)'s through \( \hat{\beta} = \frac{e^{\beta}}{1 + e^{\beta}} \) \( \beta \) terms.

\[
\text{Probit: } \frac{\partial \Pr(Y=1)}{\partial X_i} = \frac{\partial \Phi}{\partial X_i} = \Phi(X \beta) \cdot \beta
\]

\[\text{marginal p.d.f. for std normal evaluated at } X \beta\]

- Otherwise, all the substance is as above.

In sum, these are the slopes of the \( \hat{\beta}(X \beta) \) functions -- marginal changes in \( \Pr(Y=1) \) for marginal increases in \( X \).

2. The "first-difference" Method

- Another way to evaluate the effect of some increase in \( X \) on the \( \Pr(Y=1) \) is simply to manually calculate \( \Pr(Y=1) \) at two or more different levels of \( X \).

\[
\text{Logit: } \hat{\beta} = \frac{e^{X \beta}}{1 + e^{X \beta}} \Rightarrow \text{a) fix other } X \text{'s at their means, or at some other substantively revealing level.}
\]

\[
\text{Probit: } \hat{\beta} = \Phi(X \beta) \Rightarrow \text{c) } \hat{\beta} - \beta_0 \text{ is the effect of a unit increase in } X \text{ holding all other } X \text{'s fixed at } \beta_0 \text{ (whatever you hold them fixed at).}
\]

\[
\text{b) Calculate } \beta_0 \text{ at some value for the } X \text{ in question (other } X \text{'s being fixed at the level in } \hat{\beta} \text{. Calculate } \hat{\beta} \text{, again for that } X+1 \text{ (other } X \text{'s still fixed).}
\]
3. Graphical/Tabular Method
   a) Hold all other X’s fixed at some level
   b) Graph/Table the effect of moving the X in question
      from X to some other X'

4. Examples:
   Estimate: \( P(Y=1) = \frac{e^{\beta X}}{1 + e^{\beta X}} = \left[ 1 + e^{-\beta X} \right]^{-1} \)
   - \( \text{The constant \( \rightarrow X_0 \)} \)
   - \( X_1 = 1.5 \)
   - \( X_2 = 2.3 \)
   - \( X_3 = 0.7 \)
   - \( X_4 = -0.5 \)

   a) Derivative Method:
      \[ \frac{\partial P(Y=1)}{\partial X_2} = \beta_2 \left[ 1 + e^{-\beta X} \right] \left( 1 + e^{\beta X} \right)^{-1} \]
      \[ = \beta_2 \left( 1 + e^{1.5-2.3X_1-0.7X_2+0.5X_4} \right)^{-1} \]
      \[ = \left( 1 + e^{1.5-2.3X_1-0.7X_2+0.5X_4} \right)^{-1} \]
      Pick some values of \( X_1, X_2, \& X_3 \) at which to evaluate this. If nothing else, means are revealing.

   b) First-Difference:
      \[ \hat{P}_0 = P(Y=1 \mid X_0 \text{ fixed at } X_0, x_2 = x_2, x_3 \text{ fixed at } x_3) \]
      \[ \hat{P}_1 = P(Y=1 \mid X_1 \text{ fixed at } X_1, x_2 = x_2', x_3 \text{ fixed at } x_3) \]
      \[ \hat{P}_1 - \hat{P}_0 = \left[ 1 + e^{-1.5-2.3X_1-0.7X_2+0.5X_4} \right]^{-1} \]
      \[ = \left[ 1 + e^{-1.5-2.3X_1-0.7X_2+0.5X_4} \right]^{-1} \]
      \[ = \left[ 1 + e^{-1.5-2.3X_1-0.7X_2+0.5X_4} \right]^{-1} \]
      \[ = \left[ 1 + e^{-1.5-2.3X_1-0.7X_2+0.5X_4} \right]^{-1} \]

   c) Graph: fix \( X_1 \) \& \( X_3 \), set up a spreadsheet to calculate as so:

<table>
<thead>
<tr>
<th>( X_2 )</th>
<th>( \hat{P} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

   Use the spreadsheet formula, referring to the \( X_0 \) cells here.

Table:

- Unit change in \( X_1 \)
- Unit change in \( X_2 \)
- Unit change in \( X_3 \)

- Use formulas in each cell

- Or some other meaningful change

- Use an add-in, spreadsheet formula for the \( X_0 \) cells for a

- Write a spreadsheet formula for the \( X_0 \) cells for the \( X_0 \) cells here.
B. Coefficient Certainty

1. As noted, MLE spits out coefficient estimates & estimates of the Var-Covar matrix of those estimated coefficients (much more on this in a bit, but for now...)

   $H_0$: effect of $X_i$ is zero
   $\Rightarrow \beta_0 = 0$
   $\Rightarrow$ Same Test Options as before:
   a) Is $\beta$ far from zero?
   $\hat{\beta} \sim N(0, 1)$
   $\Rightarrow$ For large $N$, can compare to std. normal table. For small $N$, we don't know the exact distribution -- most use the $t$ dist. as an approximation
   b) Impose $\beta = 0$ if seen if we lose much fit. Here there's no $R^2$ measure of fit, but LR tests still apply
   
   $-2 \cdot (LL_{R} - LL_{0}) \sim \chi^2_{\text{restricted}} - \chi^2_{\text{unrestricted}}$

2. Testing whether effect is other than zero, or providing confidence intervals for effects rather than coefficients is more complicated. Just like with interactions, the variance of the effect depends on where it's evaluated:

   $V(\beta_x^*) = V\left(\frac{\exp(x)}{1 + \exp(x)} \cdot \beta\right)$

   Tricky b/c every term depends on the estimate, $\hat{\beta}$, can't be pulled out as easy as in linear case.

3. Joint Hypothesis Tests on Coefficients:
   As before: Option 1 is Wald
   $W = (R_b - \mathbf{q})' \{R(\hat{\beta})R'\}^{-1}(R_b - \mathbf{q}) \sim \chi^2_{R}$
   Option 2 is LR: $LR = -2(\text{LL}_R - \text{LL}_0) \sim \chi^2_{R}$
   * For non-linear models, LR test is often easier to implement & applies more generally than the analogous $R^2$-ratio test does in linear case.

   Example for Wald Test (it's same as for linear, so):
   $H_0: \beta_1 = \beta_2 = 0$

   $\Rightarrow \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \cdot \begin{bmatrix} V(b_1) & \text{Cov}(b_1, b_2) \\ \text{Cov}(b_1, b_2) & V(b_2) \end{bmatrix}^{-1} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \sim \chi^2_2$

   These are your coefficients
   This is your estimated $V$-Cov Mat
   By your coefficient again
\[ \frac{\partial \hat{p}}{\partial \beta_j} = \hat{p} (1 - \hat{p}) x_j \]

\[ \sqrt{\left( \frac{\partial \hat{p}}{\partial \beta_j} \right)^2} \approx \nabla_B \left( \frac{\partial \hat{p}}{\partial \beta_j} \right) \cdot \sqrt{\left( \nabla \beta \right)} \nabla_B \left( \frac{\partial \hat{p}}{\partial \beta_j} \right) \]

\[ \frac{\partial \left( \frac{\partial \hat{p}}{\partial \beta_j} \right)}{\partial \beta_k} = \frac{\partial \hat{p}}{\partial \beta_k} \cdot (1 - \hat{p}) x_j \]

\[ \frac{\partial \hat{p}}{\partial \beta_k} - \frac{\partial \hat{p}}{\partial \beta_k} x_j + \hat{p} (1 - \hat{p}) \frac{\partial x_j}{\partial \beta_k} \]

\[ \hat{p} = \Lambda (XY) \]

\[ \frac{\partial \hat{p}}{\partial \beta_j} = \left[ \Lambda (XY) \right] \left[ 1 - \Lambda (XY) \right] \frac{\partial \Lambda (XY)}{\partial \beta_j} \]

\[ \sqrt{\left( \frac{\partial \hat{p}}{\partial \beta_j} \right)^2} = \left[ \nabla_B \frac{\partial \hat{p}}{\partial \beta_j} \right] \cdot \sqrt{\left( \nabla \beta \right)} \left[ \nabla_B \frac{\partial \hat{p}}{\partial \beta_j} \right] \]

\[ \frac{\partial \frac{\partial \hat{p}}{\partial \beta_k}}{\partial \beta_j} = \frac{\partial \Lambda (XY)}{\partial \beta_k} \cdot \left[ 1 - \Lambda (XY) \right] \cdot \frac{\partial \Lambda (XY)}{\partial \beta_j} + \frac{\partial \Lambda (XY)}{\partial \beta_k} \cdot \frac{\partial \Lambda (XY)}{\partial x_j} \]

\[ - \left[ \Lambda (XY) \right] \cdot \frac{\partial \Lambda (XY)}{\partial \beta_k} \cdot \frac{\partial \Lambda (XY)}{\partial x_j} + \Lambda (XY) \left[ 1 - \Lambda (XY) \right] \cdot \frac{\partial \Lambda (XY)}{\partial \beta_k} \]

\[ \frac{\partial \Lambda (XY)}{\partial \beta_k} \]
\[ p = \frac{e^{xb}}{1 + e^{xb}} = e^{xb}(1 + e^{xb})^{-1} \]

\[ (1 - p) = \frac{1}{1 + e^{xb}} = (1 + e^{xb})^{-1} \]

\[ \frac{2p}{\partial x_k} = \frac{e^{xb}}{(1 + e^{xb})^2} \cdot x_k \]

\[ \frac{2(1 - p)}{\partial x_k} = -\frac{e^{xb}}{(1 + e^{xb})^2} \cdot x_k \]

\[ \delta(B) = \frac{\partial \beta}{\partial x_j} = \hat{p} - \hat{p} \beta_j \]

so,

\[ \frac{\partial B}{\partial \beta_k} = \frac{e^{xb}}{(1 + e^{xb})^2} \cdot x_k \cdot \frac{1}{1 + e^{xb}} \cdot \beta_j - \frac{e^{xb}}{1 + e^{xb}} \cdot \frac{e^{xb}}{(1 + e^{xb})^2} \cdot x_k \cdot \beta_j + \frac{e^{xb}}{1 + e^{xb}} \cdot \frac{1}{1 + e^{xb}} \cdot \frac{e^{xb} - (e^{xb})^3}{(1 + e^{xb})^3} \cdot x_k \cdot \beta_j + \frac{e^{xb}}{(1 + e^{xb})^2} \]

\[ \cdot \frac{e^{xb}}{(1 + e^{xb})^3} \cdot \frac{1 - e^{xb}}{(1 + e^{xb})^3} + \frac{e^{xb}}{(1 + e^{xb})^2} \cdot \frac{\partial \beta_j}{\partial \beta_k} \]

\[ \sqrt{\phi(xB) \cdot \beta_j} = \left[ \frac{\partial \phi(xB)}{\partial \beta_k} \cdot \beta_j + \phi(xB) \cdot \frac{\partial \beta_j}{\partial \beta_k} \right] \cdot V(B) \]
\[ V(f(B)) = \left[ \frac{\partial f(B)}{\partial \beta_0} \right]' V(B) \left[ \frac{\partial f(B)}{\partial \beta_0} \right] + \cdots \]

\[ \text{Estimated } V(\frac{\partial \hat{p}}{\partial x_j}) \text{ from Logit Model:} \]

\[ \frac{\partial \hat{p}}{\partial x_j} = \frac{e^{xb}}{1 + e^{xb}} \cdot \frac{1}{1 + e^{xb}} \cdot \beta \equiv f(B) \]

\[ = \hat{p} \cdot (1 - \hat{p}) \cdot \beta \]

\[ V(f(B)) = \left[ \frac{\partial f(B)}{\partial \beta_0} \right]' V(B) \left[ \frac{\partial f(B)}{\partial \beta_0} \right] = \left[ \frac{e^{xb}(1 - e^{xb})}{(1 + e^{xb})^2} \cdot x_j \beta \right]' V(\beta) \left[ \frac{\partial f(B)}{\partial \beta_0} \right] \]

\[ \text{...from Probit Model:} \]

\[ \frac{\partial \hat{p}}{\partial x_j} = \phi(XB) \cdot \beta \equiv f(B) \]

\[ V(f(B)) = \left[ \frac{\partial f(B)}{\partial \beta_0} \right]' V(B) \left[ \frac{\partial f(B)}{\partial \beta_0} \right] = \left[ \frac{\partial \phi(XB) \cdot \beta}{\partial \beta_0} \right]' V(\beta) \left[ \frac{\partial \phi(XB) \cdot \beta}{\partial \beta_0} \right] \]
I.C. Interpreting Logit/Probit: Fit of Logit/Probit Models

1. The (a) formal test for fit of the model
   a) Just like always, one hypothesis of interest is that there is nothing revealing in the model -- i.e., test predictions based on your model \( \hat{\beta}_i \) against simply predicting that \( \hat{\beta}_i \) is constant \#obs. i. In terms of predicting \( Y = 1 \) or \( Y = 0 \), this amounts to testing whether predicting on basis of \( \hat{\beta}_i = 0 \) if \( \hat{\beta}_i = f(X) \), 5 outperforms simply always predicting 1 or 0, whichever is more common.

b) In terms of more familiar testing procedures, this amounts to testing your model against the model with just a constant (which as we showed a bit ago \( \Rightarrow \hat{\beta}_i \approx \frac{\Sigma Y_i}{N} \approx \hat{\beta}_i \)). So, it's just a degradation of \( \chi^2 \) test.

c) I.e., in the context:

\[
\chi^2 = -2 \left( \ln L_r - \ln L_u \right)
\]

restricted model is model w/ just a constant
unrestricted model is your full model

\( \ln L_r \) can be obtained either by estimating the constant-only model as well as your own, or you could use:

\[
\ln \left( L_r \right) = N \ln \left( \frac{M_N}{\hat{p}_i} \right) + M \ln \left( \frac{N \hat{p}_i}{1 - \hat{p}_i} \right)
\]

where \( N \) is sample size, \( M \) is #obs. with \( Y = 0 \) & \( N - M \) is #obs. with \( Y = 1 \).

2. Measures of Fit in Logit/Probit (Binary Dependent) Models

a) \( R^2 \) does not exist, nor, strictly speaking, does M.S.E. (s.e. or s.e.e.). In Bernoulli models, there is no way to separate the estimate of the mean, \( \hat{p}_i \), from the variance, \( \hat{\sigma}_i(1 - \hat{p}_i) \).

b) Still, researchers have continued to desire some means of summarizing how well their model fits the data. Several statistics have been suggested; none is universally recognized/understood, none is without major flaws, & none is terribly useful as a test statistic in any way. Still, perhaps some are revealing, so...

c) Pseudo-\( R^2 \)'s:

1) Aldrich & Nelson:

\[
R^2 = \frac{c}{M_N}
\]

where \( c \) is \( \chi^2 \) test of model & \( N \) is # of observations

\[
E(c) = k \quad \text{if no explanatory power}
\]

\[
\Rightarrow R^2 \rightarrow 0 \quad \text{as explanatory power} \quad N \rightarrow \infty
\]

\[
E(c) = N \quad \text{if perfect explanatory power}
\]

\[
\Rightarrow R^2 \rightarrow 1 \quad \text{as explanatory power} \quad N \rightarrow \infty
\]

2) McKelvey & Zavoina:

\[
R^2 = \frac{\Sigma \left( \hat{p}_i - \hat{p}_i \right)^2}{\Sigma \left( \hat{p}_i - \hat{p}_i \right)^2 + \Sigma \left( \hat{p}_i - \hat{p}_i \right)^2}
\]

\( \hat{p}_i \) can be seen as explained sum of squares over total sum of squares.
iii) Generic Pseudo-\( R^2 \):
\[
R^2_c = 1 - \frac{\log(L_u)}{\log(L_r)} < \frac{\text{log-likelihood of full model}}{\text{log-likelihood of constant-only model}}
\]
proportionate increase in likelihood over the null model

\( \times \) Prediction Pseudo-\( R^2 \):
\[
R^2_p = \frac{\% \text{Correct Model} - \% \text{Correct Null}}{1 - \% \text{Correct Null}}
\]
e.g. a) what \( \% \) does your model correctly predict, call it \( \lambda \)
b) Take all ones or all zeros, whichever guesses better, what \( \% \) is that
c) then \( R^2_p = \frac{\lambda - \omega}{1 - \omega} \)
\( \times \) measures predictive power gain as a proportion of what was possible to be gained

A Note on Pseudo-\( R^2 \)’s:
- There is no standard. Stats packages vary on which one they report (Stata is \( \lambda \))
- All are trying to summarize the explanatory power of the model, but note that baselines differ (implicit) goals differ
- And none have “statistical justification” which is unassailable
- \( \Rightarrow \) these are heuristic devices only (which highlights how limiting
it is that there’s no agreement
on which to use (teach)).

VI. Properties of ML Estimators

A. All ML Estimators are known to have “good” asymptotically. Few small-
sample properties have been proven.

B. Asymptotic Properties:
- Parameters (such as \( \hat{\beta} \))
  1. Consistency:
  \[ \lim_{n \to \infty} \hat{\beta} = \beta \]
  2. Asymptotic Normality:
  \[ \hat{\beta} \sim N(\beta, \text{var}(\beta)) \]
  3. Asymptotic Efficiency:
  - Among consistent, asymptotically normal estimators, MLE is minimun variance
  4. Invariance: if \( \hat{\beta} \) is MLE of \( \beta \), then \( g(\hat{\beta}) \) is MLE of \( g(\beta) \)

C. So how “bias” is Asymptotic? Most authors comfortable with
100 or more for logit/probit. Many comfortable all the way
down to say 40 or so. I think the emerging intuition is that
you need larger samples the more your data is concentrated
at \( p = 0 \) or \( p = 1 \). Here’s a completely arbitrary rule. I just
dreamed up: I’d want to have at least:
\[
N \geq \left[ (p)(1-p) \right]^{-1} \cdot 10
\]
(proportion of \( p = 1 \) is in sample)

D. Issue Here is: suppose you’re concerned about sample size? Then what?
MLE or LS or some other method is still going to be “best,” you’re just
concerned “best” may not be too good. This is one of those tough luck
situations. Maybe if’d be wisest just to carry on as usual except
offer some caveats about standard hypothesis tests. Problem is, not
much info—nothing short of more info is going to solve that.
I. The maximum-likelihood approach is very flexible. It is easily applied to, in principle, models of dependent variables of any form. Its steps, in the abstract, are:

A. Determine the statistical process generating your data (e.g., for Binary Dep. Var., this is Bernoulli).

B. State your theoretical propositions as a model linking your explanatory variables, \( X \), to the parameters \( \beta \) of the distribution generating your data. (E.g., in Bernoulli's there is one parameter, \( p = P(X=1) \) & we model:

\[
p = f(X) = \begin{cases} \frac{e^{\beta}}{1 + e^{\beta}} & \text{logit} \\ \Phi(X/\beta) & \text{probit} \end{cases}
\]

C. Determine the joint likelihood function for all your observations. (E.g., if \( Y_i \) independent Bernoulli, then:

\[
P(Y = y) = p^y (1-p)^{1-y}
\]

\[
Pr(Y_1 = y_1, Y_2 = y_2, \ldots Y_n = y_n) = \prod p_i^{y_i} (1-p_i)^{1-y_i}
\]

Read this as: Probability (likelihood) \( Y_1 \) equals whatever \( y_1 \) equals, \( Y_2 \) equals whatever \( y_2 \) equals, \( \ldots \) \( Y_n \) equals whatever \( y_n \) equals.

E.g., \( Pr(Y_1 = 1, Y_2 = 0, Y_3 = 1, Y_4 = 1) = p_1 (1-p_1) p_3 p_4 \)

D. Maximize the joint likelihood with respect to your parameter estimates. That is, find the set of values for the parameters which would make your actually observed data the most likely to have been generated by the process defined in B. (According to the model defined in B, e.g., in logit, find the \( \hat{\beta} \) which maximize \( \hat{\beta} \).

\[
\ln(L) = \sum \left( y_i \ln(e^{\beta}) + (1-y_i) \ln(1 + e^{\beta}) \right)
\]

E. These are your maximum-likelihood estimates (MLEs) of the parameters. Their asymptotic Variance-Covariance is given by:

\[
V(\hat{\beta}) = - \left[ \frac{\partial^2 \ln(L) \beta}{\partial \beta \partial \beta} \right]^{-1}
\]

F. These Estimates are Then:
1. Consistent
2. Asymptotically Normal
3. Minimum Variance among consistent asymptotically normal estimates
4. Invariant
II. The Discussion/Summary just completed is thoroughly general. There is, in principle, we can model anything if we can imagine A-E steps. D-A-E can always be done by a computer if we can manage A-C. For many common statistical processes, C has been done for us already. So, really, at an intro level, all we need to know is A & B.

A. Ordered Categorical Variables

1. Suppose a dependent variable takes several possible values which can be rank-ordered, but which cannot be spaced in any meaningful way. i.e. the interval between values of Y is not meaningfully captured by the numerical distance between them. Any ranking is an example. A commonly occurring ranking is a Likert scale:
   \[ \text{Strongly Disagree - Disagree - Somewhat Disagree} \]
   \[ \text{Neither Agree nor Dis} - \text{Somewhat Agree} - \text{Agree} - \text{Strongly Agree} \]
   \[ \Rightarrow 1 - 2 - 3 - 4 - 5 - 6 - 7 \]
   but who knows ifd whether these should be taken as equally distant?

2. Option 1: Treat them as continuous interval anyway
   \[ \Rightarrow \text{We don't know how observed } Y \in \{1, 2, 3, 4, 5, 6, 7\} \]
   relates to the underlying \( Y^* \) which properly spaces the distances by the options, but perhaps it can be thought of as difference bw \( Y \& Y^* \) is measurement error.
   \[ \Rightarrow \text{Meas. Err. in Dep Var } \Rightarrow \text{Inefficient} \]
   But: absent more info we have little idea how inefficient. Could be a lot.

3. Option 2: MLE
   a) Postulate an underlying \( Y^*_i \) reflecting the actual position of obs. i on the interval scale with meaningful spacing:
   \[ Y^*_i = \mu_i + \epsilon_i \]
   b) Theorize that \[ \mu_i = X \beta \]
   c) Postulate a set of thresholds \[ \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_{m-1} < \tau_m \]
   where \( m \) is the number of categories for \( Y \) (Final Likert scale)
   \[ \Rightarrow \Pr(Y_{ij} = 1) = \Pr(\tau_{j-1} < Y^*_i < \tau_j) \]
A. Ordered Categorical Dependent Variables

d. \( P_r(Y_{ij} = 1) = \Pr(Y_{ij} < Y^*_i < \zeta_j) \)

area under the norm from \( \zeta_{j-1} \) to \( \zeta_j \)

\[ = \int_{\zeta_{j-1}}^{\zeta_j} f(x_i | \beta_i, 1) \, dx_i \]

\[ = F_n(\zeta_j | \mu_i, 1) - F_n(\zeta_{j-1} | \mu_i, 1) \]

\( \text{cumulative normal with mean } \mu_i = X_i \beta \) evaluated at \( \zeta_j \) minus same cumulative normal evaluated at \( \zeta_{j-1} \).

This is the likelihood of one observation in category \( j \).

To get the likelihood function for obs \( i \) and all \( m \) categories we get:

\[ \prod_{j=1}^{m} \left[ F_n(\zeta_j | X_i \beta, 1) - F_n(\zeta_{j-1} | X_i \beta, 1) \right]^{y_{ij}} \]

\( y_{ij} = 1 \) if obs \( i \) is in category \( j \), \( = 0 \) else.

e. The joint likelihood of all the obs (if they're independent)

product of all \( i \) from 1 to \( n \):

\[ \prod_{i=1}^{n} \prod_{j=1}^{m} \left[ F_n(\zeta_j | X_i \beta, 1) - F_n(\zeta_{j-1} | X_i \beta, 1) \right]^{y_{ij}} \]

f. Then the computer looks for the set of \( \beta \) and \( \zeta \)

which maximize \( \log \) this. Incidentally, the \( \beta \)s are always

by default and \( \zeta \)s by default so \( \zeta \) has

\( m-1 \) parameters if \( \beta \) has \( k \) parameters to be estimated.

4. So, if \( Y \) is ordered categorical, you ask for Ordered Probit

and get estimates of:

* the \( \zeta \) thresholds for the underlying \( Y^* \)

* the \( \beta \) coefficients relating \( X \) to \( Y^* \)

B. Grouped Independent, Constant Probability Events. I.e.,

there is some Bernoulli R.V. but you observe only counts

of the number of ones in some period or group. Thus also
called grouped constant probability binary variables.

1. Examples: number of elections in last 3 in which person \( i \) votes

(unobserved Bernoulli is person votes in an election; if observed

then use Logit/Probit)

* number of senators voting for some bills).

2. N.B. Assumption: \( \rightarrow \) probability \( y_i = 1 \) for any of the individual

events in the group is same

\( \Rightarrow \) probability of the individual events in the

group are independent

\[ \Rightarrow Y_i = \sum_{j} y_{ij} \]

where \( j \) indexes the individual events and \( y_{ij} = 1 \) if

individual event is 1. \( y_{ij} \) is thus the count of

# events occurring.
3. $X_i \sim f_B(p_i) = p_i^{Y_i} (1-p_i)^{N-Y_i}$
   $\Rightarrow Y_i = \sum Y_i$ is a binomial
   $\Rightarrow Y_i \sim f_B(N, p_i) = \binom{N}{Y_i} p_i^{Y_i} (1-p_i)^{N-Y_i}$
   $= \frac{N!}{Y_i! (N-Y_i)!} p_i^{Y_i} (1-p_i)^{N-Y_i}$

   This is likelihood of some single obs.

4. Model $p_i \Rightarrow \text{logit (or probit)},$ i.e. $p_i = \frac{e^{x_i \beta}}{1+e^{x_i \beta}}$
   $\Rightarrow f(Y_i \mid x_i, \beta, N) = \frac{N!}{Y_i! (N-Y_i)!} \left( \frac{e^{x_i \beta}}{1+e^{x_i \beta}} \right)^{Y_i} \left( \frac{1}{1+e^{x_i \beta}} \right)^{N-Y_i}$

5. Maximize this over choice of $\beta$
   $\Rightarrow \text{Interpretation just like logit}$
     \{$E(Y_i) = p_i$ and $E(Y_i) = Np_i$: since $p_i$ same for all individual Bernoullis in the group, just same interpretationally as logit$\}$

C. Counts of Uncorrelated Events (we don't know $N$ & $\beta$ there is no upper bound on how often the events can occur w/in one obs.)

Examples: # wars in a year
# medical consultations for each patient
# of anything in some time period or some group

Key here is: Events are occurring at some mean rate ($\lambda_i$ this is what we'll explain - the rate of event occurrence)
= we observe counts of "event occurs"

Assumptions:
- Within an observation, the rate of occurrence, $\lambda_i$, is constant
- No two events can occur in precisely the same instant

$Y_i$, the count of events, is Poisson Distributed

$\Pr(Y_i \mid \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{Y_i}}{Y_i!}$

$\lambda_i = e^{x_i \beta}$

Model: $\lambda_i = e^{x_i \beta}$

Joint Likelihood: $\prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{Y_i}}{Y_i!} = \prod_{i=1}^{n} \frac{e^{-x_i \beta} e^{x_i \beta Y_i}}{Y_i!}$

Computer finds the $\beta$ which maximizes this.
Dr. Extensions of Count Distributions

1. Grouped, correlated events w/possibly differing probabilities
   ⇒ Beta-Binomial, see King 5.5
2. Counts of uncorrelated events w/unequal observation intervals
   ⇒ Quasi-Poisson w/variable extrema, see King 5.8
3. Counts of correlated events
   ⇒ Generalized Event Count distribution, due to King, see 59

E. Duration Models (Greene 20.5) (See King, Alt, Buren, & Lawler for)

1) We are interested in explaining how long something lasts, or how long until something happens
   e.g. * Gov't Duration
       * Peace Duration (War Duration)
       * Unemployment Spell lengths (Employment spell lengths)
       * Strike lengths
       * Recovery lengths (Illness lengths)
       etc. (Comes from Physics/Chemistry/Bio on survival rates)

2) Notice that it's symmetric: can think of
   a) Hazard Rate: probability of an end to the ongoing
   b) Survival Function: probability process lasts some length
      • For Constant Hazard Rate, \( \lambda \), Expected Duration is \( \frac{1}{\lambda} \)

3) Model the Hazard Rate:
   \[ \lambda_i = \delta(X, \beta) \]
   i.e. Prob of ending in time \( t \) given by \( \psi(x) \)
   Then Survival function is:
   \[ S(t) = e^{-\lambda t} \]
   i.e. Probability of surviving to \( t \) or longer is \( e^{-\lambda t} \)

4) Model the Survival:
   p.d.f:
   \[ f(y_i = t) = \lambda e^{-\lambda t} = \lambda e^{-\lambda y_i} \]
   \[ \lambda = \text{hazard rate} \quad \text{E(Survival)} = \frac{1}{\lambda} = e^{\beta} \]
   \[ \text{p.d.f.} = \prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i} \]
   Likelihood = \[ \prod_{i=1}^{n} e^{-\lambda_i y_i} (e^{-\lambda_i y_i}) \]

⇒ This is what you interpret

⇒ This is what you maximize to find \( \beta \)
Polychotomous (Nominal, Categorical, $M \geq 2$) Dependent Variables

- $m$ Categories

- Model \( \frac{P_j}{p_m} = \frac{F(XB_j)}{1 - F(XB_j)} = G(XB) \) n.b. \( \frac{P_m}{p_m} = F(XB_m) = 1 \)

\[ \Rightarrow m-1 \text{ models of prob. category } j \text{ v. base } (m). \]

(n.b., for 2 categories \( \Rightarrow \) odds)

\[ \Rightarrow P_m = \frac{1}{1 + \sum G(XB_j)} \quad \text{b/c} \quad \sum_{j=1}^{m} \frac{P_j}{p_m} = \frac{1 - p_m}{p_m} = \frac{1}{p_m} - 1 \]

\[ \Rightarrow 1 + \sum \frac{P_j}{p_m} = \frac{1}{p_m} \]

\[ \Rightarrow P_m = \frac{1}{1 + \sum G(XB_j)} \]

\[ \Rightarrow P_j = \frac{G(XB_j)}{1 + \sum G(XB_j)} \]

let \( G(XB_j) = e^{XB_j} \) \( \Rightarrow \)

\[ P_j = \frac{e^{XB_j}}{1 + \sum e^{XB_j}} \quad \text{d}p_m = \frac{1}{1 + \sum e^{XB_j}} \]
Multinomial Logit:

\[ p_j = \frac{e^{x_{ij}}}{1 + \sum_{j=1}^{m} e^{x_{ij}}}; \quad p_m = \frac{1}{1 + \sum_{j=1}^{m} e^{x_{ij}}}, \]

Interprets set of logits cat. by cat. against base cat.

Likelihood Function:

\[ y_{ij} = \begin{cases} 1 & \text{if obs } i \text{ in category } j \\ 0 & \text{otherwise} \end{cases} \]

\[ L = \prod_{i=1}^{n} p_{i_1}^{y_{i1}} p_{i_2}^{y_{i2}} \ldots p_{im}^{y_{im}} \]

\[ \ln L = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \ln p_j \quad \text{(given above)} \]

\[ \frac{\partial \ln L}{\partial \beta_k} = \left\{ \begin{array}{l} \sum_{i=1}^{n} \frac{y_{ik} - p_{ik}(\sum_{j=1}^{m} y_{ij})}{p_{ik}} \times x_i \\ \sum_{i=1}^{n} \frac{y_{ik} - p_{ik}}{p_{ik}} \times x_i \\ \frac{1}{2} \sum_{i=1}^{n} (y_{ik} - p_{ik}) x_i \end{array} \right\} \]

\[ \frac{2 \ln L}{\partial \beta_k} \Rightarrow \text{normal equations again} \]