PS699: STATISTICAL METHODS II

The Classical (Normal) Linear Regression Model: Introduction and Basics

I. Statistical Modeling in Social Science

A. Preamble:

Positive 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. where 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 insofar as it tells us of the relationship btw y & $(X \& \varepsilon)$ i.e., it tell us the theorized (hypothesized) process producing outcomes y: The functional form.

   c. X is some (set of) variable(s) which are hypothesized to be related to, 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 $y = f(X, \cdot)$ is (generally) the systematic part of the world: The independent variable(s)

   d. ε 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: $y = f(X, \varepsilon)$

   (vote-share of incumbent) = f(state of the socio-economic world, competitor characteristics, chance)

   (duration of Parliamentary Gov't) = g(number parties in gov't, polarization of parties in gov't, scandals, chance)

B. An Alternative Preamble:

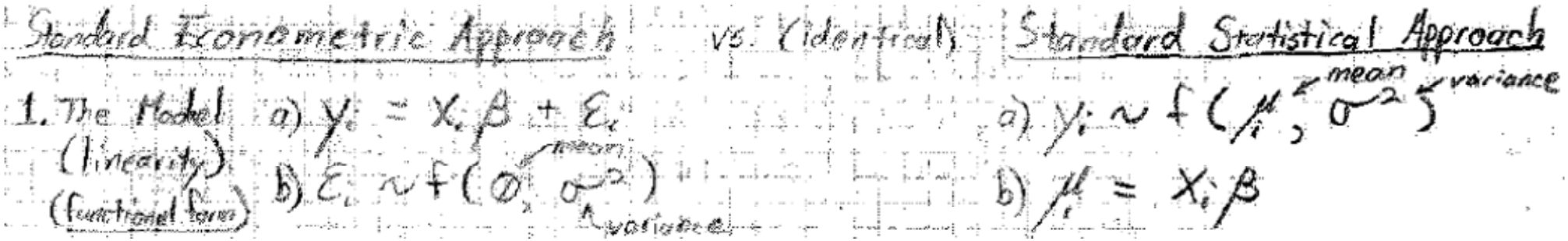
"A Language of Inference" (King, Unifying Political Methodology, 1.2)

1. Social System: the object under study (e.g. a neighborhood, a government, a market, a war, etc.)
   a. it's what the researcher wishes to know about
   b. contains features that are known or estimable & features which are unobserved & some unobservable
2. Explanatory Variable(s): measures of feature(s) 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 & are 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 or approximation to a more complex concept or social system. It is not even intended to replicate the social system in all its detail & complexity & in fact would not be terribly useful practically if it did.
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 (*sometimes* Normal) Linear Regression Model**, i.e., the **C(N)LRM**, is then a specific example of a statistical model.

A. Core Assumptions: (**X** full-column rank (no perfect colinearity) &)

1. $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$: Linearity (a.k.a., the model). Or $\mathbf{y} \sim \left(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n\right)$, $\boldsymbol{\mu} \equiv E(\mathbf{y}) = \mathbf{X\beta}$.



a) The dependent variable is a linear-additive function… (**<u>NOTE: ∀*i*</u>**).

(1) …*of the coefficients*, not necessarily linear in the regressors (leaves much flexibility)…

(2) …*with an additively separable stochastic component* (this actually the key for LS).

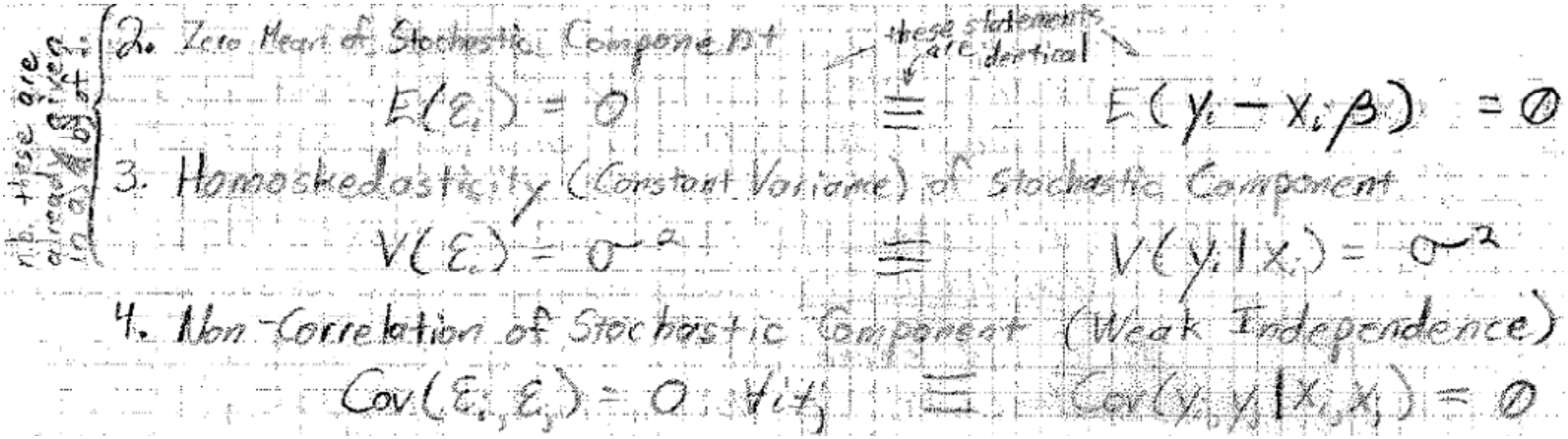
b) Note: Strong, from $y=f(\mathbf{x},\varepsilon)=f(x_1, x_2, \ldots, x_k, \varepsilon)$ to $y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k+\varepsilon$, but also considerably more flexible & capable than that may at first appear.

2. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$: Zero mean of the stochastic component. **NOTE: $\forall i$.**

3. Spherical error (*conditional*) var-covar structure: $V(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_{\varepsilon}^2 \mathbf{I}_n$.

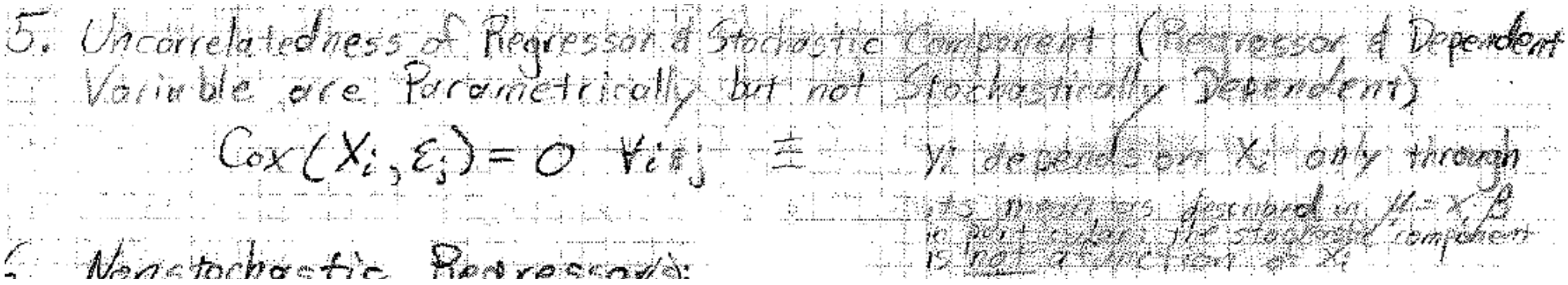
   a) Homoskedasticity constant *conditional*-variance: $V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_{\varepsilon_i}^2 = \sigma_{\varepsilon}^2 \ , \forall i$.

   b) Zero *conditional*-covariance: $C(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = \sigma_{ij} = 0 \ , \forall_{i \neq j}$.



4. Regressors exogenous (slightly diff. assumptions⇒slightly diff. properties):

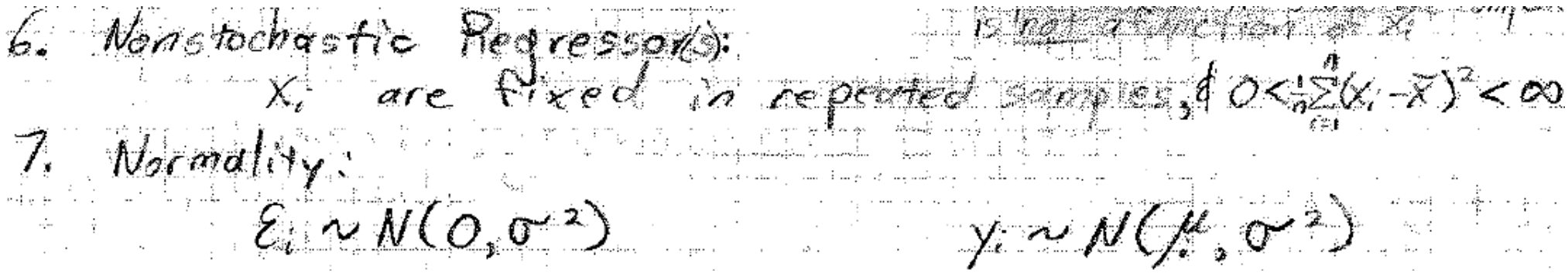
$$\text{plim } \mathbf{X}'\boldsymbol{\varepsilon}=0 \text{ or } E(\boldsymbol{\varepsilon}\,|\,\mathbf{X})=\mathbf{0} \text{ or } C(\mathbf{X},\boldsymbol{\varepsilon})=\mathbf{0}$$



B. Convenience Assumptions (Unnecessary, Add Properties):

1. **X** non-stochastic; i.e., fixed across repeated samples.

2. $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\sigma_\varepsilon^2\mathbf{I}_n)$. (If not: CLT & asymptopia! Necessary for OLS=MLE.)



III. The **C(N)LRM**:

A. Core Assumptions: (**X** full-column rank (no perfect colinearity) &)

1. $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$: Linearity (a.k.a., the model). Or $\mathbf{y} \sim \left(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n\right)$, $\boldsymbol{\mu} \equiv E(\mathbf{y}) = \mathbf{X\beta}$.

2. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$: Zero mean of the stochastic component. **NOTE: $\forall i$**.

3. Spherical error (*conditional*) var-covar structure: $V(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_\varepsilon^2 \mathbf{I}_n$.

   a) Homoskedasticity constant *conditional*-variance: $V(\varepsilon_i) = E\left(\varepsilon_i^2\right) = \sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2$, $\forall i$.

   b) Zero *conditional*-covariance: $C\left(\varepsilon_i, \varepsilon_j\right) = E\left(\varepsilon_i \varepsilon_j\right) = \sigma_{ij} = 0$, $\forall_{i \neq j}$.

4. Regressors exogenous (slightly diff. assumptions$\Rightarrow$slightly diff. properties):

$$\text{plim } \mathbf{X}'\boldsymbol{\varepsilon} = 0 \text{ or } E\left(\boldsymbol{\varepsilon} \mid \mathbf{X}\right) = \mathbf{0} \text{ or } C\left(\mathbf{X}, \boldsymbol{\varepsilon}\right) = \mathbf{0}$$

5. (*Convenience*) **X** non-stochastic; i.e., fixed across repeated samples.

6. (*Convenience*) $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n\right)$. (If not: CLT & asympt.! Nec. for OLS=MLE.)

IV. Common Issues/Challenges

A. The implicit parameter/model-homogeneity assumption in 1 (&2).

B. The spherical error-structure assumptions in 3.

C. The exogenous-regressors assumpt. 4; e.g. simultaneity, selection, etc.

1. LSDV with temporal-lag regressors has small-sample bias, of order $1/T$, (i.e., it is biased in small samples, but consistent (in T)).

2. Spatial-lag regressors correlate with residuals, so LS biased.

D. Notes:

1. In general, any misspecification (e.g., omitted variables, measurement err) can induce Cov($\mathbf{X}$,$\boldsymbol{\varepsilon}$) & so bias. I.e., key assumpt(s) for bias/consist. is (are) 1 & 4.

2. Typically, can restate violation 1 assumption also in terms of other assumpts.

E. Illustrations of assumptions & their violations:

1. Assumption 1: Linearity (really: "The Model")

a) Already noted any misspecification can induce some Cov($\boldsymbol{\varepsilon}$,**X**), implying bias.

b) More narrowly, as linearity, note that assumpt is ***linearity in parameters***: β, σ^2.

c) *Nonlinear in Variables* OK (i.e., OLS manageable w/in C(N)LRM) *vs.* Nonlin. in Parameters (Not OLS, maybe NLS; if inseparable stochastic component: MLE):

a. Non-linear in Variables) Examples: (non-linearity in X = OK)

$Y_i = \alpha + \beta X_i^2 + \varepsilon_i$ $\quad Y_i = \alpha + \beta \ln X_i + \varepsilon_i$ $\quad Y_i = A X_i^b \varepsilon_i$ (OK)

$\Rightarrow \ln Y_i = \ln A + b \ln X_i + \{\ln \varepsilon_i\}$

assumptions about residuals now refer to $\{\ln \varepsilon_i\}$

i. $\Rightarrow$ Really considerably more flexible than it first appears

ii. $\Rightarrow$ even when "truth" not linear we get "best" linear approximation to the "truth" "BLUAT"

b. Non-linear in Parameter(s) Examples: (non-linear in β = linear regression not gonna work)

$Y = \frac{\alpha}{\beta + X_i} - \frac{1}{\beta} X_i + \varepsilon_i$ $\quad \pi_i = \left[(1 + e^{-X\beta})^{-1}\right]^{Y_i} \left[1 - (1 + e^{-X\beta})^{-1}\right]^{1-Y_i}$

d) *Important Special Cases:*

(1) Log-log Models ⇔ elasticity models:

a. consider $\ln y_i = \beta \ln X_{ii} + \varepsilon_i$ (assume for the moment that the intercept is zero)

b. What is "effect" of X_i on y_i? (a total differential; I neglected to go into these in our calculus notes)

$$\frac{d \ln y_i}{d y_i} \cdot \frac{d y_i}{d x_i} = \beta \frac{d \ln x_i}{d x_i}$$

↖ Think of these as partial derivatives only everything (such as y_i) is potentially a function of x_i. (ε_i is not a function of X_i by assumption 5)

$$\Rightarrow \frac{1}{y_i} \cdot \frac{d y_i}{d x_i} = \beta \frac{1}{x_i}$$

$$\beta = \frac{d y_i / y_i}{d x_i / x_i}$$

- which in discrete notation/terms/form is like

$$\beta = \frac{\Delta y_i / y_i}{\Delta x_i / x_i}$$

} % [redacted] change of y

} % [redacted] change of x

- Thus, in a log-log model, β 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"

(2) Semi-Log Models: log 1 side only ⇒ % changes 1 side related to level changes other

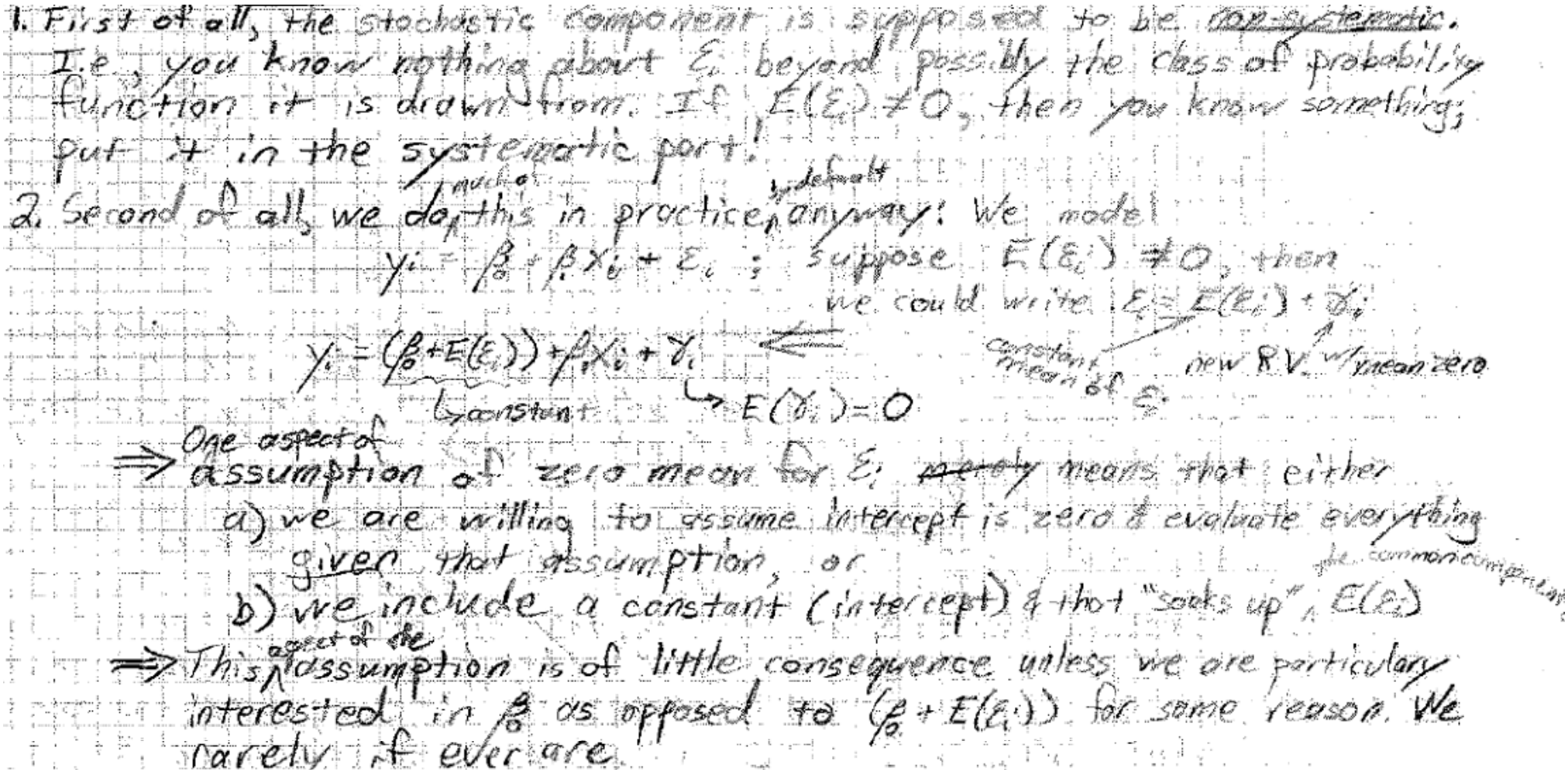
(3) Trans-Log Models (a very flexible form; 2nd-order expansion of log-log) ⇔ *ln* on both sides, but include also the interaction of all *ln*(**x**) on RHS.

(4) Many simple options, great flexibility; *key:* what does your theory imply?

2. Zero mean of stochastic component: $E(\boldsymbol{\varepsilon})=E(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})=\mathbf{0}$.

a) This subsumed by Assumption 1 in $E(\mathbf{y})=\mathbf{X}\boldsymbol{\beta}$ form. I.e., perhaps best thing to say about this assumption, if separately included, is that it just restates $E(\mathbf{y})=\mathbf{X}\boldsymbol{\beta}$.

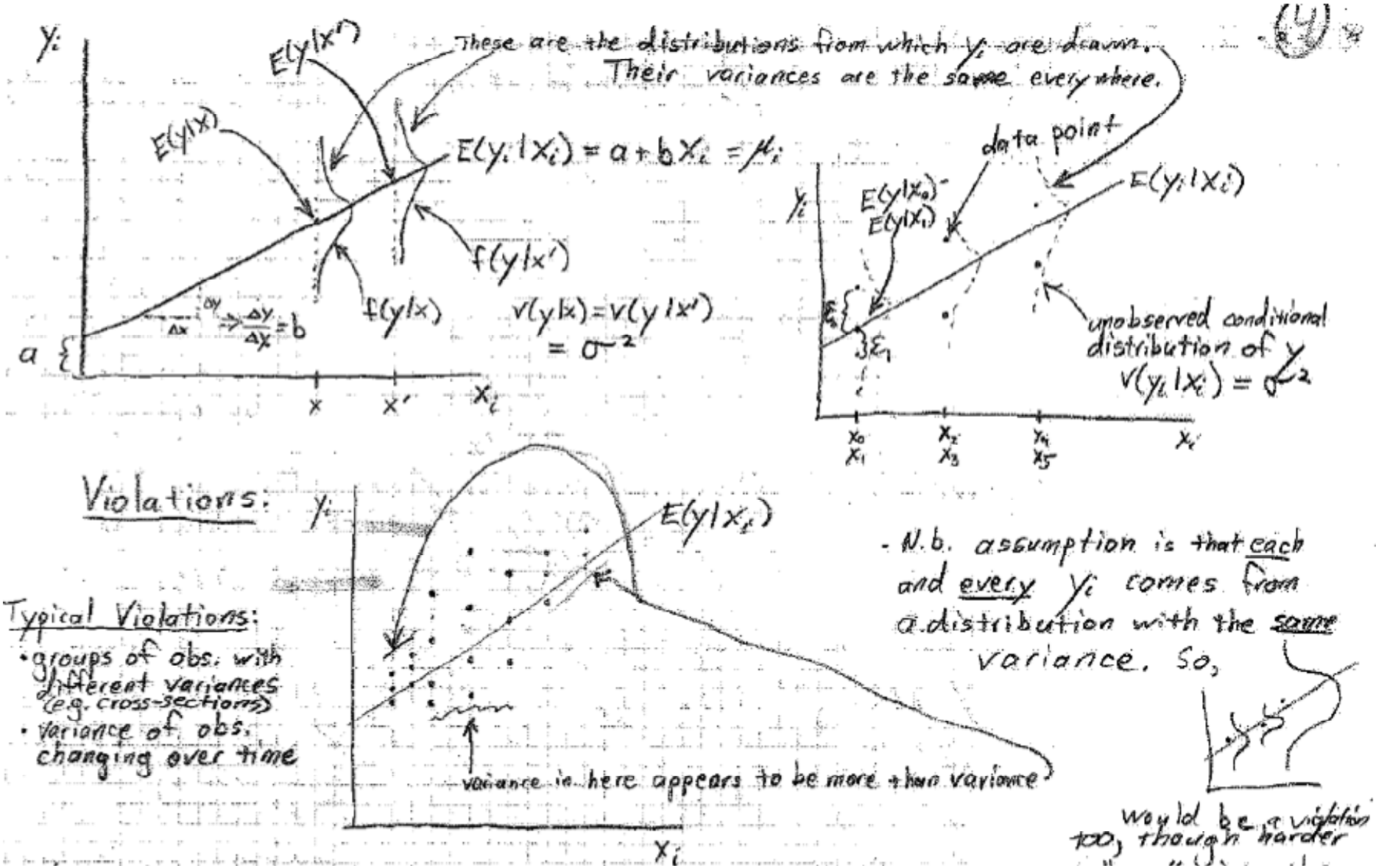
b) Some not perfectly helpful elaboration from many years ago, as adjusted over time to make it at least not-incorrect:



3. Spherical Error-Covariance Assumption & Violations: Heteroskedasticity

$$\text{Homoskedasticity: } V(y_i \mid \mathbf{X}) = V(\varepsilon_i \mid \mathbf{X}) = \sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon}$$

$$\text{Heteroskedasticity: } V(y_i \mid \mathbf{X}) = V(\varepsilon_i \mid \mathbf{X}) = \sigma^2_{\varepsilon_i} \neq \sigma^2_{\varepsilon}$$



Non-Correlation:

to "see" given the data we have

→ basic notion: knowing any ε_i would not help us predict what some other ε_j might be.

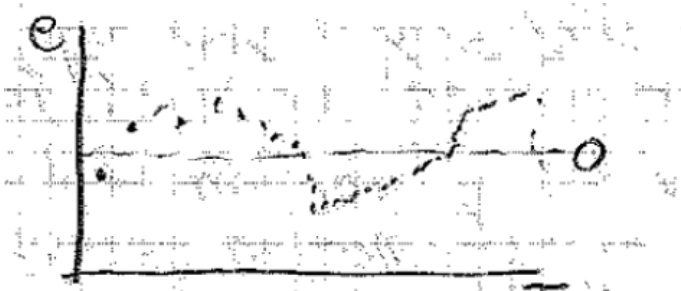
I.e., given X_i, there is no further information in $y_i - E(y_i|X_i) = y_i - X\beta = \varepsilon_i$ which would help us predict ε_j (some other observation's residual).

Typical Violations:

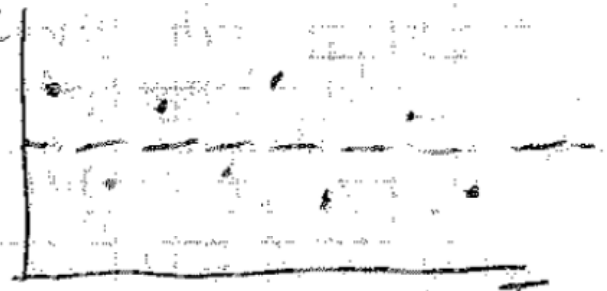
- Observations are ordered in time & something is "sticky" so that if ε_t is large then ε_{t+1} is more likely to be large, etc. (serial correlation)
- Observations are ordered in space (e.g. geographic proximity of observational units so that if ε_i is large then ε_j from an adjacent unit is likely to be large or small (contagion).

(Week Tobler's law)

Can draw (map) simple time (space) plot residuals to illustrate temporal (spatial) autocorr.:



This is strong positive correlation.



This is strong negative correlation (rare in most types of data)

E: (Weak Independence) Uncorrelatedness of X & ε / Parametric but not Stochastic Dependence of Y on X (this version encompasses C & D)

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. is constant across X_i). This includes, for example, that $V(y_i|X_i) = \sigma^2$ (constant) which is homoskedasticity. 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 (model), you'll be OK here. I.e., high X cannot be associated with high ε through Y;

Most typical Violations: a) $\{y_i \xrightarrow{\text{causes}} X_i \;\&\; X_i \rightarrow y_i\} \equiv$ "endogeneity" "simultaneity" "reverse causality"

- Forward-Looking Expectations will often imply $X = f(E(y), \cdot)$, which leads to

4. Common Violations $E(\boldsymbol{\varepsilon}|\mathbf{X})=\mathbf{0}$: Selection, omitted variables, other measure-*cum*-specify errs, autoregression of some types &/or under some conditions…
   a) E.g., *Mutual Causality*, $y \Leftrightarrow x$, & so the single-equation model is incomplete (Ass.1 violated), implying Covariance Regressor w/ Residual (violating Assumption 4):

$$\left.\begin{aligned} y &= \beta x + \gamma z + \varepsilon_y \\ x &= \theta y + \lambda w + \varepsilon_x \end{aligned}\right\} \Rightarrow \begin{cases} Cov(x,\varepsilon_y) = Cov(\varepsilon_y, \theta y + \lambda w + \varepsilon_x) = Cov(\varepsilon_y, \theta y) \\ \qquad = Cov(\varepsilon_y, \theta(\beta x + \gamma z + \varepsilon_y)) \approx Cov(\varepsilon_y, \theta\varepsilon_y) = \theta Var(\varepsilon_y) \\ Cov(y,\varepsilon_x) = Cov(\varepsilon_x, \beta x + \gamma z + \varepsilon_y) = Cov(\varepsilon_x, \beta x) \\ \qquad = Cov(\varepsilon_x, \beta(\theta y + \lambda w + \varepsilon_x)) \approx Cov(\varepsilon_x, \beta\varepsilon_x) = \beta Var(\varepsilon_y) \end{cases}$$

V. Deriving the OLS estimator for the C(N)LRM:

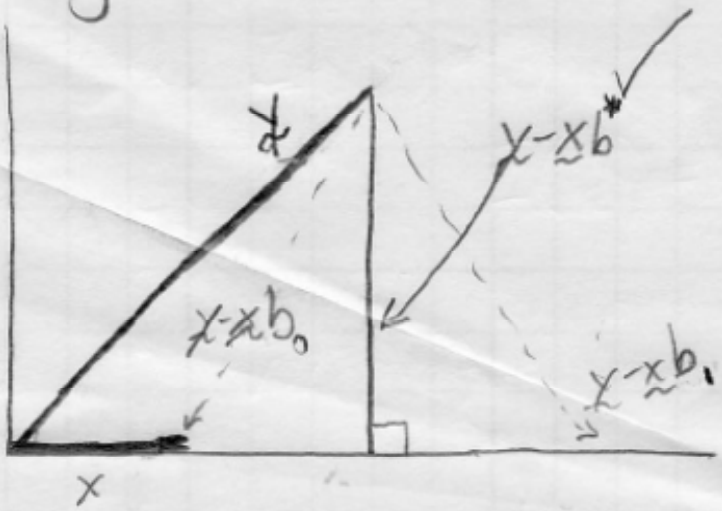
A. Recall 3 motivations give OLS estimator: Projection, MinSSE, maxLL

OLS: An Estimator for the C(N)LRM

CLRM: $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, $\underset{\sim}{\varepsilon} \sim (\underset{\sim}{0}, \sigma^2 I)$

1. Regression as a Projection Problem in Geometry:

→ Want choose β to scale information in X ↑↓ to get as close as possible to $\underset{\sim}{y}$. Like so:



$\underset{\sim}{y} = [6\ 6]'$ $\underset{\sim}{x} = [0\ 3]'$

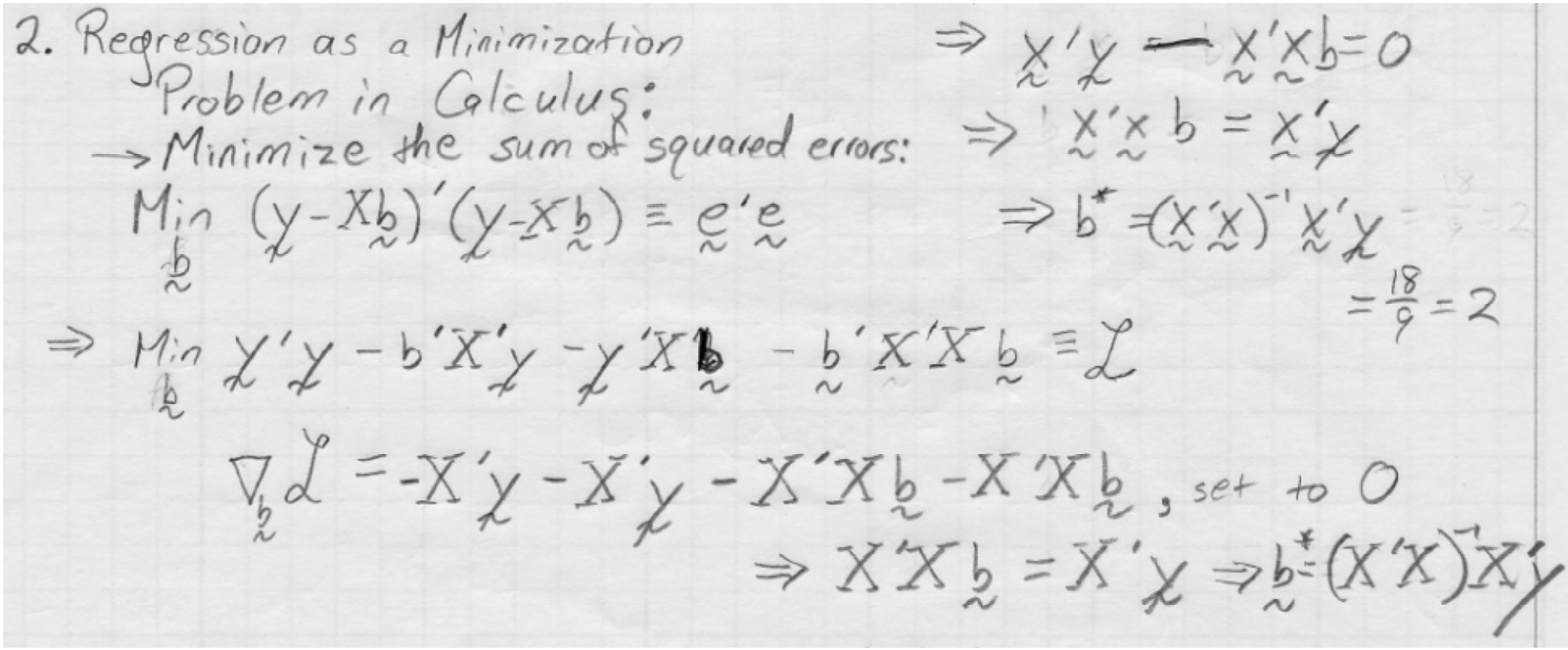
minimize $\underset{\sim}{y} - \underset{\sim}{X}b$

want:

$\Rightarrow (\underset{\sim}{y} - \underset{\sim}{X}b) \perp \underset{\sim}{X}$

$\Leftrightarrow \underset{\sim}{X}'(\underset{\sim}{y} - \underset{\sim}{X}b) = 0$

$$\Rightarrow \mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}b = 0 \Rightarrow \mathbf{x}'\mathbf{x}b = \mathbf{x}'\mathbf{y} \Rightarrow b^* = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}\{= 18/9 = 2\}$$



(Sign error: should be +**b′X′Xb**, then +**X′Xb**+**X′Xb**, to get last line.)

3. Regression as a ML problem in Statistics:

$y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = \alpha + \beta x_i$, & conditional independence

$$\Rightarrow f(y_i \mid x_i, \alpha, \beta) = (2\pi\sigma^2)^{-.5} e^{-\left\{\frac{(y_i-\alpha-\beta x_i)^2}{2\sigma^2}\right\}}$$

Maximize likelihood observing this set y_i & x_i given the model:

$$\mathscr{L}(\mathbf{y}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-.5} e^{-\left\{\frac{(y_i-\alpha-\beta x_i)^2}{2\sigma^2}\right\}} \quad ; \text{ (conditional indep. implies this)}$$

$$\ln \mathscr{L}(\mathbf{y}) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right\}$$

$$\frac{\partial \ln \mathscr{L}}{\partial \alpha} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\alpha}_{ML} - x_i\hat{\beta}_{ML}\right) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}\hat{\alpha}_{ML} = \sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}_{ML}\right) \quad \Rightarrow \hat{\alpha}_{ML} = \frac{\sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}_{ML}\right)}{n} = \overline{y} - \overline{x}\hat{\beta}_{ML}$$

$$\frac{\partial \ln \mathscr{L}}{\partial \beta} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\alpha}_{ML} - x_i\hat{\beta}_{ML}\right)x_i = 0 \quad \Rightarrow \quad \text{(using above also):}$$

$$\Rightarrow -\frac{2}{\sigma^2}\sum_{i=1}^{n}\left(y_i^* - x_i^*\hat{\beta}_{ML}\right)x_i^* = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}\left\{y_i^* x_i^* - (x_i^*)^2\hat{\beta}_{ML}\right\} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} y_i^* x_i^* = \sum_{i=1}^{n}(x_i^*)^2\hat{\beta}_{ML}$$

$$\Rightarrow \quad \hat{\beta}_{ML} = \frac{\Sigma y_i x_i}{\Sigma x_i^2} \quad = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = 18 \,/\, 9 = 2$$

$$\frac{\partial \ln \mathscr{L}}{\partial \sigma^2} = \sum_{i=1}^{n}\left\{-\frac{1}{2\hat{\sigma}^2_{ML}} + \frac{(y_i - \hat{\alpha}_{ML} - x_i\hat{\beta}_{ML})^2}{2\left(\hat{\sigma}^2_{ML}\right)^2}\right\} = 0 \quad \Rightarrow \quad \sum_i \frac{1}{2\hat{\sigma}^2_{ML}} = \sum_i \frac{(y_i - \hat{\alpha}_{ML} - x_i\hat{\beta}_{ML})^2}{2\left(\hat{\sigma}^2_{ML}\right)^2}$$

$$\Rightarrow \quad \frac{n}{2\hat{\sigma}^2_{ML}} = \sum_i \frac{(y_i - \hat{\alpha}_{ML} - x_i\hat{\beta}_{ML})^2}{2\left(\hat{\sigma}^2_{ML}\right)^2} \Rightarrow \quad n = \sum_i \frac{(y_i - \hat{\alpha}_{ML} - x_i\hat{\beta}_{ML})^2}{\hat{\sigma}^2_{ML}} \Rightarrow \quad \hat{\sigma}^2_{ML} = \frac{\Sigma(y_i - \hat{\alpha}_{ML} - x_i\hat{\beta}_{ML})^2}{n} = \frac{SSE_{ML}}{n}$$

VI. Return to simple, bivariate case in scalar algebra, more closely:

Given Assumptions 1–6 (linearity, $E(\varepsilon)=0$, $V(\varepsilon)=\sigma^2$, $E(\varepsilon_i\varepsilon_j)=0\ \forall i\neq j$, & no endogeneity, & X's fixed)

A. Least-Squares Estimation

$$y_i = \alpha + \beta X_i + \varepsilon_i$$

let's consider estimating α & β such by a & b:

$$y_i = a + b X_i + e_i$$

n.b.: e_i is an estimate of ε_i b/c a & b are only estimates of α & β

Scalar Equivalent:

$$\min_{a,b} \sum_{i=1}^{n} e_i^2 = \min_{a,b} \underbrace{\sum_{i=1}^{n} (y_i - a - b x_i)^2}_{SSR}$$

F.O.C.:

$$\frac{\partial SSR}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - b x_i)\cdot(-x_i) = 0$$

$$\Rightarrow -2\sum_{i=1}^{n} (y_i - a - b x_i)(x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} e_i x_i = 0$$

$$\frac{\partial SSR}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - b x_i)(-1) = 0$$

$$\sum_{i=1}^{n} e_i = 0$$

Scalar First-Order Conditions for Least Squares

from $\frac{\partial SSR}{\partial b} = 0$: $\sum_{i=1}^{n} e_i x_i = 0$

from $\frac{\partial SSR}{\partial a} = 0$: $\sum_{i=1}^{n} e_i = 0$

expand

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \Rightarrow \overset{(a)}{\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i}$$

expand

$$\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \Rightarrow \overset{(b)}{\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2}$$

(a) & (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 $n \Rightarrow$ $\bar{y} = a + b\bar{x}$

Conclusion 1: Least-Squares Regression Line ($\hat{y} = a + bx$) 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: $\sum x_i = n \cdot \bar{x}$ (why? $\bar{x} \equiv \frac{1}{n}\sum x_i$, so $n \cdot \bar{x} = n \cdot \frac{1}{n}\sum x_i = \sum x_i$)

$$\Rightarrow \sum x_i y_i = (\bar{y} - b\bar{x}) n\bar{x} + b \sum x_i^2$$

$$\Rightarrow \sum x_i y_i = \bar{y} n \bar{x} - b n \bar{x}^2 + b \sum x_i^2$$

$$\Rightarrow \sum x_i y_i - n\bar{x}\bar{y} = b\left(\sum x_i^2 - n\bar{x}^2\right)$$

you're done here but it's more conveniently expressed another way

$$\Rightarrow (i)\ b = \left(\sum x_i y_i - n\bar{x}\bar{y}\right) \Big/ \left(\sum x_i^2 - n\bar{x}^2\right)$$

sample covariance of x & y

$$\Rightarrow b = \sum_{i=1}^{n} (x_i y_i - \bar{x}\bar{y}) \Big/ \sum_{i=1}^{n} (x_i^2 - \bar{x}^2)$$

Conclusion 3: $$(ii)\ b = \sum_{i=1}^{n} \{(x_i - \bar{x})(y_i - \bar{y})\} \Big/ \sum_{i=1}^{n} (x_i^2 - \bar{x}^2) = S_{xy} / S_{xx}$$

S_{xx}: sample variance of x

Aside: How do we get from (i) to (ii)

$$\sum\{(x_i - \bar{x})(y_i - \bar{y})\} = \sum\{x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\bar{y}\}$$

$$= \sum\{x_i y_i\} - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y}$$

$$= \sum\{x_i y_i\} - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y}$$

$$= \sum\{x_i y_i\} - n\bar{x}\bar{y} \quad \square$$

III. B. Given Assumptions 1-6 (linearity, $E(\varepsilon)=0$, $\text{cov}(\varepsilon_i, \varepsilon_j)=0$, $V(\varepsilon)=\sigma^2$ & no endogeneity & x's fixed)

Statistical Properties of the LS Estimator

$$b = S_{xy}/S_{xx} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{S_{xx}}$$

$$= \sum_i c_i (y_i - \bar{y}) \quad \text{where } c_i \equiv \frac{x_i-\bar{x}}{S_{xx}}$$

$$= \frac{\sum (x_i-\bar{x})(\alpha+\beta x_i+\varepsilon_i-\bar{y})}{S_{xx}}$$

$$= \frac{\alpha\sum(x_i-\bar{x})}{S_{xx}} + \frac{\beta\sum\{(x_i-\bar{x})x_i\}}{S_{xx}} + \frac{\sum\{(x_i-\bar{x})\varepsilon_i\}}{S_{xx}} - \frac{\{\sum(x_i-\bar{x})\}\bar{y}}{S_{xx}}$$

aside #1:
$\sum(x_i-\bar{x}) = \sum x_i - n\bar{x} = \sum x_i - n\cdot\frac{1}{n}\sum x_i = \sum x_i - \sum x_i = 0$

$$= 0 + \frac{\beta\sum\{x_i^2-\bar{x}x_i\}}{S_{xx}} + \frac{\sum\{x_i\varepsilon_i-\bar{x}\varepsilon_i\}}{S_{xx}} - 0$$

aside #2:
$S_{xx} = \sum(x_i-\bar{x})^2 = \sum\{x_i^2 - 2x_i\bar{x}+\bar{x}^2\} = \sum\{x_i^2 - \bar{x}x_i\} = \sum\{x_i^2 - \bar{x}x_i\} = \sum\{(x_i-\bar{x})x_i\}$

$$= \beta\cdot\frac{S_{xx}}{S_{xx}} + \sum_{i=1}^{n} c_i\varepsilon_i$$

$$\Rightarrow \boxed{b = \beta + \sum_{i=1}^{n} c_i\varepsilon_i}$$

so b differs from true parameter β because ε_i are stochastic. (& varies)

$\boxed{\text{Property 1:}}$ $E(b) = E\left(\beta + \sum_{i=1}^{n} c_i \varepsilon_i\right)$

$$= \beta + \sum_{i=1}^{n} E(c_i \varepsilon_i)$$

recall, $c = f(X_i)$ & the X_i are fixed across repeated samples (i.e. constant) so may be brought out of $E(\cdot)$

$$= \beta + \sum_{i=1}^{n} c_i E(\varepsilon_i)$$

notice key step applies $Cov(X, \varepsilon) = 0$

n.b., but fixed X_i stronger than needed

. but $E(\varepsilon_i) = 0$, so

$$E(b) = \beta$$

$\boxed{\cdot\ b_{LS} \text{ is an unbiased estimate of } \beta}$

$\boxed{\text{Property 2:}}$ $V(b_{LS}) = V\left(\beta + \sum_{i=1}^{n} c_i \varepsilon_i\right)$

$$= V(\underset{\text{constant}}{\beta}) + 2\, Cov\left(\underset{\text{constant}}{\beta}, \sum_{i=1}^{n} c_i \underset{\text{R.V.}}{\varepsilon_i}\right) + V\left(\sum_{i=1}^{n} c_i \varepsilon_i\right)$$

$\Rightarrow V(\,) = 0$ $\qquad$ $\Rightarrow Cov = 0$

$$= V\left(\sum_{i=1}^{n} c_i \varepsilon_i\right)$$

X_i is fixed, $Cov(\varepsilon_i, \varepsilon_j) = 0$ both by CLRM assumptions

$$= \sum_{i=1}^{n} c_i^2 V(\varepsilon_i) = \sum_{i=1}^{n} c_i^2 \sigma^2 = \frac{\sum (X_i - \bar{X})^2}{(S_{xx})^2} \sigma^2 = \frac{S_{xx}}{(S_{xx})^2} \sigma^2$$

$$\boxed{V(b_{LS}) = \frac{\sigma^2}{S_{xx}}}$$

Given Assumptions 1-6, we have established that

$$E(b_{LS}) = \beta \iff b_{LS} \text{ is unbiased est. of } \beta$$

$$V(b_{LS}) = \sigma^2 / S_{xx}$$

Now, $E(a_{LS}) = \alpha \iff a$ is also an unbiased est. of α

Proof

$$E(a_{LS}) = E(\bar{y} - b\bar{x}) \quad \text{by def. of } a_{LS}$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} y_i - b\bar{x}\right) \quad \text{by def. of } \bar{x}$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n}\{\alpha + \beta x_i + \varepsilon_i\} - b\bar{x}\right) \quad \text{by Ass. 1: } y_i = \alpha + \beta x_i + \varepsilon_i$$

$$= \frac{1}{n}\sum_{i=1}^{n}\alpha + \frac{1}{n}\beta\sum_{i=1}^{n}E(x_i) + \frac{1}{n}\sum_{i=1}^{n}E(\varepsilon_i) - E(b\bar{x}) \quad \text{working the sum \& the } E(\,) \text{ out}$$

$$= \alpha + \beta\mu_x + 0 - \beta\mu_x = \alpha \quad \square$$

$$V(a_{LS}) = V(\bar{y} - b\bar{x})$$

$$= V(\bar{y}) - 2\,Cov(\bar{y}, b\bar{x}) + V(b\bar{x})$$

$$= \frac{\sigma^2}{n} - 2Cov(a + b\bar{x}, b\bar{x}) + \bar{x}^2 V(b)$$

b/c x's & $\therefore \bar{x}$ fixed across repeated samples

$$= \frac{\sigma^2}{n} - 2\left(Cov(a, b\bar{x}) + V(b\bar{x})\right) + \bar{x}^2 V(b)$$

by Cov. rules

$$= \frac{\sigma^2}{n} - 2\bar{x}\,Cov(a, b) - 2\bar{x}^2 V(b) + \bar{x}^2 V(b)$$

don't covary

$$= \frac{\sigma^2}{n} - 2\bar{x}\,Cov(\bar{y} - b\bar{x}, b) + \bar{x}^2 V(b)$$

$\bar{y}$ & b don't covary b/c the average d y is "netted out" by the intercept, a when the slope, b is estimated

$$= \frac{\sigma^2}{n} + 2\bar{x}^2 V(b) - \bar{x}^2 V(b)$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \cdot V(b)$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}}$$

$$\Rightarrow V(a_{LS}) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

cov(a, b.

$\Rightarrow$ 4. Three Facts About $V(B_{ls}) = V\begin{bmatrix} a_{ls} \\ b_{ls} \end{bmatrix} = \begin{bmatrix} \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) & \\ & \sigma^2/S_{xx} \end{bmatrix}$ (so far)

cov(a, b) in diagonals

(ceteris paribus) Variance of Coefficient estimates...

bigger ones:
- a) ... is higher the larger is σ^2 (the more variant the stochastic component of the world)
- b) ... is lower the larger is S_{xx} (the greater the variation of X in your sample)

not so important:
- c) ... specifically of a_{ls}, is lower as $\bar{X} \to 0$ (as sample average of X tends to zero)

Things to Note: true $V(b) = f(\text{Sample Variation in } X, \sigma^2)$ $\Rightarrow g(N)$

estimated $V(b) = h(\text{sample var. in } X, s^2)$ $\Rightarrow g(N)$ (s^2: estimate)

So, under the CLRM, we have:

$$\mathbf{b}_{ls} \equiv \begin{bmatrix} a_{ls} \\ b_{ls} \end{bmatrix} \sim \left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}}\right) & -\frac{\bar{x}}{S_{xx}} \\ -\frac{\bar{x}}{S_{xx}} & \frac{\sigma_\varepsilon^2}{S_{xx}} \end{bmatrix} \right)$$

Notice also that we have unbiased parameter estimates with their variance going to zero, i.e., we have also established consistency. So, what about efficiency (& so BLUE)…?

III. C. 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 $b_{LS} = \sum_{i=1}^{n} C_i y_i$ ← the $\bar{Y}$ part of $\sum C_i(Y_i - \bar{Y})$ drops out, remember?)

To be a valid estimator, it must be some function of the data, y_i's; it must be a function of known quantities (X_i, y_i) not unknowns (ε_i).

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 & not ε_i's.

(notational aside: b' is another way of writing "some $X_i \neq b$" from X_i here is just some other estimate of b)

$$b' = \alpha \sum q_i + \beta \sum q_i X_i + \sum q_i \varepsilon_i$$

Now, for b' to be unbiased, $E(b')$ must equal β

$$\Rightarrow^{(i)} \alpha \sum E(q_i) + \beta \sum E(q_i X_i) + \sum E(q_i \varepsilon_i) = \beta$$

$\Rightarrow q_i$'s are function of X_i's which are fixed, so $E(q_i) = q_i$ & $E(q_i X_i) = q_i X_i$

so, we must have

$$\sum q_i = 0 \quad \& \quad \sum q_i X_i = 1 \quad \& \quad E(q_i \varepsilon_i) = 0$$

for $E(b') = \beta$ to be true so that b' is a permissable alternative to b_{LS}.

So $V(b_{LS}) = \sigma^2 \sum c_i^2$

$V(b') = \sigma^2 \sum q_i^2$

& note $\sum c_i = \sum q_i = 0$ from

let $v_i \equiv q_i - c_i$ so that $q_i = c_i + v_i$ (just defining v_i as the difference between q_i & c_i - i.e. between the linear parts of the b_{LS} & b' estimators)

then $V(b') = \sigma^2 \left(\sum c_i^2 + \sum v_i^2 + 2 \sum c_i v_i \right)$

$= 0$ by orthogonality since $v_i \equiv q_i - c_i$ the c_i was already netted out

so, $V(b') = \sigma^2 \left(\sum c_i^2 + \sum v_i^2 \right)$

↳ this ≥ 0 since anything squared ≥ 0, □

i.e. $V(b') \geq V(b_{LS})$ for any

b' is a linear function of the data

& an unbiased estimator of β.

• we can do a similar but much more tedious proof for a.

This is the Big Bang, so let me re-iterate:

if:

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 $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$ (no covariance of ε's)

⑤ $Cov(X_i, \varepsilon_j) = 0 \quad \forall i \neq j$ (y_i depends on X_i only in its mean (μ))

⑥ X_i is nonstochastic (fixed in repeated samples)

(most not as restrictive as they seem & some overlap a bit (i.e. imply each other) or partially apply)

Then: $B \equiv (X'X)^{-1}(X'y) \equiv \begin{bmatrix} a_{LS} \\ b_{LS} \end{bmatrix}$ is the BLUE estimator (Best Linear Unbiased) of $\underset{\sim}{\beta} \equiv \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$

* Notice that no assumptions about the exact shape of $f(y)$ need to be made, just statements (2-5) at this point.

Statistical Inference under the CLRM:

- Gauss-Markov is true w/o Normality; however, if we also have normality, then Assumption 1: That $y_i = \alpha + \beta x_i + \varepsilon_i$, i.e. linearity is unnecessary:

Under Assumptions 2–7 (normality included, linearity unnecessary), the Best Unbiased Estimator (BUE) happens also to be linear.

- For moving on to statistical inferences regarding LS estimates, it is useful to begin with the normality assumption included.

Then, a & b, which are linear functions of non-stochastic X & of Normally distributed y (or $f(x)+\varepsilon$ as you like) must also be normally distributed. (A property of normal R.V.'s we noted before.)

So, if $\varepsilon \sim N(0, \sigma^2)$ or equivalently $f(y|x)$ is normal mean $X\beta$, variance σ^2, then: $a \sim N(\alpha, \sigma^2(\frac{1}{n} + \bar{x}^2/S_{xx}))$

$b \sim N(\beta, \sigma^2/S_{xx})$

or, more completely:

all of this we've already shown

$V(a)$

$$B \equiv \begin{bmatrix} a \\ b \end{bmatrix} \sim N\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma^2(\frac{1}{n} + \bar{x}^2/S_{xx}) & -\bar{x}/S_{xx} \\ -\bar{x}/S_{xx} & \sigma^2/S_{xx} \end{pmatrix}\right]$$

$\uparrow$ $\mathrm{Cov}(a,b)$ $\qquad$ $\uparrow$ $V(b)$

So, if we're 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:

$$T = \frac{b-\beta}{(\sigma/\sqrt{S_{xx}})} = \frac{(b-1)}{(\sigma/\sqrt{S_{xx}})} \sim N(0,1) \text{ under the null,}$$

2. but we don't know σ^2 (the true variance of the stochastic component)

$\Rightarrow$ use an estimate

One possible estimate: note: $E(e_i) = E(\varepsilon_i)$, so $\hat{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2$ seems a logical estimate of $E(\varepsilon_i^2) \equiv \sigma^2$

Problem: exactly analogous to previous results,

$$E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2 \quad \text{i.e. } \hat{\sigma}^2 \text{ biased by factor of } \left(\frac{n-2}{n}\right)$$

- where does bias come from?

$$e_i = Y_i - \underset{\text{estimated}}{a} - \underset{\text{estimated}}{b} X_i$$

$\Rightarrow$ 2 degrees of freedom used up already

$\Rightarrow$ Just as before, $S^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$ is an unbiased estimate

$\Rightarrow$ Just as before, $\left(\frac{b-\beta}{\sqrt{S^2/S_{xx}}}\right) \sim t_{n-2}$

{ why? $(b-\beta) \sim N(0,1)$

S^2/S_{xx} is χ^2 divided by its degrees of freedom

(χ^2 is sum of squared normals, $\varepsilon_i \sim N(0,\sigma^2)$

i.e. $(b-\beta)/(\text{estimated std error of } b) \sim t_{n-2}$

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) $Pr[b - t \cdot s(b) \leq \beta \leq b + t \cdot s(b)] = 1 - \alpha$

$\Rightarrow$ read t_{n-2} for $(1-\alpha)$ level

Then substitute your estimates for b & $s(b)$ into (1) & you're done.

2. What about a confidence interval (+est) for σ^2?

from before $\frac{s^2}{\sigma^2/n}$ where n = degrees of freedom of the estimate s^2 is distributed χ^2_n

so, $\frac{s^2}{(\sigma^2/n-2)} \sim \chi^2_{n-2}$

$\Rightarrow$ a 95% confidence interval:

$$\frac{(n-2)\cdot s^2}{\chi^2_{n-2}(.975)} \text{ to } \frac{(n-2)\,s^2}{\chi^2_{n-2}(.025)}$$

$\chi^2_{n-2}(.975)$: value from χ^2_{n-2} which has .975 probability to the right of it.

$\Rightarrow$ Hypoth. Test:

$T = \frac{(n-2)\,s^2}{\sigma^2}$, check T on a χ^2_{n-2} table (or computer program)

σ^2: null hypothesis

V. The "Fit" of a Regression:

A. Variance Decomposition:

1.) Total Sum of Squares: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ {the sample variation in y}

a) $y_i - \bar{y} = (\hat{y}_i + e_i) - \bar{y}$

$= x_i b + e_i - (\bar{x} b) + a - a$

$= b(x_i - \bar{x}) + e_i$

$\Rightarrow$ b) $\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SST} = \sum_{i=1}^{n} [(\hat{y}_i - \bar{y}) + e_i]^2 = \underbrace{\sum (\hat{y} - \bar{y})^2}_{SSR} + \underbrace{\sum e_i^2}_{SSE}$

regression sum of squares

error sum of squares

$= \sum (b(x_i - \bar{x}))^2 + \sum e_i^2$

$= b^2 \sum (x_i - \bar{x})^2 + \sum e_i^2$

2. $SST = SSR + SSE$

B. 1.) $R^2 = \{\text{Variation in } Y \text{ fit by the Regression}\} / \{\text{Total Variation in } Y\} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

n.b. Notation across authors of Total Sum Squares (TSS) etc. is hopelessly erratic. Too late to do anything about that now. So, be sure you understand don't try to memorize. (Full use Greene's notation)

R^2 (Continued)

2) • $R^2 = \frac{SSR}{SST} = \frac{b^2 S_{xx}}{S_{yy}} = \frac{(S_{xy}/S_{xx})^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{yy}S_{xx}} = r_{x,y}^2$

($b_{LS} = S_{xy}/S_{xx} = \text{SampleCov}(X,Y)/S.Var(X)$ by def of S_{xx}; by def of S_{yy}; sample correlation^2)

3) • More Generally: $R^2 = (r_{\hat{y},y})^2$

R^2 F-Tests:

$SSR = b^2 S_{xx}$ has 1° of freedom (only b varies across samples)

$SSE = \sum_{i=1}^{n} e_i^2$ has $(n-2)$° of freedom (each e_i varies across samples, but estimating a & b ties two of them down)

$SST = S_{yy}$ has $(n-1)$° of free. (each y_i varies across samples, but estimating $\bar{y}$ ties one of them down)

) $\Rightarrow \frac{(SSR/1)}{(SSE/n-2)} \sim F_{1,n-2}$

) $\therefore \frac{R^2/1}{(1-R^2)/n-2} = \frac{(SSR/SST)}{(SSE/SST)/n-2} = \frac{SSR}{(SSE/n-2)} \sim F_{1,n-2}$

is a test of the regression model, $y = a + bX + e$

against the alternative $y = a + e$

$\Rightarrow$ in the bivariate case, exactly equivalent to a t-test on $\beta = 0$

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? No

1. $R^2 = SSR/SST$ & is therefore an estimate of the proportion of the world (as 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 cumbrous a model to carry around that would account for all of it. That is, some portion of the world is stochastic, & we don't know how much

$\Rightarrow R^2$ is something you estimate not something you maximize

2. $S = \left(\frac{1}{n-2}\sum e_i^2\right)^{.5}$ is also called the "standard error of the estimate" or "... of the regression" (SEE or SER) b/c it is an estimate of σ^2 the (true) error of the regression model.

- Obviously, this too is something you estimate not something you minimize. & Notice, $R^2 \approx 1 - \frac{S^2}{(SST/n)}$ $\left(\bar{R}^2 = 1 - \frac{S^2}{(SST/(n-1))}\right)$

R^2 (continued) So what good is it?

1. First of all, 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 β (the relationships b/w the X's & the dependent variable(s)).

2. Second, as we've just seen, we can test hypotheses about coefficients (it's more useful when the hypoth. 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

All of these tests & confidence intervals, etc. assumed normality, what if $f(y|x)$ is not normal (i.e. $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$)?

- Answer: Central Limit Theorem

→ b is linear combo of (large) (hopefully) numbers of random variables (y_i's), & so...

$$b \overset{A}{\sim} N(\beta, \sigma^2/S_{xx})$$ (asymptotically)

whatever the actual distribution of y_i

⟹ All of our discussion about "how large is large" & "when 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 O.L.S.

Prediction: Sometimes, we use OLS to make predictions about y_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 (a & b).

A. We have estimated $y_i = a + bX_i + e_i$ & now we get some new data X_j we want to use to "predict" the $y_{j\text{th}}$ observation.

1. What's the prediction error? ($\hat{a}, \hat{b}$ estimated from the i's)

$$e_j = y_j - \hat{y}_j = y_j - a - bX_j$$

$$= (\alpha + \beta X_j + \varepsilon_j) - a - bX_j$$

$$= (\alpha - a) + (\beta - b)X_j + \varepsilon_j$$

2. Thus two sources of error:
   a) One because a & b are only estimates of α & β
   b) One because there is a natural stochastic component to the world

Prediction Intervals: $V(e_j) = V\{(\alpha - a) + (\beta - b)X_j + \varepsilon_j\}$

$$= V(a) + X_j^2 V(b) + V(\varepsilon_j) + 2\,Cov(a,b)\cdot X_j$$

OLS assumptions $\Rightarrow$ $Cov(a,\varepsilon)$ & $Cov(b,\varepsilon) = 0$

$$= \sigma^2\left(1 + \frac{1}{n} + (X_j - \bar{X})^2 / S_{xx}\right)$$

So, we can test statements such as "given our model, the outcome in this new Y_j observation was 'surprising'" where that last is taken to mean beyond some specified confidence interval.

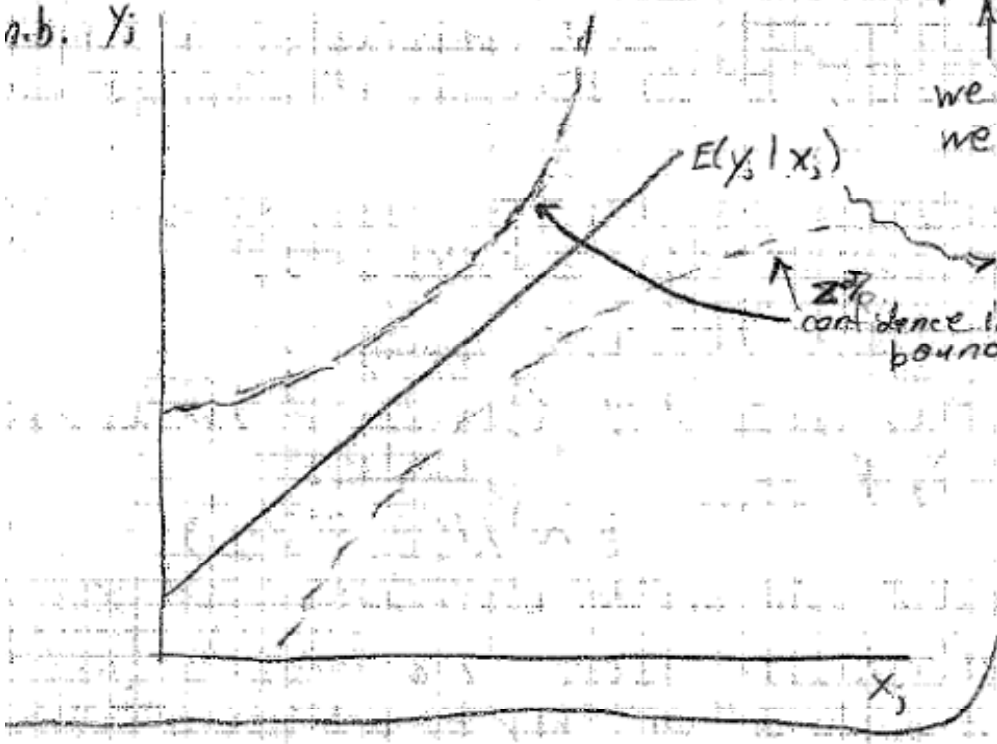
Conf. Intervals:

for $\hat{Y}_j = (a + bX_j)$

critical level corresponding to the size of the confidence interval

$$(a + bX_j) \pm t_{n-2} \cdot \sqrt{s^2\left(1 + \frac{1}{n} + \frac{(X_j - \bar{X})^2}{S_{xx}}\right)}$$

we don't know σ^2, so we use its estimate



(ceteris paribus)

n.b. Just like with coefficients

$\uparrow S_{xx}$ (sample variation in X) $\Rightarrow \downarrow$ forecast error variance

$\uparrow s^2$ (sample variance of e_i) $\Rightarrow \uparrow$ forecast error variance

$\uparrow n$ (sample size used to est. a & b) $\Rightarrow \downarrow$ forecast error variance

& one new thing (which we'll see again)

$\uparrow (X_j - \bar{X}) \Rightarrow \uparrow$ forecast error variance

(farther the new observation is from the mean of those used to estimate a & b)

E. 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, & we often simply do not know enough to go beyond linear approximation & bounded ($< \infty$) variance.
   - Still, sometimes we do know more, sometimes a lot more, & sometimes even the best linear approximation is still a crummy one b/c the relationship is very non-linear.

C. When does OLS do poorly?

all of these can be dealt with

1. Since $\min_{\underset{\sim}{\beta}} (y - x\underset{\sim}{\beta})^2$ ← squared, it 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. When things are highly non-linear
3. When assumptions 3-5 & esp. last two of these violated

VII. Multivariate Regression:

A. Preamble…

A Refresher on the Point of All This... A Positivist Philosophy... Some Thoughts on Evidentiary Analysis...

General Background: A. As always, our goal is to use history (what has happened) as a data base from which to infer the utility & (partial) validity of our theories & models about relationships b/w 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. (The only decidedly rare exception is the occasional experimental analysis.)

1. Further, 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 logical issues regarding this measurement & this 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 and all the empirical issues regarding measurement that relate to that statement.

3. Statistical Inference, then, has the advantage of forcing greater explicitness about these necessary assumptions & of providing consistent, proven, & logically immutable rules for making evidential inference given the data. A large part of its "weakness" is often said to be this very rigidity in its rules & the "over-simplicity" of its usually maintained assumptions, but...

a) "rigidity" in the rules of inference (& in parallel in the rules of logic) is paramount to objectivity & self-evidently crucial to the scientific use of evidence. We may suggest different sets of rules but no one could seriously propose a set of rules that will not be followed; rather, it is said, the "problem" is 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. If, for example, we are unwilling to assume that each "observation" or event in some slice of history are 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 those events & 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 present & usually more simplistic (e.g. of the form: this case is representative & conclusive — the latter implying no other case is representative &/or 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 inference is useless without accompanying interpretation. Correctly establishing an empirical relationship that has held helps 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 & 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.

Up Shots: ① There is no quantitative/qualitative divide (may be a positivist/interpetivist 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 fruitful.)

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 inferred relationships.

--→ subpoint: interpretation is a theoretical exercise; as such, what is interpreted in one analysis as regards some inferential results may itself be subjected in subsequent study to evidentiary analysis

⑤ We see then that this course fits into our broader training at any number of points

→ most obviously, we are getting tools for (statistical) inference

→ we are thereby learning indirectly about the broader topic of making correct, & efficient use of evidentiary analysis
↳ unbiased, consistent ↳ min. Var, min. M.S.E.

these lessons apply to quant. & qual. positive analysis alike (see King, Rebhare, & Verba for example)

→ we are also (or should be) learning by doing as regards interpretting our inferences for theoretical import.

A parallel course might, for example, focus on substantive interpretation, but in so doing shed light on the exercise of evidentiary analysis

(Sorry, got side-tracked a while here, enough of this & back to the business at hand)

A Refresher on the Point of All This... A Positivist Philosophy... Some Thoughts on Evidentiary Analysis...

C. Multiple Regression is simply the straightforward extension of the C(N)LRM in the bi-variate case (last week) to the case where several independent variables (X's) are related to (a) dependent variable(s), (Y).

1. As before, our model of the part of the world in which we're interested begins with

$$Y = f(X, \varepsilon)$$

Y: some outcome of interest; is; f: some function of...; X: some set of variables (explanatory), the systematic part of the world; ε: some (set of) random elements, the stochastic part 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, a) 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, & d) (most important) how to interpret these estimates -- what do they mean.

4. We then "clean up" with a few remaining issues.

The (Multivariate) Classical (Normal) Linear Regression Model

A. Assumptions (all are perfectly analogous to last week's) CLRM

1. $\underset{(n\times 1)}{\underset{\sim}{Y}} = \underset{(n\times k)(k\times 1)}{X\underset{\sim}{\beta}} + \underset{(n\times 1)}{\underset{\sim}{\varepsilon}}$
2. $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$
3. $V(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$
   $(\equiv E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'))$
4. $E(\underset{\sim}{\varepsilon}|X) = \underset{\sim}{0}$
5. X is non-stochastic & of full column-rank
   (columns of X not perfectly linearly dependent)

1. $\underset{\sim}{Y} \sim f(X, \underset{\sim}{\varepsilon})$
2. $E(\underset{\sim}{Y}|X) = X\beta$
3. $V(\underset{\sim}{Y}|X) = \sigma^2 I_n$
4. $\underset{\sim}{Y} = f(\cdot)$ depends on X only in its mean
   $\{f_{Y|X} - E(Y|X) = f_Y - E(Y|X)\}$
5. Same

B. What do they mean? (n = # of observations, 1 to n; k = # of indivs., 1 to k)

1. We are assuming that Y is distributed somehow, $Y = f(X, \varepsilon)$ & that the expected value of Y (or best prediction) given X is a linear function of X & ε:

$$\underset{(n\times 1)}{\tilde{Y}} = \beta_0 + \beta_1 \underset{(n\times 1)}{\tilde{X}_1} + \beta_2 \underset{(n\times 1)}{\tilde{X}_2} + \ldots + \beta_k \underset{(n\times 1)}{\tilde{X}_k} + \underset{(n\times 1)}{\tilde{\varepsilon}} = \underset{(n\times k)}{X} \underset{(k\times 1)}{\tilde{\beta}} + \underset{(n\times 1)}{\tilde{\varepsilon}}$$

or, equivalently, $E(\tilde{Y}|X) = \beta_0 + \beta_1 \tilde{X}_1 + \beta_2 \tilde{X}_2 + \ldots + \beta_k \tilde{X}_k + E(\tilde{\varepsilon})$

2. $E(Y|X) = X\beta$ equivalently $E(\tilde{\varepsilon}) = \tilde{0}$
- the stochastic term, $\tilde{\varepsilon}$, has mean zero (for all its elements)

$$E(\tilde{Y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} (X\beta)_1 \\ \vdots \\ (X\beta)_n \end{bmatrix} \quad \& \quad E(\tilde{\varepsilon}) = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \tilde{0}$$

3. $$V(\tilde{\varepsilon}) = \begin{bmatrix} Cov(\varepsilon_1,\varepsilon_1) & \cdots & Cov(\varepsilon_1,\varepsilon_n) \\ \vdots & & \vdots \\ Cov(\varepsilon_n,\varepsilon_1) & \cdots & Cov(\varepsilon_n,\varepsilon_n) \end{bmatrix} = E(\tilde{\varepsilon}\tilde{\varepsilon}') = \begin{bmatrix} E(\varepsilon_1^2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ \vdots & \ddots & \vdots \\ E(\varepsilon_n\varepsilon_1) & \cdots & E(\varepsilon_n^2) \end{bmatrix}$$
$(n\times n)$ $\qquad$ $(n\times n)$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I_n \quad (n\times n)$$

$\Rightarrow$ Variance of each ε_i (of each $\underset{(1\times 1)}{[Y_i|\underset{(1\times k)}{\tilde{X}_i}]}$) = some constant σ^2 (homoskedasticity)

& $\Rightarrow Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall\ i \neq j$

(loosely) no ε_i tells us anything about ε_j or given $\underset{(1\times k)}{\tilde{X}_i}$ & $\underset{(1\times k)}{\tilde{X}_j}$ no Y_i tells us anything more about Y_j (equally loosely) (uncorrelatedness)

4. $E(\underset{(n\times 1)}{\varepsilon} \mid \underset{(n\times k)}{X}) = 0$: once the systematic relationship between $E(Y)$ & X is netted out, X has nothing more to tell us about ε.

5. X is non-stochastic (this can be relaxed) & 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 & X is non-linear in the parameters, β.

typ. violations: ① wrong funct. form ② wrong regressors ③ changing parameters

recall that we can include in X & or transform Y in a large manner of ways to make the equation non-linear in X & still be able to apply the CLRM. It is β 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 $E(\varepsilon_i) = 0$. This is not problematic unless (1) is or unless we fail to include a constant.

3. $V(\varepsilon) = \sigma^2 I_n$ → heteroskedasticity: σ_i^2 is not a constant;
(usual data type of plagues) → (cross-section) → group-wise hetero.
(time-series) → time-wise hetero.
$V(\varepsilon) = g(\cdot)$ → other sources
} we'll go into this much more when we get to heterosked.

→ correlated residuals:
(time-series) → the shadow of the past
(cross-section) → geographic or other proximity
} more when we do autocorrelation

4. $E(\underset{\sim}{\varepsilon} | X) = 0$ The most typical violations involve ("selection bias")

a) self-selection of observations into the sample
b) Reverse Causality (Endogeneity, Simultaneity)
c) Some omitted factor causes both Y & X

these are not disjoint sets

• autoregression • measurement error or errors in variables

Basically, the assumption is violated whenever X contains information about the distribution of y_i which goes beyond information about $E(y_i)$ {Thanks to Nancy Burns for this additional way of looking at it.}

5. X non-stoch. & no $\underset{(n\times 1)}{\underset{\sim}{X}_j}$ lin. funct. of $\underset{(n\times(k-1))}{X_{\text{not } j}}$: non-stochastic X means we treat it as fixed in repeated samples {Thanks again to Nancy: another way to view 4 is as saying that so treating X does not exclude in a systematic way any relevant obtainable info.}

• No variable, $\underset{\sim}{X}$ (a column of X) is an exact linear function of others.

Why not? Suppose $x_1 = 3x_2$, then suppose $b_1 = 1$ and $b_2 = 0$ is an optimal solution, so too then is $b_2 = 3$ & $b_1 = 0$ & in fact, so is any pair of b_1 & b_2 along the entire line ($-\infty$ to $+\infty$) given by the points $(1,0)$ & $(0,3)$.

Also (why not?) (Same thing by geometry of Matrix Algebra):

if some $\underset{\sim}{X}_j$ is an exact lin. function of $\{\underset{\sim}{X}_{\text{not } j}\}$'s then

$$\text{Rank}(X) < k \Rightarrow \text{Rank}(X'X) < k \Rightarrow \underset{=0}{\text{determinant}(X'X)}$$

$$\Rightarrow (X'X)^{-1} \text{ does not exist.}$$

• most common violations (all are researcher dumb-mistake type stuff, usu. around 4am)

a) Type same variable name twice when giving computer program indep. var. list.
b) Some set of X's sum to 1 (e.g. set of dummies, % breakdowns of something) & (∴ are perfectly colinear with the constant)
c) Some X or X's do not vary in the sample &

solution: leave one out (more on this in a couple weeks)

it's funny, it's also by far the most common source of problem

OLS Estimation of the CLRM in Multivariate Case (Again)

A. As always we want to estimate $\underset{(k\times 1)}{\beta}$ (the relationships between X's & Y) by finding $\underset{(k\times 1)}{b}$ that minimizes the distance from the hyperplanes (=planes, only in many dimensions) given by X & the lines given by Y. (This is "projection.") Equivalently, we find the $\underset{(k\times 1)}{b}$ that minimizes the sum of squared distances from the $\underset{(1\times k)}{X_i}\underset{(k\times 1)}{b}$'s to the $\underset{(1\times 1)}{Y_i}$'s. (Iff $Y \sim N(X\beta, \sigma^2 I)$ these can be shown equivalent to maximum likelihood estimates)

① So, $\min_{b} \sum_{i=1}^{n} e_i^2$ where $\underset{(1\times 1)}{e_i} \equiv \underset{(1\times 1)}{Y_i} - \underset{(1\times k)}{X_i}\underset{(k\times 1)}{b}$

② by definition of inner products $\equiv \min_{b} e'e$ where e is the $n\times 1$ vector of the e_i's

③ writing out e using definition $= \min_{b}\{(Y - Xb)'(Y - Xb)\} = \min_{b}\left[\underset{(1\times n)}{(Y' - b'X')}\underset{(n\times 1)}{(Y - Xb)}\right]$

$\{(n\times 1) - (n\times k)(k\times 1)\}'$ $\{(n\times 1) - (n\times k)(k\times 1)\}$

$\Rightarrow (n\times 1)'$ $\Rightarrow (n\times 1)$

$\Rightarrow (1\times n)$ $\Rightarrow (1\times 1)$

$(1\times k)(k\times n) = (1\times n)$

④ taking the transpose inside

⑤ multiplying out $= \min_{b}\{Y'Y - b'X'Y - Y'Xb + b'X'Xb\}$ (n.b.: all the dimensions still work)

⑤ multiplying out $= \min_{\underset{\sim}{b}} \{ \underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{b}'X'\underset{\sim}{y} - \underset{\sim}{y}'X\underset{\sim}{b} + \underset{\sim}{b}'X'X\underset{\sim}{b} \}$ (n.b.: all the dimensions still work)

⑥ call this γ & minimize with respect to $\underset{\sim}{b}$

⑥ First-order condition $\Rightarrow \frac{\partial \gamma}{\partial \underset{\sim}{b}} = -\frac{\partial(\underset{\sim}{b}'X'\underset{\sim}{y})}{\partial \underset{\sim}{b}} - \frac{\partial(\underset{\sim}{y}'X\underset{\sim}{b})}{\partial \underset{\sim}{b}} + \frac{\partial(\underset{\sim}{b}'X'X\underset{\sim}{b})}{\partial \underset{\sim}{b}} = 0$

⑦ by our rules of multivariable calculus $\Rightarrow -X'\underset{\sim}{y} - (\underset{\sim}{y}'X)' + (X'X)\underset{\sim}{b} + [\underset{\sim}{b}'X'X]' = 0$ (* multirule)

⑧ by transpose rules $\Rightarrow -X'\underset{\sim}{y} - X'\underset{\sim}{y} + (X'X)\underset{\sim}{b} + (X'X)\underset{\sim}{b} = 0$

⑨ combine & divide both sides by 2 $\Rightarrow -2X'\underset{\sim}{y} + 2(X'X)\underset{\sim}{b} = 0 \Rightarrow -X'\underset{\sim}{y} + (X'X)\underset{\sim}{b} = 0$

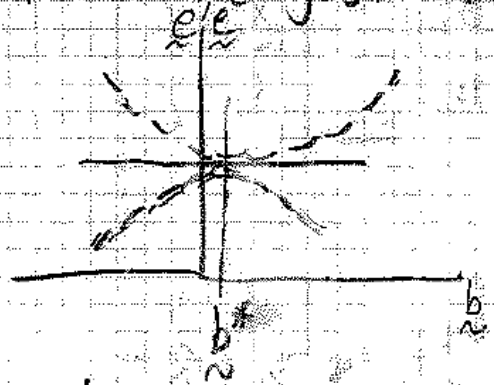
⑩ put $\underset{\sim}{b}$ term on one side, else on other $\Rightarrow (X'X)\underset{\sim}{b} = X'\underset{\sim}{y}$

⑪ pre-multiply both sides by inverse to get F.O.C. solution:

$$\boxed{\underset{\sim}{b}^* = (X'X)^{-1}(X'\underset{\sim}{y})}$$

$$\boxed{\underset{\sim}{b}^* = (X'X)^{-1}(X'\underset{\sim}{y})}$$

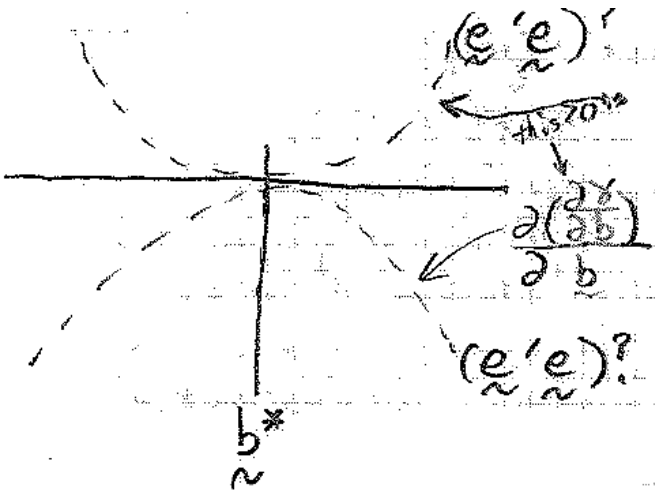
B. What did we just do? We had some function of $\underset{\sim}{b}$, namely: $\underset{\sim}{e}'\underset{\sim}{e}$



(← this is loose picture only b/c $\underset{\sim}{b}$ is a vector so can't put it on a single axis, but, never mind that, just suppose $\underset{\sim}{b}$ a scalar for purposes of this figure)

We found some $\underset{\sim}{b}^*$ for which $\underset{\sim}{e}'\underset{\sim}{e}$ (the sum of squared residuals) had slope (derivative) equal to zero, but this could be a maximum or min for $\underset{\sim}{e}'\underset{\sim}{e}$ (see figure) depending on curvature

specifically, does $\underset{\sim}{e}'\underset{\sim}{e}$ curve up or down at b^*? that's the same question as $\frac{\partial^2 Y}{\partial b^2} > \text{or} < 0$



III. A. Estimating $\underset{\sim}{b}$

$$\underset{\sim}{b}^* = (X'X)^{-1}X'\underset{\sim}{Y}$$

comes from $\frac{\partial(\underset{\sim}{e}'\underset{\sim}{e})}{\partial\underset{\sim}{b}} = 0$

B. but is it a maximum or minimum?

$$\left.\frac{\partial(\frac{\partial\gamma}{\partial\underset{\sim}{b}})}{\partial\underset{\sim}{b}}\right|_{\underset{\sim}{b}^*} < 0 \Rightarrow \text{maximum}$$

rate of change of the rate of change, evaluated at $\underset{\sim}{b}^*$

$$\left.\frac{\partial(\frac{\partial\gamma}{\partial\underset{\sim}{b}})}{\partial\underset{\sim}{b}}\right|_{\underset{\sim}{b}^*} > 0 \Rightarrow \text{minimum}$$

$$\frac{\partial\gamma}{\partial\underset{\sim}{b}} = -X'Y + (X'X)\underset{\sim}{b} \Rightarrow \frac{\partial(\frac{\partial\gamma}{\partial\underset{\sim}{b}})}{\partial\underset{\sim}{b}'} \equiv \frac{\partial^2\gamma}{\partial\underset{\sim}{b}\partial\underset{\sim}{b}'} = (X'X)$$

notational definition

$$\frac{\partial \gamma}{\partial \underset{\sim}{b}} = -X'Y + (X'X)\underset{\sim}{b} \quad \Rightarrow \frac{\partial\left(\frac{\partial \gamma}{\partial \underset{\sim}{b}}\right)}{\partial \underset{\sim}{b}'} \equiv \frac{\partial^2 \gamma}{\partial \underset{\sim}{b}\partial \underset{\sim}{b}'} = (X'X)$$

notational definition

is this ">0"? In matrices that translates to is $(X'X)$ positive definite? Which means is

$v' \cdot (X'X) v > 0$ for any conformable vector, v?

let $q \equiv Xv$, then this can be written $q'q$ which equals $\sum_{i=1}^{n} q_i^2$ 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)

$$\left.\frac{\partial^2 \gamma}{\partial \underset{\sim}{b}\partial \underset{\sim}{b}'}\right|_{\underset{\sim}{b}^*} > 0 \quad \text{so}$$

$$\underset{\sim}{b}_{LS} = (X'X)^{-1} X'\underset{\sim}{Y}$$

minimizes $\underset{\sim}{e}'\underset{\sim}{e}$ (the sum of squared residuals)

C. As usual, we could solve all this for the "normal equations":

$$(X'X)\underset{\sim}{b} = X'\underset{\sim}{Y}$$

what are elements of $X'X$? →

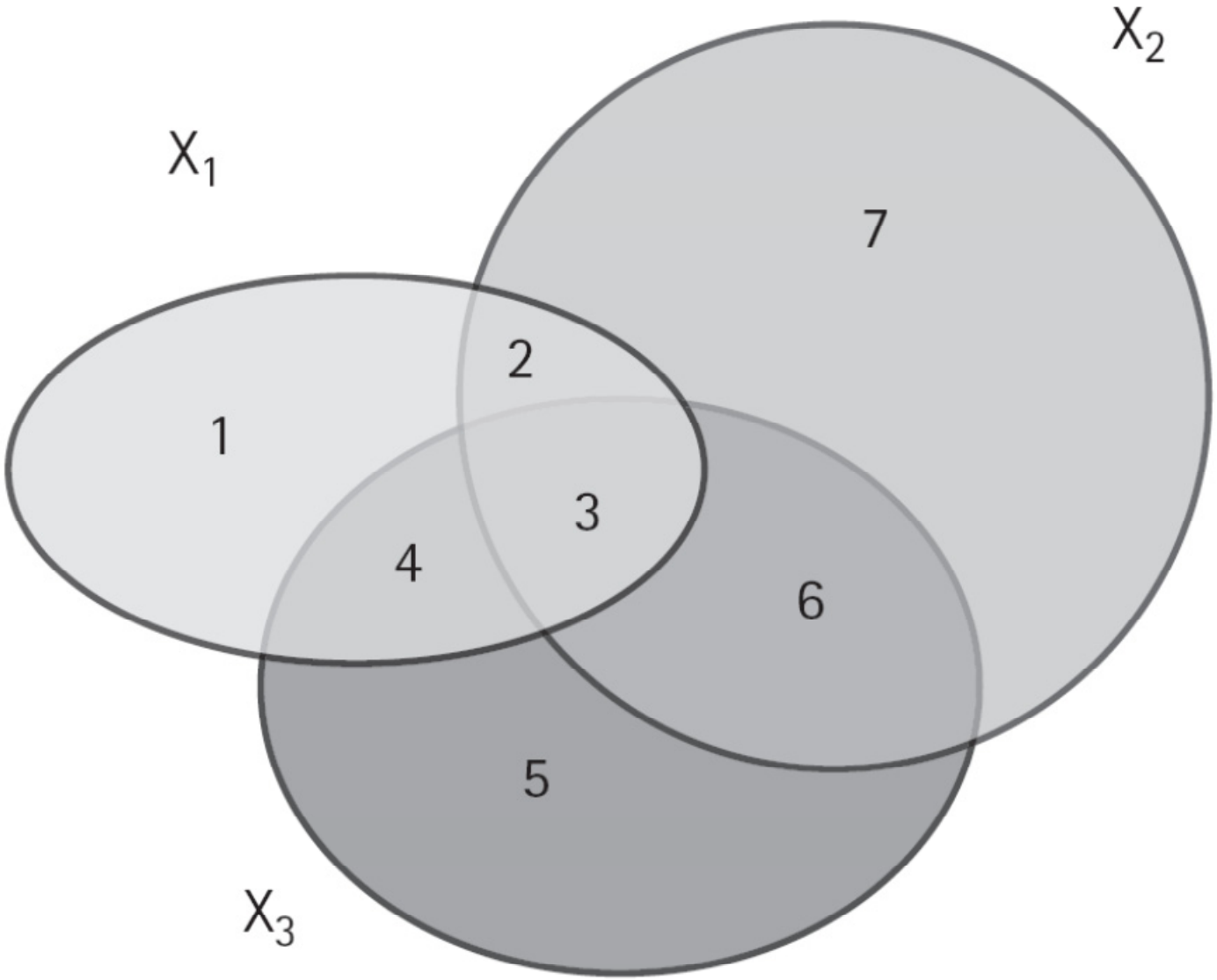
$$\begin{bmatrix} \underset{\sim}{X}_1'\underset{\sim}{X}_1 & \underset{\sim}{X}_1'\underset{\sim}{X}_2 & \cdots & \underset{\sim}{X}_1'\underset{\sim}{X}_k \\ \underset{\sim}{X}_2'\underset{\sim}{X}_1 & & & \\ \vdots & & & \vdots \\ \underset{\sim}{X}_k'\underset{\sim}{X}_1 & & & \underset{\sim}{X}_k'\underset{\sim}{X}_k \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}_1'\underset{\sim}{Y} \\ \underset{\sim}{X}_2'\underset{\sim}{Y} \\ \vdots \\ \underset{\sim}{X}_k'\underset{\sim}{Y} \end{bmatrix}$$

$(k \times k)$ $(k \times 1)$ $(k \times 1)$

each of these is a column of X (i.e. a vector)

B. Addressing *multiple causes*=what mult. regression does; 3 ways see:

1. Working from normal equations, can see how "partial out" shared variation & covariation, so estimates based only on unique variation & covariation of each **x**, *y*, and **x** with *y*. First, though, graphically:



2. Now, algebraically, we can work from the normal equations to see how this "partialing out" of shared variation & covariation is actually achieved:

$$\underset{\hat{\alpha},\hat{\beta}_1,\ldots,\hat{\beta}_k}{Min} \sum_i e_i^2 \equiv \underset{\hat{\boldsymbol{\beta}}}{Min}\ \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}} \equiv \underset{\hat{\boldsymbol{\beta}}}{Min}\ \left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

$$\Rightarrow \frac{\partial \underset{\sim}{\mathbf{e}}'\underset{\sim}{\mathbf{e}}}{\partial \hat{\boldsymbol{\beta}}} = 0 \Rightarrow \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

now, notice that, for **X** 3x3, for example, we have:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_1 & \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_2 & \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_3 \\ \underset{\sim}{\mathbf{x}}_2'\underset{\sim}{\mathbf{x}}_1 & \underset{\sim}{\mathbf{x}}_2'\underset{\sim}{\mathbf{x}}_2 & \underset{\sim}{\mathbf{x}}_2'\underset{\sim}{\mathbf{x}}_3 \\ \underset{\sim}{\mathbf{x}}_3'\underset{\sim}{\mathbf{x}}_1 & \underset{\sim}{\mathbf{x}}_3'\underset{\sim}{\mathbf{x}}_2 & \underset{\sim}{\mathbf{x}}_3'\underset{\sim}{\mathbf{x}}_3 \end{bmatrix}, \text{ so, we can rewrite:}$$

$$\begin{bmatrix} \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_1 & \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_2 & \underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_3 \\ \underset{\sim}{\mathbf{x}}_2'\underset{\sim}{\mathbf{x}}_1 & \underset{\sim}{\mathbf{x}}_2'\underset{\sim}{\mathbf{x}}_2 & \underset{\sim}{\mathbf{x}}_2'\underset{\sim}{\mathbf{x}}_3 \\ \underset{\sim}{\mathbf{x}}_3'\underset{\sim}{\mathbf{x}}_1 & \underset{\sim}{\mathbf{x}}_3'\underset{\sim}{\mathbf{x}}_2 & \underset{\sim}{\mathbf{x}}_3'\underset{\sim}{\mathbf{x}}_3 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\mathbf{x}}_1'\mathbf{y} \\ \underset{\sim}{\mathbf{x}}_2'\mathbf{y} \\ \underset{\sim}{\mathbf{x}}_3'\mathbf{y} \end{bmatrix}, \text{ which in scalar notation is:}$$

$$(1) \qquad \mathbf{x}_1'\mathbf{x}_1\hat{\beta}_1 + \mathbf{x}_1'\mathbf{x}_2\hat{\beta}_2 + \mathbf{x}_1'\mathbf{x}_3\hat{\beta}_3 = \mathbf{x}_1'\mathbf{y}$$

$$(2) \qquad \mathbf{x}_2'\mathbf{x}_1\hat{\beta}_1 + \mathbf{x}_2'\mathbf{x}_2\hat{\beta}_2 + \mathbf{x}_2'\mathbf{x}_3\hat{\beta}_3 = \mathbf{x}_2'\mathbf{y}$$

$$(3) \qquad \mathbf{x}_3'\mathbf{x}_1\hat{\beta}_1 + \mathbf{x}_3'\mathbf{x}_2\hat{\beta}_2 + \mathbf{x}_3'\mathbf{x}_3\hat{\beta}_3 = \mathbf{x}_3'\mathbf{y}$$

Now, usually first column (variable) is $\underset{\sim}{\mathbf{x}}_1 = \begin{bmatrix} \cdots & 1 & \cdots \end{bmatrix}'$, which implies:

$\underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_1 = \sum_{i=1}^{n}(1\times 1) = n$, and $\underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_2 = \sum_{i=1}^{n}\left(1\times \underset{\sim}{\mathbf{x}}_{2_i}\right) = \sum_{i=1}^{n} x_2$ and $\underset{\sim}{\mathbf{x}}_1'\underset{\sim}{\mathbf{x}}_3 = \sum_{i=1}^{n}\left(1\times \underset{\sim}{\mathbf{x}}_{3_i}\right) = \sum_{i=1}^{n} x_3$

Using this, we can simplify (1),(2),(3) as…

$$(4) \qquad n\hat{\beta}_1 + \hat{\beta}_2\sum x_2 + \hat{\beta}_3\sum x_3 = \sum y$$

$$(5) \qquad \hat{\beta}_1\sum x_2 + \mathbf{x}_2'\mathbf{x}_2\hat{\beta}_2 + \mathbf{x}_2'\mathbf{x}_3\hat{\beta}_3 = \mathbf{x}_2'\mathbf{y}$$

$$(6) \qquad \hat{\beta}_1\sum x_3 + \mathbf{x}_3'\mathbf{x}_2\hat{\beta}_2 + \mathbf{x}_3'\mathbf{x}_3\hat{\beta}_3 = \mathbf{x}_3'\mathbf{y}$$

Now, since we starting converting to scalar notation, let's finish:

$$(7) \qquad n\hat{\beta}_1 + \hat{\beta}_2 \sum x_2 + \hat{\beta}_3 \sum x_3 = \sum y$$

$$(8) \quad \hat{\beta}_1 \sum x_2 + \hat{\beta}_2 \sum x_2^2 + \hat{\beta}_3 \sum x_2 x_3 = \sum x_2 y$$

$$(9) \quad \hat{\beta}_1 \sum x_3 + \hat{\beta}_2 \sum x_2 x_3 + \hat{\beta}_3 \sum x_3^2 = \sum x_3 y$$

Now, divide (7) through by n to get the solution for $\hat{\beta}_1$:

$$(10) \qquad \hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}_2 - \hat{\beta}_3 \overline{x}_3$$

The intercept, $\hat{\beta}_1$, is thus the conditional (sample) mean of *y*.

This result also shows that 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):

$$\left(\bar{y}-\hat{\beta}_2\bar{x}_2-\hat{\beta}_3\bar{x}_3\right)\sum x_2+\hat{\beta}_2\sum x_2^2+\hat{\beta}_3\sum x_2x_3=\sum x_2y \quad \text{...substituting in}$$

$$\Rightarrow\left(\bar{y}\sum x_2-\hat{\beta}_2\bar{x}_2\sum x_2-\hat{\beta}_3\bar{x}_3\sum x_2\right)+\hat{\beta}_2\sum x_2^2+\hat{\beta}_3\sum x_2x_3=\sum x_2y \quad \text{...multiplying out}$$

$$\Rightarrow\left(\bar{y}n\bar{x}_2-\hat{\beta}_2\bar{x}_2n\bar{x}_2-\hat{\beta}_3\bar{x}_3n\bar{x}_2\right)+\hat{\beta}_2\sum x_2^2+\hat{\beta}_3\sum x_2x_3=\sum x_2y \quad \text{...because } \sum x_2=n\bar{x}_2$$

$$\Rightarrow\hat{\beta}_2\left(\sum x_2^2-n\bar{x}_2^2\right)+\hat{\beta}_3\left(\sum x_2x_3-n\bar{x}_3\bar{x}_2\right)=\sum x_2y\text{-}n\bar{x}_2\bar{y} \quad \text{...rearranging}$$

$$\Rightarrow\hat{\beta}_2\times S_{x_2x_2}+\hat{\beta}_3\times S_{x_2x_3}=S_{x_2y} \quad \text{...by definitions of sample variations \& covariations}$$

(11) $$\Rightarrow\hat{\beta}_2=\frac{S_{x_2y}-\hat{\beta}_3\times S_{x_2x_3}}{S_{x_2x_2}} \quad \text{...solving for } \hat{\beta}_2\text{, which can also be rewritten:}$$

$$\Rightarrow\hat{\beta}_{2.3}=\frac{S_{x_2y}}{S_{x_2x_2}}-\frac{\hat{\beta}_{3.2}\times S_{x_2x_3}}{S_{x_2x_2}}=\hat{\beta}_2-\hat{\beta}_{3.2}\frac{S_{x_2x_3}}{S_{x_2x_2}}=\hat{\beta}_2-\hat{\beta}_{3.2}\times\hat{\beta}_{(x_3 \text{ on } x_2)}$$

And, by analogy: (12) $\hat{\beta}_{3.2}=\hat{\beta}_3-\hat{\beta}_{2.3}\times\hat{\beta}_{(x_2 \text{ on } x_3)}$.

Alternative, identical expressions are these:

$$(1^*) \qquad b_1 = \bar{y} - b_2\bar{x}_2 - b_3\bar{x}_3$$

$$(2^*) \qquad b_{2.3} = \frac{\left(\Sigma x_2^* y^*\right)\left(\Sigma (x_3^*)^2\right) - \left(\Sigma x_3^* y^*\right)\left(\Sigma x_2^* x_3^*\right)}{\left(\Sigma (x_2^*)^2\right)\left(\Sigma (x_3^*)^2\right) - \left(\Sigma x_2^* x_3^*\right)^2} = \frac{S_{x_2 y} S_{x_3}^2 - S_{x_3 y} S_{x_2 x_3}}{S_{x_2}^2 S_{x_3}^2 - \left(S_{x_2 x_3}\right)^2}$$

$$(3^*) \qquad b_{3.2} = \frac{\left(\Sigma x_3^* y^*\right)\left(\Sigma (x_2^*)^2\right) - \left(\Sigma x_2^* y^*\right)\left(\Sigma x_2^* x_3^*\right)}{\left(\Sigma (x_3^*)^2\right)\left(\Sigma (x_2^*)^2\right) - \left(\Sigma x_2^* x_3^*\right)^2} = \frac{S_{x_3 y} S_{x_2}^2 - S_{x_2 y} S_{x_2 x_3}}{S_{x_2}^2 S_{x_3}^2 - \left(S_{x_2 x_3}\right)^2}$$

To get these, you solve the right-hand-side $\hat{\beta}$ out of (11) and (12). Doing so, you get the above (2*) and (3*), and then, recall:

$$b_2 = \frac{Cov(X_2, Y)}{Var(X_2)} = \frac{S_{x_2 y}}{S_{x_2}^2}$$

So, divide both numerator & denom of (2*) & (3*) by V(X_2) & by V(X_3), respectively:

$$(2') \quad b_{2.3} = \frac{b_2 - b_3 b_{x_3 x_2}}{1 - r^2_{x_2 x_3}}$$

$$(3') \quad b_{3.2} = \frac{b_3 - b_2 b_{x_2 x_3}}{1 - r^2_{x_2 x_3}}$$

where $\begin{cases} b_{2.3} \text{ is coeff reg } y \text{ on } x_2 \text{ in presence } x_3; \\ b_{x_2 x_3} \text{ is coeff reg } x_2 \text{ on } x_3; \\ \text{and } b_2 \text{ is bivariate reg coeff .} \end{cases}$

All the intuitions about how bivariate coefficients relate to partial coefficients remain the same, whether by (11) & (12) or by (2') and (3').**

(** Notice the difference between first term after the minus sign in the numerator on the right-hand side of the two pairs of expressions! b_2 & b_3 vs. $b_{2.3}$ & $b_{3.2}$)

Namely…

bivariate reg. coeff, y on X_2

call that b_{yx_2}, call the b_2 from OLS of y on X_1, X_2, & X_3: $b_{yx_2 \cdot x_3}$, then:

bivar, y on X_3

bivar, X_3 on X_2

$$* \quad b_{yx_2 \cdot x_3} = \frac{b_{yx_2}}{1 - r^2_{x_2x_3}} - \frac{b_{yx_3} \cdot b_{x_3x_2}}{1 - r^2_{x_2x_3}}$$

coeff. on X_2 in regression of y on X_2 in presence of X_3 (& constant

squared correlation of X_2 & X_3

$\Longrightarrow$ coeff. on X_2 in regression with an X_3 differs from bivariate regression coeff. by additive two factors:

① relationship b/w X_3 & X_2

② relationship b/w y & X_3

- by itself, greater relationship b/w X_2 & X_3 (either sign) drives the coeff on X_2 (& on X_3) in the big regression (the $1 - r^2$ effect)

↑↓ dep. on relation, but "inflates" both

- also, though, jointly $b(y, X_3) \cdot b(X_3, X_2)$ change the coeff. on X_2 in the big regression.

e.g. if both positive, then X_2 pos. corr. w/ X_3 which pos. corr. w/ y. So some of effect on y in bivariate regression attributed to X_2 may have really been due to X_3: more on this in a bit & much more when we get to "omitted variable bias"

$\Longrightarrow$ iff variables in Mult. Reg. are uncorrelated ($\Rightarrow r^2_{x_2x_3} = 0$ & $b_{x_2x_3} = 0$) will bivariate & Mult. regression coefficients be the same

3. Let's push this basic math further to see what "partialing" entails; first though:

VIII. Some Useful Linear-Regression Matrices:

A. *Sample Var-Cov* **X**: $\mathbf{Q} \equiv \mathbf{X}'\mathbf{X}$

1. **Q** is matrix of products & cross-products of columns of **X**, i.e., of **x**, i.e., of IndVars: sample variation-covariation matrix of **X**.

B. *Coefficient Maker*: $\mathbf{A} \equiv \mathbf{Q}^{-1}\mathbf{X}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

1. **A** is a matrix expression, i.e., a linear-additive/multiplicative function, of **X**, i.e., it is that particular combination of the information in **X** that returns the coefficients on **X** from a regression on **X** of whatever **A** multiplies.

C. *Fit Maker*: $\mathbf{N} \equiv \mathbf{X}\mathbf{A} = \mathbf{X}\mathbf{Q}^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

1. **N** is a matrix expression, a linear…function, of **X**, i.e., it is that particular combo of the information in **X** that returns the linear-regression prediction of whatever **N** multiplies from a regression of whatever **N** multiplies on **X**.

D. *Residual Maker*: $\mathbf{M} \equiv \mathbf{I} - \mathbf{N} = \mathbf{I} - \mathbf{X}\mathbf{A} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

1. **M** is a matrix express…function, of **X**, i.e., it is that particular combination of the information in **X** that returns the linear-regression prediction errors for whatever **M** multiplies from a regression of whatever **M** multiplies on **X**.

E. (Some) Useful Properties of these Matrices (that should eventually become intuitive to you):

1. $\mathbf{Q}$ is $k \times k$, symmetric, and positive-definite. So is $\mathbf{Q}^{-1}$.
2. $\mathbf{Ay}=\mathbf{b}_{ls}$.
   a) $\mathbf{AX}=\mathbf{I}_k$: (proof)$=(\mathbf{X'X})^{-1}\mathbf{X'X}=\mathbf{I}_k$.
   b) $\mathbf{AN}=\mathbf{A}$: (proof) $\mathbf{AN}=(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1}\mathbf{X'}=\mathbf{IA}=\mathbf{A}$.
   c) $\mathbf{AM}=\mathbf{0}$: (proof) $\mathbf{AM}=\mathbf{A}(\mathbf{I}-\mathbf{N})=\mathbf{A}-\mathbf{A}=\mathbf{0}$.
3. $\mathbf{Ny}=\hat{\mathbf{y}}_{ls}$.
   a) $\mathbf{N}$ is symmetric & idempotent. (right? Write $\mathbf{NNy}$; say that phrase about what $\mathbf{N}$ does aloud & think it through; reduce that expression)
   b) $\mathbf{NM}=\mathbf{0}$. (right? Say words & think through; write & reduce the expression)
4. $\mathbf{My}=\mathbf{e}_{ls}$.
   a) $\mathbf{M}$ is symmetric & idempotent. (right?)
   b) $\mathbf{MX}=\mathbf{0}$. (right?)

2. Recall the "normal equations" the $\underset{\sim}{\beta}_1$ & $\underset{\sim}{\beta}_2$ are the vectors (variables) of their coefficients.

$$(X'X)\underset{\sim}{b} = X'\underset{\sim}{Y}$$, now these can be (are) "broken up" into

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \underset{\sim}{b}_1 \\ \underset{\sim}{b}_2 \end{bmatrix} = \begin{bmatrix} X_1'\underset{\sim}{Y} \\ X_2'\underset{\sim}{Y} \end{bmatrix}$$

(a)
(b)

, which, solving by pre-multiplying both sides by $[X'X]^{-1}$ gives

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'\underset{\sim}{Y} \\ X_2'\underset{\sim}{Y} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{b}_1 \\ \underset{\sim}{b}_2 \end{bmatrix}$$

3. By rules of inverting partitioned matrices (~~chapt. 2~~ App. A)

$$\underset{\sim}{b}_1 = (X_1'X_1)^{-1}X_1'\underset{\sim}{Y} - (X_1'X_1)^{-1}X_1'X_2\underset{\sim}{b}_2$$

$$= A_1\underset{\sim}{Y} - A_1X_2\underset{\sim}{b}_2$$

$$= \boxed{\underset{\sim}{b}_1} - \text{some adjustment + which involves coeffs from } X_2 \text{ on } X_1 \ (A_1X_2)$$

↳ from regression of $\underset{\sim}{Y}$ on X_1 only

times coeffs. on X_2 in whole regression

subst. this into equation (b) & you get: whole regression

$$X_2'[X_1(A_1\underset{\sim}{Y} - A_1X_2\underset{\sim}{b}_2) + X_2\underset{\sim}{b}_2] = X_2'\underset{\sim}{Y}$$

$$X_2'X_1A_1\underset{\sim}{Y} - X_2'X_1A_1X_2\underset{\sim}{b}_2 + X_2'X_2\underset{\sim}{b}_2 = X_2'\underset{\sim}{Y}$$

(b/c $X_1A_1 \equiv N_1$)

$$X_2'N_1\underset{\sim}{Y} - X_2'N_1X_2\underset{\sim}{b}_2 + X_2'X_2\underset{\sim}{b}_2 = X_2'\underset{\sim}{Y}$$

(gathering up $\underset{\sim}{b}_2$ terms on LHS & $\underset{\sim}{Y}$ terms on RHS)

$$X_2'(I - N_1)X_2\underset{\sim}{b}_2 = X_2'(I - N_1)\underset{\sim}{Y}$$

$$(X_2'M_1X_2)\underset{\sim}{b}_2 = X_2'M_1\underset{\sim}{Y}$$

$$\underset{\sim}{b}_2 = (X_2'M_1X_2)^{-1}X_2'M_1\underset{\sim}{Y}$$

$$\underset{\sim}{b}_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 \underset{\sim}{Y}$$

now, M_1 is symmetric & idempotent, so $M_1' M_1 = M_1 M_1 = M_1$;
replace M_1 in above by $M_1' M_1$:

$$\underset{\sim}{b}_2 = (X_2' M_1' M_1 X_2)^{-1} X_2' M_1' M_1 \underset{\sim}{Y}$$

(c) $$\underset{\sim}{b}_2 = [(M_1 X_2)'(M_1 X_2)]^{-1} (M_1 X_2)' M_1 \underset{\sim}{Y}$$

M_1 is "residual-maker" for regression of thing it multiplies on X_1,

and $\underset{\sim}{b}_2$ is coefficient from Regression of $(M_1 \underset{\sim}{Y})$ on $(M_1 X_2)$

(equation (c) is in $(Z'Z)^{-1}Z'y$ form, right? so it's an A matrix times $\underset{\sim}{Y}$, thus regression coefficients)

but $M_1 \underset{\sim}{Y}$ are the residuals from $\underset{\sim}{Y}$ regressed on X_1 & $M_1 X_2$ are the residuals from (each of) the X_2's regressed on (all of) the X_1's.

Thus, $\underset{\sim}{b}_2$, the coefficient(s) on X_2 are also the coefficients on X_2^* when regressed on $\underset{\sim}{Y}^*$ where the *'s mean that all (linear) relationship with X_1 has been netted out. This is what it means to say "the effect(s) of X_2, controlling for X_1".

Third way to see what "controlling [for]" means: it is indeed "partialing", i.e., it gives the partial linear-association, i.e., the partial derivative:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$\beta_1 = \frac{\partial Y}{\partial X_1}$$ i.e. change in Y produced by a change in X_1 holding all else constant (ceteris paribus)

↳ all else here is X_2 & ε (& α & β's of course)

$$\beta_2 = \frac{\partial Y}{\partial X_2}$$ analogous

****

Some More Byproducts of OLS Estimation:

1. $(X'X)\underset{\sim}{b} - X'\underset{\sim}{y} = -X'(\underset{\sim}{y} - X\underset{\sim}{b}) = \boxed{-X'\underset{\sim}{e} = \underset{\sim}{0}}$

distrib. prop. mult. commut. prop. add. — by def of $\underset{\sim}{e}$

why? b/c if $X'\underset{\sim}{e} \neq 0$, then some linear relat. b/w X & $\underset{\sim}{e}$ remains & $\underset{\sim}{b}$ couldn't have been optimal

$$\Rightarrow \boxed{X'\underset{\sim}{e} = 0 \quad \forall \text{ column's of } X}$$ (i.e. this is true for every variable)

since X_1 = constant, this $\Rightarrow \underset{\sim}{i}'e = 0 \Rightarrow \boxed{\Sigma e_i = 0}$

OLS residuals sum to zero (average mistake is zero)

2. $\underset{(1\times 1)}{\bar{y}} = \underset{(k\times 1)'}{\bar{X}'}\underset{(k\times 1)}{b}$ i.e., the regression line passes through the sample means of the X's & y

$\Rightarrow \bar{\hat{y}} = \bar{y}$ i.e. the average predicted value equals the average observed value

IX. Properties of OLS Estimator under C(N)LRM:

A. The C(N)LRM (one more time!):

ASSUMPTIONS:

1. Linearity (a.k.a. "the model"): the dependent variable is a linear additive function of the regressors, with an additively separable stochastic component:

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \quad \text{or} \quad \underset{\sim}{y} \sim (\underset{\sim}{\mu}, \sigma^2 I), \; E(\underset{\sim}{y}) = \underset{\sim}{\mu} = \ldots$$

(n.b. $\forall i$)

2. Zero mean of the stochastic component:

$$E(\underset{\sim}{\varepsilon}) = 0 \quad (\text{n.b. } \forall_i)$$

3. Homoskedasticity (i.e. constant conditional variance) of the stochastic component:

&

4. Zero (conditional) covariance of the stochastic component:

$$V(\underset{\sim}{\varepsilon}) \equiv E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I \quad = V(\ldots Y$$

"spherical errors" not "nonspherical error-structure"

5. The regressors are (weakly) exogenous, i.e., the regressors do not covary with the stochastic components:

$$Cov(X, \underset{\sim}{\varepsilon}) = 0 \quad \text{i.e. } E(X'\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$$

UNNECESSARY ASSUMPTIONS THAT ADD CONVENIENCE &/OR ENHANCE PROPERTIES

6. X is non-stochastic, i.e. fixed across repeated (hypothetical) samples

7. Normality:

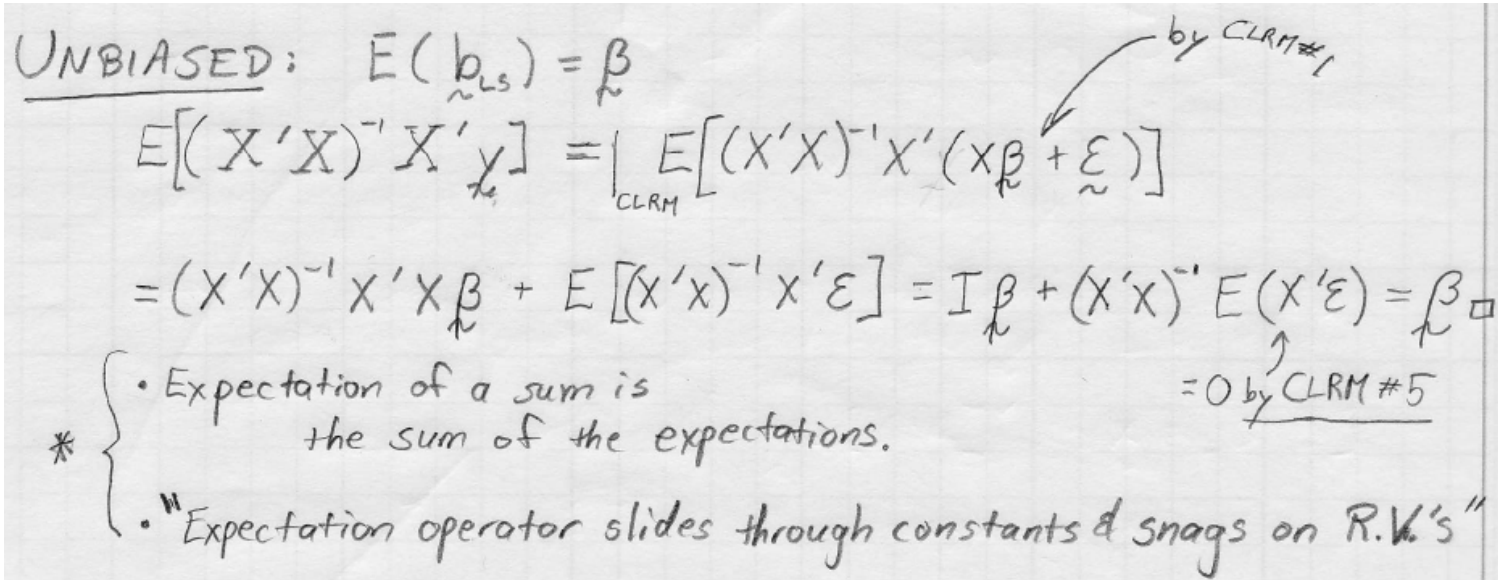
$$\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \sigma^2 I) \quad \text{or} \quad \underset{\sim}{y} \sim N(X\underset{\sim}{\beta}, I)$$

B. Coefficient estimates deviate from true coefficients only by $+\mathbf{A}\boldsymbol{\varepsilon}$:

$$\hat{\boldsymbol{\beta}}_{LS} \equiv \mathbf{b}_{LS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \equiv \mathbf{A}\mathbf{y}$$

$$= \mathbf{A}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

C. Unbiasedness for OLS applied to C(N)LRM rests on 1-2,4, not 3:



$$E\left(\hat{\boldsymbol{\beta}}_{LS}\right) = E\left(\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + E(\mathbf{A}\boldsymbol{\varepsilon})$$

$$= \boldsymbol{\beta} + \mathbf{A}E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}, \text{ if } \mathbf{X} \text{ non-stoch.}$$

$$= \boldsymbol{\beta} + E(\mathbf{A}\boldsymbol{\varepsilon}) = \boldsymbol{\beta}, \text{ if } \mathbf{X} \text{ stoch. \& } E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$$

D. True variance (across repeated samples) of OLS est's under CLRM:

VARIANCE: $V(\underset{\sim}{b}_{LS}) = V[(X'X)^{-1}X'(X\underset{\sim}{\beta}+\varepsilon)]$

$= V[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon]$

* {
- 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 Cov's
}

$V(\underset{\sim}{b}_{LS}) = 0 + V[(X'X)^{-1}X'\varepsilon] = (X'X)^{-1}X'V(\varepsilon)X(X'X)^{-1}$ } *

now, under CLRM, given sphericity in particular:

$= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1}$

$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$

$= \sigma^2(X'X)^{-1}$

$$V\left(\hat{\boldsymbol{\beta}}_{LS}\right) = V\left(\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}\right) = \mathbf{A}V(\boldsymbol{\varepsilon})\mathbf{A}'$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[V(\boldsymbol{\varepsilon})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma_{\varepsilon}^{2}\mathbf{I}_{n}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \begin{cases} \sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}, \text{ if } \mathbf{X} \text{ non-stoch.} \\ \sigma_{\varepsilon}^{2}E\left\{(\mathbf{X}'\mathbf{X})^{-1}\right\}, \text{ if } \mathbf{X} \text{ stoch.} \end{cases}$$

$$\Rightarrow V(b_k) = f\left(\overset{+}{\sigma}_{\varepsilon}^{2}, \overset{-}{V}(x_k), \overset{+}{R}_{x_k,\mathbf{x}_{\sim k}}^{2}\right)$$

Highlight lines 2 & 3 especially… Important intuitions in that last line…

$$s.e.(b_j) = \sqrt{s_e^2 (X'X)^{-1}_{jj}} = g(\overset{+}{s_e^2}, \overset{-}{s_{x_j}^2}, \overset{+}{s_{x_j x_{not\,j}}})$$

that is, $s.e.(b_j) = g\left(\overset{+}{\text{estimated variance of stochastic term}}, \overset{-}{\text{sample variation of } X_j}, \overset{+}{\text{sample covariance of } X_j \text{ with all the } X_{not\,j} \text{ (specifically, } R^2 \text{ from } X_j \text{ on all } X_{not\,j}}\right)$

1. $s.e.(b_j) = f(\overset{+}{s_e^2}, \overset{-}{s_{x_j}^2}, \overset{+}{R^2_{x_j \text{ on } x_{not\,j}}})$

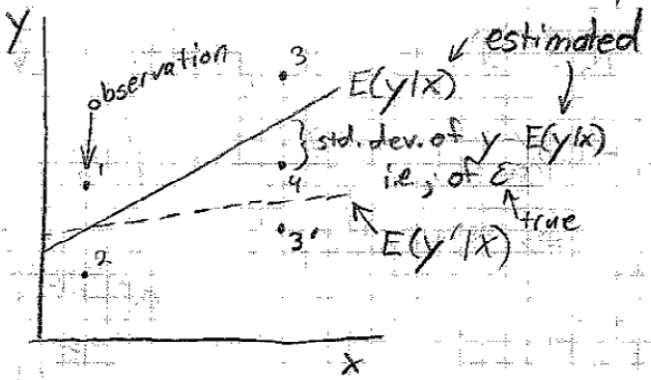
a) s.e.'s bigger when estimate of stochastic component bigger

b) smaller when (meaningful) sample variation in X is larger ↑

c) larger when X_j covaries more with the other X's

2. Graphical Intuition

a) Why does fact that ε varies imply that b varies?



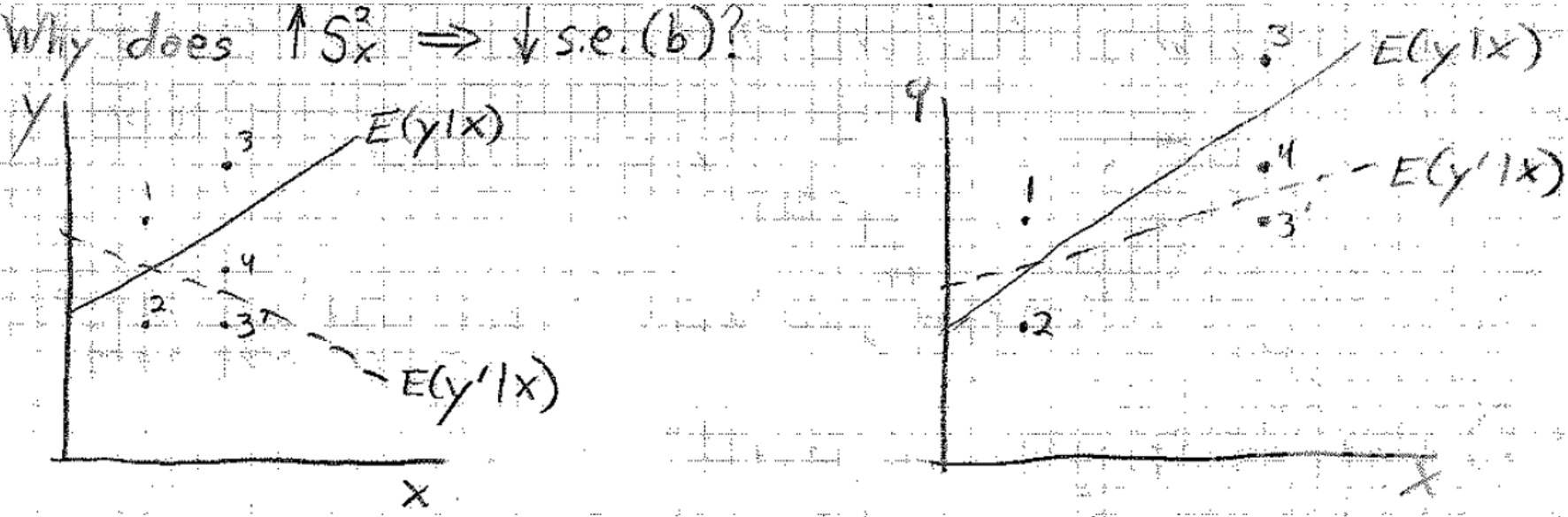
Suppose instead of being 1 std. dev. above regression line, observation 3 were two standard deviations below at obs. 3'.

Then regression line would be as the dotted line case

Thus movements in ε_i change our estimated slopes. The larger is $V(\varepsilon_i)$, the more b will move for a "typical" movement in ε_i, & the larger such a "typical" move will be! Thus, s.e.(b) which measures how b estimates change across repeated samples correctly increases as $V(\varepsilon_i)$ increases.

Unfortunately, we only have an estimate of $V(\varepsilon_i)$ namely s_e^2 so our estimate of s.e.(b) can only respond to this estimate.

b) Why does $\uparrow S_x^2 \Rightarrow \downarrow$ s.e.(b)?



These two examples are same in that $b = 2/3$ in 1st case & in that obs 1, 2, 3, 4 are 2 units from the regression line & 3' is 4 units below the original line. Only difference is the X's are more spread out (S_x^2 higher in second case). Larger sample variation in X produces less variance in b estimates across samples for a given $V(\varepsilon)$ (represented by the equal movements from 3 to 3'.

c) Why does ↑ sample covariance among X's increase variance of b_j?

b_j measures $\Delta Y / \Delta X_j$ holding all other X's fixed, ie. netting out covariance of X_j with other X's. When X_j more highly correlated with other X's the more is netted out & the less information we have left to estimate the effect of X_j per se. Less Info = More Variance

Relaxing the fixed **X** assumption:

E. Relaxing the fixed X assumption:

1) $\underset{\sim}{b} = \underset{\sim}{\beta} + A\underset{\sim}{\varepsilon}$ still, only now $A \equiv (X'X)^{-1}X'$ made up of stochastic X's (they vary across repeated samples)

$\underset{\sim}{b}_{LS}$ ... no change

2) $E(\underset{\sim}{b}_{LS}|X) = \underset{\sim}{\beta} + E[A\underset{\sim}{\varepsilon}|X] = \underset{\sim}{\beta} + AE(\varepsilon|X)$ (b/c, giving the X's means fixing them)

$= \underset{\sim}{\beta} + 0 = \beta$

$\Rightarrow$ conditional on X, $\underset{\sim}{b}$ unbiased

$E(\underset{\sim}{b}_{LS}) = \beta$ no change

· but now, we can take expectation of $E(\underset{\sim}{b}|X)$ across all possible X's

$$E_X[E(\underset{\sim}{b}|X)] = E_X[\underset{\sim}{\beta}] = \underset{\sim}{\beta}$$ (law of iterated expectations)

so unbiasedness of $\underset{\sim}{b}_{LS}$ rests only on $E(\varepsilon|X) = 0$ (assumption 4)

3) $V(\underset{\sim}{b}_{LS}|X) = \sigma^2 Q^{-1}$ (giving the X's fixes them)

$V(b_{LS}) =$ slight change

This step below by variance decomposition:

$$V(\underset{\sim}{b}_{LS}) = E_X[Var(\underset{\sim}{b}_{LS}|X)] + Var_X[E(\underset{\sim}{b}_{LS}|X)]$$

$$= E_X[V(\underset{\sim}{b}_{LS}|X)] + Var_X[\beta]$$ → $Var(\text{constant}) = 0$

$$= E_X[\sigma^2 Q^{-1}]$$

$$\boxed{V(\underset{\sim}{b}_{LS}) = \sigma^2 E(Q^{-1})}$$

now have $E[(X'X)^{-1}]$ rather than raw $(X'X)^{-1}$

OLS is BLUE

4) $V(\underset{\sim}{b}_{LS}) \leq V(\underset{\sim}{b}^*)$ $\forall$ $\underset{\sim}{b}^*$ linear & unbiased

I.e., difference 2nd minus 1st is pos. def.

I.e., $\left[V(\tilde{\mathbf{b}}) - V(\mathbf{b}_{ls})\right]$ is positive definite; i.e., $\mathbf{x}'\left[V(\tilde{\mathbf{b}}) - V(\mathbf{b}_{ls})\right]\mathbf{x} > 0 \quad \forall \mathbf{x} \neq \mathbf{0}$

E. EFFICIENT: "no unused useful information".

With normality, $V\left(\hat{\boldsymbol{\beta}}_{LS}\right)=\sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}$ is Cramer-Rao lower-bound;

i.e., $\hat{\boldsymbol{\beta}}_{LS}$ is efficient; i.e., $\hat{\boldsymbol{\beta}}_{LS}$ is BLUE.

In fact, if $\boldsymbol{\varepsilon}\sim MVN$, then $\hat{\boldsymbol{\beta}}_{LS}$ is BUE.

[can also prove as before, w/o normality]

F. If $\boldsymbol{\varepsilon}\sim MVN\left(\mathbf{0},\sigma_{\varepsilon}^{2}\mathbf{I}_{n}\right)$, then:

$\mathbf{b}_{LS}=\boldsymbol{\beta}+\mathbf{A}\boldsymbol{\varepsilon}=$ constant + linear-sum of normals

$$\Rightarrow \mathbf{b}_{LS}\sim MVN\left(\boldsymbol{\beta},\sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}\right) \quad ;\text{ if not, then}$$

$\mathbf{b}_{LS}=\boldsymbol{\beta}+\mathbf{A}\boldsymbol{\varepsilon}=$ constant + linear-sum of some **?**

$$\Rightarrow \mathbf{b}_{LS}\sim^{A} MVN\left(\boldsymbol{\beta},\sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}\right) \quad .$$

1. For instance, in the bivariate-regression case, as we already noted:

$$\mathbf{b}_{ls}\equiv\begin{bmatrix}a_{ls}\\ b_{ls}\end{bmatrix}\sim\left(\begin{bmatrix}\alpha\\ \beta\end{bmatrix},\begin{bmatrix}\sigma_{\varepsilon}^{2}\left(\frac{1}{n}+\frac{\overline{x}}{S_{xx}}\right) & -\frac{\overline{x}}{S_{xx}}\\ -\frac{\overline{x}}{S_{xx}} & \frac{\sigma_{\varepsilon}^{2}}{S_{xx}}\end{bmatrix}\right) \Rightarrow \mathbf{b}_{ls_1}=a_{ls}\sim N\left(\alpha,\sigma_{\varepsilon}^{2}\left(\frac{1}{n}+\frac{\overline{x}}{S_{xx}}\right)\right) \Rightarrow \mathbf{b}_{ls_2}=b_{ls}\sim N\left(\beta,\frac{\sigma_{\varepsilon}^{2}}{S_{xx}}\right)$$

X. Estimating σ_ε^2 using $\frac{1}{n-k}\mathbf{e}'\mathbf{e}$:

$$\mathbf{e} \equiv \mathbf{My} = \mathbf{M}(\mathbf{X\beta} + \boldsymbol{\varepsilon}) = \mathbf{MX\beta} + \mathbf{M\varepsilon} = \mathbf{0\beta} + \mathbf{M\varepsilon} = \mathbf{M\varepsilon}$$

$$\Rightarrow E(\mathbf{e}'\mathbf{e}) = E\left\{(\mathbf{M\varepsilon})'\mathbf{M\varepsilon}\right\} = E\{\boldsymbol{\varepsilon}'\mathbf{M}'\mathbf{M\varepsilon}\} = E\{\boldsymbol{\varepsilon}'\mathbf{M\varepsilon}\}$$

$$= E\left\{trace(\boldsymbol{\varepsilon}'\mathbf{M\varepsilon})\right\} = E\left\{trace(\mathbf{M\varepsilon\varepsilon}')\right\}$$

$$= trace\left\{\mathbf{M}E(\boldsymbol{\varepsilon\varepsilon}')\right\} = trace\left\{\mathbf{M}\sigma_\varepsilon^2\mathbf{I}_n\right\}$$

$$= \sigma_\varepsilon^2 \times trace\left\{\mathbf{I}_n - \mathbf{N}\right\} = \sigma_\varepsilon^2 \times trace\left\{\mathbf{I}_n\right\} - trace\left\{\mathbf{N}\right\}$$

$$= \sigma_\varepsilon^2 \times \{n - trace\left\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right\}$$

$$= \sigma_\varepsilon^2 \times \{n - trace\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right\} = \sigma_\varepsilon^2\left[n - trace(\mathbf{I}_k)\right]$$

$$= \sigma_\varepsilon^2(n-k)$$

$$\Rightarrow E\left(\frac{\mathbf{e}'\mathbf{e}}{n-k}\right) = \sigma_\varepsilon^2 \Rightarrow \text{use } s_e^2 \equiv \frac{\mathbf{e}'\mathbf{e}}{n-k} \text{ as LS (unbiased) est. } \sigma_\varepsilon^2.$$

A. NOTE: $\frac{\mathbf{e}'\mathbf{e}}{n-k}$ is sum of squared (asymptotic) normals, divided by degrees of freedom, so s_e^2 is (asymptotically) $\chi^2_{n-k} \Big/ n-k$.

B. Therefore, standard **Wald** *t*-tests & confidence intervals obtained by:

1. $$T = \frac{b_j - c_0}{s.e.(b_j)} = \frac{b_j - c_0}{\sqrt{s_e^2 \left\{ (X'X)^{-1} \right\}_{jj}}} \sim^{(A)} t_{n-k}$$

$$b_j \pm t^{\alpha}_{n-k} \times s.e.(b_j) = b_j \pm t^{\alpha}_{n-k} \times \sqrt{s_e^2 \left\{ (X'X)^{-1} \right\}_{jj}}$$

2. $$\Rightarrow \{1-\alpha\}\% \text{ (asympt.) conf. int.}$$

3. More generally, tests of linear restrictions (by **Wald strategy**):

   a) For instance, H_o: $\beta_1+\beta_2=1$. ***Wald-Test Logic:*** **If null-hypothesis were true, then estimate should not be far (in std-error units) from null**:

$$T = \frac{(b_1 + b_2) - 1}{s.e.(b_1 + b_2)} = \frac{(b_1 + b_2) - 1}{\sqrt{\hat{V}(\hat{b}_1 + \hat{b}_2)}}$$

$$= \frac{(b_1 + b_2) - 1}{\sqrt{\hat{V}(\hat{b}_1) + \hat{V}(\hat{b}_2) + 2 \times \hat{C}(\hat{b}_1, \hat{b}_2)}} \sim^{A} t_{n-k}$$

b) But note how we could use matrix algebra to generalize:

$H_o: \beta_1 + \beta_2 = 1 \Rightarrow \mathbf{r}'\boldsymbol{\beta} = q$; for instance, if $\boldsymbol{\beta}$ is (4×1), then:

$$\mathbf{r}' = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix} \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$\Rightarrow H_o: \mathbf{r}'\boldsymbol{\beta} = q; \; e.g., \; H_o: \mathbf{r}'\boldsymbol{\beta} = 1 \Rightarrow T = \frac{\mathbf{r}'\mathbf{b} - 1}{\sqrt{\hat{V}(\mathbf{r}'\mathbf{b})}} = \frac{\mathbf{r}'\mathbf{b} - 1}{\sqrt{\mathbf{r}'\hat{V}(\mathbf{b})\mathbf{r}}} \sim^{A} t_{n-k}$$

Notice how $\mathbf{r}' \ldots \mathbf{r}$ plucks $\hat{V}(\mathbf{b}_1), \hat{V}(\mathbf{b}_2), \hat{C}(\mathbf{b}_1, \mathbf{b}_2), \hat{C}(\mathbf{b}_1, \mathbf{b}_2)$, just as in scalar!

[Let's do some more examples (illustrative cases) of these simple (above) and joint (below) Wald-strategy hypothesis tests on board.]

4. Test of joint hypotheses (by **Wald Strategy**):

$H_o: \beta_1 = q_1, \beta_2 = q_2 \Rightarrow \mathbf{R\beta} = \mathbf{q}$;

for instance, $H_o: \beta_1 = 0, \beta_2 = 1$ and $\boldsymbol{\beta}$ is (4×1), then:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \text{ and } \mathbf{q} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$H_o: \beta_1 = q_1, \beta_2 = q_2 \Rightarrow \mathbf{R\beta} = \mathbf{q}$; for instance, $H_o: \beta_1 = 0, \beta_2 = 1$ and $\boldsymbol{\beta}$ is (4×1), then:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \text{ and } \mathbf{q} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Rightarrow H_o: \mathbf{R\beta} = \mathbf{q} \Rightarrow C = (\mathbf{Rb} - \mathbf{q})' \left[\hat{V}(\mathbf{Rb} - \mathbf{q}) \right]^{-1} (\mathbf{Rb} - \mathbf{q})$$

is a ratio of chi-squares. "Denominator" has n-k deg. free, and has n-k in its denom. "Numerator" is square 2×1 normals. So, C/2, or, more generally, C/J where J=rows(**R**), is $F_{J,n-k}$!

$$F = (\mathbf{Rb} - \mathbf{q})' \left[\hat{V}(\mathbf{Rb} - \mathbf{q}) \right]^{-1} (\mathbf{Rb} - \mathbf{q}) \Big/ J \sim^{A} F_{J,n-k}$$

1. (Sets of) Linear Restrictions:

$$R\underset{\sim}{\beta} = \underset{\sim}{q}$$

examples: a) $\beta_1 = 0$
b) $\beta_4 = 7$
c) $\beta_2 = 3$ & $\beta_4 = 0$
d) $\beta_2 = 2\beta_5 + 3$
e) $\beta_1 = 0,\ \beta_2 = 3\beta_3,$ & $\beta_4 = -\beta_6$

2. As we saw last week, you construct R & $\underset{\sim}{q}$ so that $R\underset{\sim}{\beta} = \underset{\sim}{q}$ states your hypotheses:

these examples assume your $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$

e.g. a) $R = \underset{(1\times k)}{[0\ 1\ 0\ 0\ \cdots]} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = 0 \leftarrow \underset{\sim}{q}$

b) $R = \underset{(1\times k)}{[0\ 0\ 0\ 0\ 1\ 0\ 0\ \cdots]} \underset{(k\times 1)}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}} = 7 \leftarrow \underset{\sim}{q}$

c) $R = \underset{(2\times k)}{\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & 0 & \cdots \end{bmatrix}} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \vdots \\ \beta_k \end{bmatrix} = \underset{(2\times 1)}{\begin{bmatrix} 3 \\ 0 \end{bmatrix}} \leftarrow \underset{\sim}{q}$

d) $R = \underset{(1\times k)}{[0\ 0\ 1\ 0\ 0\ -2\ 0\ 0\ \cdots]} \underset{(k\times 1)}{\begin{bmatrix} \beta \\ \sim \end{bmatrix}} = 3$

$\Rightarrow \beta_2 - 2\beta_5 = 3$

$\Rightarrow \beta_2 = 2\beta_5 + 3$ ✓ which is what we wanted

e) $R = \underset{(3\times k)}{\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -3 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & \cdots \end{bmatrix}} \underset{(k\times 1)}{\begin{bmatrix} \beta \\ \sim \end{bmatrix}} = \underset{(3\times 1)}{\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}} \leftarrow \underset{\sim}{q}$

3. Check to see whether $(R\underset{\sim}{b} - q)$ differs a "lot" from $(R\beta - q)$ (which latter equals zero by construction.)
($\underset{\sim}{b}$ estimates of β)
where "a lot" is measured in std. deviation units, or, equivalently, whether $(R\underset{\sim}{b} - q)^{"2"}$ differs a "lot" from $(R\beta - q)^{"2"}$ where "a lot" is measured in variance units

4. Which gives us the test statistic:

$$\frac{1}{J}(R\underset{\sim}{b} - q)'\{R[s^2(X'X)^{-1}]R'\}^{-1}(R\underset{\sim}{b} - q) = F\text{-stat}$$

is F "big" in an $F_{J,n-k}$ distribution?
i.e. is there little probability of an F this large or larger if null were true?

5. Notes: a) this is the $V(\underset{\sim}{b})$ estimated; thus, if $\underset{\sim}{b}$ estimated with "nice" properties & $V(\underset{\sim}{b})$ estimated w/ "nice" properties (BLUE, or consistent or ...) then F (Wald) test will also have "nice" properties

b) in cases (like examples a & b) where we're testing single hypotheses, the F-stat reduces to $\frac{(R\underset{\sim}{b} - q)^2}{RR'V(b)} = \left(\underset{\sim t_{n-k}}{T\text{-stat}}\right)^2 = \underset{\sim F_{1,n-k}}{F\text{-stat}}$

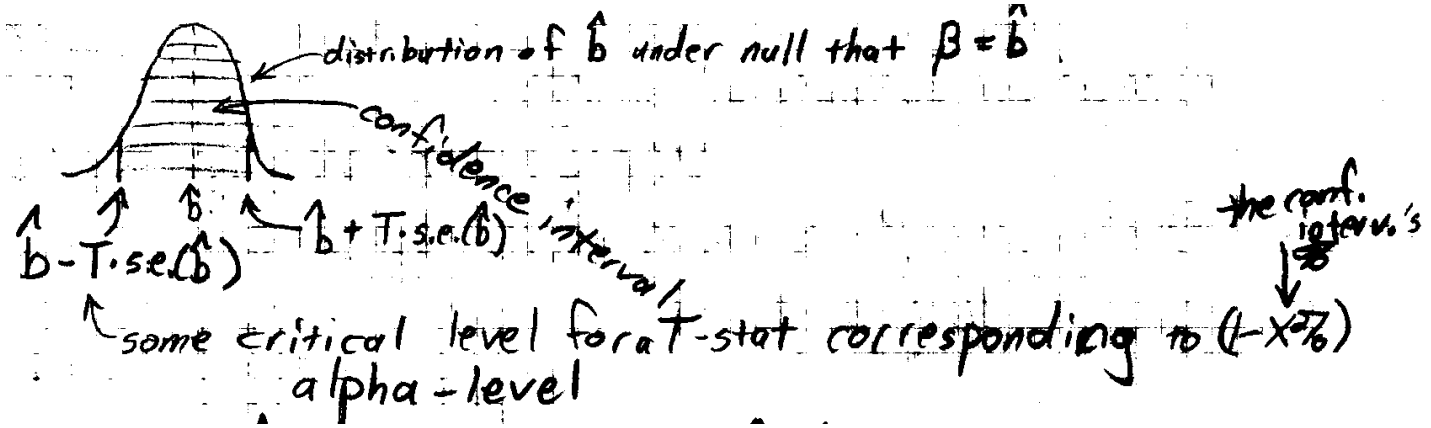
aside fact: $\forall\, T \sim t_{n-k}$, $T^2 \sim F_{1,n-k}$

C. Elaboration on Confidence Intervals & Confidence Regions

1. Recall the simple formula for confidence intervals:

$$(1-\alpha)\% \text{ confidence interval } (c.i.) \text{ for } \beta:$$

$$\hat{\beta} \pm T_{crit} \times s.e.(\hat{\beta})$$



2. Recall 1-for-1 correspondence b/w hypothesis tests @ level α & (1-α)% *c.i.*:

   a) *c.i.* overlaps 0 ⇔ fail-to-reject; 0 lies outside c.i. ⇔ reject

   b) The (1-α)% *c.i.* also corresponds to set of hypothesized β that one would fail-to-reject at level α given estimated *b*.

3. By latter, can see how might construct (1-α)% ***confidence region*** for (β_1, β_2) as set of (β_1, β_2) that we would fail-to-reject at α given est's (b_1, b_2):

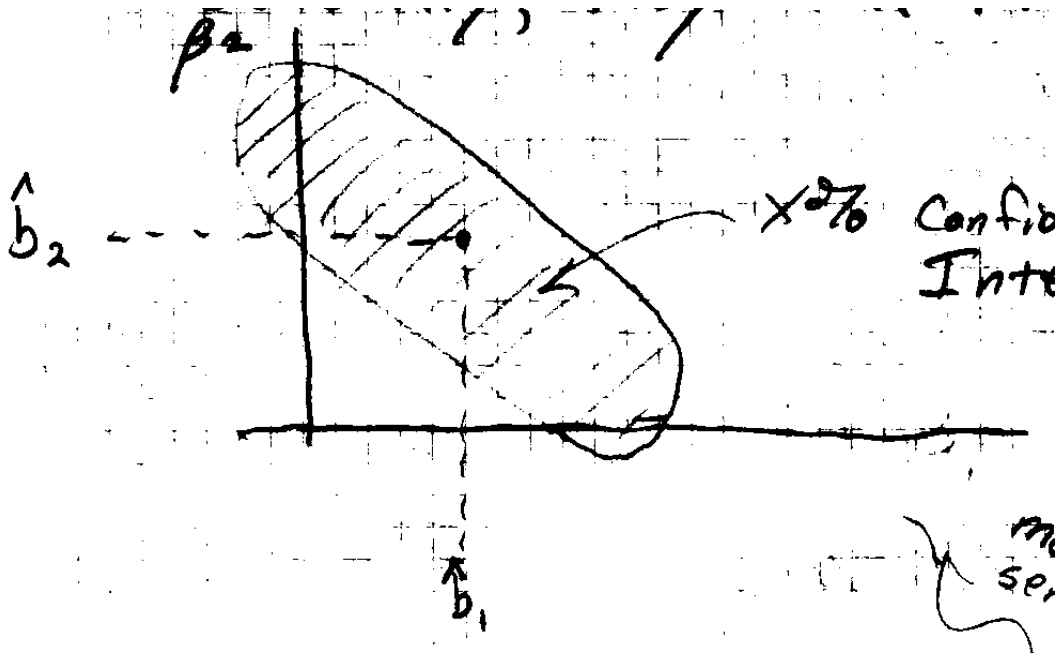
$$\frac{1}{J}\begin{bmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \end{bmatrix}' \begin{bmatrix} \hat{V}(b_1) & \hat{C}(b_1, b_2) \\ \hat{C}(b_1, b_2) & \hat{V}(b_2) \end{bmatrix}^{-1} \begin{bmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \end{bmatrix} \leq F_{J,n-k}$$

a) Using $F_{J,n-k}$ critical value for α from desired (1-α)% *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 (β_1, β_2) that just-satisfy inequality.

4. They generally *look* like this:



5. Are *properties* of conf. reg. intuitive to you?

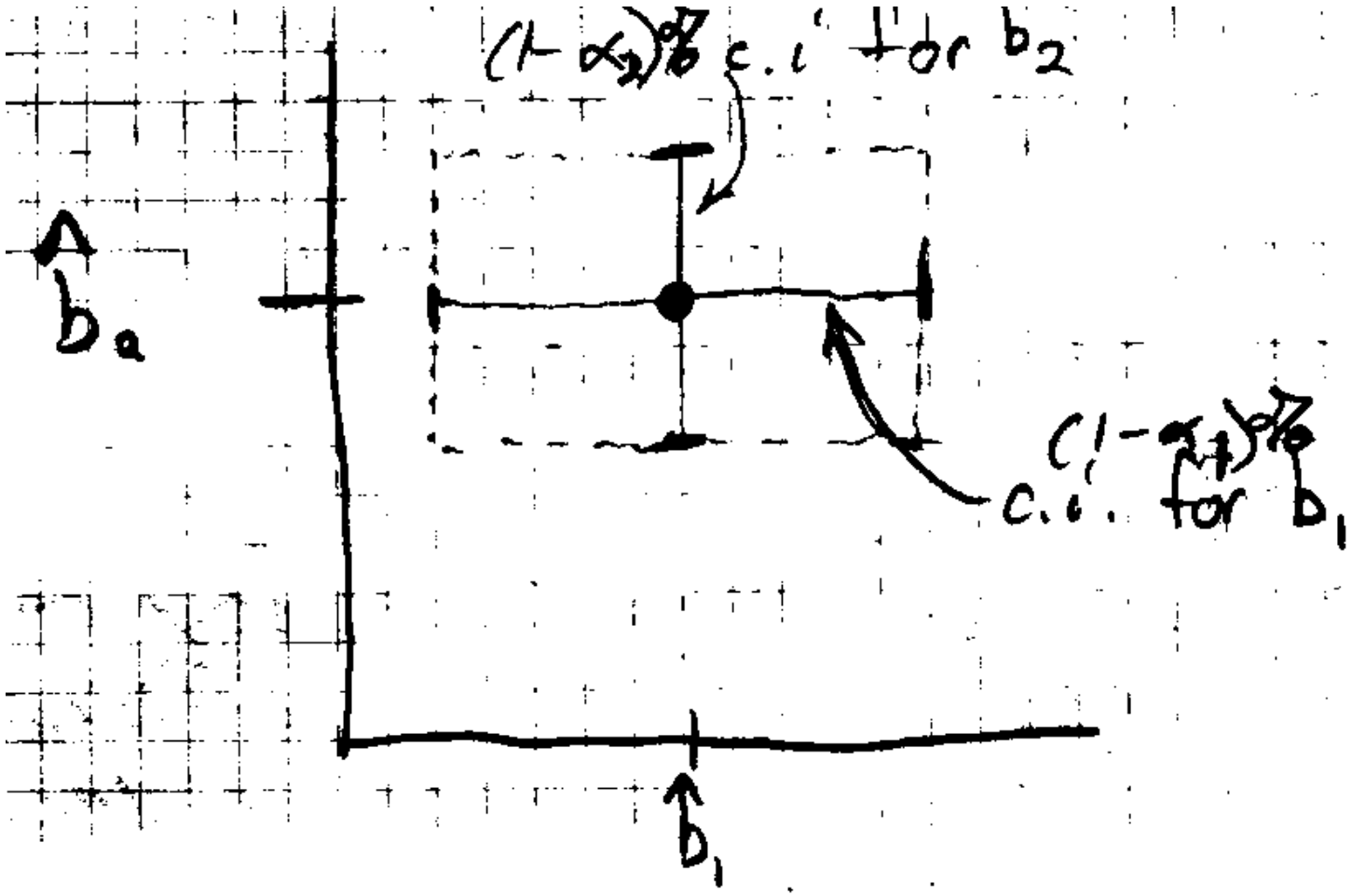
   a) Ellipsoidal and centered on (b_1,b_2).

   b) Go top-left to bottom-right if $C(b_1,b_2)<0$, & this will be when (partial) $C(x_1,x_2)>0$; go bottom-right to top-left if $C(b_1,b_2)>0$, & 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 Approx.*: Rectang. region of k univar $(1-\alpha_k)\%$ *c.i.* contain $\geq(1-\Sigma\alpha_k)\%$:



a) Ex: two 95% *c.i.*'s ε 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.):

$$s_e = \sqrt{s_e^2} = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{n-k}} \qquad \text{note: If } \boldsymbol{\varepsilon} \sim^{A} \text{MVN, then } s_e \sim^{A} \sqrt{\frac{\chi^2_{n-k}}{n-k}}$$

a) Sometimes also denoted σ or $\hat{\sigma}$, w/ or w/o sub *e* or ε, but perhaps best to reserve $\hat{\sigma}_e$ for ML est., using s_e for LS est., and definitely to use the hat & *e* not ε:

$$\hat{\sigma}_e = \sqrt{\sigma_e^2} = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{n}}$$

b) Notes:

(1) *Kinda* measure of typical or avg error or mistake. (Act'ly, measures square root of average squared mistake…)

(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 variation in **y** 'explained' (linearly accounted) by the model (**Xb**).

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{\sum[(y-e)-\bar{y}]^2}{\sum(y-\bar{y})^2} = \frac{\sum[(y-\bar{y})-e]^2}{\sum(y-\bar{y})^2}$$

$$= \frac{\sum(y-\bar{y})^2 - \sum(2(y-\bar{y})e) + \sum e^2}{\sum(y-\bar{y})^2}$$

$$= 1 - \frac{2\sum(\hat{y}e + e^2 - \bar{y}e) - \sum e^2}{\sum(y-\bar{y})^2}$$

$$= 1 - \frac{2\sum e^2 - \sum e^2}{\sum(y-\bar{y})^2} = 1 - \frac{\sum e^2}{\sum(y-\bar{y})^2} = 1 - \frac{SSE}{SST}$$

a) R^2 is also the square of the correlation of $\mathbf{y}$ & $\hat{\mathbf{y}}$, i.e., $r^2_{\mathbf{y},\hat{\mathbf{y}}}$.

b) If $\boldsymbol{\varepsilon} \sim^A MVN$, then $R^2 \sim^A \frac{\chi^2}{\chi^2} = F$.

3. Adjusted R^2, Adj. R^2, R-bar squared, $\bar{R}^2$:

   a) Can always increase R^2 just by add variables. Want some penalty for lack parsimony. Common adj. to R^2 is to replace numerator & denominator w/ unbiased estimates.

$$\bar{R}^2 = 1 - \frac{unbiased(SSE)}{unbiased(SST)} = 1 - \frac{\sum e^2 \big/ (n-k)}{\sum (y-\bar{y})^2 \big/ n-1} = 1 - \frac{s_e^2}{s_y^2}$$

   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^{kaike}IC, B^{ayesian}/S^{chwartz}IC, … http://en.wikipedia.org/wiki/Akaike%27s_information_criterion .

   c) Although not directly used (that I'm aware), notice that:

$$\text{If } \boldsymbol{\varepsilon} \sim^{A} MVN, \text{ then } \bar{R}^2 \sim^{A} \frac{\chi^2_{n-k} \big/ (n-k)}{\chi^2_{n-1} \big/ (n-1)} = F_{n-k,n-1}$$

4. (Log) Likelihood from ML est. is also measure of fit.

5. Use & Abuse of Fit Statistics/Measures:

   a) Can use to compare model perform *in same sample*; use to compare across samples only w/ great attention & care how much V(**y**) to explain varies across samples.

   b) At end, (relative) fit of model much more something to estimate given a model, than something to model to maximize.

E. "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-R^2" or "delta-R^2" or "ΔR^2" Test:**

   a) Determine how to *impose null hypothesis*. Example (pooled v. unit-by-unit TSCS):

$$\left.\begin{aligned} H_o&: \boldsymbol{\beta}_i = \boldsymbol{\beta} \quad \forall i \\ H_1&: \boldsymbol{\beta}_i \neq \boldsymbol{\beta} \quad \forall i \end{aligned}\right\} \Rightarrow \begin{cases} H_o: y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it} \\ H_1: y_{it} = \beta_0^i + \beta_1^i x_{it} + \varepsilon_{it} \end{cases}$$

b) Measure loss of explanatory power relative to gap from big-model explanatory power to 1, & divide each numerator & denominator by its degrees of freedom:

$$\text{loss of fit: } \Delta R^2 = R_1^2 - R_0^2 = \left(1 - \frac{SSE^1_{n-k_1}}{SST_{n-1}}\right) - \left(1 - \frac{SSE^0_{n-k_0}}{SST_{n-1}}\right) = \frac{SSE^0_{n-k_0} - SSE^1_{n-k_1}}{SST_{n-1}}$$

$$\text{fit-gap: } 1 - R_1^2 = \frac{SSE^1_{n-k_1}}{SST_{n-1}} \quad \Rightarrow \quad \text{ratio: } \frac{SSE^0_{n-k_0} - SSE^1_{n-k_1}}{SSE^1_{n-k_1}}$$

$$\Rightarrow \frac{\chi^2_{NT-k_0} - \chi^2_{NT-k_1}}{\chi^2_{NT-k_1}} \Rightarrow \ \circ\text{free: } \frac{k_1 - k_0}{NT - k_1} = \frac{\Delta k}{NT - k_1} = \frac{Nk - k}{NT - k_1}$$

$$\text{So: } \mathrm{F} = \frac{\Delta R^2 \big/ \Delta k}{(1 - R_1^2)\big/(NT - k_1)} \overset{(A)}{\sim} F_{\Delta k, NT - k_1}$$

4. Another example:

$$\left.\begin{array}{c} H_o: \beta_3 = \beta_4 = 0 \\ H_1: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \end{array}\right\} \Rightarrow \left\{\begin{array}{c} H_o: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ H_1: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \end{array}\right.$$

a) Determine how to impose hypothesis (this case, just drop x_3 & x_4).

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:

$$\text{loss of fit: } \Delta R^2 = R_1^2 - R_0^2 = \left(1 - \frac{SSE^1_{n-k_1}}{SST_{n-1}}\right) - \left(1 - \frac{SSE^0_{n-k_0}}{SST_{n-1}}\right) = \frac{SSE^0_{n-k_0} - SSE^1_{n-k_1}}{SST_{n-1}}$$

$$\text{fit-gap: } 1 - R_1^2 = \frac{SSE^1_{n-k_1}}{SST_{n-1}} \quad \Rightarrow \quad \text{ratio: } \frac{SSE^0_{n-k_0} - SSE^1_{n-k_1}}{SSE^1_{n-k_1}}$$

$$\Rightarrow \frac{\chi^2_{n-k_0} - \chi^2_{n-k_1}}{\chi^2_{n-k_1}} \Rightarrow \ ^\circ\text{free: } \frac{(n-k_0)-(n-k_1)}{n-k_1} = \frac{k_1 - k_0}{n-k_1} = \frac{\Delta k}{n-k_1}$$

$$\text{So: } F = \frac{\Delta R^2 / \Delta k}{(1-R_1^2)/(n-k_1)} \overset{(A)}{\sim} F_{\Delta k, n-k_1} \qquad \text{Equivalently, } \frac{\Delta SEE / \Delta k}{SEE^1/(n-k_1)}$$

5. For instance, “F-test of model”, i.e., loss-of-fit, by ΔR^2, test of null hypothesis all coefficients, except intercept, =0, i.e., that model provides 0 explan. pow.:

$$\left.\begin{array}{l} H_o: \beta_j = 0 \quad \forall j > 0 \\ H_1: \exists j \text{ s.t. } \beta_j \neq 0 \end{array}\right\} \Rightarrow \begin{cases} H_o: y = \beta_0 + \varepsilon \\ H_1: y = \beta_0 + \ldots + \beta_k x_k + \varepsilon \end{cases}$$

loss of fit: $\Delta R^2 = R_1^2 - R_0^2 = R_1^2 - 0 = R_1^2$; fit gap: $1 - R_1^2$; $\Delta^\circ \, free = k$

$$\text{So: } \mathrm{F} = \frac{R_1^2 / k}{(1 - R_1^2)/(n - k - 1)} \overset{(A)}{\sim} F_{k, n-k-1}$$

So, for this particular hypothesis, we only need to estimate 1 regression b/c we know the restricted regression has $R^2 = 0$.

More generally, the restrictions will reduce R^2 but not to zero:

e.g. model big $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + e$ $\quad R^2 = .80$

(* ← different result, right?)

model small $Y = b_0 + b_1 X_1 + e$ $\quad R^2 = .55$, $n = 150$

$\Delta R^2 = .25$ $\quad \Delta k = 2$ $\quad \Rightarrow$ F-stat. If this "big", then restriction really bit. How big is "big". Check p-level in the tail (to right))

$R^2_{big} = .80$ $\quad n - k_{big} = 146$

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 (those) variable(s) from the regression

e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$

$H_0: \beta_2 = 0$ & $\beta_3 = 0 \Rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \ldots + \beta_k X_k + \varepsilon$

2.) Suppose, instead:

a) $H_0: \beta_2 = 3$ & $\beta_3 = 1$, then $(Y - 3X_2 - X_3) = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \ldots$ embodies the restriction

b) $H_0: \beta_2 = 2\beta_3$, then $Y = \beta_0 + \beta_1 X_1 + \beta^*(X_2 + \frac{1}{2}X_3) + \beta_4 X_4 + \ldots$ embodies the restriction

c) in case (a), you "impose the restriction" by netting out $3X_2$ & X_3 from Y before regressing on the rest. This forces $b_2 = 3$ & $b_3 = 1$ to be fixed while we estimate the rest of $\underset{\sim}{b}$. In case (b), the estimate of b^*, the coefficient on $(X_2 + \frac{1}{2}X_3)$ will produce a coefficient on X_2 that is twice that of X_3 by construction.

d) Practically, then, in case (a) you create a variable, Y^*, which is $(Y - 3X_2 - X_3)$ & use that as dependent variable. In (b) you create $X_{23}^* = (X_2 + \frac{1}{2}X_3)$ & use that in place of X_2 & X_3

3.) Forcing OLS to labor under restrictions like this, you are certain that the restricted S^2 will be higher & (equivalently) that restricted R^2 will be lower. Note: these are residuals & 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: $$\left[\frac{(\Delta R^2/\Delta k)}{(1-R^2_{1,j})/(n-k)_j}\right] = F \sim F_{\Delta k, n-k}$$

* As before, less area there is to the right of F in a $F_{\Delta k, n-k}$ distribution, the less likely we are to have seen so much degradation of fit simply by chance.

6. Loss-of-fit tests using other measures of fit, s_e^2 or *ln(L)*, also:

   a) **Chow test** [of parameter stability; e.g.,] of fully pooled v. unit-by-unit regression in TSCS context [would be same for any grouping of data sub-samples vs. same model across whole sample]:

   b) $$\frac{\left(\mathbf{e}'\mathbf{e}_{pooled} - \sum_{i=1}^{N}\mathbf{e}_i'\mathbf{e}_i\right) \Big/ \Delta k_{=Nk-k}}{\sum_{i=1}^{N}\mathbf{e}_i'\mathbf{e}_i \Big/ \left(NT - \sum_{i=1}^{N} k_i\right)} \sim^{A} F_{Nk-k,\,NT-\Sigma_i k_i}$$

   c) **Likelihood-Ratio test:** $2[ln(\mathrm{L}_1)\text{-}ln(\mathrm{L}_0)] \sim \chi_r^2$.

F. Third logic, ***Lagrange-Multiplier* (LM) *Tests*:** if null hypoth were true, then to impose it as constraint on max{*ln(L)*} or min{*SSE*} should not bind, implying Lagrange multipliers, $\boldsymbol{\lambda}=\mathbf{0}$, and also implying that constrained $\{\partial \ln(L) \text{or} \partial SSE\}/\partial \boldsymbol{\beta}_{\text{at}\boldsymbol{\beta}\text{null}}=\mathbf{0}$.

1. Often, someone has shown how some properly configured post-estimation auxiliary regression has coefficients that are the Lagrange multipliers, & so can conduct LM test simply by Wald or Loss-of-Fit test on those auxiliary coefficient-estimates.

   a) E.g., LM test for s^{th}-order (time-)serial correlation remaining in residuals of a linear regression model simply regresses e_t on whatever e_{t-s} are desired, controlling for all **x** from original model (including lagged **y** or **x** if any): $T \times R^2 \sim^{A} \chi_r^2$, where r=# lags.

2. We can get to the Lagrange-multiplier statistics by directly ***restricted*** or ***constrained least-squares*** (i.e., min SSE s.t. hypothesized constraints):

$$\underset{\mathbf{b}}{\text{Min}}\ (\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb}) \text{ s.t. } \mathbf{Rb}=\mathbf{q} \ \Rightarrow$$

$$\underset{\mathbf{b},\boldsymbol{\lambda}}{\text{Min}}\ (\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})+2\boldsymbol{\lambda}(\mathbf{Rb}-\mathbf{q})\equiv\mathcal{L} \ \Rightarrow$$

(1) $\nabla_{\mathbf{b}}\mathcal{L}=\mathbf{0} \ \Rightarrow \ -2\mathbf{X}'(\mathbf{y}-\mathbf{Xb}^*)+2\mathbf{R}'\boldsymbol{\lambda}=\mathbf{0}$

(2) $\nabla_{\boldsymbol{\lambda}}\mathcal{L}=\mathbf{0} \ \Rightarrow \ 2(\mathbf{Rb}^*-\mathbf{q})=\mathbf{0}$

$$\Rightarrow \ (1)\ \mathbf{X}'\mathbf{Xb}^*+\mathbf{R}'\boldsymbol{\lambda}=\mathbf{X}'\mathbf{y} \ \Rightarrow \ (2)\ \mathbf{Rb}^*+\mathbf{0}'\boldsymbol{\lambda}=\mathbf{q} \Rightarrow \underbrace{\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{R}'\\ \mathbf{R} & \mathbf{0}'\end{bmatrix}}_{\mathbf{A}}\underbrace{\begin{bmatrix}\mathbf{b}^*\\ \boldsymbol{\lambda}\end{bmatrix}}_{\mathbf{d}^*}=\underbrace{\begin{bmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{q}\end{bmatrix}}_{\mathbf{v}}$$

(Notice (1) would be LS, except modified by constraints, $\mathbf{R}'\boldsymbol{\lambda}$.)

(Therefore, is unconstrained LS if $\boldsymbol{\lambda}=\mathbf{0}$., i.e., constraints not bind.)

So, $\mathbf{d}^*=\mathbf{A}^{-1}\mathbf{v}$, which, using formula for inverse partitioned matrix

$\Rightarrow$ (3) $\mathbf{b}^*=\mathbf{b}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})$, and

$\Rightarrow$ (4) $\boldsymbol{\lambda}^*=[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})$.

Can show from (3) that: $V(\mathbf{b}^*)=\sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}-\sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$,

i.e., that $V(\mathbf{b}^*)$=$V(\mathbf{b})-$(some positive semi-definite matrix), so $V(\mathbf{b}^*)$"<"$V(\mathbf{b})$, so imposing constraints, structure, is to add some information to the estimation (to ask less of the data, because you provide some parameter values, not ask to estimate).

Lagrange-Multiplier Test uses (4). If $\boldsymbol{\lambda}^*=\mathbf{0}$, restrictions not binding, i.e., they imposed no penalty on fit and, from (3) above, they improved the precision of the estimates, **b**.

3. Comparing the Logics: Why would we want 3 testing strategies? On what bases ought we to choose between them?

a) In general, asymptotically equivalent; however, small-sample properties may differ.

b) Based on estimates of different quantities—e.g., Wald based on estimates of (particular) **b** & V(**b**) vs. Loss-Fit based on estimates of fit (& so on set or sum of **b**)—so need confidence different aspects of your estimation results.

c) Feasibility: sometimes easier implement on or other of tests. E.g., Chow parameter-stability by Loss-of-Fit generally easier implement than Wald of same hypothesis.

d) Tests require estimation of different models: Wald=only unconstrained; LM=only constrained; LR=both. In figs below, C='curvature'; S='slope'.

$$LR = 2\left(\log L(\hat{\theta}) - \log L(\theta_0)\right)$$

$$W = (\hat{\theta} - \theta_0)^2 C(\hat{\theta})$$

$$LM = \left[S(\theta_0)\right]^2 C(\theta_0)^{-1}$$

Not unrelated to feasibility. E.g., tests spatial v. non-spatial models have latter as constraint on former. Test by LM requires only estimate non-spatial model, which far easier.

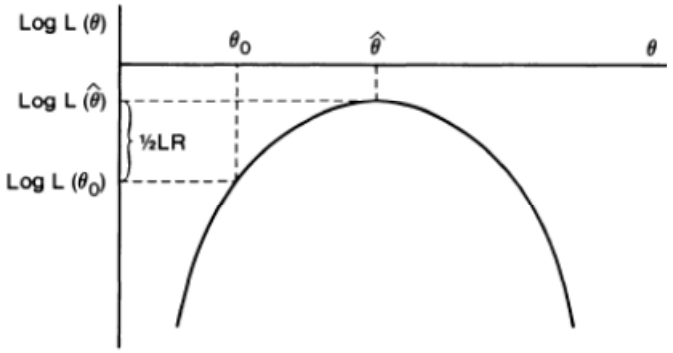


Figure 1. The Likelihood Ratio Test

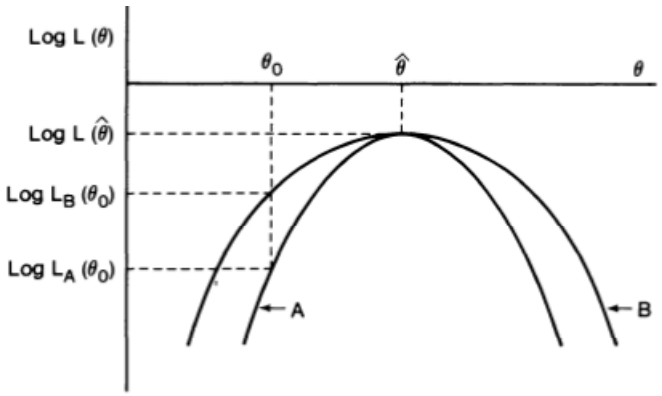


Figure 2. The Wald Test

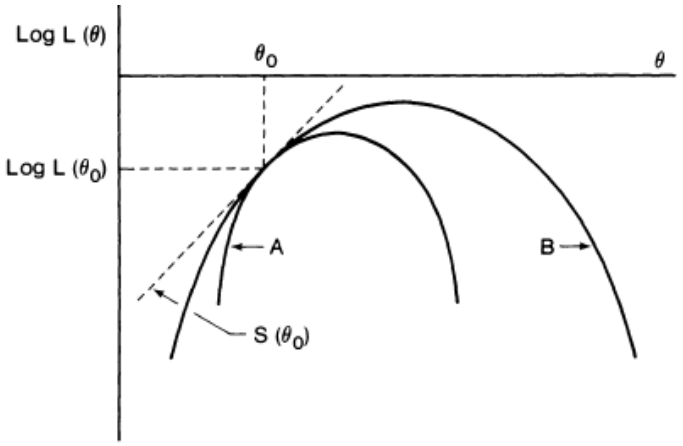


Figure 3. The Lagrange Multiplier Test

Prediction: You can look over this on your own. We are generally less interested in predicting & forecasting per se than in evaluating theory which is generally about relationships b/w variables (i.e. coefficients).

Still, note:

A. Predicting some y_j using the vector of explan. vars. $\underset{\sim}{X}_j$

$$\Rightarrow \hat{y}_j \equiv \underset{\sim}{X}_j' \hat{\underset{\sim}{b}}$$

($\underset{\sim}{X}_j$: data for obs. j; $\hat{\underset{\sim}{b}}$: column, estimated from some sample (not including j usually))

Then $e_j \equiv y_j - \hat{y}_j$ is the prediction error

$$V(e_j) = V(y_j - \hat{y}_j) = \sigma^2 + V(\hat{y}_j)$$

$$= \sigma^2 + \underset{\sim}{X}_j' [\sigma^2 (\underset{\sim}{X}'\underset{\sim}{X})^{-1}] \underset{\sim}{X}_j$$

($(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$: from the sample used to predict j)

$\Rightarrow$ Forecast Interval for y_j: (confidence)

$$\hat{y}_j \pm T \cdot s.e.(e_j) = \hat{y}_j \pm T \cdot \sqrt{s_e^2 + \underset{\sim}{X}_j' [s_e^2 (\underset{\sim}{X}'\underset{\sim}{X})^{-1}] \underset{\sim}{X}_j}$$

(s_e^2: estimate of σ^2)

(T: critical value from t_{n-k} in sample as usual)