PS699: STATISTICAL METHODS II

The Generalized (Normal) Linear-Regression Model: The G(N)LRM

I. The G(N)LRM: Summary Overview

A. *Formal Assumptions*

1. $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, and $E(\boldsymbol{\varepsilon}) = \mathbf{0}$;

2. $V(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ some symmetric p.d. matrix $\neq \mathbf{I}$;

3. $E(\mathbf{X}'\boldsymbol{\varepsilon}) = \mathbf{0}$, and $\mathbf{X}$ of full-column rank;

---

4. optional convenience: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Omega})$;

5. optional: $\mathbf{X}$ non-stochastic; if not, plim $\mathbf{X}'\mathbf{X}$ defined, finite, & p.d..

B. *Intuiting what changes from C(N)LRM to G(N)LRM:*

For What did we use $V(\underset{\sim}{\varepsilon}) = \sigma^2 I$ from the CLRM?

① Estimating σ^2; To show OLS efficient, & what it's standard errors are.

② Distribution of $\underset{\sim}{b}_{LS}$; & so for hypoth. tests, etc.

i.e., to show $V(\underset{\sim}{b}_{LS}) = \sigma^2 (X'X)^{-1}$

We did <u>not</u> use it to show:

③ $\underset{\sim}{b}_{LS}$ unbiased, consistent, or asymptotically normal

…so OLS on data gen'd by G(N)LRM will retain these properties (proofs to come).

C. Visualizing the variance-covariance (of stoch. comp.) matrix

What does $\sigma^2\Omega$ look like in multilevel data? (for example)

$$V(\underset{\sim}{\varepsilon}) = \begin{bmatrix} \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1n_1} \\ \vdots & \ddots & & \\ \sigma_{n_1 1} & \cdots & & \sigma_{1n_1}^2 \end{pmatrix}_1 & \cdots & \begin{pmatrix} \sigma_{11,J1} & \cdots & \\ \vdots & \ddots & \\ \sigma_{n_1 1,J1} & \cdots & \sigma_{1n_1,Jn_J} \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} \sigma_{J1,11} & \cdots & \\ \vdots & \ddots & \\ \cdots & & \sigma_{Jn_J 1 n_1} \end{pmatrix} & \cdots & \begin{pmatrix} \sigma_{J,J1}^2 & & \\ & \ddots & \\ & & \sigma_{J,n_J}^2 \end{pmatrix}_J \end{bmatrix}$$

- What does OLS assume?
- What might be more reasonable to assume for ML data?
- Why do we care if:

$\rightarrow \sigma_{ij} \neq 0$ for $i \neq j$?

$\rightarrow \sigma_i^2 = \sigma \ \forall i$

- Relax $\sigma^2 I$, but how many unique elements in $V(\varepsilon)$?

Note: $\varepsilon_{ij} = \eta_j + \omega_{ij} \Rightarrow Cov(\varepsilon_{ij}, \varepsilon_{kj}) = Cov(\eta_j + \omega_{ij}, \eta_j + \omega_{kj}) = Var(\eta_j) = \sigma_\eta^2$

Ω=

$$
\begin{bmatrix}
\omega_{11}^2 & \omega_{1,12} & \omega_{1,13} & \cdots & \omega_{1,1T} & \omega_{1,21} & \omega_{1,22} & \omega_{1,23} & \cdots & \omega_{1,2T} & \omega_{1,31} & \omega_{1,32} & \omega_{1,33} & \cdots & \omega_{1,3T} & & \omega_{1,N1} & \omega_{1,N2} & \omega_{1,N3} & \cdots & \omega_{1,NT} \\
\omega_{12,11} & \omega_{12}^2 & & & \vdots & \omega_{12,21} & \omega_{12,22} & & & \vdots & \omega_{12,31} & \omega_{12,32} & & & \vdots & & \omega_{12,N1} & \omega_{12,N2} & & & \vdots \\
\omega_{13,11} & & \omega_{13}^2 & & \vdots & \omega_{13,21} & & \omega_{13,23} & & \vdots & \omega_{13,31} & & \omega_{13,33} & & \vdots & \cdots & \omega_{13,N1} & & \omega_{13,N3} & & \vdots \\
\vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & & \vdots & & & \ddots & \vdots \\
\omega_{1T,11} & \cdots & \cdots & \cdots & \omega_{1T}^2 & \omega_{1T,21} & \cdots & \cdots & \cdots & \omega_{1T,2T} & \omega_{1T,31} & \cdots & \cdots & \cdots & \omega_{1T,3T} & & \omega_{1T,N1} & \cdots & \cdots & \cdots & \omega_{1T,NT} \\
\omega_{21,11} & \omega_{21,12} & \omega_{21,13} & \cdots & \omega_{21,1T} & \omega_{21}^2 & \omega_{21,22} & \omega_{21,23} & \cdots & \omega_{21,2T} & & & & & & & & & & & \\
\omega_{22,11} & \omega_{22,12} & & & \vdots & \omega_{21,22} & \omega_{22}^2 & & & \vdots & & & & & & & & & & & \\
\omega_{23,11} & & \omega_{23,13} & & \vdots & \omega_{21,23} & & \omega_{23}^2 & & \vdots & & & & & & & & & \vdots & & \\
\vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots & & & & & & & & & & & \\
\omega_{2T,11} & \cdots & \cdots & \cdots & \omega_{2T,1T} & \omega_{21,2T} & \cdots & \cdots & \cdots & \omega_{2T}^2 & & & & & & & & & & & \\
\omega_{31,11} & \omega_{31,12} & \omega_{31,13} & \cdots & \omega_{31,1T} & & & & & & \omega_{31}^2 & \omega_{31,32} & \omega_{31,33} & \cdots & \omega_{31,3T} & & & & & & \\
\omega_{32,11} & \omega_{32,12} & & & \vdots & & & & & & \omega_{32,31} & \omega_{32}^2 & & & \vdots & & & & & & \\
\omega_{33,11} & & \omega_{33,13} & & \vdots & & & & & & \omega_{33,31} & & \omega_{33}^2 & & \vdots & & & & \vdots & & \\
\vdots & & & \ddots & \vdots & & & & & & \vdots & & & \ddots & \vdots & & & & & & \\
\omega_{3T,11} & \cdots & \cdots & \cdots & \omega_{3T,1T} & & & & & & \omega_{3T,31} & \cdots & \cdots & \cdots & \omega_{3T}^2 & & & & & & \\
 & & \vdots & & & & & & & & & & & & & \ddots & & & \vdots & & \\
\omega_{N1,11} & \omega_{N1,12} & \omega_{N1,13} & \cdots & \omega_{N1,1T} & & & & & & & & & & & & \omega_{N1}^2 & \omega_{N1,N2} & \omega_{N1,N3} & \cdots & \omega_{N1,NT} \\
\omega_{N2,11} & \omega_{N2,12} & & & \vdots & & & & & & & & & & & & \omega_{N2,N1} & \omega_{N2}^2 & & & \vdots \\
\omega_{N3,11} & & \omega_{N3,13} & & \vdots & & & \cdots & & & & & \cdots & & & \cdots & \omega_{N3,N1} & & \omega_{N3}^2 & & \vdots \\
\vdots & & & \ddots & \vdots & & & & & & & & & & & & \vdots & & & \ddots & \vdots \\
\omega_{NT,11} & \cdots & \cdots & \cdots & \omega_{NT,1T} & & & & & & & & & & & & \omega_{NT,N1} & \cdots & \cdots & \cdots & \omega_{NT}^2
\end{bmatrix}
$$

Note: $\varepsilon_{it} = \rho\varepsilon_{it-1} + \omega_{it} \Rightarrow Cov(\varepsilon_{it}, \varepsilon_{it-1}) = Cov(\varepsilon_{it}, \rho\varepsilon_{it-1}) = \rho Var(\varepsilon)$

D. Properties of OLS under the G(N)LRM:

① $\underset{\sim}{b}_{LS}$ UNBIASED:

$$\underset{\sim}{b}_{LS} = \overset{(X'X)^{-1}X'}{A}\underset{\sim}{y} = AX\underset{\sim}{\beta} + A\underset{\sim}{\varepsilon} = \underset{\sim}{\beta} + A\underset{\sim}{\varepsilon}$$

$$\Rightarrow E(\underset{\sim}{b}_{LS}) = \underset{\sim}{\beta}$$

Unbias/Consistent continues to rely on E(**x′ε**)=0 or plim= **x′ε**.

② $V(\underset{\sim}{b}_{LS}) = V(Ay) = A\,V(\underset{\sim}{y})A' = A\,V(\underset{\sim}{\varepsilon})A' = A\sigma^2\Omega A'$

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

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

Notice that V(**b**) continues →0 as $n\to\infty$ (as long as off-diag elements of **Ω** (covar's) dampen at sufficient rate as get farther from diag (i.e., as $n\to\infty$): **<u>b consistent also</u>**.

$\Rightarrow$ OLS standard-errors wrong (biased, inconsistent, ineff.)

OLS coefficients inefficient (b/c useful info unexploited)

(Perhaps better to say OLS V(**b**) "not unbiased"; could be unbiased, not generally so though.)

E. Strategies Address & Redress deficiencies OLS under G(N)LRM:

1. "Sandwich Estimators":

- Since OLS $\underset{\sim}{b}$ "only" inefficient, could fix s.e.'s:
- Difference OLS $\widehat{V(\underset{\sim}{b})}$ from CLRM to GLRM: I v. Ω

$$\underbrace{[X'\sigma^2 I X]}_{Q} \quad \text{vs.} \quad \underbrace{[X'\sigma^2 \Omega X]}_{Q^*}$$

Q^* is $k \times k$, where as Ω $N \times N$

$$Q^* = \sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon_i \varepsilon_j \underset{\sim}{X}_i \underset{\sim}{X}_j'$$

$\therefore$ only need estimate the $\frac{1}{2}k(k+1)$ unique elements of $\mathbf{Q}$, not the $\frac{1}{2}n[n+1]$ (which $>n$) unique elements of $\mathbf{\Omega}$.

a) Need estimate bracketed term: $V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{X}'\mathbf{\Omega}\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1}$ **. An *OMEGA sandwich*! Get it?**

(1) A correction only need concern itself, therefore, with how:

ε_i^2 and ε_{ij}, i.e. the variations & covariations,

associate with corresponding $\mathbf{x}_i$, $\mathbf{x}_i^2$, and $\{x_i\} \cdot \{x_j\}$,

(2) …that is, with this sort of term: $\sum_i \sum_j e_i e_j \{\mathbf{x}_i \mathbf{x}_j' + \mathbf{x}_j \mathbf{x}_i'\}$.

b) Examples:

Sandwich Estimators: First one offered: "Pure Heteroskedast."

$$V(\varepsilon) = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$$

$$Q^* = X'\Omega X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_k & - \end{bmatrix} \begin{bmatrix} \omega_{11} & & & 0 \\ & \omega_{22} & & \\ & & \ddots & \\ 0 & & & \omega_{nn} \end{bmatrix} \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_k \\ | & | & & | \end{bmatrix}$$

$$= \sum_{i=1}^{n} \omega_{ii} \underset{\sim}{x}_i \underset{\sim}{x}_i'$$

$$\Rightarrow \hat{Q}^* = \sum_{i=1}^{n} e_i^2 x_i x_i'$$

$$= \sum_{i=1}^{n} \omega_{ii} \underset{\sim}{X}_i \underset{\sim}{X}_i'$$

$$\Rightarrow \hat{Q}^* = \sum_{i=1}^{n} e_i^2 X_i X_i'$$

$$\Rightarrow \widehat{V_s(b_{LS})} = (X'X)^{-1} X' \hat{Q}^* X (X'X)^{-1}$$

- Many extensions since, essentially versions of:

$$\hat{Q}^* = \sum_i \sum_j e_i e_j \underset{\sim}{X}_i \underset{\sim}{X}_j' \quad ,$$

but not all necessarily "well behaved" so 'tricks' added

$\hat{Q}^*$ converges to constant as $N \to \infty$

$\hat{Q}^*$ positive definite

c) Properties of OLS+Sandwich (i.e., rightly constituted sandwich) under GLRM:

(1) ***b*** *unbiased, consistent, but inefficient.*

(2) *V(**b**) "not unbiased" and inefficient, but at least consistent.*

(a) "not unbiased" here regards a small-sample issue of estimating $\hat{Q}^*$ versus known Q^*.

2. (Asymptotically) Efficient Least-Squares Estimation under the G(N)LRM: (Feasible) Generalized Least-Squares, (F)GLS

<u>(Feasibly) Efficient Estimation</u>: Need use the wasted info:

Stipulate: $\exists$ matrix $\Omega^{-1/2}$, "square root of the inverse of Ω"

Then: $\Omega^{-1/2} \underset{\sim}{y} = \Omega^{-1/2} X \underset{\sim}{\beta} + \Omega^{-1/2} \underset{\sim}{\varepsilon}$ under GLRM

follows the CLRM, so OLS on this BLUE

$$\underset{\sim}{\varepsilon}^* \equiv \Omega^{-1/2} \underset{\sim}{\varepsilon} \Longrightarrow V(\underset{\sim}{\varepsilon}^*) \equiv E(\boldsymbol{\varepsilon}^* \boldsymbol{\varepsilon}^{*\prime})$$

$$= E(\Omega^{-1/2} \underset{\sim}{\varepsilon} \underset{\sim}{\varepsilon}' \Omega^{-1/2}) = \Omega^{-1/2} E(\underset{\sim}{\varepsilon} \underset{\sim}{\varepsilon}') \Omega^{-1/2}$$

$$= \Omega^{-1/2} \sigma^2 \Omega \, \Omega^{-1/2} = \sigma^2 \Omega \Omega^{-1/2} \Omega^{-1/2}$$

$$= \sigma^2 \Omega \Omega^{-1} = \sigma^2 I \; \square$$

a) Pure Heteroscedasticity example:

"Pure Het" Example:

in Multilevel Context

$$V(\underset{\sim}{\varepsilon}) = \underset{\sim}{\Omega} = \begin{bmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & 0 & \\ & & \sigma_2^2 & & & \\ & & & \ddots & & \\ & 0 & & & \sigma_3^2 & \\ & & & & & \ddots \end{bmatrix}$$

e.g.

$$V(\underset{\sim}{\varepsilon}) = \underset{\sim}{\Omega} = \begin{bmatrix} 2 & & & & & \\ & \ddots & & & 0 & \\ & & 2 & & & \\ & & & 6 & & \\ & 0 & & & \ddots & \\ & & & & & 6 \\ & & & & & & 3 \\ & & & & & & & \ddots \\ & & & & & & & & 3 \end{bmatrix}$$

$$\Rightarrow \Omega^{-1/2} = \begin{bmatrix} 1/\sqrt{2} & & & & & \\ & \ddots & & & 0 & \\ & & 1/\sqrt{2} & & & \\ & & & 1/\sqrt{6} & & \\ & 0 & & & \ddots & \\ & & & & & 1/\sqrt{6} \\ & & & & & & 1/\sqrt{3} \\ & & & & & & & \ddots \\ & & & & & & & & 1/\sqrt{3} \end{bmatrix}$$

$\Rightarrow$ Using the Wasted info:

→ More reliable units weighed more heavily

b) AR(1) (temporal) correlation example:

$$y_t = \underset{\sim}{x}_t \beta + \varepsilon_t \quad , \quad \varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$$

$$V(\varepsilon) = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1 & \\ & & & & \ddots \\ & & & & & 1 \end{bmatrix} \Rightarrow \Omega^{-1/2} = \begin{bmatrix} \sqrt{1-\rho^2} & & & & 0 \\ -\rho & 1 & & & \\ 0 & -\rho & 1 & & \\ 0 & 0 & -\rho & \ddots & \\ & 0 & & \ddots & \\ & & & -\rho & 1 \end{bmatrix}$$

$$\Omega^{-1/2} y \Rightarrow \begin{bmatrix} (\sqrt{1-\rho^2}) y_1 \\ y_2 - \rho y_1 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}$$

analogous for X & for ε

i.e., partial differencing

Recall: issue with $\sigma_{ij} \neq 0$ was that each obs not wholly new info. This effects a weighing of only the "new" info

II. Elaborated General Introduction to Generalized (Normal) Linear-Regression Model

NON-SPHERICAL DISTURBANCES (i.e. $V(\varepsilon) \neq \sigma^2 I$) ①

I. Background Discussion → begin with C(N)LRM:

① $Y = X\beta + \varepsilon$
② $E(\varepsilon) = 0$
③ $V(\varepsilon) = \sigma^2 I$
④ $Cov(X, \varepsilon) = 0$
⑤ X of full-column rank
(⑥ $\varepsilon_i \sim N(0, \sigma^2)$ iid)

Greene Sections: [redacted]
Chapter 8
(6th ed, 2008)

- Have been dealing with ①, in particular: incorrect set of X's, non-linear in X
  Now we move on to consider violations of ③

$V(\varepsilon) = \sigma^2 \Omega$, since each obs. can have diff. variance (σ_{ii}) think of the σ^2 here as pulling out some base-line obs' variance to which others are scaled. It's in that sense arbitrary, usually it's taken to be the mean variance

A. What did we use $V(\varepsilon) = \sigma^2 I$ for?
   A. Estimating σ^2
   B. Distribution of b_{LS} (i.e. hypoth. tests, conf. intervals, etc.)

1) I.e. did not use it to prove OLS: { • Unbiased • Consistent } (nor that b_{LS} asymptotically normal)

2) Did use it to prove: OLS efficient & that OLS std Errors are unbiased, consistent, & efficient

B. What does $V(\varepsilon) = \sigma^2\Omega$ mean?

$$V(\varepsilon) = E\begin{bmatrix} \varepsilon_1\varepsilon_1 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_N \\ \varepsilon_2\varepsilon_1 & \varepsilon_2\varepsilon_2 & \cdots & \varepsilon_2\varepsilon_N \\ \vdots & & \ddots & \vdots \\ \varepsilon_N\varepsilon_1 & \varepsilon_N\varepsilon_2 & \cdots & \varepsilon_N\varepsilon_N \end{bmatrix}$$

then

$$V(\varepsilon) = \sigma^2\Omega = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & & \\ \sigma_{31} & & \ddots & \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

$\cdots\rightarrow$ We're interested in Cov(obs i w/ obs j)

Var(obs$_i$), defined

$\sigma_{ij} \equiv Cov(\varepsilon_i, \varepsilon_j)$

& $\sigma_{ij} = \sigma_{ji}$

Why do we care if $\sigma_{ij} \neq 0$ for some (or some set of) (i,j):

→ $\sigma_{12} \neq 0$ implies that observations 1 & 2 (i.e. their stochastic part) are not independent. This means that, given obs 1, obs 2 is not completely new info, & vice versa. Accordingly, we would be lying to say "I have 200 observations" if it were assumed, as OLS does, that these are 200 observations not stochastically related (uncorrelated). Thus, we're exaggerating how much info we actually have.

→ $\sigma_{ii} \neq \sigma_{jj}$ means that observations i & j have differing amounts of stochastic relative to systematic info, $(y = f(X, \varepsilon))$ if obs i more or less systematic relative to stochastic info. ⟹ this is itself something we know, but OLS & the CNLRM has imposed otherwise. Thus, we're not using all the info we have

more (still) bogey

→ also, $\hat{\sigma}_{12} \neq 0$ &/or $\hat{\sigma}_{ii} \neq \hat{\sigma}_{jj}$ might well imply that we've left something out that correlates across obs &/or explains different amounts across obs ⟹ might indicate omitted variables of a sort

- $\sigma_{ij} \neq 0 \Rightarrow$ each obs not wholly new info; exaggerate info in data, & so misstate, likely understate, uncertainty. Also inefficient to have ignored this info. Also, worse, possibly indicative omitted systematic features.
- $\sigma_i^2 \neq \sigma_j^2 \Rightarrow$ obs have differing shares stochastic & systematic info, so inefficient to have ignored this info & misstate uncertainty. Also, worse, possibly indicative omitted systematic features.

C. A More Concrete [redacted] Illustration: Suppose we have observations on some y, say votes for incumbent, & some X explanatory variable, say the unemployment rate. We observe (y, X) in 20 countries over 40 years. Then

$$y_{ct} = X_{ct}\beta + \varepsilon_{ct}$$ is the model for country $c \in [1,2,\ldots 20]$ & $t \in [1950, 1951, \ldots, 1990]$ or $t = [1,2,\ldots 40]$

$$V(\varepsilon_{ct}) \equiv \sigma^2_{ct} \quad \& \quad Cov(\varepsilon_{ct}, \varepsilon_{du}) \equiv \sigma_{ct,du}$$

just some other ctry-year

$$\Rightarrow V(\underset{\sim}{\varepsilon}) = \begin{bmatrix} \sigma^2_{11} & \sigma_{11,21} & \cdots & \\ \sigma^2_{12} & \sigma^2_{21} & & \\ \vdots & \sigma^2_{22} & & \\ \sigma^2_{140} & \sigma_{140,240} & & \\ & \sigma^2_{240} & & \\ & & & \sigma^2_{201} \\ & & & \sigma^2_{2040} \end{bmatrix}$$

Cov of each ctry 2 obs with each ctry 1 obs

ctry 1, years 1 → 40

- This is a time-series-cross-section (TSCS) sample

- nb, this just an example, could be individuals organized by states, provinces by ctry, individuals over time, etc.

$NT \times NT$

N → # of individs

T → # of time periods

1) The C(N)LRM assumes $V(\varepsilon) = \sigma^2 I$

$\Rightarrow$ No obs. in any ctry-year is correlated with any other obs, within & across both ctry & year.

$\Rightarrow$ The variance of every obs. (w/in & across ctry & year) is the same.

1) The C(N)LRM assumes $V(\varepsilon) = \sigma^2 I$

$\Rightarrow$ No obs. in any ctry-year is correlated with any other obs, within & across both ctry & year.

$\Rightarrow$ The variance of every obs. (w/in & across ctry & year) is the same.

2) Now this might well be overly restrictive. We might well suppose

a) Obs corr. over time w/in ctry (usually called <u>auto-</u> or <u>serial correlation</u>)

b) Obs. corr. across ctry's at same times (usually called <u>contemporaneous correlation</u>)

c) Obs. diff. variance by ctry (ctry-wise <u>heteroskedasticity</u>)

aside: notice that we can distinguish spatial & temporal correlation like

[redacted]

3) Now suppose we we unwilling to make any restrictions on the parameters in $V(\varepsilon)$. We wish to leave it open that Cov(1950, 1951) differs from Cov(1970, 1972) that Cov(US, JA) differs from Cov(US, UK) that V(US1950) differs from V(GE1972) & so on. How many unique σ_{ij} are there to estimate, then? Square: $NT \times NT$ elements. Only Diagonal & one of triangles are unique ($\sigma_{ij} = \sigma_{ji}$). So:

$$\underbrace{\tfrac{1}{2}(NT)^2}_{\text{½ the off-diagonals}} + \underbrace{\tfrac{1}{2}NT}_{\text{the remaining ½ the diagonal}} = \tfrac{1}{2}NT(NT+1) \text{ parameters!}$$

I.C.3. TSCS Example

a) So, if we wanted to estimate all of the $E(\varepsilon_i \varepsilon_j)$ separately we'd have $\frac{1}{2}n(n+1)$ parameters to estimate & only n observations. "Is this even possible?" amounts to

$$\left\{ \max_n \; n - \tfrac{1}{2}(n^2+n) \right\} > 0$$

$$\Rightarrow \text{F.O.C. } \frac{\partial}{\partial n}\{\;\} = 1 - n - \tfrac{1}{2} = 0$$

max is at $n = \frac{1}{2}$. Best we could have is $\frac{1}{2}$ an observation?! At which degrees of freedom are

$$\tfrac{1}{2} - \tfrac{1}{2}\left[\left(\tfrac{1}{2}\right)^2 + \tfrac{1}{2}\right] = \tfrac{1}{2} - \tfrac{1}{8} - \tfrac{1}{4} = \tfrac{1}{8}$$

b) All this is just more proof than perhaps you needed that it is <u>logically impossible</u> to garner enough info to identify (&: begin to estimate) that many parameters

c) Upshot: however you come to evaluate evidence you have <u>no</u> <u>choice</u> but to make some assumptions about how your observations relate & vary which will limit the number of things you don't know so you can gain leverage (degrees of freedom) with which to estimate relations you care about.

- You should try to keep in mind what sort of implicit assumptions you're making when you infer from data & also how much or little "new info" each observation actually is given those assumptions. (We just showed that without any assumptions you have no new info -- that should perhaps have been logically obvious.)

4. Some Possible Assumptions limiting the number of parameters in $V(\varepsilon)$ to estimate:

a) C(N)LRM & OLS : $V(\varepsilon) = \sigma^2 I$ ← 1 parameter

b) Time-Series:
$Cov(\varepsilon_t, \varepsilon_{t-i})$ = some simple declining function of temporal distance b/w the obs. One very common one is $\varepsilon_t = \rho \varepsilon_{t-1} + \gamma_t$ (ρ: serial corr. parameter; γ_t: iid)
$\Rightarrow$ 2 parameters: σ_γ^2 & ρ

c) Cross-Sections: $V(\sigma_j^2)$ different for each "group" j → groups would be theoretically defined in bestrase scenario.

d) TSCS: a ρ_i for each ctry i; a σ_i^2 for each ctry & a ρ_{ij} for each pair of countries
$\Rightarrow 2N + N(N-1)$ parameters

e) Panel Data (same as TSCS, but (many) more C's than T's)
- a single ρ for all individuals
- perhaps a σ^2 for each T

$\Rightarrow T + 1$ parameters

These are just plausible specifications which permit estimation

- In general, there will be a trade-off between how realistic the maintained restrictions on Ω are & how well the necessary parameters of Ω may be estimated (& ∴ how valid results based on estimating Ω are likely to be across repeated samples).

Elaborating the Properties of OLS under the G(N)LRM

II. Generalized Linear Regression Model (GLRM) (4)

can generalize further to non-linear also (GRM) but for simplicity we haven't here

① $y = X\beta + \varepsilon$

② $E(\varepsilon) = 0$

③ $V(\varepsilon) = \sigma^2 \Omega$

④ $Cov(X, \varepsilon) = 0$

⑤ X of full rank

(⑥ $\varepsilon \sim N(0, \sigma^2 \Omega)$)

A. Suppose This is Model, but we estimate by OLS

$\Rightarrow$ FINITE SAMPLE PROPERTIES OF OLS

1. Still Unbiased: exactly as before

$$b_{LS} = Ay = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \varepsilon)$$

$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon \quad \text{(distributive property)}$$

$$\boxed{b_{LS} = \beta + (X'X)^{-1}X'\varepsilon} \quad \text{Just as before}$$

$$\Rightarrow E(b_{LS}) = \beta + E[(X'X)^{-1}X'\varepsilon] = \beta \text{ if } E(A\varepsilon) = 0$$

$\Rightarrow$ so long as $Cov(X, \varepsilon) = 0$, b_{LS} unbiased

2. Still Consistent (for "well behoved" Ω)

$$V(b_{LS}) = V(\beta + A\varepsilon) = 0 + \underbrace{A}_{A}V(\varepsilon)\underbrace{A'}_{A'} = (X'X)^{-1}X'V(\varepsilon)X(X'X)^{-1}$$

$$\boxed{V(b_{LS}) = (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1}}$$

this usually rearranged to: $\sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$

Now, unbiased + $\lim_{n\to\infty} V(\cdot) = 0$ is sufficient (not necessary) for consistency

loose demonstration of consistency:

① n.b. $(X'X)^{-1}$ twice, $(X'X)$ once (with $\sigma^2\Omega$ in middle of it) leaves $(X'X)$ "in the denominator"

② $\sigma^2\Omega$ in "numerator"

⇒ ③ if $(X'X) \to \infty$ faster than $\sigma^2\Omega$ does as $n\to\infty$

- In general, this means that off-diag's of $\sigma^2\Omega$ must decline sooner or later as we get away from diag (return to this 2 weeks hence)
- And that X's continue to exhibit variation as increase sample (this usually taken as self-evident for any reasonably defined X)

Intuition for $\mathbf{V(b)\to 0}$ as $n\to\infty$ is $(\mathbf{X'X})(\mathbf{X'X})$ in 'denominator' whereas just one $(\mathbf{X'X})$, with $\mathbf{\Omega}$ in middle, $(\mathbf{X'\Omega X})$ in 'numerator', so, as long as extreme off-diagonal elements of $\mathbf{\Omega}$, which are covariances of 1st observations with last ones, are not growing at equal or faster pace than the additional $\mathbf{x}^2$ term added to 'denominator' with each new obs, then numerator is growing in n more slowly than is denominator, so $\mathbf{V(b)\to 0}$ as $n\to\infty$.

3. Still Normal or Asymptotically Normal:

a) If $\varepsilon \sim N(0, \sigma^2\Omega)$ then b_{LS} just a linear combo of β's (fixed), X's, & ε's (normal) so b_{LS} is also normal: $b_{LS} \sim N(\beta, A\sigma^2\Omega A')$

b) If $\varepsilon \sim$ same now $(0, \sigma^2\Omega)$ then $b_{LS} \overset{A}{\sim} N(\beta, A\sigma^2\Omega A')$ (asymptotically)

provided basically the same conditions as in ② above hold.

4. However OLS is ⓐ Inefficient --> have some info, namely $V(\varepsilon) = \sigma^2\Omega$ not $\sigma^2 I$ we're not using

ⓑ $V(b_{LS}) \neq \sigma^2(X'X)^{-1}$

but rather

$$V(b_{LS}) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

{so std. errors & everything based on them (t's, F's) wrong}

That is, in general, estimating **V(b)** by $\sigma^2(\mathbf{X'X})^{-1}$ is not going to produce good estimate of **V(b)** that is actually $\sigma^2(\mathbf{X'X})^{-1}(\mathbf{X'\Omega X})(\mathbf{X'X})^{-1}$. Could be unbiased & consistent under certain conditions (see below), but not even so would still be inefficient (because **b** estimates inefficient).

III. Robust Estimation under the GLRM

→ Since OLS is unbiased and consistent but inefficient in its estimate of β, but produces the wrong standard errors, one possible fix would be to estimate by OLS, do nothing about the inefficiency but address the standard-error problem.

⇒ need to estimate $V(b) = (X'X)^{-1}X'\sigma^2\Omega X(X'X)^{-1}$

A. As stated before, if we try to estimate Ω directly, without any restrictions on its form, there are too many parameters.

B. However, $[X'\sigma^2\Omega X]$ which is the problem here is not $n \times n$ (like Ω) but $k \times k$ (like $X'X$). As we've seen above what we need for consistency involves convergence of this whole thing. Similarly, to obtain a consistent estimate of standard errors, what we need is not element-by-element estimation of Ω, but rather a consistent estimate of $(X'\sigma^2\Omega X)$. For consistent standard errors, then, we need only consistent estimates of those $K(K+1)/2$ elements of this matrix.

1) Pure Heteroskedasticity: $V(\varepsilon) = \begin{bmatrix} \sigma_{11} & & & 0 \\ & \sigma_{22} & & \\ & & \sigma_{33} & \\ 0 & & & \ddots \; \sigma_{nn} \end{bmatrix}$ $\Rightarrow$ N obs. & N parameters

but $\{X'\sigma^2\Omega X\} =$

$$\begin{bmatrix} - & X_1 & - \\ - & X_2 & - \\ & \vdots & \\ - & X_k & - \end{bmatrix} \begin{bmatrix} \omega_{11} & & & 0 \\ & \omega_{22} & & \\ & & \ddots & \\ 0 & & & \omega_{nn} \end{bmatrix} \begin{bmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_k \\ | & | & & | \end{bmatrix}$$

$$= \sum_{i=1}^{n} \omega_{ii} X_i X_i'$$

X_i: i^{th} obs of X as a column vector

ω_{ii}: i^{th} residual-**variance**

- Rather than try to estimate ω_{ii} directly, just use $(e_i)^2$ & concern yourself with relation b/w $(e_i)^2$ & $X_i X_i'$ i.e. b/w $(e_i)^2$ & $X_i X_i'$ over the sample.

This produces a consistent estimate of the <u>sum</u> (but not of each ω_{ii}). This is all we need b/c what produces inconsistency of standard errors is how $\varepsilon_i \varepsilon_j$ tends to vary with X_i, X_i^2, & X_j, X_j^2, & $X_i X_j$.

$\Rightarrow$ White's heteroskedasticity-consistent standard errors (Stata calls them "Huber/White/Sandwich estimators"

• So, use OLS to estimate $\underset{\sim}{b}$ $\Rightarrow$ unbiased, consistent, inefficient
using White's "robust" s.e.'s: $V(\underset{\sim}{b}) = (X'X)^{-1}\underbrace{(\sum e_i^2 X_i X_i')}_{(\hat{Q} \text{ in Greene})}(X'X)^{-1}$

$\Rightarrow$ "Robust" "fixes" your standard errors in that they are now <u>consistent</u> for any heteroskedasticity which is a function of (the X's, their squares, & their cross-products.) It turns out (as we saw at the beginning) these are the only forms that produce <u>inconsistency</u> of the standard errors. It should be noted, perhaps, that White's s.e.'s & related forms are <u>not</u> unbiased (nor efficient of course) -- however, so far, Monte Carlo simulations has shown them to work surprisingly well

C. Extensions to White's when Ω not simple heteroskedastic:

$$\underset{(k\times k)}{X'\sigma^2\Omega X} = \sum_{i=1}^{n}\sum_{j=1}^{n} \underset{(k\times k)}{\sigma_{ij}\, X_i X_j'}$$ which could be estimated by

$$\hat{Q}_* = \sum_{i=1}^{n}\sum_{j=1}^{n} e_i e_j X_i X_j'$$ we would think

Two Problems: ① if $\sum e_i e_j X_i X_j'$ not "well behaved" in the same sense as before, this sum may not converge to a constant as $n\to\infty$ $\left(\frac{1}{n}\hat{Q}_*\right)$ but may instead grow without bound.

② Whereas with true Ω we are sure it's positive definite so that $X'\Omega X$ is p.d. so that $V(b) > 0$, with $\hat{Q}_*$ substituted for $(X'\Omega X)$ we cannot be sure of p.d. any more

⇒ Newey-West Autocorrelation-Consistent Standard-Errors

$$\hat{\mathbf{Q}}^*_{nw} = \hat{\mathbf{Q}}^*_{w} + \frac{1}{T}\sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-l}(\mathbf{x}_t\mathbf{x}'_{t-l} + \mathbf{x}_{t-1}\mathbf{x}'_t)$$

which takes care of the $e_i \cdot e_i$ as it correlates to X's, X^2's, & cross-products

Newey-West addition which takes care of $e_i \cdot e_j$ as it correlates to cross-products

these two assure that $\hat{Q}^*$ p.d. so that $V(b) > 0$

- setting $w_l = 1 - \frac{l}{L+1}$ weights $e_i e_j$ by (inversely) the temporal distance b/w observations
  → assumes/forces relation of e_i & e_j to fade as time from i to j increases
- L is unfortunately fairly arbitrary maximum "lag-length" time separation at which you wish to consider $E(e_i e_j)$ non-negligible
  → There are rules out there, but no good theory (some simulations)
  → I don't know of any tests or Monte Carlo simulations showing N-W works as well as White's seems to.

- My reading is that jury is out on N-W as opposed to more direct fixes for serial correlation.

Beck & Katz (APSR 1995) Panel-Consistent Standard Errors

- In exactly analogous spirit, B&K consider that if we are concerned with contemporaneous correlation across cross-sectional units in TSCS data, then we can estimate $\Sigma_i e_i e_j X_i X_j'$ appropriately to reflect this concern:

① Re-configure errors into $E \equiv \underset{\text{time-periods}}{T} \times \underset{\text{cross-sections}}{N}$ shape : $\begin{bmatrix} e_{c_1 t_1} & \cdots & e_{c_N t_1} \\ e_{c_1 t_2} & \ddots & \vdots \\ e_{c_1 t_T} & \cdots & e_{c_N t_T} \end{bmatrix}$

② Then $A\left(\frac{E'E}{T} \otimes I_T\right)A' = V(b)$ is consistent in same way as White & N-W

(See B&K for more; See my diss. appendix on PCSE's for brief intro & extension)

So, in the presence of OLS + Robust, we have: $\underset{\sim}{b}_{OLS}$ unbiased consistent inefficient & s.e.$(\underset{\sim}{b}_{OLS})$ biased consistent inefficient

We may wish to get more efficient estimates & or may be unsatisfied with merely consistent std errs.

$\Rightarrow$ Generalized Least Squares (GLS)

⇒ Generalized Least Squares (GLS)

A. Suppose we happen to know Ω (we never will, but this is just to get the "in principle" facts out on the table)

① Ω is positive definite (it's the Var-Cov matrix which has to be p.d.)

$\Rightarrow \Omega = C\Lambda C'$ where C is the matrix of characteristic vectors of Ω & Λ are the corresponding characteristic roots arrayed in a diagonal matrix

② Define $\Lambda^{1/2} \equiv \{\lambda_i^{1/2}\}$ (square root of Λ element by element)

& Def. $T \equiv C\Lambda^{1/2}$

③ Then $\Omega = TT' \{= (C\Lambda^{1/2}\Lambda^{1/2}C') = (C\Lambda C')\}$

this, W_c $C' = C^{-1}$ see 2.7 above

& $P' = C\Lambda^{-1/2}$ so $\Omega^{-1} = P'P$ $\{= C\Lambda^{-1/2}\Lambda^{-1/2}C' = C\Lambda^{-1}C' = C^{-1}\Lambda^{-1}(C')^{-1}\}$

- Loosely speaking, then, <u>P is the square root of the inverse of the variance-covariance matrix, Ω</u>. We'll see in a moment why it is absolutely key.

B. Take the GLRM: ① $Y = X\beta + \varepsilon$ (⑤ X full rank)
② $E(\varepsilon) = 0$ (⑥ ε normal)
③ $V(\varepsilon) = \sigma^2 \Omega$
④ $Cov(X, \varepsilon) = 0$

1. & transform the model by pre-multiplying everything by P:

$$Py = PX\beta + P\varepsilon \quad , \text{let } y^* = Py,\ X^* \equiv PX, \text{ \& } \varepsilon^* = P\varepsilon$$

$$y^* = X^*\beta + \varepsilon^*$$

2. Now what's $V(\varepsilon^*)$?

$$V(\varepsilon^*) = E(\varepsilon^* \varepsilon^{*\prime}) = E(P\varepsilon\varepsilon' P') = PE(\varepsilon\varepsilon')P'$$

$$= P\underbrace{\sigma^2 \Omega}_{\text{def of } E(\varepsilon\varepsilon') = V(\varepsilon)} P' = \sigma^2 \Omega \underbrace{P'P}_{\text{Greene 211}} = \sigma^2 \Omega \Omega^{-1} = \sigma^2 I$$

3. Thus, $V(\varepsilon^*) = \sigma^2 I$ $\Rightarrow$ The C(N)LRM applies to the transformed model

$\Rightarrow$ (Gauss-Markov) OLS is BLUE for the transformed model (Applied to GLRM by transformation, this is called Aitken's Theorem: GLS is BLUE.

$\Rightarrow$ (Gauss-Markov) OLS is BLUE for the transformed model (Applied to GLRM by transformation, this is called Aitken's Theorem: GLS is BLUE.

4. So, if we know Ω, we can find P & transform the variables & then <u>everything</u> is as before except R^2 & perhaps S.E.E. (M.S.E., S.E.R.)

- Problem with these things is that GLS doesn't minimize the sum of squared errors but a weighted sum of squared errors.
  - Thus R^2 is purely descriptive. We usually prefer R^2 based on b_{GLS} but original data (This no longer $0, 1$ bound)
  - Similarly, S.E.E. is estimate of σ^2 in $\sigma^2 \Omega$ & so is no longer exactly an estimate of $V(\varepsilon_i)$ for all i

5. Otherwise, though, once we have the appropriate weighting matrix, we transform & then estimate & carry on tests or whatever exactly as before.

Easy Example: • $V(\varepsilon) = \begin{bmatrix} 2 & & & 0 \\ & 3 & & \\ & & 1 & \\ 0 & & & 5 \\ & & & & 7 \end{bmatrix}$ (pure hetero)

• $Y = X\beta + \varepsilon$ use $P = \underbrace{\sqrt{[V(\varepsilon)^{-1}]}}_{\text{this loose notation}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & & & 0 \\ & \frac{1}{\sqrt{3}} & & \\ & & 1 & \\ 0 & & & \frac{1}{\sqrt{5}} \\ & & & & \frac{1}{\sqrt{7}} \end{bmatrix}$

• $PY = PX\beta + P\varepsilon$

$$\Rightarrow V(P\varepsilon) = \underbrace{P^2}_{\text{more loose notation}} V(\varepsilon) = \begin{bmatrix} 1/2 & & & 0 \\ & 1/3 & & \\ & & 1 & \\ 0 & & & 1/5 \\ & & & & 1/7 \end{bmatrix} \begin{bmatrix} 2 & & & 0 \\ & 3 & & \\ & & 1 & \\ 0 & & & 5 \\ & & & & 7 \end{bmatrix} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix}$$

Basically, we have

$V(\varepsilon_i) = \sigma_i^2$, then weight obs. by $\frac{1}{\sqrt{\sigma_i^2}}$

$$V\left[\frac{1}{\sqrt{\sigma_i^2}}\varepsilon_i\right] = \frac{1}{\sigma_i^2} V(\varepsilon_i) = \frac{1}{\sigma_i^2}\sigma_i^2 = 1$$

• N.b. One more thing: notice that if we know σ_{ij} $\forall i \neq j$ then we can weight appropriately so that $V(\varepsilon^*) = I$ (all variances are 1)

But for OLS = BLUE all we need is $V(\varepsilon^*) = \sigma^2 I$, σ^2 need not be 1. This implies we don't need to know Ω exactly but rather proportionally. I.e. if we know $V(\varepsilon_i) = k \cdot \sigma_i^2$, we don't need to know k: weighting by $\frac{1}{\sigma_i} \Rightarrow V(\varepsilon_i^*) = k$.

C. Maximum Likelihood Estimation of the GLRM (loose intro)

1. MLE is actually a remarkable intuitive approach to statistical estimation (it's merely that it can look technical on paper)

   Basic Idea: Each y_i are independent -- given the (controlling for the) X's -- draws from some probability distribution

$$f(\underbrace{X}_{\text{systematic part}}, \underbrace{\varepsilon}_{\text{stochastic part}})$$

   b) Indep. means that the "joint probability" (actually likelihood) of all the data are the product of all the individual ones

$$\underbrace{Pr(y_i) \propto}_{\text{this means "is proportional to"}} \prod_{i=1}^{n} f(X_i, \varepsilon_i)$$

   c) So maximize this likelihood with respect to the parameter estimates (in GLRM, parameters are β & σ^2)

2. The MLE (which as with CNLRM) turns out to be the GLS estimator

   If Ω known then GLS (= MLE) is simple weighted version of OLS (= MLE)

$$\rightarrow \hat{\beta}_{ML} = \hat{\beta}_{GLS} = (X^{*\prime}X^*)X^{*\prime}y^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

$$\rightarrow S^2_{ML} = \frac{n-k}{n} \cdot S^2_{GLS} = \frac{1}{n}\sum_{i=1}^{n}(y^* - X^*\hat{\beta})'(y^* - X^*\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y - X\hat{\beta})'\Omega^{-1}(y - X\hat{\beta})$$

V. More usually, we do not know Ω $\Rightarrow$ estimate it
(Feasible GLS or FGLS)

$\Rightarrow$ but we've seen Ω has $\frac{1}{2}n(n+1) > n$ parameters

$\Rightarrow$ can't estimate without putting structure on Ω which limits the number of parameters to be estimated

A) <u>Asymptotically</u>, if $\hat{\Omega}$ is a consistent estimate of Ω (converges to it as $n \to \infty$) then $\hat{\beta}_{GLS}$ is asymptotically efficient as is s.e. $(\hat{\beta}_{GLS})$.

B) Some take this as an amazing & powerful result: we don't need that $\hat{\Omega}$ is an efficient estimate of Ω, only that it is consistent, to get asymptotically efficient results for $\hat{\beta}$ & s.e. $(\hat{\beta})$.

C) The worry, of course, is that in limited samples there's no guarantee that $\hat{\beta}_{GLS}$ performs better than $\hat{\beta}_{LS}$. The crux of the problem, of course, is that we introduce some instability by estimating Ω. My reading of the literature is that the general prescription is not to try to estimate too many parameters in $\hat{\Omega}$ relative to the number of observations

too many parameters in $\hat{\Omega}$ relative to the number of observations

Some guidelines/examples

① Always try to use your head about what is likely to be true about $V(\varepsilon)$. E.g. if each obs. i is an average of n_i random variables, then $V(\varepsilon_i) \propto \frac{1}{n_i^2}$

"is proportional to"

Theory should not be put in the closet just because we're now talking about "residual" properties.

② Don't try to capture every conceivable nuance in the structure of Ω in estimating it. Monte Carlo usually reveals that using fewer-parameter forms of $\hat{\Omega}$ (where fewer is relative to # obs) does better across repeated samples in practice

③ Examples: a) Simple AR(1): $\varepsilon_t = \rho \varepsilon_{t-1} + \gamma_t$ ← i.i.d. $V(\gamma_t) = \sigma^2$

last period's shock related to this period's

$$\Rightarrow V(\varepsilon) = \left(\frac{\sigma^2}{1-\rho^2}\right) \cdot \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \vdots \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \vdots & \rho^2 & \rho & 1 & \rho & \rho^2 \\ & & \rho^2 & \rho & 1 & \rho \\ \rho^{T-1} & \cdots & & \rho^2 & \rho & 1 \end{bmatrix}$$

$\Rightarrow$ only one more parameter: ρ

• n.b. $|\rho|$ must be less than one for $V(\varepsilon)$ to be well behaved (more later)

b) $V(\varepsilon_i) = \sigma^2 f(Z_i)$

---> i.e., variance can be modeled just like mean ($X\beta$) was as a function of some variable or set of variables.

e.g. $V(y_i) = f(\text{population}, \text{diversity of economy}, \text{etc.})$ (GDP)

c) $V(\varepsilon_j) = \sigma_j^2$ where j's are groups of observations.

So, Variance is same w/in some group (e.g. Michigan survey responders) but possibly varies across groups (states)

d) • Sometimes you know something for sure about variance:

• e.g. Average of n_i obs

"Quality of Data" indexed by data-recorder or somehow else

$\Rightarrow$ Always use info. that you know "for sure" or reasonably surely

e) "Robust" solutions & FGLS are not necessarily exclusive.

- It's usually permissable to use some info. you know "for sure" & also use robust b/c you also suspect V-Cov matrix to be related to $X, X^2, X'X$ in some undefined way.

Example (my diss, chpts 3 & 4): Chpt 3: $Var(\varepsilon_i)$ different by country $\Rightarrow$ FGLS

+ Possible contemporaneous correlation $\Rightarrow$ Beck-Katz

Chapt 4: Obs. are averages over differing # of years $\Rightarrow$ FGLS

+ Possible to Likely other hetero. &/or contemp. corr

$\Rightarrow$ White's &/or Beck-Katz

VI. Testing When $V(\varepsilon) = \sigma^2 \Omega$: Final Note

A. if FGLS only, then exactly as always — be sure to use transformed data (for F-tests especially)

B. if Robust or FGLS + Robust, then you have done nothing to "fix" the fit. Accordingly "degradation of fit" tests are not valid $\Rightarrow$ use Wald-types (including T-tests)

n.b. Asymptotic tests b/c s.e.'s only consistent

VII. Estimation in Stages:

A. Stage 1: Estimate by OLS

B. Stage 2: Use estimated Residuals to estimate $\hat{\Omega}$ (i.e. to estimate the reduced set of parameters for your structured $\hat{\Omega}$)

C. Stage 3: FGLS $\quad b_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$

$$V(b_{FGLS}) = \sigma^2(X'\hat{\Omega}^{-1}X)^{-1}$$

D. Stage 4: Optional Iteration --> some times we repeat this until b_{FGLS} doesn't change much from iteration to iteration. I.e. use residuals from $(y - Xb_{FGLS})$ to re-estimate $\hat{\Omega}$. Calculate b_{FGLS} using this new $\hat{\Omega}$. Repeat until b_{FGLS} doesn't change noticeably (can be set arbitrarily small) from iteration to next.

Examples of this Procedure you May have Heard of Before: Mostly for Time Series:

- Cochrane-Orcutt (Corc)
- iterated C-O
- Hildreth-Lu (HiLu)
- iterated HiLu
- Prais-Winston (no nickname, aww...)
- iterated Prais-Winston

Alternatively, at estimation stage 3, could weight original data by estimated $\Omega^{-1/2}$ & estimate by OLS. An analogous equivalent pair alternative strategies will arise in instrumental variables by 2SLS—worth showing that math:

$$\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\mathbf{y} = \hat{\mathbf{\Omega}}^{-\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \;\Rightarrow\; \mathbf{b}_{FGLS} = \left[(\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\mathbf{X})'(\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\mathbf{X})\right]^{-1}(\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\mathbf{X})'\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\mathbf{y} = \left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\hat{\mathbf{\Omega}}^{-\frac{1}{2}}\mathbf{y} = \left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}$$

$$V(\mathbf{b}_{FGLS}) = V\left[\left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}\right] = \left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}V(\mathbf{y})\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1} = \left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\sigma^2\hat{\mathbf{\Omega}}\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}$$

$$= \sigma^2\left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}$$

I. Within the GLRM:
① $Y = X\beta + \varepsilon$
② $E(\varepsilon) = 0$
③ $V(\varepsilon) = \sigma^2 \Omega$
④ $Cov(X, \varepsilon) = 0$
⑤ $Rank(X) = k$

We now focus in a particular form of violation of CLRM assumption 3. I.e. we focus in on a particular form of Ω, a particular set of properties of the stochastic part of the world.

(Pure) Heteroskedasticity

A. Def: Refers to situation where $V(\varepsilon_i) = \sigma_i^2$, but there is no covariance across observations

$\Longrightarrow$ We have independent observations, but it's possible each observation comes from a distribution with different variance.

B. This Means $V(\underset{\sim}{\varepsilon})$ is diagonal:

$$V(\varepsilon) = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \\ & & & \sigma_n^2 \end{bmatrix}$$

→ We could also write:

$$V(\varepsilon_i) = \sigma^2 \omega_i$$

1) • That is, we could pull a common factor, σ^2, out of the variance of each, σ_i^2, and write it $\sigma_i^2 = \sigma^2 \omega_i$

2) • But now, it would be somewhat arbitrary what σ^2 to "pull out" like this. The common practice is to normalize it like so: $trace(\Omega) \equiv \sum_{i=1}^{n} \text{diagonal elements} = \sum_{i=1}^{n} \omega_i$

we normalize $\sum_{i=1}^{n} \omega_i = n$

→ Note: Ω is "omega" (capital), ω is "omega" (lower case). Recall that elements of matrix A are denoted a_{ij}.

3) Now $\sigma^2\Omega$ can be thought of as so:

a) → Ω captures the variety of the variances of ε_i's, i.e., the ω_i reflect those relativitely

b) → σ^2 provides the overall scaling of the variance

c) By the way: if you have a heteroskedastic model and you estimate it as such, what M.S.E. (s.e.r. or s.e.e.—i.e. the variance (e)) do you get?

· depends: · some software provide simply $\sum_{i=1}^{n} \frac{e_i^2}{n-k}$

· others give estimates of the σ^2 in $\sigma^2\Omega$ with Ω normalized this way → always check what a software package is actually reporting before you use it—don't report anything you don't know what it is (obviously)

C. <u>Common Perceptions</u> <u>(Possibly Misleading)</u>:

1. Heteroskedasticity is commonly thought of as a cross-sectional phenomenon—and it's true it often arises there—but there is no necessary link one way or the other.

2. It is also commonly thought of as "a problem" that must be "corrected." I would argue, however, we might better think of it as a "feature" of the social world which we want to "estimate," in some cases it may well be a "feature" we can take advantage of.

D. Examples: ① Cross-section of ctry's, states, etc. etc. -- basically cross-sections of observations on anything that involve input from multiple decision-makers

$\Longrightarrow$ often be quite reasonable to expect variance to be inversely proportional to the number of such decision-makers

→ Most certainly, this is so when the dependent variable is directly an average or a group lower-level aggregation outputs.

→ In many settings, especially economic ones but not exclusively, "diversification" of multiple units aggregated into a single observation will also affect variance. E.g. Variance (Ctry Output) or V(firm profit) are typically decreasing in diversity of the "portfolio".

$\Longrightarrow$ Illustrates a point: Existence or Not / Degree of Heterosked. is itself no less a theoretical proposition to which data can speak than are statements about means ($E(Y) = X\beta$ is no less or more a theoretical proposition than is $V(\varepsilon) = f(Z)$)

② For certain survey questions it will often be reasonable to expect more or less variant answers the more educated is the respondent.

⟹ It is important to note that these are propositions about Variances (2nd Moments) not Means (1st moments). That is, the above example is not a statement about what more educated people will say (that's in the $E(Y) = X\beta$ part) but rather that the distribution of educated people's answers will have greater or lesser variance ($V(\varepsilon_i) = f(Ed_i)$) -- spread of educated answers around $X\beta$ will be larger or smaller than spread of less educated around the same regression line: $E(Y) = X\beta$.

③ Variances certainly can change in Time-Series Data as well:

a) Structural Shift: → Prior to Voting Rights Act & After
→ Populations becoming wealthier, more educated, more ..., less...
→ Capital Mobility moves over time
• legal regime changes: Pre-Post Bretton Woods

Remember

b) ARCH (GARCH, ARCH-M & ARCH-M): New-fangled seemingly esoteric stuff -- basic notion is simple though:
• Variance produces more (or less) variance
→ In other words, fluctuations (unpredictable movements) trigger other unpredictable movements (or force their own corrections).
→ These the rave in finance these days

Basic Model: $Y = X\beta + \varepsilon$

$$V(\varepsilon_t) = f(\varepsilon_{t-1}) + \gamma$$

④ WRAP-UP: • No reason our theorizing ought to be limited to 1st moments -- i.e. we can make predictions about variances as well as expected values.
• That said, it's a good idea to make really sure your prediction is about the spread of outputs not about what the outputs themselves are.

II. Estimation by OLS & White's "Robust" V-Cov($\underset{\sim}{b}$):

A. Coefficient estimates, $\underset{\sim}{b}$, $\underset{\sim}{b} = Ay = (X'X)^{-1}X'y$ are ...

1) Unbiased — across repeated samples they would be right on avg.
2) Consistent — as sample size $\to \infty$, $\underset{\sim}{b}$ converges to exactly $\underset{\sim}{\beta}$
3) Inefficient — we have more information we could have used so $V(\underset{\sim}{b})$ is higher than it could have been

B. OLS estimate of $V(\underset{\sim}{b})$, the usual $\sigma^2(X'X)^{-1}$ would be biased, inconsistent & garbage b/c

$$V(\underset{\sim}{b}) = A\,V(y)\,A' = A\,V(X\beta + \varepsilon)\,A'$$

$$= A\,V(\varepsilon)\,A' \quad \text{b/c } X\beta \text{ is fixed — no variance \& } Cov(X,\varepsilon)=0 \text{ both by assumption of GLRM}$$

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

b/c means "because"

C. Defining $Q \equiv (X'X)/n$ & Q^* as $\frac{1}{n}\sum_{i=} \omega_i X_i X_i'$ (the White's estimator of the middle term $X'\Omega X$)

then, the following: $\frac{\sigma^2}{n} Q^{-1} Q^* Q^{-1}$ is the "asymptotic variance" of $\underset{\sim}{b}$

notice that all the n's cancel out. There usually is no "proper" variance

D. OLS estimation, then, has

$$\underset{\sim}{b} \overset{A}{\sim} N\left(\underset{\sim}{\beta}, \frac{\sigma^2}{n} Q^{-1} Q^* Q^{-1}\right)$$

to write it out long-hand at least once:

$$\underset{\sim}{b}_{OLS} \overset{A}{\sim} N\left(\underset{\sim}{\beta}, \frac{\sigma^2}{n} \cdot \left(\frac{X'X}{n}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \omega_i X_i X_i'\right)\left(\frac{X'X}{n}\right)^{-1}\right)$$

E. <u>Upshot</u>: Asymptotically (i.e., for "large" samples), OLS estimation with White's standard errors Produces: Unbiased & Consistent $\underset{\sim}{b}$ estimates & consistent $V(\underset{\sim}{b})$ estimates & so is a decent way to proceed, "just" inefficient.

→ However, how "large" is "large"? Cannot be said unambiguously, but my read of MC's is "not very". By around 50 or so This tends not to do noticeably badly — inefficient (so if you could get more info from data could have smaller $V(\underset{\sim}{b})$ but tests & C.I.'s and such tend not to be overly optimistic.

F. How inefficient? Of course, it again depends, but for a simple white type of heteroskedasticity: $V(\varepsilon_i) = \sigma^2 X_i^2$, Greene 12.2.1 shows that degree of inefficiency is a function of the kurtosis of X — i.e. of how fat the tails of the empirical distribution of X is relative to a Normal distribution. (Formally, kurtosis can be measured by $k = E(X_i^4)/E(X_i^2)^2$.) Typical for typical economic data k = something between 2 & 4 which implies confidence intervals based on OLS are 70% or more wider than the proper GLS estimate would produce.

→ I read the general conclusion as follows (& it's simple and abstract and maybe obvious): the more variant are your variances (i.e. further from $\sigma^2 I$ you are, proportionately) the less efficient you are. For $V(\omega_i) \doteq 2$ to 4 times σ^2, you get Greene's "70% or more" too wide confidence intervals.

III. OK, Now Suppose you use OLS without White's. How bad are things? That is, how far is $\sigma^2(X'X)^{-1}$ from $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$? (loosely)

A. The σ^2 in $\sigma^2(X'X)^{-1}$ is estimated by $s^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-k}$ as we know. If $V(\varepsilon_i) = \sigma_i^2$ though, what does this estimate produce?

B. $s^2 = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2$ is like estimating $\overline{\sigma_i^2}$ — the average (or mean) of all the different variances. In fact, it's an unbiased & consistent estimate of that "mean of the variances" under fairly general circumstances.

C. So now the question how far is the estimate of $\sigma^2(X'X)^{-1}$ from $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$ reduces to how far is $X'\Omega X(X'X)^{-1}$ from I. Why? Because if that last thing is I, then $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$ differs from $\sigma^2(X'X)^{-1}$ only in their estimates of σ^2 & which we've just "shown" the OLS estimate is OK & not great.

D. So, how far is $X'\Omega X(X'X)^{-1}$ from I? Note: this is like saying how far is $X'\Omega X$ from $X'X$ because if $X'\Omega X = X'X$, then we have some matrix times it's inverse which is I.

E. So, key question is how much does $X'\Omega X$ differ from $X'X$? Remember what Ω is. It's that matrix of the relative variances:

$$\Omega = \begin{bmatrix} \omega_1 & & 0 \\ & \omega_2 & \\ 0 & & \ddots \; \omega_n \end{bmatrix}$$

Now, if the ω_i are not a function of the X_i — that is if the heteroskedasticity is not related to the X's so that (loosely) $Cov(X_i, \omega_i) = 0$, then on average, $X'\Omega X$ is not different from $X'X$.

across repeated samples

$\Longrightarrow$ If heteroskedasticity is unrelated to the X's (i.e. the X's, the X^2's & the crossproducts X_iX_j) then the OLS standard errors are "right on average" — i.e. (more or less)

they are "unbiased".

Which, as Greene concludes (page 547), means loosely speaking: if the heteroskedasticity is not related to $(X'X)$ then the OLS computations are not "wrong on average" though they are, of course, inefficient.

$\Longrightarrow$ Much of the motivation for using White's

IV. Testing for Heteroskedasticity

A. We already explored some graphical methods for rooting out heteroskedasticity. Recall that they amounted to looking for patterns in the "spread" of residuals. The trick in going from these informal "tests" to more formal ones is two-fold

1) Realizing that e_i^2 is a reasonable estimate of $E(\varepsilon_i^2) = V(\varepsilon_i)$
2) Figuring out the distribution that applies to the test stat, which we, as always, will leave to the "experts"

B. <u>White's General Test</u>

1) Recall: OLS std. err.'s inconsistent <u>iff</u> $V(\varepsilon_i)$ is some $f(X'X)$ (loose notation)
2) So, White's suggestion:

Regress e_i^2 on each X, each X^2, & each $X_i \cdot X_j$ (i.e., the unique elements of this — e.g. if X_3 is a dummy, then $X_3 = X_3^2$ — enter it just once, also $X_1 \cdot X_{N1} = X_{N1}$ so enter that just once)

3) If $V(\varepsilon_i)$ is not of function of the X's, X^2's, & cross-product of the X's, then this auxilary regression should be insignificant.

4) Two Stats: a) $n \cdot R^2 \overset{A}{\sim} \chi^2_{k-1}$

where n is # obs. in aux. reg.
R^2 is R^2 in aux. reg.
$\overset{A}{\sim}$ is "asymptotically distributed"
χ^2 is "chi-squared"
$k-1$ is # of regressors in aux. reg. minus 1 incl. const.

b) Could also calculate the F-stat of the aux. reg. Exact distribution of this is not known though

$\Rightarrow$ if $n \cdot R^2$ bigger than some critical level from χ^2_{k-1}, then you likely have heteroskedasticity which is related to $X'X$.

$\Rightarrow$ Use "White's Robust" or WLS (more later)

C. <u>Goldfeld-Quandt Test:</u>

1) Intuition: under homoskedasticity, however you group your data, $(\Sigma e_i^2)/n$ as an estimate of $V(\varepsilon_i)$ in that group should be the same.

2) Group / Order your data somehow. E.g., you think $V(\varepsilon_i) = f(X_i)\gamma$
$\Rightarrow$ Sort your data by X_i.

3) Separate a high X_i group and a low X_i group (some suggest leaving a small (less than 1/3 the obs) middle group out.

4) Run your regression model separately in each sample. Then...

5) $$\frac{(e_1'e_1)/(n_1-K)}{(e_2'e_2)/(n_2-K)} = F \sim F_{n_1-K,\, n_2-K}$$

- if $\varepsilon \sim$ Normally then this exactly F distribution
- if not, distribution is unkown.

D. Breusch-Pagan / Godfrey Test:

1) Intuition $V(\varepsilon_i) = f(\underset{\sim}{Z}) + \gamma$ — variance is some function of a (set of) variable(s) Z

2) Normalize your squared residuals $g_i = e_i^2 / (\underset{\sim}{e}'\underset{\sim}{e})/n$
(i.e., all e_i^2 measured relative to the average e^2)

3) Then $LM = \frac{1}{2}\left[\underset{\sim}{g}' Z(Z'Z)^{-1} Z' \underset{\sim}{g} - n\right]$ (basically this is a Wald test on the auxilary regression of g_i's on Z's)

$LM \overset{A}{\sim} \chi^2_p$ ← # of variables in Z

4) Again, if LM large, then likely have heteroskedasticity which is a function of the Z's.

(n.b. if Z is same as variables in White's test, these two tests are identical)

5) Again, there's some concern about what the distribution of LM actually is if ε not normal. Koenker & Bassett have suggested: let $\underset{\sim}{u} \equiv (e_1^2, e_2^2, \ldots, e_n^2)'$ and $\underset{\sim}{i} = (1, 1, \ldots, 1)'$ $(n \times 1)$

Then, $\bar{u} = \underset{\sim}{e}'\underset{\sim}{e}/n$. let $V = \frac{1}{n}\sum_{i=1}^{n}\left[e_i^2 - \frac{e'e}{n}\right]^2$

Then, $LM^* = \frac{1}{V}(\underset{\sim}{u} - \bar{u}\underset{\sim}{i})' Z(Z'Z)^{-1} Z'(\underset{\sim}{u} - \bar{u}\underset{\sim}{i})$

$\Rightarrow$ asymptotically same as LM above, but absent normality it seems to work better in small samples. → (Monte Carlo's)

E. <u>Likelihood Ratio Tests</u>: 1) Maximum likelihood will estimate $\hat{\beta}$ & $\hat{\Omega}$ simultaneously/directly. Accordingly we can estimate $(\hat{\beta}, \hat{\Omega})$ in one go, then estimate $(\hat{\beta}^*, \hat{\sigma}^2 I)$ in another go (constraining Ω to be $\sigma^2 I$).

2) Then, as usual,

$$-2(\ln L_0 - \ln L_1) \sim \chi^2_{\Delta p}$$

L_0 ← restrained, L_1 ← unrestrained

where Δp is the number of parameters in $\hat{\Omega}$ less number of parameters in $\hat{\sigma}^2 I$ (which is 1)

3) An Example: Groupwise Heteroskedasticity

Restrained: $V(\varepsilon_i) = \sigma^2 \quad \forall i$

Unrestrained: $V(\varepsilon_i) = \sigma_g^2 \quad \forall i$ in group g

let $s^2 \equiv \left(\frac{e'e}{n}\right)$ from OLS regression on whole sample

let $s_g^2 \equiv (e_g' e_g / n_g)$ from OLS regression [redacted] group g one at a time.

Then

$$\left(n \cdot \ln s^2 - \sum_{g=1}^{G} n_g \cdot \ln s_g^2\right) = LR \sim \chi^2_{G-1}$$

4) As always, if this too big, then likely have groupwise hetero.

F. Glesjer's Test: The obvious regression tests are basically valid.

- for any $V(\varepsilon_i) = f(\underset{\sim}{Z}) + \gamma$

- Just write the appropriate regression model & conduct the appropriate (Wald) test.

E.g. $V(\varepsilon_i) = f(Z_i) + \gamma$

$$\Rightarrow e_i^2 = \beta_0 + \beta_1 Z_i + \gamma$$

t-Test on b_1 (using White's V-Cov matrix is recommended)

E.g. White's test is an example. So is Breusch-Pagan/Godfrey.

E.g. $V(\varepsilon_i) = \sigma^2 \exp\{\underset{\sim}{\alpha}' \underset{\sim}{Z}\} + \gamma$ (Variances must be positive, this model assures that all predicted values are positive).

$$\Rightarrow \ln V(\varepsilon_i) = \ln \sigma^2 + \underset{\sim}{\alpha}' \underset{\sim}{Z} + \gamma$$

$$\Rightarrow \ln e_i^2 = \beta_0 + \underbrace{\beta_1 Z_1 + \beta_2 Z_2 + \ldots \beta_k Z_k}_{\text{Wald Test all of these}} + \gamma$$

not this, b/c $\ln e_i^2 = \beta_0$ would be homoskedasticity

- I find this "Glesjer" approach by far the most intuitive. It is also the most powerful way to test the specific hypothesis embodied in the variance model. Some of the other's (e.g. White's version of the "Glesjer") have an advantage of generality.

V. Efficient Estimation: Weighted Least Squares (WLS)

A. The Pure Heteroskedasticity Model:

1) $Y = X\beta + \varepsilon$
2) $E(\varepsilon) = 0$
3) $V(\varepsilon) = \sigma^2 \Omega = \sigma^2 \begin{bmatrix} \omega_1 & & 0 \\ & \omega_2 & \\ 0 & & \ddots \; \omega_n \end{bmatrix}$
4) $Cov(X, \varepsilon) = 0$
5) X of full rank

(6) ε (& thus Y) normal)

B. Ω is known (Base-line case, which never obtains of course)

$$\Omega^{-1} = \begin{bmatrix} 1/\omega_1 & & 0 \\ & 1/\omega_2 & \\ 0 & & \ddots \; 1/\omega_n \end{bmatrix} \quad ; \left(\text{loose notation}\right) \underbrace{\sqrt{\Omega^{-1}}}_{\text{call this } P} = \begin{bmatrix} 1/\sqrt{\omega_1} & & 0 \\ & 1/\sqrt{\omega_2} & \\ 0 & & \ddots \; \frac{1}{\sqrt{\omega_n}} \end{bmatrix}$$

C. Then $(PY) = (PX)\beta + (P\varepsilon)$

just by premultiplying everything by same P matrix

$$\Rightarrow V(P\varepsilon) = \underset{\text{loose notation}}{P^2} V(\varepsilon) = \begin{bmatrix} 1/\omega_1 & & 0 \\ & \ddots & \\ 0 & & 1/\omega_n \end{bmatrix} \cdot \sigma^2 \begin{bmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_n \end{bmatrix} = \sigma^2$$

D. So, if we know what $V(\varepsilon_i)$ is proportional to; i.e. if we know the relative variances of the observations, i.e. the ω_i's, then multiply Y_i by $\frac{1}{\sqrt{\omega_i}}$ & all X_i by $\frac{1}{\sqrt{\omega_i}}$, call those Y^* & X^*

- OLS of Y^* on X^* is BLUE because Y^* & X^* & ε^* follow the C(N)LRM assumptions

E. Intuition: Why do we want to "weight" some observations more or less than others?

$$Y_i = f(X_i, \varepsilon_i)$$

systematic — stochastic

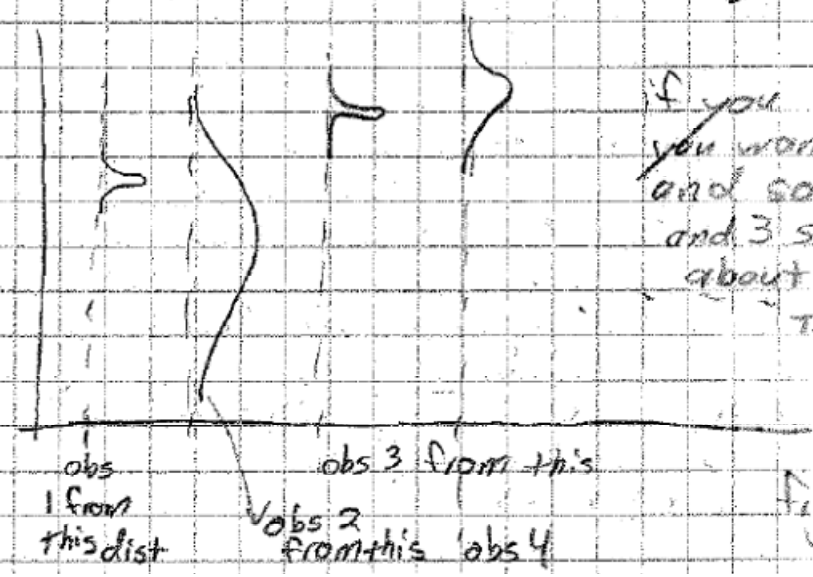
- We are trying to estimate the systematic relationship b/w X & Y

- With heteroskedasticity, the proportion of information in obs. i that is systematic & proportion that is stochastic (random) varies.

$\Longrightarrow$ We want to weight more systematic observations more, or, equivalently, we want to weight more stochastic (less systematic) obs. less.

- Thought experiment: You ask 100 people the temperature. You happen to know how accurate each person is in terms of their expected squared error. What do you do? You take a weighted average of their statements, weighing more accurate people more. Same thing here.

- Visual Picture



If you knew this, wouldn't you want to discount obs 2? and somewhat obs 4? Obs 1 and 3 seem more reliable about where mean for that obs is. I.e., where $E(y_i)$ is which is what we're trying to figure out: $E(y_i) = X_i\beta$

V. F. Efficient Estimation when Ω known:

① WLS by transformed variables:

a) $Y^* \equiv PY \equiv \Omega^{-1/2} Y$

b) $X^* \equiv PX \equiv \Omega^{-1/2} X$

c) OLS of Y^* on X^* $\Rightarrow$

d) $\begin{cases} (X^{*\prime} X^*)^{-1} X^{*\prime} Y = \hat{\beta}_{WLS} \text{ is BLUE} \\ V(\hat{\beta}_{WLS}) = \sigma_*^2 (X^{*\prime} X^*)^{-1} \end{cases}$

② WLS in one step:

$$\hat{\beta}_{WLS} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} Y$$

$$V(\hat{\beta}_{WLS}) = \sigma_*^2 (X'\Omega^{-1}X)^{-1}$$

VI. Estimation of Ω: Feasible GLS (or WLS), FGLS (or FWLS)

A. Two-Step FWLS:

step 1:
① Estimate $Y = X\beta + \varepsilon$ by OLS
② Use e_i estimates from this to model $V(\varepsilon_i)$ as we did in testing

step 2:
③ Use estimated $V(\varepsilon_i)$ to construct $\hat{\Omega}^{-1/2}$
④ Do WLS using $\hat{\Omega}^{-1/2}$

B. Example: ① $Y_i = \beta_0 + \beta_1' X_i + \varepsilon_i$

Y_i: GDP growth in ctry i

X_i: some vector of stuff that matters

estimate this by OLS & save residuals

② Hypoth. $\ln V(\varepsilon_i) = \alpha_0 + \alpha_1 Z + \gamma$

Z = sectoral diversity of country's output ($\alpha_1 < 0$ I'd guess)

model: $(\ln e_i^2) = \hat{\alpha}_0 + \hat{\alpha}_1 Z + \hat{\gamma}$

this is done so that all $\hat{e}_i^2$ from the model will be positive

model: $(\ln e_i^2) = \hat{\alpha}_0 + \hat{\alpha}_1 Z + \hat{\delta}$

$\rightarrow$ this is done so that all $\hat{e}_i^2$ from the model will be positive

estimate by OLS; fitted values are $\widehat{(\ln e_i^2)}$

$\Rightarrow$ so, exp{fitted values} are $\hat{e}_i^2$

$\Rightarrow$ so, weight by $\frac{1}{\hat{e}_i}$

i.e. $\hat{P}$, the weighting matrix, is

$$\hat{P} = \begin{bmatrix} 1/\hat{e}_1 & & 0 \\ & 1/\hat{e}_2 & \\ & & \ddots \\ 0 & & \frac{1}{\hat{e}_n} \end{bmatrix}$$

③ Estimate WLS, using $\hat{Y}^* = \hat{P}Y$ & $\hat{X}^* = \hat{P}X$

c. Properties of $\hat{\beta}_{WLS}$:
① Unbiased
② Consistent
③ Asymptotically Efficient

Properties of $s.e.(\hat{\beta}_{WLS})$
① unbiased if variance model unbiased
② consistent, same "if"
③ Asymptotically efficient

VII. Upshot:

A. White's Robust S.E's or some more general version thereof are almost always a good idea.

① They're consistent precisely in the case where OLS s.e.'s are not: when $V(\varepsilon) \neq f(X'X)$

② They're not much less efficient than OLS s.e.'s except in very small samples perhaps

③ Properties in very small samples are hard to tie down — but then so are properties of nearly anything in such samples

B. White's General Test & Glesjer's Specific Tests are the most direct ways to go about considering heteroskedastic possibilities. I like the following process:

① Think about $V(\varepsilon_i) = f(Z)$ model & test it

② Consider White's General Test in comparison to your specific one as revealing whether there's more hetero. than your Glesjer model speaks to: of the "dangerous type"

③ Consider a) WLS + White's or b) WLS or c) White's or d) OLS depending on outcome of ①&②

But: → be wary of over-fitting. My "rule" ~~of thumb~~ you want many times as many obs. as parameters in $\hat{\Omega}$ you're going to estimate. I like 10 as a bare minimum.

E.g. 10 obs ⇒ OLS

20 obs ⇒ maybe start considering two parameters in $\hat{\Omega}$ (one + one)

etc.

This is the sludgest of guesstimates

C. In principle, no less theorizing should go into the probability distributions from which you consider your data to have been drawn than goes into your theories about the mean ($X\beta$) of those distributions. (Unfortunately, usually a lot less goes in.)

D. However, also in principle, it is generally a good idea to check first whether some Z variable belongs in the $Y = X\beta + \varepsilon$ equation (the theory about $E(Y)$) then proceed to consider $V(Y) \equiv V(\varepsilon) = f(Z)$.

→ Reason is that if you have Z out of mean model but it belonged in (omitted variable) then ε_i related to Z_i which assures that ε^2 is too, but this has nothing to do with heteroskedasticity. Really this rather indicative of mis-specification.