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. \( y = X\beta + \varepsilon \), and \( E(\varepsilon) = 0 \);

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

3. \( E(X'\varepsilon) = 0 \), and \( X \) of full-column rank;

4. optional convenience: \( \varepsilon \sim N(0, \sigma^2 \Omega) \);

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

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

For what did we use \( V(\varepsilon) = \sigma^2 I \) from the CLRM?

1. Estimating \( \sigma^2 \); To show: OLS efficient, i.e., what it’s standard errors are.

2. Distribution of \( \hat{\beta}_{LS} \); & so for hypoth. tests, etc.
   i.e., to show \( V(\hat{\beta}_{LS}) = \sigma^2 (X'X)^{-1} \)

We did not use it to show:

3. \( \hat{\beta}_{LS} \) 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 $\mathbf{\Omega}$ look like in multilevel data? (for example)

Note: $\mathbf{\epsilon}_{ij} = \eta_j + \omega_{ij} \implies \text{Cov}(\mathbf{\epsilon}_{ij}, \mathbf{\epsilon}_{kj}) = \text{Cov}(\eta_j + \omega_{ij}, \eta_j + \omega_{kj}) = \text{Var}(\eta_j) = \sigma_{\eta}^2$
\[ \Omega = \]

\[
\begin{bmatrix}
\omega_{1,1} & \omega_{1,2} & \omega_{1,3} & \ldots & \omega_{1,11} \\
\omega_{2,1} & \omega_{2,2} & \omega_{2,3} & \ldots & \omega_{2,11} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\omega_{T,1} & \omega_{T,2} & \omega_{T,3} & \ldots & \omega_{T,11}
\end{bmatrix}
\]

\[ = \rho \varepsilon_{it-1} + \omega_{it} \Rightarrow Cov(\varepsilon_{it}, \varepsilon_{it-1}) = Cov(\varepsilon_{it}, \rho \varepsilon_{it-1}) = \rho Var(\varepsilon) \]
D. Properties of OLS under the G(N)LRM:

1. \( b_{LS} \) UNBIASED:
\[
\hat{b}_{LS} = X'\hat{\beta} = X\beta + \Sigma \varepsilon = \beta + \Sigma \varepsilon
\]
\[
\Rightarrow E(\hat{b}_{LS}) = \beta
\]

2. \( V(\hat{b}_{LS}) = V(X'\hat{\beta}) \)
\[
= A V((\Sigma)^{1/2} \varepsilon) A' = A V(\varepsilon) A' = A \sigma^{-2} \Omega A'
\]
\[
= \sigma^{-2} (X'X)^{-1} [X' \Omega X] (X'X)^{-1}
\]

Notice that \( V(\beta) \) continues →0 as \( n \to \infty \) (as long as off-diag elements of \( \Omega \) (covar’s) dampen at sufficient rate as get farther from diag (i.e., as \( n \to \infty \)): \( \beta \) consistent also.

\( \Rightarrow \) OLS standard-errors wrong (biased, inconsistent, inefficient)

OLS coefficients inefficient (\% useful info unexploited)

(Perhaps better to say OLS \( V(\beta) \) “not unbiased”; could be unbiased, not generally so though.)
E. Strategies Address & Redress deficiencies OLS under G(N)LRM:

1. "Sandwich Estimators":

- Since OLS b "only" inefficient, could fix s.e.'s:

- Difference OLS $\hat{V}(\frac{1}{b})$ from CLRM to GLRM: $\Omega$

$$[X'\Omega^{-2}I\Omega X] \text{ vs. } [X'\Omega^{-2}\Omega X]$$

$$Q^{*} \rightarrow Q$$

- $Q^{*}$ is $k \times k$, whereas $\Omega \text{ N x N}$

$$Q^{*} = \sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon_i \varepsilon_j X_i X_j'$$

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

a) Need estimate bracketed term: $V(b) = (X'X)^{-1}[X'\Omega X](X'X)^{-1}$. An OMEGA sandwich! Get it?
(1) A correction only need concern itself, therefore, with how:

\[ \varepsilon_i^2 \text{ and } \varepsilon_{ij}, \text{ i.e. the variations & covariations, } \]

associate with corresponding \( x_i, x_i^2 \), and \( \{x_i\} \cdot \{x_j\} \).

(2) ...that is, with this sort of term: \[ \sum_i \sum_j e_i e_j \{ x_i x_j' + x_j x_i' \} \].

b) Examples:

Sandwich Estimators: First one offered: “Pure heteroskedastic.”

\[ V(\varepsilon) = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_n^2 \end{bmatrix} \]

\[ Q^* = X' \Omega X = \begin{bmatrix} -X_1 \\ -X_2 \\ \vdots \\ -X_k \end{bmatrix} \begin{bmatrix} \omega_1 & \omega_{12} & 0 \\ \omega_{12} & \omega_{22} & 0 \\ 0 & 0 & \omega_{nn} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \]

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

\[ \Rightarrow \hat{Q}^* = \sum_{i=1}^{n} e_i^2 X_i X_i' \]
c) Properties of OLS+Sandwich (i.e., rightly constituted sandwich) under GLRM:

(1) \( \hat{b} \) unbiased, consistent, but inefficient.

(2) \( V(\hat{b}) \) "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

(Feasibly) Efficient Estimation: Need use the wasted info:

Stipulate: \( \exists \) matrix \( \Omega^{-\frac{1}{2}} \), “square root of the inverse of \( \Omega \)

Then: \( \Omega^{-\frac{1}{2}} y = \Omega^{-\frac{1}{2}} X \beta + \Omega^{-\frac{1}{2}} \xi \) under GLRM

follows the CLRM, so OLS on this BLUE

\[
\mathbb{E}(\xi^* \xi^*) = \Omega \mathbb{E}(\xi \xi') \Omega^{-\frac{1}{2}}
\]

\[
\mathbb{E}(\Omega^{-\frac{1}{2}} \xi \xi' \Omega^{-\frac{1}{2}}) = \Omega^{-\frac{1}{2}} \mathbb{E}(\xi \xi') \Omega^{-\frac{1}{2}}
\]

\[
= \Omega^{-\frac{1}{2}} \sigma^2 \Omega \Omega^{-\frac{1}{2}} = \sigma^2 \Omega \Omega^{-1} \Omega^{-\frac{1}{2}}
\]

\[
= \sigma^2 \Omega \Omega^{-1} = \sigma^2 I
\]
a) Pure Heteroscedasticity example:

"Pure Het" Example:

\[
\begin{bmatrix}
\omega_1^2 & 0 \\
0 & \omega_2^2 \\
0 & 0 & \omega_3^2
\end{bmatrix}
\]

\[\Omega = \begin{bmatrix}
2 & 2 & 2 \\
2 & 6 & 6 \\
2 & 6 & 3 & 3
\end{bmatrix}\]

\[\Rightarrow \Omega^{-\frac{1}{2}} = \begin{bmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\
\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}
\end{bmatrix}\]

\[\Rightarrow \text{Using the wasted info:}\]

\[\Rightarrow \text{More reliable units weighed more heavily}\]
b) AR(1) (temporal) correlation example:

\[ y_t = x_t \beta + \epsilon_t \quad , \quad \epsilon_t = \rho \epsilon_{t-1} + \nu_t \]

\[
\sqrt{\mathbb{V}(\epsilon)} = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots \\ \rho & 1 & \rho & \ldots \\ \rho & \rho & 1 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad \Rightarrow \quad \Omega^{-\frac{1}{2}} = \begin{bmatrix} 1 \\ -\rho \\ -\rho \rho \\ \vdots \end{bmatrix}
\]

\[
\Omega^{-\frac{1}{2}} y \Rightarrow \begin{bmatrix} \sqrt{\mathbb{V}(Y)} \\ y \end{bmatrix} = \text{if analogous for } x
\]

i.e., partial differencing

Recall: issue with \( O_{ij} \neq 0 \) 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. $V(\varepsilon) \neq \sigma^2 I$)

**I. Background**

- $Y = X\beta + \varepsilon$
- $E(\varepsilon) = 0$
- $V(\varepsilon) = \sigma^2 I$

- $\text{Cov}(X_{i}\varepsilon) = 0$
- $X$ not full-column rank
- $E_{i} \sim N(0, \sigma^2)$

- Have been dealing with $\text{C}(\varepsilon)$ in particular: incorrect set of $X$s, non-linear in $X$

Now we move on to consider violations of $\text{C}(\varepsilon)$

- $V(\varepsilon) = \sigma^2 I$
- Since each obs can have different variance ($\sigma^2$), think of the $\sigma^2$ here as pulling out some base-line obs' variance to which others are added. It's in that sense arbitrary, usually taken to be the mean variance

A. Estimation of $\sigma^2$

B. Distribution of $\hat{\beta}$

I.e. did not use OLS:
- Unbiased
- Consistent (nor that $\hat{\beta}$ is asymptotically normal)

2) Did use $t$-test to prove: OLS efficient in that

OBS: $\hat{\beta}$sd Errors are unbiased, consistent, efficient
B. What does \( \mathbf{\text{v}}(\mathbf{e}) = \sigma^2 \Omega \) mean?

\[
\mathbf{\text{v}}(\mathbf{e}) = E \begin{bmatrix}
\varepsilon_1 \varepsilon_1 & \varepsilon_1 \varepsilon_2 & \cdots & \varepsilon_1 \varepsilon_n \\
\varepsilon_2 \varepsilon_1 & \varepsilon_2 \varepsilon_2 & \cdots & \varepsilon_2 \varepsilon_n \\
\vdots & \vdots & \ddots & \vdots \\
\varepsilon_n \varepsilon_1 & \varepsilon_n \varepsilon_2 & \cdots & \varepsilon_n \varepsilon_n
\end{bmatrix}
\]

\[
\Rightarrow \quad \text{Cov} (\text{obs}_i \ v \ \text{obs}_j) = \text{Var} (\text{obs}_i), \ \text{defined}
\]

\[
\Sigma_{ij} = \text{Cov} (\varepsilon_i, \varepsilon_j)
\]

\[
\Sigma_{ii} = \text{Var} (\varepsilon_i)
\]
\[ \sigma_{ij} \neq 0 \Rightarrow \text{each obs not wholly new info; exaggerate info in data, \& so misstate, likely understate, uncertainty. Also inefficient to have ignored this info. Also, worse, possibly indicative omitted systematic features.} \]

\[ \sigma_i^2 \neq \sigma_j^2 \Rightarrow \text{obs have differing shares stochastic \& systematic info, so inefficient to have ignored this info \& misstate uncertainty. Also, worse, possibly indicative omitted systematic features.} \]
C. A More Concrete Illustration

Suppose we have observations on some \( y \), say votes for incumbent, and some \( X \) explanatory variable, say the unemployment rate. We observe \((y, X)\) in 20 countries over 10 years. Then

\[
y_{ct} = X_{ct}' \beta + \varepsilon_{ct}
\]

is the model for country \( c \in [1, 2, \ldots, 20] \) and \( t \in [1950, 1951, \ldots, 1960] \) or \( t = [1, 2, \ldots, 10] \).

\[
V(\varepsilon_{ct}) = \sigma_{ct}^2 \quad \text{and} \quad \text{Cov}(\varepsilon_{ct}, \varepsilon_{du}) = 0_{10}
\]

\[
\Rightarrow V(\varepsilon) = \begin{bmatrix} \sigma_{11}^2 & \cdots & 0_{10} \\ \vdots & \ddots & \vdots \\ 0_{10} & \cdots & \sigma_{10}^2 \end{bmatrix}
\]

- This is a time-series-cross-section (TSCS) sample.
- ab: this is just an example, could be individuals organized by states, provinces by ctry, individuals over time, etc.

1. The Q(N) LRM assumes \( V(\varepsilon) = \sigma^2 I \)

\[
\Rightarrow \text{No obs, in any ctry-year is correlated with any other obs, within or across both ctry & year.}
\]

\[
\Rightarrow \text{The variance of every obs, within or across ctry & year is the same.}
\]
1) The (N) LRM assumes \( V(\varepsilon) = \sigma^2 I \)

\[ \Rightarrow \text{No obs. in any ctry-year is correlated with any other obs. within & across both ctry & year} \]

\[ \Rightarrow \text{The variance of every obs. (within across ctry & year) is the same.} \]

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

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

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

c) Obs. diff. variance by ctry (ctry-wise heteroskedasticity)

3) Now suppose we are unwilling to make any restrictions on the parameters in \( V(\varepsilon) \). We wish to leave it open that \( \text{Cov}(1950, 1951) \) differs from \( \text{Cov}(1970, 1972) \) that \( \text{Cov}(\text{U.S., JA}) \) differs from \( \text{Cov}(\text{US, UK}) \) that \( V(\text{US}1950) \) differs from \( V(\text{GER}1972) \) & so on. How many unique \( \sigma^2 \) are there to estimate then? Square: \( NT \times NT \) elements. Only Diagonal \& one of triangles are unique \((\sigma^2)\). So:

\[ \frac{1}{2} (NT)^2 + \frac{1}{2} NT = \frac{1}{2} NT(NT+1) \] parameters!
E. Example

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

$$\begin{align*}
\frac{\max}{\min} n - \frac{1}{2} (n^2 + n) \frac{3}{2} &> 0 \\
\Rightarrow \text{E.o.C. } 2n \{ \frac{3}{2} = 1 - n - \frac{1}{2} = 0 \\
\max \text{ is } 0 + 1 = \frac{1}{2}. \text{ Best we could have is } \frac{1}{2} \text{ an observation?!}
\end{align*}$$

At which degrees of freedom are

$$\frac{1}{2} - \frac{1}{2} \left( \frac{1}{3} \right)^2 + \frac{1}{3} = \frac{1}{2} - \frac{1}{8} - \frac{1}{4} = \frac{1}{8}$$

(Do begin to estimate) that many parameters.

b) All this is just more proof than perhaps you needed that it is logically impossible to garner enough info to identify (a.k.a. begin to estimate) that many parameters.

c) Upshot: however you come to evaluate evidence, you have no choice but to make some assumptions about how your observations relate to 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 and 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 \( \Omega(\varepsilon) \) to Estimate:

(a) \( C(N) \) LRM & OLS: \( \Omega(\varepsilon) = \sigma^2 I \)

(b) Time-Series:

\[ \text{Cov}(\varepsilon_t, \varepsilon_{t-i}) = \text{some simple, declining function of temporal distance between obs. One very common one is } \varepsilon_t = \rho \varepsilon_{t-1} + \varepsilon_{t-1} \]

\[ \Rightarrow 2 \text{ parameters: } \sigma^2, \rho \]

(c) Cross-Sections: \( \Omega(0, 0^2) \) different for each "group" \( \Rightarrow \) groups theoretically defined in best-case scenario.

1) TSCS: a \( \rho_i \) for each country \( i \); a \( \sigma_i^2 \) for each country \( i \); and a \( \rho_{ij} \) for each pair of countries

\[ \Rightarrow 2N + N(N-1) \text{ parameters} \]

2) Panel Data (same as TSCS, but (many) more C's than T's)

\[ \Rightarrow T + 1 \text{ parameters} \]

In general, there will be a trade-off between how realistic the maintained restrictions on \( \Omega \) are and how well the necessary parameters of \( \Omega \) may be estimated (i.e., how valid results based on estimating \( \Omega \) are likely to be across repeated samples).
II. Generalized Linear Regression Model (GLRM)

\[ y = X\beta + \varepsilon \]
\[ E(\varepsilon) = 0 \]
\[ V(\varepsilon) = \sigma^2 \Omega \]
\[ \text{Cov}(X, \varepsilon) = 0 \]
\[ X \text{ of full rank} \]
\[ \varepsilon \sim N(0, \sigma^2 \Omega) \]

A. Suppose this is model, but we estimate by OLS

\[ \Rightarrow \text{Finite Sample Properties of OLS} \]

1. Still unbiased: exactly as before

\[ b_{\hat{\beta}} = Ay = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \varepsilon) \]
\[ = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon \]

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

Just as before

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

\[ \Rightarrow \text{so long as} \text{ Cov}(X, \varepsilon) = 0, \text{ } b_{\hat{\beta}} \text{ unbiased} \]
Intuition for $V(b) \to 0$ as $n \to \infty$ is $(X'X)(X'X)$ in ‘denominator’ whereas just one $(X'X)$, with $\Omega$ in middle, $(X'\Omega X)$ in ‘numerator’, so, as long as extreme off-diagonal elements of $\Omega$, which are covariances of 1st observations with last ones, are not growing at equal or faster pace than the additional $x^2$ term added to ‘denominator’ with each new obs, then numerator is growing in $n$ more slowly than is denominator, so $V(b) \to 0$ as $n \to \infty$. 
That is, in general, estimating \( V(b) \) by \( \sigma^2(X'X)^{-1} \) is not going to produce good estimate of \( V(b) \) that is actually \( \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \). 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, 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. We need to estimate $\mathbf{V}(\mathbf{b}) = (\mathbf{X}^\prime \mathbf{X})^{-1} \mathbf{X}^\prime \Omega \mathbf{X} (\mathbf{X}^\prime \mathbf{X})^{-1}$.

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

B. However, $[\mathbf{X}^\prime \Omega^{-2} \mathbf{X}]$ which is the problem here is not $n \times n$ (like $\Omega$) but $k \times k$ (like $\mathbf{X}^\prime \mathbf{X}$). As we've seen above, what we need for consistency involves convergence of this whole thing. Similarly, to obtain a consistent estimate of standard errors, what we need is not element-by-element estimation of $\Omega$, but rather a consistent estimate of $(\mathbf{X}^\prime \Omega^{-2} \mathbf{X})$. For consistent standard errors, then, we need only consistent estimates of those $K(1+K)/2$ elements of this matrix.
1) Pure Heteroskedasticity: $\text{Var}(\mathbf{\epsilon}) = \begin{bmatrix} \sigma_1^2 & 0 \\ \vdots & \sigma_3^2 \\ \sigma_n^2 \end{bmatrix} \Rightarrow N \text{ obs.}$

but $X' \Omega X \begin{bmatrix} X_1 & X_2 & \cdots & X_k \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nn} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$

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

with obs of $X$ as a column vector

$i^{th}$ residual variance

Rather than try to estimate $\omega_{ii}$ directly, just use $\text{Var}(\mathbf{\epsilon})$ & concern yourself with relation b/w $\epsilon_i^2$ & $X_i X_i'$

i.e. b/w $(\epsilon_i)^2$ & $X_i X_i'$ over the sample.

This produces a consistent estimate of the sum (but not of $\omega_{ii}$). This is all we need & what produces inconsistency of standard errors is how $\epsilon_i \epsilon_j$ tends to vary with $X_i X_j$ & $X_i X_j'$

$\Rightarrow$ White's heteroskedasticity-consistent standard errors (aka Huber/White/Sandwich estimators).
So, use OLS to estimate $b \rightarrow$ unbiased, consistent, inefficient<br>using White’s “robust” s.e.’s: $v(b) = (X'X)^{-1}(\Sigma e_i^2 x_i x_i') (X'X)^{-1}$

→ "Robust" "fixes" your standard<br>errors in that they are now consistent for any heteroskedasticity<br>which is a function of (the $x_i$’s, their squares, & their cross-products).

It turns out (as we saw at the beginning) these are the only<br>forms that produce inconsistency of the standard errors. It<br>should be noted perhaps, that White’s s.e.’s & related forms<br>are not unbiased (nor efficient of course) -- however, so far,<br>Monte Carlo simulations has shown them to work surprisingly well.

C. Extensions to White’s when $\Omega$ not simple heteroskedastic:

\[ X' \Omega X = \sum_{j=1}^{n} \sum_{j=1}^{n} \Omega_{ij} x_i x_j \]  which could be estimated by

\[ \hat{\Omega} = \sum_{j=1}^{n} \sum_{j=1}^{n} e_i e_j x_i x_j \]

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 $n \to \infty$ (i.e.
   but may instead grow without bound.

2. whereas with true $\Omega$ we are sure it’s
   positive definite so that $X' \Omega X$ is p.d. so
   that $v(b) > 0$, with $\hat{\Omega}$ substituted for
   $(X' \Omega X)$, we cannot be sure of p.d. anymore
Newey-West: Autocorrelation Consistent Standard Errors

\[ \hat{Q}_{mw} = \hat{Q}_w + \frac{1}{T} \sum_{l=1}^{L} \sum_{t=l+1}^{T} w_t e_{t-1} (x_t x'_{t-1} + x_{t-1} x'_t) \]

which takes care of

the \( e_0, e_1 \) as it correlates
to \( X^t, X^2, e_t \)
cross-products

\[ \sum_{t=1}^{T} \sum_{r=1}^{R} \]

Newey-West addition

which takes care of

\( e_i, e_j \) as it correlates
to cross-products

setting \( w_t = 1 - \frac{1}{L+1} \)
weights \( e_i, e_j \) by (inversely) the
temporal distance \( \tau \) between observations

\( \tau \) increases as \( \tau \) increases

so that \( \tau \) is an arbitrarily maximum "lag-length"

time separation at which you wish to consider

\[ V(\tau) \geq 0 \]

\[ E(e_i e_j) \text{ non-negligible.} \]

\[ \rightarrow \text{There are some thoughts out there, but} \]

\[ \rightarrow \text{no good theory (some simulations).} \]

\[ \rightarrow \text{I don't know of any tests or Monte Carlo} \]

\[ \text{simulations showing N-W works as well as White's seems to.} \]

\[ \text{My reading is that jury is out on N-W as opposed to} \]

\[ \text{more direct fixes for serial correlation.} \]
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 Z.C. E_i X_i, appropriately to reflect this concern.

@ Re-configure errors into $E = T \times N$ shape:

$A (E' E \otimes I_T) A' = V(b)$

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

is consistent in some way as white & N-W

So, in the Presence of OLS + Robust, we have: bias unbiased & s.e.($b_{OLS}$) biased consistent inefficient

We may wish to get more efficient estimates for may be unsatisfied with merely consistent still errs.

$\Rightarrow$ Generalized Least Squares (GLS)
Generalized Least Squares (GLS)

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

1) $\Omega$ is positive definite (it's the Var-Cov matrix which has to be
   \text{found})

   $\Rightarrow \Omega = C \Lambda C'$ where $C$ is the matrix of characteristic vectors of $\Omega$
   and $\Lambda$ are the corresponding characteristic roots arranged
   \text{in a diagonal matrix} $\Lambda$

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

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

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

   Loosely speaking, then, $P$ is the square root of the inverse
   of the variance-covariance matrix $\Omega$. We'll see in a
   moment why it is absolutely key.
B. Take the GLRM: \[ y = X\beta + \varepsilon \]
1. Transform the model by pre-multiplying everything by \( P \):
\[
\begin{align*}
py &= PX\beta + PE \\
\gamma^* &= X^*\beta + \varepsilon^*
\end{align*}
\]
\[ y^* = \gamma^* = PX, \quad \gamma^* = P, \quad \varepsilon^* = PE \]
2. Now what is \( V(\varepsilon^*) \)?
\[
V(\varepsilon^*) = E(\varepsilon^*\varepsilon^{*\prime}) = E(PE\varepsilon\varepsilon^{\prime}P^\prime) = PE(\varepsilon\varepsilon^{\prime})P^\prime \\
= P\sigma^2P^\prime = \sigma^2 I \\
\text{(def. of } E(\varepsilon\varepsilon) = V(\varepsilon))
\]
3. Thus, \( V(\varepsilon^*) = \sigma^2 I \)
\[
\Rightarrow \text{The (N)GLM applies to the transformed model}
\]
\[
\Rightarrow \text{(Gauss-Markov) OLS is BLUE for the transformed model. (Applied to GLRM, by transformation, this is called Arthan's Theorem. GLS is BLUE.}
\]
\[ \Rightarrow \text{ (Gauss-Markov) } \text{ OLS is BLUE for the transformed model (Applied to GLRM by transformation, this is called Artken's Theorem). GLS is BLUE.} \]

4. So, if we know \( \Omega \), we can find \( P \) and transform the variables, and then everything is as before except \( R^2 \) and perhaps S.E.E. (R.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 basis but original data (This is no longer 0,1 band)

- Similarly, S.E.E. is estimate of \( \sigma^2 \) in \( \Omega \) and so is no longer exactly an estimate of \( V(\varepsilon) \) for all \( i \).
5. Otherwise, though, once we have the appropriate weighting matrix, we transform & then estimate & carry on.

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

\[ y = X \beta + \varepsilon \]

- \( P \varepsilon = P \varepsilon \)

\( \Rightarrow V(P \varepsilon) = P^T V(\varepsilon) P \)

\[ P = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \]

\[ V(\varepsilon) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 5 \sigma^2 \end{bmatrix} \]

 Basically, we have

\[ V(e_i) = \frac{1}{\sigma_i^2} \]

\[ \sqrt{\frac{\sigma_i^2}{\sigma_i^2}} V(e_i) = \frac{1}{\sigma_i^2} \sigma_i^2 = 1 \]

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

\[ V(\varepsilon^*) = \begin{bmatrix} \sigma_i^2 \\ \sqrt{\frac{\sigma_i^2}{\sigma_i^2}} \end{bmatrix} \]

\[ V(\varepsilon^*) = \begin{bmatrix} \sigma_i^2 \\ \sqrt{\frac{\sigma_i^2}{\sigma_i^2}} \end{bmatrix} \]

\[ V(\varepsilon^*) = k \cdot \sigma_i^2 \]

\[ \frac{1}{\sigma_i^2} \Rightarrow V(\varepsilon^*) = k \]
C. Maximum Likelihood Estimation of the G-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)
   $x_i$'s -- draws from some probability distribution
   $f(x_i, \varepsilon)$, stochastic part.
   
   b) Indep. means that the joint probability (actually likelihood)
   of all the data are the product of all the individual ones
   
   $p(y_i) \propto \prod_{i=1}^{n} f(x_i, \varepsilon_i)$
   
   c) So maximize this likelihood with respect to the
   parameter $\beta$ (in G-GLRM, parameters are $\beta$ etc.)
   estimates.

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

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

   \Rightarrow s_{ML}^2 = n-K, \quad s_{GLS}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \hat{\beta}) (y_i - x_i \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \hat{\beta}) \Omega (y_i - x_i \hat{\beta})$
More usually, we do not know $\Omega$. (Feasible GLS or FGLS) 
(\Rightarrow \text{ estimate } \theta) 
\Rightarrow \text{ but we've seen } \Omega \text{ has } \frac{1}{2} n (m+1) > n \text{ parameters} 
\Rightarrow \text{ can only estimate } \theta \text{ without putting structure on } \Omega \text{ which limits the number of parameters to be estimated.} 

A) Asymptotically, if $\hat{\Omega}$ is a consistent estimate of $\Omega$ (converges to $\Omega$ as $n \to \infty$) then $\hat{\beta}_{\text{GliS}}$ is asymptotically efficient as is $\text{s.e.} (\hat{\beta}_{\text{GliS}})$.

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

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

Some guidelines/examples

1. Always try to use your head about what is likely to be true about $\sqrt{\text{Var}(\epsilon)}$. E.g. if each obs. $\epsilon$ is an average of $m$ random variables, then $\sqrt{\text{Var}(\epsilon)} \propto \sqrt{m}$.

$\sqrt{\text{Var}(\epsilon)} \propto \sqrt{m}$

"is proportional to"

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

2. Don't try to capture every conceivable nuance in the structure of $\Omega$ in estimating it. Monte Carlo usually reveals that using fewer-parameter forms of $\Omega$ (where fewer is relative to $\#$ obs) does better across repeated samples in practice.
Always use the data. Sometimes you know something for sure, or possibly vary across groups. If variance in some topics.
e) "Robust" solutions & FGLS are not necessarily exclusive.
- It's usually permissible to use some info. you know "for
sure" & also use robust & you
- suspect V(cov matrix) to be related to X, X',X in some undefined way.

Example: (my diss, chpts 3&4): Chpt 3: Var (ε_i): different by country ➞ FGLS

- Possible contemporaneous correlation ➞ Beck-Katz

Chpt 4: Obs. are averages over differing # of years ➞ FGLS
- Possible to likely other hetero. for contempor. corr

⇒ White's & Beck-Katz

VI. Testing When V(ε) = α^2 I: 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 ➞ use Wald-types (including F-tests).

n.b. Asymptotic tests w/ s.e.'s only consistent
Alternatively, at estimation stage 3, could weight original data by estimated \( \Omega^{-\frac{1}{2}} \) & estimate by OLS. An analogous equivalent pair alternative strategies will arise in instrumental variables by 2SLS—worth showing that math:

\[
\hat{\Omega}^{-\frac{1}{2}} Y = \hat{\Omega}^{-\frac{1}{2}} X \beta + \varepsilon \Rightarrow b_{FGLS} = \left[ (\hat{\Omega}^{-\frac{1}{2}} X) (\hat{\Omega}^{-\frac{1}{2}} X) \right]^{-1} (\hat{\Omega}^{-\frac{1}{2}} X) \hat{\Omega}^{-\frac{1}{2}} Y = (X' \hat{\Omega}^{-\frac{1}{2}} \hat{\Omega}^{-\frac{1}{2}} X)^{-1} X' \hat{\Omega}^{-\frac{1}{2}} \hat{\Omega}^{-\frac{1}{2}} Y = (X' \hat{\Omega}^{-\frac{1}{2}} X)^{-1} X' \hat{\Omega}^{-\frac{1}{2}} Y
\]

\[
V(b_{FGLS}) = V \left[ \left( X' \hat{\Omega}^{-\frac{1}{2}} X \right)^{-1} X' \hat{\Omega}^{-\frac{1}{2}} Y \right] = (X' \hat{\Omega}^{-\frac{1}{2}} X)^{-1} X' \hat{\Omega}^{-\frac{1}{2}} Y \hat{\Omega}^{-\frac{1}{2}} X (X' \hat{\Omega}^{-\frac{1}{2}} X)^{-1} = (X' \hat{\Omega}^{-\frac{1}{2}} X)^{-1} X' \hat{\Omega}^{-\frac{1}{2}} \sigma^2 \hat{\Omega}^{-\frac{1}{2}} X (X' \hat{\Omega}^{-\frac{1}{2}} X)^{-1}
\]

\[
= \sigma^2 (X' \hat{\Omega}^{-\frac{1}{2}} X)^{-1} X' \hat{\Omega}^{-\frac{1}{2}} X (X' \hat{\Omega}^{-\frac{1}{2}} X)^{-1} = \sigma^2 (X' \hat{\Omega}^{-\frac{1}{2}} X)^{-1}
\]
I. Within the CLRM:

1. \( y = X\beta + \epsilon \)
2. \( E(\epsilon) = 0 \)
3. \( V(\epsilon) = \sigma^2\Omega \)
4. \( \text{Cov}(X, \epsilon) = 0 \)
5. \( \text{Rank}(X) = p \)

...We now focus on a particular form of violation of CLRM assumption.

3. i.e., we focus in on a particular form of \( \Omega \), a particular set of properties of the stochastic part of the world.

(Pure) Heteroskedasticity

A. Definition: Refers to situation where

\[ V(\epsilon_i) = \sigma_i^2 \]but there is no covariance across observations.

\[ \Rightarrow \text{We have independent observations, but it's possible each observation comes from a distribution with different variance.} \]

B. This means \( V(\epsilon_i) \) is diagonal:

\[ V(\epsilon) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \]

\[ \Rightarrow \text{We could also write:} \]

\[ V(\epsilon_i) = \sigma_i^2 \Omega_i \]

1. That is, we could pull a common factor, \( \sigma_i^2 \), out of the variance of each \( \sigma_i^2 \) and write it \( \sigma_i^2 = \sigma^2 \).

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

we normalize \( \sum_{i=1}^{n} \sigma_i^2 = n \)
3) Now $\sigma^2\Omega$ can be thought of as so:
   a) $\rightarrow \Omega$, captures the variety of the variances of $E_i$s,
      i.e., the $\Omega_i$s reflect those relatively.
   b) $\rightarrow \sigma^2$ provides the overall scaling of the variance.

   c) By the way: if you have a heteroskedastic model and you estimate it as such, what M.S.E. (s.e.r. or s.e.e. -- i.e. the variance ($\sigma^2$)) do you get?

   *depends: some software provide simply $\sum \epsilon^2 / n$
   *others give estimates of the $\sigma^2_i$s
   $\sigma^2_i \rightarrow \sigma^2 \Omega_i$, with $\Omega_i$ 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. Common Perceptions *(Possibly Misleading)*:
1. Heteroskedasticity is commonly thought of as a cross-sectional phenomenon -- and it's true it often arises there -- but there is no necessary link one way or the other.
2. It is also commonly thought of as "a problem" that must be "corrected." I would argue, however, we might better think of it as a "feature" of the social world which we want to "estimate." In some cases it may well be a "feature" we can take advantage of.
D. Examples: ① Cross-section of counties, states, etc. etc. -- basically cross-sections of observations on anything that involve input from multiple decision-makers
    ⇒ often be quite reasonable to expect variance to be inversely proportional to the number of such decision-makers
    ⇒ Most certainly, this is so if the dependent variable is directly on average of a group lower level aggregation outputs.
    ⇒ In many settings, especially economic ones but not exclusively, “diversification” of multiple units aggregated into a single observation will also affect variance. E.g. Variance (CityOutput) or V(GrossProfit) are typically decreasing in diversity of the “portfolio”.

⇒ Illustrates a point: Existence or Not / Degree of Heteroskedasticity is itself no less a theoretical proposition to which data can speak than are statements about means (E(Y) = xp is no less or more a theoretical proposition than is V(x) = E(x^2))
For certain survey questions it will often be reasonable to expect more or less valid 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 \( E(Y|X_B) \) part) but rather that the distribution of educated people's answers will have greater or lesser variance. \( \text{var}(E(Y|X_B)) = \text{var}(E(Y|X_B)) \) — spread of educated answers around \( X_B \) will be larger or smaller than spread of less educated around the same regression line, \( E(Y) = X_B \).

(2) Variances certainly can change in time-series data as well:

a) Structural Shift:

- Prior to Voting Rights Act & Affirmative Action
- Populations becoming wealthier, more educated, more... less...
- Capital mobility moves over time
- Legal regime changes: Pre-Post Vietnam War

b) ARCH: (GARCH, ARCH-M & ARCH-M): New-fangled seeminglly esoteric stuff — basic notion is simple enough:

- Variance produces more (or less) variance
- In other words, fluctuations (unpredictable movements) trigger other unpredictable movements (or force the own corrections).

- These... the rave in finance these days

Basis Model: \( Y_t = X_B + \varepsilon_t \)

\[ \text{var}(\varepsilon_t) = E(\text{var}(\varepsilon_{t+1})) + \varepsilon_t \]

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 the outputs themselves are.
E. Estimation by OLS & White's "Robust" V-Cov(\(\hat{\beta}\)):

A. Coefficient estimates, \(\hat{\beta}\), \(\hat{\gamma} = Ay = (X'X)^{-1}X'y\) are:

1) Unbiased — across repeated samples they would be right on avg.
2) Consistent — as sample size \(\rightarrow \infty\), \(\hat{\beta}\) converges to exactly \(\beta\).
3) Inefficient — we have more information we could have used.

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

\[
V(\hat{\beta}) = A V(y)A' = A V(\tilde{\beta} + \varepsilon)A' = A V(\varepsilon)A', \text{ b/c } \tilde{\beta} \text{ is fixed — no variance if } Cov(x, \varepsilon) = 0.
\]

Both by assumption of GLRM:

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

C. Defining \(Q = (X'X)/n\) b/c \(Q^*\) 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 \(\hat{\beta}\).

D. OLS estimation then, has:

\[
\hat{\beta} \sim N(\beta, \frac{\sigma^2}{n} Q^{-1} Q^* Q^{-1})
\]

to write it out long-hand at least once:

\[
\hat{\beta}_{OLS} \sim N(\beta, \frac{\sigma^2}{n} (X'X)^{-1} \left( \frac{1}{n} \sum_i \omega_i x_i x_i' \right) (X'X)^{-1})
\]
E: Upshot: Asymptotically (i.e., for "large" samples), OLS estimation with White's standard error produces: Unbiased & Consistent.

- Estimates consistent. \( \hat{V}(\hat{b}) \) estimates \( \sigma^2 \) so is a decent way to proceed; "just" inefficient.

However, how "large" is "large"? Cannot be said unambiguously, but my read of NC'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 \( \hat{V}(\hat{b}) \), but tests \( \hat{b} \)'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: \( \hat{V}(\epsilon_i) = \sigma^2 x_i^2 \), 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 \( k = \text{something between } 2 \text{ and } 4 \) 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: (A it's simple and abstract and maybe obvious): the more variant are your variances (i.e. farther from \( \sigma^2 \) if you are, proportionately) the less efficient you are. For \( \hat{V}(\hat{w}_i) \approx 2 \text{ to } 4 \text{ times } \sigma^2 \) 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 $\sigma^{-2}$ in $\sigma^{-2}(X'X)^{-1}$ is estimated by $\hat{\sigma}^{-2} = \sum \frac{e_i^2}{n-k}$ as we know. If $\text{Var}(e_i) = \sigma_i^2$, though, what does this estimate produce?

B. $\hat{\sigma}^2 = \frac{1}{n-k} \sum e_i^2$ is like estimating $\overline{\sigma_i^2}$ — the average (or mean)" of all the different variances. In fact, it's

an unbiased 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 $\hat{\sigma}^{-2}(X'X)^{-1}X'\Omega X(X'X)^{-1}$? Reduces to how far is $X'\Omega X(X'X)^{-1}$ from $I$. Why? Because if that last thing is $I$, then $\hat{\sigma}^{-2}(X'X)^{-1}X'\Omega X(X'X)^{-1}$ differs from $\sigma^{-2}(X'X)^{-1}$ only in their estimates of $\sigma_i^2$, which we've just "shown" the OLS estimate is OK, not great.
D. So, how far is $X' \Omega X(X'X)^{-1}$ from $I$? Note: this is like saying how far is $X'\Omega X$ from $X'X$ because if $X'\Omega X = X'X$, then we have some matrix times its inverse which is $I$.

E. So, key question is: how much does $X'\Omega X$ differ from $X'X$? Remember what $\Omega$ is. It's that matrix of the relative variances: $\Omega = \begin{bmatrix} \omega_1 & 0 \\ 0 & \omega_2 \end{bmatrix}$.

Now, if the $\omega_i$ are not a function of the $x_k$—that is, if the heteroskedasticity is not related to the $x_k$'s so that (loosely) $\text{Cov}(X', \omega_i) = 0$, then on average, $X'\Omega X$ is not different from $X'X$ across repeated samples.

If heteroskedasticity is unrelated to the $x_k$'s (i.e., the $x_k$'s, the $x_k$'s & the cross-products $X'X$), then the OLS standard errors are "right on average"—i.e., they are "unbiased".

Which, as Greene concludes (page 517), means: loosely speaking, if the heteroskedasticity is not related to $(X'X)$, then the OLS computations are not "wrong on average" though they are, of course, inefficient.

$\Rightarrow$ 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 is given from these informal "tests" to more formal ones is two-fold:

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

B. White's General Test

1) Recall: OLS std. err.'s inconsistent iff $V(e)$ is some $f(X'X)$
2) So, White's suggestion:

Regress $e^2$ on each $X$, each $X^2$, & each $X_i \cdot X_j$ (i.e., the unique elements of this

3) If $V(e)$'s not a function of the $X$, $X^2$s, & cross product of the $X$'s, then
   this auxiliary regression should be insignificant ($X = X^2$ -- enter it just once,

4) Two Stats: 1) $n \cdot R^2 \sim \chi^2_{k-1}$, where $n$ is # obs in aux. reg.
   $R^2$ is $R^2$ in aux. reg.
   $\sim \chi^2_{k-1}$ is "asymptotically distributed"
   $k-1$ is # of regressors in aux. reg., minus 1 indep. const.
   $\chi^2_{k-1}$ is # of regressors in aux. reg., minus 1 indep. const.

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

→ if $n \cdot R^2$ bigger than some critical level from $\chi^2_{k-1}$, then
   you likely have heteroskedasticity which is related to $XX$.
C. Goldfeld-Quandt Test:

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

2) Group your data somehow. E.g., you think \(V(e) = f(x_i)\).
   \[\Rightarrow \text{Sort your data by } x_i.\]

3) Separate a high \(x_i\) group and a low \(x_i\) group (some suggest leaving a small (less than 5%) middle group out).

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

5) \[\frac{(e_i'e_i)(n-K)}{(e_i'e_i)(n-K)} \sim F \sim F_{n-K, n-K} \quad \text{if } E \sim \text{Normally then this is exactly } F \text{ distribution.}\]
   \[\text{If not, distribution is unknown.}\]
D. Breusch-Pagan/ Godfrey Test:

1) Intuition: $V(\varepsilon_i) = \text{a function of a (set of) variables(s)} Z$

2) Normalize your squared residuals: $q_i = \frac{\varepsilon_i^2}{(\varepsilon_i')\hat{\beta}_0}$

(i.e., all $\varepsilon_i^2$ measured relative to the average $\bar{\varepsilon}^2$)

3) Then: $LM = \frac{1}{2} [q_i'Z(Z'Z)^{-1}Z(q_i - n)]$

   $LM \sim \chi^2_{k}$

   $k =$ number of variables in $Z$

   (basically this is a Wald test on the auxiliary regression of $q_i$ on $Z$)

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

(Perhaps if $Q$ is some as variables in White's test, these two tests are identical)

5) Again, there's some concern about what the distribution of $LM$ actually is if $\varepsilon$ not normal: Koenker & Bassett have suggested: $u = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ and $\tilde{u} = (1, 1, \ldots, 1)'$

   Then, $\tilde{u} = \varepsilon_i' \hat{\beta}_0$.

   $LM' = \frac{1}{2} \tilde{u}' ( \sum_{i=1}^{n} \varepsilon_i^2 - \tilde{u}' \hat{\beta}_0 )^2$

   Then, $LM = \sqrt{(\tilde{u} - \tilde{u}_k)' Z (Z'Z)^{-1} Z'(\tilde{u} - \tilde{u}_k)}$

   $\Rightarrow$ asymptotically same as $LM$ above, but absent normality it seems to work better in small samples.
E. Likelihood Ratio Tests:

1) Maximum likelihood will estimate $\hat{\Sigma}$ & $\hat{\Omega}$ simultaneously, directly. Accordingly, we can estimate $\hat{\Sigma}$ if we go in another go (constraining $\hat{\Omega}$ to be $\sigma^2 I$).

2) Then, as usual, $-2 (\ln L_0 - \ln L_1) \sim \chi^2_{n_p}$ where $n_p$ is the number of restrained unrestrained $\Omega$ less number of parameters in $\Omega$.

3) An Example: Groupwise Heteroskedasticity

\[
\text{Restrained: } V(E_i) = \sigma^2 \quad \forall i, \quad \text{Unrestrained: } V(E_i) = \sigma^2 \quad \forall i \text{ in group } g
\]

Let $s^2 = \frac{(e_i'e_i)}{n}$ from OLS regression on whole sample

le $S_g^2 = (e_i'e_i_{g} / n_g)$ from OLS regression 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.
E. Glesjer's Test: The obvious regression tests are basically valid.

For any \( V(\varepsilon_i) = f \left( \frac{Z_i}{\sigma^2} \right) + \gamma \)

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

\[ V(\varepsilon_i) = f(\frac{Z_i}{\sigma^2}) + \gamma \]

\[ \Rightarrow \varepsilon_i^2 = \beta_0 + \beta_1 Z_i + \gamma \]

- Test on \( \beta_1 \) (using White's V-Cov matrix is recommended)

- **Example:** White's test is an example. So is Breusch-Pagan/Godfrey

\[ V(\varepsilon_i) = \sigma^2 \exp\left[ \alpha \sum \frac{Z_i}{\sigma^2} \right] + \gamma \]  

\[ \Rightarrow \ln V(\varepsilon_i) = \ln \sigma^2 + \alpha \sum \frac{Z_i}{\sigma^2} + \gamma \]

\[ \Rightarrow \ln \varepsilon_i^2 = \beta_0 + \beta_1 Z_i + \beta_2 Z_2 + \cdots + \beta_k Z_k + \gamma \]

Wald Test all of these

\[ \text{not this, but } \ln \varepsilon_i^2 = \beta_0 \text{ would be homoskedasticity} \]

I find this "Glesjer" approach by far the most intuitive.

It is also the most powerful way to test the specific hypothesis embodied in the variance model. Some of the others (e.g., White's version of the "Glesjer") have an advantage of generality.
V. Efficient Estimation: Weighted Least Squares (WLS)

A. The Pure Heteroskedasticity Model:
1. \( Y = X\beta + \varepsilon \)
2. \( E(\varepsilon) = 0 \)
3. \( V(\varepsilon) = \sigma^2 \Omega = \sigma^2 \begin{bmatrix} \omega_1 & \omega_2 & 0 \\ 0 & \omega_3 & \omega_4 \\ \end{bmatrix} \)
4. \( \text{Cov}(X, \varepsilon) = 0 \)
5. \( X \) of full rank
6. \( \varepsilon \) (thus \( Y \) normal)

B. \( \Omega \) is known (Baseline case, which never obtains of course)

\[
\Omega^{-1} = \begin{bmatrix} \frac{1}{\omega_1} & 0 & 0 \\ 0 & \frac{1}{\omega_2} & 0 \\ 0 & 0 & \frac{1}{\omega_3} \end{bmatrix} \text{ (loose notation)} \\
\sqrt{\Omega}^{-1} = \begin{bmatrix} \frac{1}{\sqrt{\omega_1}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{\omega_2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{\omega_3}} \end{bmatrix}, \text{ call this } P
\]

C. Then \((PY) = (PX)\beta + (PE)\)

\[
\Rightarrow V(PE) = P^2 V(\varepsilon) \text{ (loose notation)} = \begin{bmatrix} \frac{1}{\omega_1} & 0 & 0 \\ 0 & \frac{1}{\omega_2} & 0 \\ 0 & 0 & \frac{1}{\omega_3} \end{bmatrix}, 0^2 \begin{bmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ \omega_3 & 0 & \omega_4 \end{bmatrix} = 0^2
\]

D. So, if we know what \( V(\varepsilon) \) is proportional to, i.e. if we know the relative variances of the observations, i.e. the \( \omega_i \)'s, then multiply \( Y_i \) by \( \sqrt{\omega_i} \) and \( X_i \) by \( \sqrt{\omega_i} \), call those \( Y^* \) and \( X^* \)

* OLS of \( Y^* \) on \( X^* \) is BLUE because \( Y^* \Delta X^* \& \varepsilon^* \) follow the \( CN(0) \) LRM assumptions
E: Intuition: Why do we want to "weight" some observations more of less than others?

\[ y_i = f(X_i, \varepsilon_i) \]  
\[ \text{systematic} \quad \text{stochastic} \]  
We are trying to estimate the systematic relationship \( \text{why?} \)

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

\[ \Rightarrow \] We want to weight more systematic observations more, or, equivalently, we want to weight more stochastic (less systematic) obs. less.

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

*Visual Picture*

If you knew this, would it you want to discount obs. 2 and somewhat obs. 1 and 3, since more reliable about where mean (or that obs. is, i.e., \( E(y_i) \)), which is what we're trying to figure out?

1 from this dist. \( y_i \) from this obs. 4
2 \( \text{obs. 2 from this} \)
3 \( \text{obs. 3 from this} \)
2. Efficient Estimation when \( \Omega \) known:

(1) WLS: a) \( Y^* \equiv PY = \Omega^{-1/2} Y \)

By transformed variables:

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

c) OLS of \( Y^* \) on \( X^* \) \( \Rightarrow (X^*, X^*)^{-1} X^*^T Y = \beta \) is BLUE

(2) WLS in one step:

\[ \beta = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y \]

\[ V(\beta) = \Omega^{-1} (X^T \Omega^{-1} X)^{-1} \]
VI. Estimation of $\Omega$: Feasible GLS (or WLS), FGLS (or FWLS)

A. Two-Step FWLS:
1. Estimate $Y = XB + \varepsilon$ by OLS
2. Use $\hat{\varepsilon}_i$ estimates from this to model $\Sigma = \text{V}(\varepsilon_i)$ as we did in testing $\hat{\Sigma}_{-i}$.
3. Use estimated $\Sigma(\varepsilon_i)$ to construct $\hat{\Sigma}_{-i}$.
4. Do WLS using $\hat{\Sigma}_{-i}$.

B. Example:
1. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
2. GDP growth in city $i$; some vector of stuff that matters; residuals
3. Hypoth. $\text{V}(\varepsilon_i) = \alpha_0 + \alpha_1 Z + \gamma$
4. $Z =$ sectoral diversity of country's output ($\alpha_1 < 0$ I'd guess)
5. Model: $\ln(\varepsilon_i^2) = \hat{\alpha}_0 + \hat{\alpha}_1 Z + \hat{\gamma}$
6. This is done so that all $\varepsilon_i^2$ from the model will be positive.
model: \( (\alpha e_i^2) = \alpha_0 + \alpha_1 Z + \varepsilon \)

\( \Rightarrow \)

this is done so that all \( e_i^2 \) from the model will be positive.

estimate by OLS; fitted values are \( (\alpha e_i^2) \)

\( \Rightarrow \)

so, expected values are \( e_i^2 \)

\( \Rightarrow \)

so, weight by \( e_i^2 \)

i.e. \( \mathbf{A} \), the weighting matrix, is

\[
\mathbf{A} = \begin{bmatrix}
\frac{1}{e_1^2} & 0 & \cdots & 0 \\
0 & \frac{1}{e_2^2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{1}{e_n^2}
\end{bmatrix}
\]

\( \mathbb{3} \) Estimate WLS, using \( \hat{\mathbf{Y}} = \mathbf{A} \hat{\mathbf{Y}} \) and \( \hat{\mathbf{X}} = \mathbf{A} \hat{\mathbf{X}} \)

\( c. \) Properties of \( \hat{\mathbf{\beta}} \) : 0. Unbiased

Properties of s.e. (\( \hat{\mathbf{\beta}} \)) 0. Unbiased

0. Consistent

0. Asymptotically efficient

0. Consistent, same “if”

0. Asymptotically efficient
VII. Upshot:
A. White's Robust S.E.'s are almost always a good idea.

1. They're consistent precisely in the case where OLS s.e.'s are not: when \( V(\hat{\theta}) = f(X'X) \)

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 & Glejser’s Specific Tests are the most direct ways to go about considering heteroskedastic possibilities. I like the following process:

1. Think about \( V(e_i) = f(Z_i) \) model & test it
2. Consider White’s General Test in comparison to your specific one as revealing whether there’s more heteroskedasticity than your model speaks to, of the “dangerous type.”
3. Consider of WLS + White’s or HWLS or GLS or White’s or DOLS, depending on outcome of above.

But: \( \rightarrow \) be wary of over-fitting. My “rule” is that you want many times as many obs. as parameters in \( \mathcal{D} \). You’re going to estimate \( I \) like 10 as a bare minimum.

E.g., \( 10 \text{ obs.} \rightarrow \text{ OLS} \)

20 obs. \( \rightarrow \) maybe start considering two parameters in \( \mathcal{D} \), one + one.
C. In principle, no theorizing should go into the probability distributions from which you consider your data to have been drawn than goes into your theories about the mean (\(\mu\)) 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 \(Y = X \beta + \epsilon\) equation (the theory about \(E(Y)\)) then proceed to consider \(V(Y) = V(\epsilon) = f(Z)\).

Reason is that if you have \(Z\) out of \(\epsilon\) mean model but it belonged in \(\epsilon\) (related \(\epsilon_1\) to \(Z\), which assures that \(E_\epsilon^2\) is too, but this has nothing to do with heteroskedasticity really. It's rather indicative of mis-specification.