Lecture 11
GLS

CLM: Review

- Recall the CLM Assumptions
  (A1) DGP: \( y = X \beta + \varepsilon \) is correctly specified.
  (A2) \( E[\varepsilon | X] = 0 \)
  (A3) \( \text{Var}[\varepsilon | X] = \sigma^2 I_T \)
  (A4) \( X \) has full column rank \(-\text{rank}(X) = k-\) where \( T \geq k \).

- OLS estimation:
  \( b = (XX')^{-1}X'y \)
  \( \text{Var}[b | X] = \sigma^2 (XX')^{-1} \)
  \( \Rightarrow b \) unbiased and efficient (MVUE)

- If \((A5) \varepsilon | X \sim N(0, \sigma^2 I_T) \) \( \Rightarrow b | X \sim N(\beta, \sigma^2 (X'X)^{-1}) \)

Now, \( b \) is also the MLE (consistency, efficiency, invariance, etc). \((A5)\) gives us finite sample results for \( b \) (and for tests: \( t \)-test, \( F \)-test, Wald tests).
CLM: Review - Relaxing the Assumptions

• Relaxing the CLM Assumptions:

(1) (A1) – Lecture 5. Now, we allow for some non-linearities in the DGP.
⇒ as long as we have intrinsic linearity, $b$ keeps its nice properties.

(2) (A4) and (A5) – Lecture 7. Now, $X$ stochastic: \{${x_i, \epsilon_i}$\} $i=1, 2, \ldots, T$ is a sequence of independent observations. We require $X$ to have finite means and variances. Similar requirement for $\epsilon$, but we also require $E[\epsilon]=0$. Two new assumptions:

(A2) plim ($X'\epsilon$/$T$) = 0.

(A4) plim ($X'X$/$T$) = $Q$.
⇒ We only get asymptotic results for $b$ (consistency, asymptotic normality). Tests only have large sample distributions. Bootstrap or simulations may give us better finite sample behavior.

(3) (A2') – Lecture 8. Now, a new estimation is needed: IVE/2SLS. We need to find a set of $l$ variables, $Z$ such that

(1) plim($Z'X$/$T$) $\neq 0$ (relevant condition)
(2) plim($Z'\epsilon$/$T$) = 0 (valid condition –or exogeneity)

$$b_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

$$b_{IV} = (Z'X)^{-1}Z'y$$
⇒ We only get asymptotic results for $b_{2SLS}$ (consistency, asymptotic normality). Tests only have asymptotic distributions. Small sample behavior may be bad. Problem: Finding $Z$.

Generalized Regression Model

• Now, we go back to the CLM Assumptions:
  (A1) DGP: \( y = X \beta + \epsilon \) is correctly specified.
  (A2) \( \mathbb{E}[\epsilon | X] = 0 \)
  (A3) \( \text{Var}[\epsilon | X] = \sigma^2 I_T \)
  (A4) \( X \) has full column rank – \( \text{rank}(X) = k \) –, where \( T \geq k \).

• We will relax (A3). The CLM assumes that observations are uncorrelated and all are drawn from a distribution with the same variance, \( \sigma^2 \). Instead, we will assume:
  (A3') \( \text{Var}[\epsilon | X] = \Sigma = \sigma^2 \Omega \), where \( \Omega \neq I_T \)

• The generalized regression model (GRM) allows the variances to differ across observations and allows correlation across observations.

Generalized Regression Model: Implications

• From (A3) \( \text{Var}[\epsilon | X] = \sigma^2 I_T \) \( \Rightarrow \text{Var}[b | X] = \sigma^2 (X'X)^{-1} \).

• The true variance of \( b \) under (A3') should be:
  \[
  \text{Var}_T[b | X] = \mathbb{E}[(b - \beta)(b - \beta)^' | X] \\
  = (X'X)^{-1} \mathbb{E}[X'\epsilon\epsilon'X | X] (X'X)^{-1} \\
  = (X'X)^{-1} X'\Sigma X (X'X)^{-1}
  \]

• Under (A3'), the usual estimator of \( \text{Var}[b | X] \) – i.e., \( s^2 (X'X)^{-1} \) – is biased. If we want to use OLS, we need to estimate \( \text{Var}_T[b | X] \).

• To avoid the bias of inference based on OLS, we would like to estimate the unknown \( \Sigma \). But, it has \( Tx(T+1)/2 \) parameters. Too many to estimate with only \( T \) observations!

Note: We used (A3) to derive our test statistics. A revision is needed!
• The generalized regression model:
  (A1) DGP: \( y = X \beta + \varepsilon \) is correctly specified.
  (A2) \( E[\varepsilon | X] = 0 \)
  (A3') \( \text{Var}[\varepsilon | X] = \Sigma = \sigma^2 \Omega \).
  (A4) \( X \) has full column rank – \( \text{rank}(X) = k \), where \( T \geq k \).

• Leading Cases:
  – Pure heteroscedasticity: \( E[\varepsilon_i \varepsilon_j | X] = \sigma_{ij} = \sigma_i^2 \) if \( i=j \)
    \( = 0 \) if \( i \neq j \)
    \( \Rightarrow \text{Var}[\varepsilon_i | X] = \sigma_i^2 \)
  – Pure autocorrelation: \( E[\varepsilon_i \varepsilon_j | X] = \sigma_{ij} \) if \( i \neq j \)
    \( = \sigma^2 \) if \( i=j \)

Generalized Regression Model – Pure cases

Relative to pure heteroscedasticity, LS gives each observation a weight of \( 1/T \). But, if the variances are not equal, then some observations (low variance ones) are more informative than others.
Generalized Regression Model – Pure cases

• Relative to pure autocorrelation, LS is based on simple sums, so the information that one observation (today’s) might provide about another (tomorrow’s) is never used.

Note: Heteroscedasticity and autocorrelation are different problems and generally occur with different types of data. But, the implications for OLS are the same.

GR Model: OLS Properties

• Unbiased
Given assumption (A2), the OLS estimator \( \mathbf{b} \) is still unbiased. (Proof does not rely on \( \Sigma \)):
\[
E[\mathbf{b} | \mathbf{X}] = \beta + (\mathbf{X}'\mathbf{X})^{-1} E[\mathbf{X}'\epsilon | \mathbf{X}] = 0.
\]

• Consistency
We relax (A2). Now, we assume use (A2*). To get consistency, we need \( \text{Var}_T[\mathbf{b} | \mathbf{X}] \to \infty \) as \( T \to \infty \):
\[
\text{Var}_T[\mathbf{b} | \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\Sigma \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}
= (1/T) (\mathbf{X}'\mathbf{X}/T)^{-1} (\mathbf{X}'\Sigma \mathbf{X}/T) (\mathbf{X}'\mathbf{X}/T)^{-1}
\]
Assumptions:
- \( \text{plim} (\mathbf{X}'\mathbf{X}/T) = Q_{XX} \) a pd matrix of finite elements
- \( \text{plim} (\mathbf{X}'\Sigma \mathbf{X}/T) = Q_{XX} \) a finite matrix.

Under these assumptions, we get consistency for OLS.
GR Model: OLS Properties

- *Asymptotic normality*
  \[ \sqrt{T} (\hat{b} - \beta) = (X'X/T)^{-1} (X'e/\sqrt{T}) \]

Asymptotic normality for OLS followed from the application of the CLT to \( X'e/\sqrt{T} \):

\[ b \xrightarrow{a} N(\beta, \frac{Q_{xx}^{-1}Q_{xe}Q_{ee}^{-1}}{T}) \]

where \( Q_{xx} = \lim_{T \to \infty} \text{Var}[T^{1/2} \Sigma x e] \).

In the context of the GR Model:

- Easy to do for heteroscedastic data. We can use the Lindeberg-Feller (assuming only independence) version of the CLT.

- Difficult for autocorrelated data, since \( X'e/\sqrt{T} \) is not longer an independent sum. We will need more assumptions to get asymptotic normality.

GR Model: Robust Covariance Matrix

- \( \Sigma = \sigma^2 \Omega \) is unknown. It has \( T x (T+1)/2 \) elements to estimate. Too many! A solution? Be explicit about (A3'): we model \( \Sigma \).

- But, models for autocorrelation and/or heteroscedasticity may be incorrect. The robust approach estimates \( \text{Var}_T[\hat{b} | X] \), without specifying (A3') –i.e., a covariance robust to misspecifications of (A3').

- We need to estimate \( \text{Var}_T[\hat{b} | X] = (X'X)^{-1} X' \Sigma X (X'X)^{-1} \)

- It is important to notice a distinction between estimating
  \( \Sigma \), a \((T x T)\) matrix \( \Rightarrow \) difficult with \( T \) observations.
  & estimating
  \( X' \Sigma X = \Sigma \Sigma \sigma_{ij} x_i x_j' \), a \((k x k)\) matrix \( \Rightarrow \) easier!
GR Model: Robust Covariance Matrix

- We will not be estimating $\Sigma = \sigma^2 \Omega$. That is, we are not estimating $T \times (T+1)/2$ elements. Impossible with $T$ observations!

- We will estimate $X' \Sigma X = \sum_{i} \sum_{j} \sigma_{ij} x_i x_j'$, a $(k \times k)$ matrix. That is, we are estimating $[k \times (k+1)]/2$ elements.

- This distinction is very important in modern applied econometrics:
  - The White estimator
  - The Newey-West estimator

- Both estimators produce a consistent estimator of $\text{Var}_Y[b | X]$. To get consistency, they both rely on the OLS residuals, $e$. Since $b$ consistently estimates $\beta$, the OLS residuals, $e$, are also consistent estimators of $\epsilon$. We use $e$ to consistently estimate $X' \Sigma X$.

GR Model: $X' \Sigma X$

- Q: How does $X' \Sigma X$ look like? Time series intuition.

We look at the simple linear model, with only one regressor (in this case, $x_i \epsilon_i$ is just a scalar). Assume $x_i \epsilon_i$ is covariance stationary (see Lecture 13) with autocovariances $\gamma_j$. Then, we derive $X' \Sigma X$:

$$X' \Sigma X = \text{Var}[X' e / \sqrt{T}] = \text{Var}[(1/\sqrt{T}) (x_1 \epsilon_1 + x_2 \epsilon_2 + \ldots + x_T \epsilon_T)]$$
$$= (1/T) [T \gamma_0 + (T-1)(\gamma_1 + \gamma_{-1}) + (T-2)(\gamma_2 + \gamma_{-2}) + \ldots + 1(\gamma_{T-1} + \gamma_{1-T})]$$
$$= \gamma_0 + \frac{1}{T} \sum_{j=1}^{T-1} (T-j)(\gamma_j + \gamma_{-j})$$
$$= \sum_{j=-T+1}^{T-1} \gamma_j - \frac{1}{T} \sum_{j=1}^{T-1} j(\gamma_j + \gamma_{-j})$$

where $\gamma_j$ is the autocovariance of $x_i \epsilon_i$ at lag $j$ ($\gamma_0 = \sigma^2 = \text{variance of } x_i \epsilon_i$).
GR Model: $X'\Sigma X$

Under some conditions (autocovariances are \textit{\`{a}-summable}, so $\sum |\gamma_j| < \infty$), then

$$X'\Sigma X = \text{var}\left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (x_t e_t) \right) \rightarrow \sum_{j=-\infty}^{\infty} \gamma'_j$$

Note: In the frequency domain, we define the spectrum of $xe$ at frequency $\omega$ as:

$$S(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \Gamma_j e^{-ij}$$

Then, $Q^* = 2\pi S(0)$ \quad ($Q^*$ is called the long-run variance.)

---

Covariance Matrix: The White Estimator

- The White estimator simplifies the estimation since it assumes heteroscedasticity only \textendash;i.e., $\gamma_j = 0$ (for $j \neq 0$). That is, $\Sigma$ is a diagonal matrix, with diagonal elements $\sigma_i^2$. Thus, we need to estimate:

$$Q^* = (1/T) X'\Sigma X = (1/T) \Sigma_i \sigma_i^2 x_i x'_i$$

- The OLS residuals, $e$, are consistent estimators of $\varepsilon$. This suggests using $e_i^2$ to estimate $\sigma_i^2$.

That is, we estimate $(1/T) X'\Sigma X$ with $S_0 = (1/T) \Sigma_i e_i^2 x_i x'_i$.

Note: The estimator is also called the \textit{sandwich estimator} or the \textit{White estimator} (also known as \textit{Eicker-Huber-White estimator}).
Covariance Matrix: The White Estimator

• White (1980) shows that a consistent estimator of \( \text{Var}[b \mid X] \) is obtained if the squared residual in observation \( i \) --i.e., \( e_i^2 \) -- is used as an estimator of \( \sigma_i^2 \). Taking the square root, one obtains a heteroscedasticity-consistent (HC) standard error.

• Sketch of proof. Suppose we observe \( \epsilon_i \). Then, each element of \( Q' \) would be equal to \( E[\epsilon_i^2 x_i x_i' \mid x_i] \).

Then, by LLN

\[
\text{plim} \left( \frac{1}{T} \sum \epsilon_i^2 x_i x_i' \right) = \text{plim} \left( \frac{1}{T} \sum \epsilon_i^2 x_i x_i' \right)
\]

Q: Can we replace \( \epsilon_i^2 \) by \( e_i^2 \)? Yes, since the residuals \( e \) are consistent. Then, the estimated HC variance is:

\[
\text{Est. Var}_T[b \mid X] = \left( \frac{1}{T} \right) \left( X'X / T \right)^{-1} \left[ \sum \epsilon_i^2 x_i x_i' / T \right] \left( X'X / T \right)^{-1}
\]

Covariance Matrix: The White Estimator

• Note that (A3) was not specified. That is, the White estimator is robust to a potential misspecifications of heteroscedasticity in (A3).

• The White estimator allows us to make inferences using the OLS estimator \( b \) in situations where heteroscedasticity is suspected, but we do not know enough to identify its nature.

• Since there are many refinements of the White estimator, the White estimator is usually referred as HC0 (or just “HC”):

\[
HC0 = \left( X'X \right)^{-1} X' \text{Diag}[e_i^2] X \left( X'X \right)^{-1}
\]
The White Estimator: Some Remarks

(1) The White estimator is consistent, but it may not perform well in finite samples—see, MacKinnon and White (1985). A good small sample adjustment, HC3, following the logic of analysis of outliers:

\[ HC3 = (X'X)^{-1} X' \text{Diag}[e_i^2/(1-h_{ii})] X (X'X)^{-1} \]

where \( h_{ii} = x_i(X'X)^{-1} x_i' \).

HC3 is also recommended by Long and Ervin (2000).

(2) The White estimator is biased (show it!). Biased corrections are popular—see above & Wu (1986).

(3) In large samples, SEs, \( \ell \)-tests and \( F \)-tests are asymptotically valid.

(4) The OLS estimator remains inefficient. But inferences are asymptotically correct.

(5) The HC standard errors can be larger or smaller than the OLS ones. It can make a difference to the tests.

---

The White Estimator: Some Remarks

(6) It is used, along the Newey-West estimator, in almost all papers. Included in all the packaged software programs. In R, you can use the library “sandwich,” to calculate White SEs. They are easy to program:

```r
# White SE in R
White_f <- function(y,X,b) {
  T <- length(y); k <- length(b);
  yhat <- X%*%b
  e <- y-yhat
  bhat <- t(X)*as.vector(t(e))
  G <- matrix(0,k,k)
  za <- bhat[1:k]%*%t(x)(bhat[1:k])
  G <- G + za
  F <- t(X)%*%X
  V <- solve(F)%*%G%*%solve(F)
  white_se <- sqrt(diag(V))
  ols_se <- sqrt(diag(solve(F)*drop((t(e)%*%e))/(T-k)))
  l_se = list(white_se,olse_se)
  return(l_se) }
```
Application: 3 Factor Fama-French Model

• We estimate the 3 factor F-F model for IBM returns, using monthly data Jan 1990 – Aug 2016 (T=320):

\[
\begin{align*}
\text{IBM} \text{Ret}_t &= \beta^0 + \beta_1 (\text{MktRet}_t - \text{r}_f) + \beta_2 \text{SMB} + \beta_4 \text{HML} + \epsilon
\end{align*}
\]

\[
> \text{b} \leftarrow \text{solve}(t(x) \times y) \quad \# \text{OLS regression}
\]

\[
> \text{t(b)}
\]

\[
\begin{array}{ccc}
\text{x1} & \text{x2} & \text{x3} \\
-0.2331471 & 0.01018722 & 0.0009802843 -0.004445901
\end{array}
\]

\[
> \text{SE}_b
\]

\[
\begin{array}{ccc}
\text{x1} & \text{x2} & \text{x3} \\
0.011286100 & 0.002671466 & 0.003727686 0.003811820
\end{array}
\]

⟹ OLS SE

\[
> \text{library(sandwich)}
\]

\[
> \text{reg} \leftarrow \text{lm(y~x -1)}
\]

\[
> \text{VCHC} \leftarrow \text{vcovHC(reg, type = "HC0")}
\]

\[
> \text{sqrt(diag(VCHC))}
\]

\[
\begin{array}{ccc}
\text{x} & \text{xx1} & \text{xx2} & \text{xx3} \\
0.011389299 & 0.002724617 & 0.004054315 & 0.004223813
\end{array}
\]

⟹ White SE HC0

\[
> \text{VCHC} \leftarrow \text{vcovHC(reg, type = "HC3")}
\]

\[
> \text{sqrt(diag(VCHC))}
\]

\[
\begin{array}{ccc}
\text{x} & \text{xx1} & \text{xx2} & \text{xx3} \\
0.011583411 & 0.00288216 & 0.004522115 & 0.004416238
\end{array}
\]

⟹ White SE HC3
Baltagi and Griffin’s Gasoline Data (Greene)

World Gasoline Demand Data, 18 OECD Countries, 19 years
Variables in the file are

COUNTRY = name of country
YEAR = year, 1960-1978
LGASPCAR = log of consumption per car
LINCOME = log of per capita income
LRPMG = log of real price of gasoline
LCARPCAP = log of per capita number of cars

See Baltagi (2001, p. 24) for analysis of these data. The article on which the analysis is based is Baltagi, B. and Griffin, J., "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures," European Economic Review, 22, 1983, pp. 117-137. The data were downloaded from the website for Baltagi’s text.

Groupwise Heteroscedasticity: Gasoline (Greene)

Countries are ordered by the standard deviation of their 19 residuals.

Regression of log of per capita gasoline use on log of per capita income, gasoline price and number of cars per capita for 18 OECD countries for 19 years. The standard deviation varies by country. The “solution” is “weighted least squares.”
White Estimator vs. Standard OLS (Greene)

BALTAGI & GRIFFIN DATA SET

| Variable  | Coefficient       | Standard Error  | t-ratio | P[|T|>t] |
|-----------|-------------------|-----------------|---------|---------|
| Constant  | 2.39132562        | .11693429       | 20.450  | .0000   |
| LINCOME   | .88996166         | .03580581       | 24.855  | .0000   |
| LRPMG     | -.89179791        | .03031474       | -29.418 | .0000   |
| LCARPCAP  | -.76337275        | .01860830       | -41.023 | .0000   |

White heteroscedasticity robust covariance matrix

| Variable  | Coefficient       | Standard Error  | t-ratio | P[|T|>t] |
|-----------|-------------------|-----------------|---------|---------|
| Constant  | 2.39132562        | .11794828       | 20.274  | .0000   |
| LINCOME   | .88996166         | .04429158       | 20.093  | .0000   |
| LRPMG     | -.89179791        | .03890922       | -22.920 | .0000   |
| LCARPCAP  | -.76337275        | .02152888       | -35.458 | .0000   |

Autocorrelated Residuals: Gasoline Demand

\[ \log G = \beta_1 + \beta_2 \log P_g + \beta_3 \log Y + \beta_4 \log P_{nc} + \beta_5 \log P_{uc} + \epsilon \]
Now, we also have autocorrelation. We need to estimate $Q^* = (1/T) \Sigma \Sigma \sigma_{ij} x_i x_j'$

Newey and West (1987) follow White (1980) to produce a HAC (Heteroscedasticity and Autocorrelation Consistent) estimator of $Q^*$, also referred as long-run variance (LRV): Use $e_i e_j$ to estimate $\sigma_{ij}$

$\Rightarrow$ natural estimator of $Q^*$: $(1/T) \Sigma \Sigma x_i e_i e_j x_j'$

Or using time series notation, estimator of $Q^*$: $(1/T) \Sigma \Sigma x_i e_i e_j x_j'$

That is, we have a sum of the estimated autocovariances of $x_i e_i$, $\Gamma_j$:

$$\Gamma_T(j) = \sum_{j=-(T-1)}^{T-1} E[x_i e_i e_{i-j} e_{i-j}']$$

Whitney Newey, USA
Kenneth D. West, USA

Newey-West Estimator

- Natural estimator of $Q^*$: $S_T = (1/T) \Sigma \Sigma x_i e_i e_j x_j'$.

Note: If $x_i e_i$ are serially uncorrelated, the autocovariances vanish. We are left with the White estimator.

Under some conditions (autocovariances are “l-summable”), then

$Q^* = \text{var} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (x_i e_i) \right) \xrightarrow{p} \sum_{j=-\infty}^{\infty} \Gamma_T(j)$

- Natural estimator of $Q^*$: $S_T = \sum_{j=-\infty}^{\infty} \hat{\Gamma}_T(j)$

- We can estimate $Q^*$ in two ways:
  1. parametrically, assuming a model to calculate $\gamma_i$.
  2. non-parametrically, using kernel estimation.

Note: (1) needs a specification of ($A3'$); while (2) does not.
The parametric estimation uses an ARMA model—say, an AR(2)—to calculate $\gamma_j$. The non-parametric estimation uses:

- Issues:
  - Order of ARMA in parametric estimation or number of lags ($L$) in non-parametric estimation.
  - Choice of $k(j)$ weights—i.e., kernel choice.
  - The estimator, $S_T$, needs to be psd.

NW propose a robust—no model for (A3') needed—non-parametric estimator.

**Newey-West Estimator**

- Natural estimator of $Q^*$: $S_T = \sum_{j=-\infty}^{\infty} \hat{\Gamma}_T(j)$

**Issue 1**: This sum has $T^2$ terms. It is difficult to get convergence.

**Solution**: We need to make sure the sum converges. Cutting short the sum is one way to do it, but we need to careful, for consistency the sum needs to grow as $T \to \infty$ (we need to sum infinite $\Gamma_l$'s).

- **Trick**: Use a truncation lag, $L_n$ that grows with $T$ but at a slower rate—i.e., $L=L(T)$; say, $L=0.75\times(T)^{1/3}$. Then, as $T \to \infty$ and $L/T\to 0$:

$$Q^*_T = \sum_{j=-L(T)}^{L(T)} \Gamma_T(j) \quad \overset{p}{\rightarrow} \quad Q^*$$

- Replacing $\Gamma_l(j)$ by its estimate, we get $S_T$, which would be consistent for $Q^*$ provided that $L(T)$ does not grow too fast with $T$. 

Newey-West Estimator

• **Issue 2 (＆3):** $S_T$ needs to be psd to be a proper covariance matrix.

• Newey-West (1987): Based on a quadratic form and using the *Bartlett kernel* produce a consistent psd estimator of $Q^*$:

$$S_T = \sum_{j=-(T+1)}^{T-1} k\left(\frac{j}{L(T)}\right) \hat{\Gamma}_T(j)$$

where $k\left(\frac{j}{L(T)}\right) = 1 - \frac{|j|}{L + 1}$ is the *Bartlett kernel or window*; and $L(T)$ is its *bandwidth*.

• **Intuition for Bartlett kernel:** Use weights in the sum that imply that the process becomes less autocorrelated as time goes by –i.e, the terms have a lower weight in the sum as the difference between $i$ and $s$ grows.

Newey-West Estimator

• Other kernels work too. Typical requirements for $k(.)$:
  - $|k(x)| \leq 1$;
  - $k(0) = 1$;
  - $k(x) = k(-x)$ for all $x \in \mathbb{R}$,
  - $\int |k(x)| \ dx < \infty$;
  - $k(.)$ is continuous at 0 and at all but a finite number of other points in $\mathbb{R}$, and

$$\int_{-\infty}^{\infty} k(x) e^{-i\omega x} \ dx \geq 0, \ \forall \omega \in \mathbb{R}$$

The last condition is bit technical and ensures psd, see Andrews (1991).
Newey-West Estimator

• Two components for the NW HAC estimator:
  (1) Start with Heteroscedasticity Component:
  \[ S_0 = \frac{1}{T} \sum_i c_i^2 x_i x_i' \]
  – the White estimator.
  (2) Add the Autocorrelation Component
  \[ S_T = S_0 + \frac{1}{T} \sum_j k_j(T) \sum_{i=1}^{T-j} \left( x_i c_i x_i' + x_i c_i x_{i+j}' \right) \]
  where
  \[ k(\frac{j}{L(T)}) = 1 - \frac{|j|}{L+1} \]
  – The Bartlett kernel
  \[ \Rightarrow \] linearly decaying weights.

Then,
\[ \text{Est. Var}[\beta] = \frac{1}{T} \left( X'X / T \right)^{-1} S \left( X'X / T \right)^{-1} \]
– NW’s HAC Var.

• Under suitable conditions, as \( L, T \to \infty \) and \( L/T \to 0 \), \( S_T \to Q^* \).

Asymptotic inferences on \( \beta \), based on OLS \( \beta \), can be done with \( t \)-test and \( Wald \) tests using \( N(0,1) \) and \( \chi^2 \) critical values, respectively.

NW Estimator: Alternative Computation

• The sum-of-covariance estimator can alternatively be computed in the frequency domain as a weighted average of periodogram ordinates (an estimator of the spectrum at frequency \( 2\pi j/T \)). To be discussed in Time Series lectures):
  \[ S_T^{WP} = 2\pi \sum_{j=1}^{T-1} K_T(2\pi j/T) I_{xex} (2\pi j/T) \]
  where
  \[ K_T = T^{-1} \sum_{u=0}^{T-1} K_T(u/L)e^{iu\omega} \]
  and \( I_{xex} \) is the periodogram of \( x_i c_i \) at frequency \( \omega \):
  \[ I_{xex} (\omega) = d_{xe} (\omega) d_{xe} (\omega)' \]
  where
  \[ d_{xe} (\omega) = 2\pi \sum_{i=1}^{T} (x_i c_i) e^{-i\omega x} \]

• Under suitable conditions, as \( L & T \to \infty \) and \( L/T \to 0 \),
  \[ S_T^{WP} \to Q^* \].
NW Estimator: Kernel Choice

- Other kernels, $k_{L}(l)$, besides the Bartlett kernel, can be used:

    \[ k_{L}(l) = 1 - 6l^2 + 6|l|^3 \quad \text{for} \quad 0 \leq |l| \leq 1/2 \]
    \[ = 2(1 - |l|^3) \quad \text{for} \quad 0 \leq |l| \leq 1/2 \]
    \[ = 0 \quad \text{otherwise} \]

    \[ k_{L}(l) = \frac{25}{12\pi^2 l^2}[\sin(6\pi l/5)/(6\pi l) - \cos(6\pi l/5)] \]

  - Daniell kernel – Ng and Perron (1996):
    \[ k_{L}(l) = \frac{\sin(\pi l)}{\pi l} \]

- These kernels are all symmetric about the vertical axis. The Bartlett and Parzen kernels have a bounded support $[-1, 1]$, but the other two have unbounded support.

NW Estimator: Kernel Choice

- Q: In practice – i.e., in finite samples – which kernel to use? And $L(T)$? Asymptotic theory does not help us to determine them.

- Andrews (1991) finds optimal kernels and bandwidths by minimizing the (asymptotic) MSE of the LRV. The QS kernel is 8.6% more efficient than the Parzen kernel; the Bartlett kernel is the worst one. (BTW, different kernels have different optimal $L$.)

### Diagram

![Graph of different kernels](image)

- $k_{L}(x)$

Q: In practice – i.e., in finite samples – which kernel to use? And $L(T)$? Asymptotic theory does not help us to determine them.

- Andrews (1991) finds optimal kernels and bandwidths by minimizing the (asymptotic) MSE of the LRV. The QS kernel is 8.6% more efficient than the Parzen kernel; the Bartlett kernel is the worst one. (BTW, different kernels have different optimal $L$.)
NW Estimator: Remarks

• Today, the HAC estimators are usually referred as NW estimators, regardless of the kernel used if they produce a psd covariance matrix.

• All econometric packages (SAS, SPSS, Eviews, etc.) calculate NW SE. In R, you can use the library “sandwich,” to calculate NW SEs:

   ```
   > NeweyWest(x, lag = NULL, order.by = NULL, prewhite = TRUE, adjust = FALSE,
   diagnostics = FALSE, sandwich = TRUE, ar.method = "ols", data = list(), verbose = FALSE)
   ```

Example:

````
## fit investment equation using the 3 factor Fama French Model for IBM returns,
fit <- lm(y ~ x -1)

## NeweyWest computes the NW SEs. It requires lags=L & suppression of prewhitening
NeweyWest(fit, lag = 4, prewhite = FALSE)
```

Note: It is usually found that the NW SEs are downward biased.

NW Estimator: Remarks

• You can also program the NW SEs yourself. In R:

```r
NW_f <- function(y,X,b,lag)
    {
    T <- length(y);
    k <- length(b);
    yhat <- X%*%b
    e <- y - yhat
    hhat <- t(X)*as.vector(t(e))
    G <- matrix(0,k,k)
    a <- 0
    w <- numeric(T)
    while (a <= lag) {
        Ta <- T - a
        ga <- matrix(0,k,k)
        w[lag+1+a] <- (lag+1-a)/(lag+1)
        za <- hhat[,a+1:Ta] %*% t(hhat[1:Ta])
        ga <- ga + za
        G <- G + w[lag+1+a]*ga
        a <- a+1
    }
    F <- t(X)%*%X
    V <- solve(F)%*%G%*%solve(F)
    nw_se <- sqrt(diag(V))
    ols_se <- sqrt(diag(solve(F)*drop((t(e)%*%e))/(T-k)))
    l_se = list(nw_se,ols_se)
    return(l_se)
    }
NW_f(y,X,b,lag=4)
```
NW Estimator: Example in R

Example: We estimate the 3 factor F-F model for IBM returns:

```r
> library(sandwich)
> reg <- lm(y~x -1)
> reg$coefficients
 x    xx1    xx2    xx3
-0.2331471  0.0101872  0.0009803 -0.0044460
⟹ OLS b

> SE_HC <- diag(sqrt(abs(vcovHC(reg, type= "HC3"))))
> SE_HC
 x    xx1    xx2    xx3
0.0115834  0.0028082  0.0043221  0.0044162
⟹ White SE HC3

>NW <- NeweyWest(reg, lag = 4, prewhite = FALSE)
> SE_NW <- diag(sqrt(abs(NW)))
> SE_NW
 x    xx1    xx2    xx3
0.0236296  0.0027982  0.0038953  0.0054311
⟹ NW SE
```

NW Estimator: Remarks

- Parametric estimators of $Q'$ are simple and perform reasonably well. But, we need to specify the ARMA model. Thus, they are not robust to misspecification of $(A3')$. This is the appeal of White & NW.

- NW SEs perform poorly in Monte Carlo simulations:
  - NW SEs tend to be downward biased.
  - The finite-sample performance of tests using NW SE is not well approximated by the asymptotic theory.
  - Tests have serious size distortions.

- A key assumption in establishing consistency is that $L \to \infty$ as $T \to \infty$, but $L/T \to 0$. But, in practice, the fraction $L/T$ is never equal to 0, but approaches some positive fraction $b (b \in (0,1])$. Under this situation, we need new asymptotics to derive properties of estimator.
NW Estimator: Remarks

- There are estimators of $Q^*$ that are not consistent, but with better small sample properties. See Kiefer, Vogelsang and Bunzel (2000).

- The SE based on these inconsistent estimators of $Q^*$ that are used for testing are referred as Heteroskedasticity-Autocorrelation Robust (HAR) SE.

- More on this topic in Lecture 13.

References: Müller (2014) & Sun (2014). There is a recent review (not that technical) paper by Lazarus, Lewis, Stock & Watson (2016) with recommendations on how to use these HAR estimators.

Autocorrelated Residuals: Gasoline Demand

$logG = \beta_1 + \beta_2 logPg + \beta_3 logY + \beta_4 logPnc + \beta_5 logPuc + \varepsilon$
### NW Estimator vs. Standard OLS (Greene)

**BALTAGI & GRIFFIN DATA SET**

| Variable | Coefficient | Standard Error | t-ratio | P[|T|>t] |
|----------|-------------|----------------|---------|---------|
| **Standard OLS** |
| Constant | -21.2111*** | .75322         | -28.160 | .0000   |
| LP       | -.02121     | .04377         | -.485   | .6303   |
| LY       | 1.09587***  | .07771         | 14.102  | .0000   |
| LPNC     | -.37361**   | .15707         | -2.379  | .0215   |
| LPUC     | .02003      | 1.0303         | .194    | .8471   |
| **Robust VC Newey-West, Periods = 10** |
| Constant | -21.2111*** | 1.33095        | -15.937 | .0000   |
| LP       | -.02121     | .06119         | -.347   | .7305   |
| LY       | 1.09587***  | .14234         | 7.699   | .0000   |
| LPNC     | -.37361**   | .16615         | -2.249  | .0293   |
| LPUC     | .02003      | 1.41176        | .141    | .8882   |

### Generalized Least Squares (GLS)

- **Assumptions** *(A1), (A2), (A3’), & (A4)* hold. That is,
  - *(A1)* DGP: \( y = X\beta + \epsilon \) is correctly specified.
  - *(A2)* \( E[\epsilon | X] = 0 \)
  - *(A3’)* \( \text{Var}[\epsilon | X] = \sigma^2 \Omega \) (recall \( \Omega \) is symmetric \( \Rightarrow T' \Omega = \Omega \))
  - *(A4)* \( X \) has full column rank –i.e., \( \text{rank}(X) = k \), where \( T \geq k \).

**Note:** \( \Omega \) is symmetric \( \Rightarrow \) exists \( T \in \mathbb{R} \) \( T' \Omega = \Omega \) \( \Rightarrow T^{-1} \Omega T^{-1} = I \)

- We transform the linear model in *(A1)* using \( P = \Omega^{1/2} \).
  \[
P = \Omega^{1/2} \quad \Rightarrow P'P = \Omega
  \]
  \[
P \epsilon = P \epsilon^* \]
  \[
y^* = X^* \beta + \epsilon^*
  \]
  \[
  E[\epsilon^* \epsilon^* | X^*] = PE[\epsilon \epsilon' | X^*]P' = PE[\epsilon \epsilon' | X]P' = \sigma^2 P \Omega P' = \sigma^2 \Omega^{1/2} \Omega \Omega^{1/2} = \sigma^2 I_T \quad \Rightarrow \text{back to (A3)}
  \]
Generalized Least Squares (GLS)

• The transformed model is homoscedastic:
  \[ E[\varepsilon^*\varepsilon^*'|X^*] = PEP\varepsilon^*\varepsilon^*'|X^*]P' = \sigma^2P\Omega P' = \sigma^2I_T \]

• We have the CLM framework back \(\Rightarrow\) we can use OLS!

• Key assumption: \(\Omega\) is known, and, thus, \(P\) is also known; otherwise we cannot transformed the model.

• Q: Is \(\Omega\) known?

Alexander C. Aitken (1895 –1967, NZ)

Generalized Least Squares (GLS)

  \[ P\gamma = PX\beta + Pe \text{ or} \]
  \[ y^* = X^*\beta + \varepsilon^*. \]
  \[ E[\varepsilon^*\varepsilon^*'|X^*] = \sigma^2I_T \]
  We can use OLS in the transformed model. It satisfies G-M theorem.
  Thus, the GLS estimator is:
  \[ b_{GLS} = b^* = (X^*X^*^{-1})X^*y^* = (X'P'PX)^{-1}X'P'y \]
  \[ = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \]

Note I: \( b_{GLS} \neq b \). \( b_{GLS} \) is BLUE by construction, \( b \) is not.

Note II: Both unbiased and consistent. In practice, both estimators will be different, but not that different. If they are very different, something is rotten in Denmark.
Generalized Least Squares (GLS)

• Check unbiasedness:
  \[ b_{GLS} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} y = \beta + (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} \epsilon \]
  \[ E[b_{GLS} | X] = \beta \]

• Efficient Variance
  \[ \text{Var}[b_{GLS} | X] = E[(b_{GLS} - \beta)(b_{GLS} - \beta)' | X] \]
  \[ = E[(X'\Omega^{-1}X)^{-1} X'\Omega^{-1} \epsilon \epsilon' X'\Omega^{-1} (X'\Omega^{-1}X)^{-1} | X] \]
  \[ = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} E[\epsilon \epsilon' | X] \Omega^{-1} X(X'\Omega^{-1}X)^{-1} \]
  \[ = \sigma^2 (X'\Omega^{-1}X)^{-1} \]

Note: \( b_{GLS} \) is BLUE. This “best” variance can be derived from

\[ \text{Var}[b_{GLS} | X] = \sigma^2 (X^*X^*)^{-1} = \sigma^2 (X'\Omega^{-1}X)^{-1} \]

Then, the usual variance of the OLS estimator is biased and inefficient!

Generalized Least Squares (GLS)

• If we relax the CLM assumptions \((A2)\) and \((A4)\), as we did in Lecture 7, we only have asymptotic properties for GLS:
  – Consistency - “well behaved data.”
  – Asymptotic distribution under usual assumptions.
    (easy for heteroscedasticity, complicated for autocorrelation.)
  – Wald tests and \(F\)-tests with usual asymptotic \(\chi^2\) distributions.
Consistency (Green)

Use Mean Square
\[ \text{Var}[\hat{\beta}_i|x] = \frac{\sigma^2}{n} \left( \frac{x' \Sigma^{-1} x}{n} \right)^{-1} \rightarrow 0? \]

Requires to be \( \left( \frac{x' \Sigma^{-1} x}{n} \right) \) "well behaved"

Either converge to a constant matrix or diverge.

Heteroscedasticity case:
\[ \frac{x' \Sigma^{-1} x}{n} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\omega_{ij}} \text{x}_i \text{x}_j' \]

Autocorrelation case:
\[ \frac{x' \Sigma^{-1} x}{n} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\omega_{ij}} \text{x}_i \text{x}_j', \text{n}^2 \text{ terms. Convergence is unclear.} \]

Consistency – Autocorrelation case (Green)

\[ \frac{x' \Omega^{-1} x}{T} = \frac{1}{T} \sum_{j=1}^{T} \sum_{i=1}^{T} \frac{1}{\omega_{ij}} \text{x}_i \text{x}_j' = \frac{\sigma_0^2}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} \rho_{i-j} \]

• If the \( \{X_t\} \) were uncorrelated — i.e., \( \rho_k = 0 \) —, then \( \text{Var}[\hat{\beta}_{\text{GLS}}|X] \rightarrow 0. \)

• We need to impose restrictions on the dependence among the \( X_t \)’s. Usually, we require that the autocorrelation, \( \rho_k \), gets weaker as \( T \) grows (and the double sum becomes finite).
Asymptotic Normality (Green)

\[ \sqrt{n}(\hat{\beta} - \beta) = \sqrt{n} \left( \frac{\mathbf{x}' \Omega^{-1} \mathbf{x}}{n} \right)^{-1} \frac{1}{n} \mathbf{x}' \Omega^{-1} \epsilon \]

Converge to normal with a stable variance O(1)?

\( \left( \frac{\mathbf{x}' \Omega^{-1} \mathbf{x}}{n} \right)^{-1} \rightarrow \) a constant matrix?

\( \frac{1}{n} \mathbf{x}' \Omega^{-1} \epsilon \rightarrow \) a mean to which we can apply the central limit theorem?

Heteroscedasticity case?

\( \frac{1}{n} \mathbf{x}' \Omega^{-1} \epsilon = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \left( \frac{e_i}{\sqrt{\lambda_i}} \right) \) \( \text{Var} \left( \frac{e_i}{\sqrt{\lambda_i}} \right) = \sigma^2, \frac{x_i}{\sqrt{\lambda_i}} \) is just data.

Apply Lindeberg-Feller. (Or assuming \( x_i / \sqrt{\lambda_i} \) is a draw from a common distribution with mean and fixed variance - some recent treatments.)

Autocorrelation case?

Asymptotic Normality – Autocorrelation case

For the autocorrelation case

\[ \frac{1}{n} \mathbf{x}' \Omega^{-1} \epsilon = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \Omega_{ij} \mathbf{x}_i e_i \]

Does the double sum converge? Uncertain. Requires elements of \( \Omega^{-1} \) to become small as the distance between \( i \) and \( j \) increases. (Has to resemble the heteroscedasticity case.)

• The dependence is usually broken by assuming \{\( \mathbf{x}_i, e_i \}\} form a mixing sequence. The intuition behind mixing is simple; but, the formal details and its application to the CLT can get complicated.

• Intuition: \{\( Z_i \}\} is a mixing sequence if any two groups of terms of the sequence that are far apart from each other are approximately independent -- and the further apart, the closer to being independent.
Brief Detour: Time Series

• With autocorrelated data, we get dependent observations. Recall, 
  \[ \varepsilon_t = \rho \varepsilon_{t-1} + u_t \]

• The independence assumption (A2') is violated. The LLN and the CLT cannot be easily applied, in this context. We need new tools and definitions.

• We will introduce the concepts of stationarity and ergodicity. The ergodic theorem will give us a counterpart to the LLN.

• To get asymptotic distributions, we also need a CLT for dependent variables, using the concept of mixing and stationarity. Or we can rely on the martingale CLT. We will leave this as “coming attractions.”

Brief Detour: Time Series - Stationarity

• Consider the joint probability distribution of the collection of RVs:
  \[ F(z_{t_1}, z_{t_2}, \ldots, z_{t_n}) = P(Z_{t_1} \leq z_{t_1}, Z_{t_2} \leq z_{t_2}, \ldots, Z_{t_n} \leq z_{t_n}) \]

Then, we say that a process is

1st order stationary if \[ F(z_{t_1}) = F(z_{t_1+k}) \] for any \( t_1, k \)

2nd order stationary if \[ F(z_{t_1}, z_{t_2}) = F(z_{t_1+k}, z_{t_2+k}) \] for any \( t_1, t_2, k \)

Nth-order stationary if \[ F(z_{t_1}, \ldots, z_{t_n}) = F(z_{t_1+k}, \ldots, z_{t_n+k}) \] for any \( t_1, t_n, k \)

• Definition. A process is strongly (strictly) stationary if it is a Nth-order stationary process for any \( N \).
Brief Detour: Time Series – Moments

• The moments describe a distribution. We calculate the moments as usual.

\[ E(Z_t) = \mu_t = \int Z_t f(z_t) dz_t \]

\[ \text{Var}(Z_t) = \sigma^2_t = E(Z_t - \mu_t)^2 = \int (Z_t - \mu_t)^2 f(z_t) dz_t \]

\[ \text{Cov}(Z_{t_1}, Z_{t_2}) = E[(Z_{t_1} - \mu_{t_1})(Z_{t_2} - \mu_{t_2})] \]

\[ \rho(t_1, t_2) = \frac{\text{cov}(Z_{t_1}, Z_{t_2})}{\sqrt{\sigma^2_{t_1}} \sqrt{\sigma^2_{t_2}}} \]

• Stationarity requires all these moments to be independent of time.

Brief Detour: Time Series – Moments

• For strictly stationary process: \( \mu_t = \mu \) and \( \sigma^2_t = \sigma^2 \)

because \( F(z_{t_i}) = F(z_{t_i+k}) \rightarrow \mu_{t_i} = \mu_{t_i+k} = \mu \)

provided that \( E(|Z_t|) < \infty, \ E(Z^2_t) < \infty \)

Then, \( F(z_{t_1}, z_{t_2}) = F(z_{t_1+k}, z_{t_2+k}) \Rightarrow \text{cov}(z_{t_1}, z_{t_2}) = \text{cov}(z_{t_1+k}, z_{t_2+k}) \Rightarrow \rho(t_1, t_2) = \rho(t_1 + k, t_2 + k) \)

let \( t_1 = t - k \) and \( t_2 = t \), then \( \rho(t_1, t_2) = \rho(t - k, t) = \rho(t, t + k) = \rho_k \)

The correlation between any two RVs depends on the time difference.
Brief Detour: Time Series – Weak Stationarity

• A process is said to be \( N \)-th order weakly stationary if all its joint moments up to order \( N \) exist and are time invariant.

• A covariance stationary process (or 2nd order weakly stationary) has:
  – constant mean
  – constant variance
  – covariance function depends on time difference between RV.

Brief Detour: Time Series – Ergodicity

• We want to allow as much dependence as the LLN allows us to do it.

• But, stationarity is not enough, as the following example shows:

• Example: Let \( \{U_t\} \) be a sequence of i.i.d. RVs uniformly distributed on \([0, 1]\) and let \( Z \) be \( N(0, 1) \) independent of \( \{U_t\} \).

Define \( Y_t = Z + U_t \). Then, \( Y_t \) is stationary (why?), but

\[
\bar{Y}_n = \frac{1}{n} \sum_{t=1}^n Y_t \xrightarrow{\text{as}\ n\ \to\ \infty} E(Y_t) = \frac{1}{2} \\
\bar{Y}_n - Z \xrightarrow{\text{P}} \frac{1}{2}
\]

The problem is that there is too much dependence in the sequence \( \{Y_t\} \).
In fact the correlation between \( Y_1 \) and \( Y_t \) is always positive for any value of \( t \).
We want to estimate the mean of the process \( \{Z_t\} \), \( \mu(Z_t) \). But, we need to distinguishing between ensemble average and time average:

- Ensemble Average \( \overline{Z} = \frac{1}{m} \sum_{i=1}^{m} Z_i \)
- Time Series Average \( \overline{Z} = \frac{1}{n} \sum_{t=1}^{n} Z_t \)

Q: Which estimator is the most appropriate?
A: Ensemble Average, the average of re-runs of identical experiments. But, it is impossible to calculate. We only observe one experiment: \( Z_t \).

Q: Under which circumstances we can use the time average (only one realization of \( \{Z_t\} \))? Is the time average an unbiased and consistent estimator of the mean? The Ergodic Theorem gives us the answer.

**Brief Detour: Time Series – Ergodicity of mean**

- Recall the sufficient conditions for consistency of an estimator: the estimator is asymptotically unbiased and its variance asymptotically collapses to zero.

1. Q: Is the time average asymptotically unbiased? Yes.
   \[
   E(\overline{Z}) = \frac{1}{n} \sum_{t} E(Z_t) = \frac{1}{n} \sum_{t} \mu = \mu
   \]

2. Q: Is the variance going to zero as \( T \) grows? It depends.
   \[
   \text{var}(\overline{Z}) = \frac{1}{n^2} \sum_{t=1}^{n} \sum_{s=1}^{n} \text{cov}(Z_t, Z_s) = \frac{\gamma_0}{n^2} \sum_{t=1}^{n} \sum_{s=1}^{n} \rho_{t-s} = \frac{\gamma_0}{n^2} \sum_{t=1}^{n} (\rho_{t-1} + \rho_{t-2} + \cdots + \rho_{t-n}) = \\
   = \frac{\gamma_0}{n} [\rho_0 + \rho_1 + \cdots + \rho_{n-1}] + (\rho_1 + \rho_0 + \rho_1 + \cdots + \rho_{n-2}) + \\
   + \cdots + (\rho_{-(n-1)} + \rho_{-(n-2)} + \cdots + \rho_0)]
   \]
**Brief Detour: Time Series – Ergodicity of mean**

\[
\text{var}(\bar{z}) = \frac{\gamma_0}{n^2} \sum_{k=-(n-1)}^{n-1} (n-|k|) \rho_k = \frac{\gamma_0}{n} \sum_k (1-\frac{|k|}{n}) \rho_k
\]

\[
\lim_{n \to \infty} \text{var}(\bar{z}) = \lim_{n \to \infty} \frac{\gamma_0}{n} \sum_k (1-\frac{|k|}{n}) \rho_k \to 0
\]

- If \( Z_t \) were uncorrelated, the variance of the time average would be \( O(n^{-1}) \). Since independent random variables are necessarily uncorrelated (but not vice versa), we have just recovered a form of the LLN for independent data.

Q: How can we make the remaining part, the sum over the upper triangle of the covariance matrix, go to zero as well?

A: We need to impose conditions on \( \rho_k \). Conditions weaker than "they are all zero;" but, strong enough to exclude the sequence of identical copies.

---

**Brief Detour: Time Series – Ergodicity of mean**

- We use two inequalities to put upper bounds on the variance of the time average:

\[
\sum_{t=1}^{n-1} \sum_{k=1}^{n-1} \rho_k \leq \sum_{t=1}^{n-1} \sum_{k=1}^{n-1} |\rho_k| \leq \sum_{t=1}^{\infty} \sum_{k=1}^{\infty} |\rho_k|
\]

Covariances can be negative, so we upper-bound the sum of the actual covariances by the sum of their magnitudes. Then, we extend the inner sum so it covers all lags. This might of course be infinite (sequence-of-identical-copies).

- **Definition**: A covariance-stationary process is *ergodic* for the mean if

\[
p \lim \bar{z} = E(Z_t) = \mu
\]

Ergodicity Theorem: Then, a sufficient condition for ergodicity for the mean is

\[\rho_k \to 0 \text{ as } k \to \infty\]
Brief Detour: Time Series – Ergodicity of 2nd moments

- A sufficient condition to ensure ergodicity for second moments is:
  \[ \sum_k |\rho_k| < \infty \]

A process which is ergodic in the first and second moments is usually referred as \textit{ergodic in the wide sense}.

- \textit{Ergodicity under Gaussian Distribution}
  If \{Z_t\} is a stationary Gaussian process, \[ \sum_k |\rho_k| < \infty \]
is sufficient to ensure ergodicity for all moments.

\textbf{Note:} Recall that only the first two moments are needed to describe the normal distribution.

Test Statistics (Assuming Known \(\Omega\)) (Green)

- Back to GLS. From (A1)-(A4), we get:
  \[ b_{\text{GLS}} = (X^*X^*)^{-1}X^*y^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \]
  \[ \text{Var}[b_{\text{GLS}}|X] = \sigma^2(X^*X^*)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1} \]

- With known \(\Omega\), apply all familiar results to the transformed model:
  - With normality, (A5) holds, \(t\)- and \(F\)-statistics apply to least squares based on \(Py\) and \(PX\)

  - Without normality, we rely on asymptotic results, where we get asymptotic normality for \(b_{\text{GLS}}\). We use Wald statistics and the chi-squared distribution, still based on the transformed model.

- Key step to do GLS: Derive the transformation matrix \(P = \Omega^{1/2}\).
(Weighted) GLS: Pure Heteroscedasticity

- Key step to do GLS: Derive the transformation matrix \( P = \Omega^{1/2} \).

\[
\begin{bmatrix}
\omega_1 & 0 & \ldots & 0 \\
0 & \omega_2 & \ldots & 0 \\
0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & \omega_n \\
\end{bmatrix}
\]

\[
\Omega^{1/2} = \begin{bmatrix}
1/\sqrt{\omega_1} & 0 & \ldots & 0 \\
0 & 1/\sqrt{\omega_2} & \ldots & 0 \\
0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 1/\sqrt{\omega_n} \\
\end{bmatrix}
\]

\[
\hat{\beta} = (X' \Omega^{-1} X)^{-1} (X' \Omega^{-1} y) = \left( \sum_{i=1}^{n} \frac{1}{\omega_i} x_i x'_i \right)^{-1} \left( \sum_{i=1}^{n} \frac{1}{\omega_i} x_i y_i \right)
\]

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left( y_i - x_i \hat{\beta} \right)^2}{n-K} \Rightarrow \text{WLS: Think of } [\omega_i]^{-1/2} \text{ as weights. We do OLS with the weighted data.}
\]

GLS: First-order Autocorrelation -AR(1)- Case

- Let \( \epsilon_t = \rho \epsilon_{t-1} + u_t \) (a first order autocorrelated process). Let \( u_t \) = non-autocorrelated, white noise error \( \sim D(0, \sigma_u^2) \)

- Then, \( \epsilon_t = \rho \epsilon_{t-1} + u_t \) (the autoregressive form)
  \[
  = \rho (\rho \epsilon_{t-2} + u_{t-1}) + u_t \\
  = \cdots \\
  = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + \cdots \\
  = \sum_{i=0}^{\infty} \rho^i u_{t-i} \\
  \text{(a moving average)}
  \]

- \( \text{Var}[\epsilon_t] = \sum_{i=0}^{\infty} \rho^i \text{Var}[\epsilon_{t-i}] = \sum_{i=0}^{\infty} \rho^i \sigma_u^2 = \sigma_u^2 / (1 - \rho^2) \) \( \Rightarrow \text{we need to assume } |\rho| < 1. \)

- Easier:
  \[
  \text{Var}[\epsilon_t] = \rho^2 \text{Var}[\epsilon_{t-1}] + \text{Var}[u_t] \Rightarrow \text{Var}[\epsilon_t] = \sigma_u^2 / (1- \rho^2)
  \]
GLS: AR(1) Case - Autocovariances

Continuing...
\[
\text{Cov}[\varepsilon_t, \varepsilon_{t-1}] = \text{Cov}[\rho \varepsilon_{t-1} + u_t, \varepsilon_{t-1}]
\]
\[
= \rho \text{Cov}[\varepsilon_{t-1}, \varepsilon_{t-1}] + \text{Cov}[u_t, \varepsilon_{t-1}]
\]
\[
= \rho \text{Var}[\varepsilon_{t-1}] = \rho \text{Var}[\varepsilon_t]
\]
\[
= \frac{\rho \sigma_u^2}{(1 - \rho^2)}
\]
\[
\text{Cov}[\varepsilon_t, \varepsilon_{t-2}] = \text{Cov}[\rho \varepsilon_{t-1} + u_t, \varepsilon_{t-2}]
\]
\[
= \rho \text{Cov}[\varepsilon_{t-1}, \varepsilon_{t-2}] + \text{Cov}[u_t, \varepsilon_{t-2}]
\]
\[
= \rho \text{Cov}[\varepsilon_t, \varepsilon_{t-1}]
\]
\[
= \frac{\rho^2 \sigma_u^2}{(1 - \rho^2)} \text{ and so on.}
\]

GLS: AR(1) Case - Autocorrelation Matrix

• Now, we get \( \Sigma = \sigma^2 \Omega \)

\[
\sigma^2 \Omega = \left( \frac{\sigma_u^2}{1 - \rho^2} \right) \begin{bmatrix}
1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\
\rho & 1 & \rho & \ldots & \rho^{T-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\
\vdots & \vdots & \vdots & \ldots & \vdots \\
\rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1
\end{bmatrix}
\]

(Note, trace \( \Omega = n \) as required.)
GLS: First-order Autocorrelation Case

• Then, we can get the transformation matrix $P = \Omega^{1/2}$:

$$\Omega^{1/2} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \ldots & 0 \\ -\rho & 1 & 0 & \ldots & 0 \\ 0 & -\rho & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & -\rho \end{bmatrix}$$

$$\Omega^{1/2}y = \begin{pmatrix} (\sqrt{1-\rho^2})y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{pmatrix} \Rightarrow \text{GLS: Transformed } y^*.$$ 

GLS: The Autoregressive Transformation

• With AR models, sometimes it is easier to transform the data by taking pseudo differences.

• For the AR(1) model, we have,

$$y_t = x_t'\beta + \varepsilon_t \quad \varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

$$\rho y_{t-1} = \rho x_{t-1}'\beta + \rho \varepsilon_{t-1}$$

$$y_t - \rho y_{t-1} = (x_t - \rho x_{t-1})'\beta + (\varepsilon_t - \rho \varepsilon_{t-1})$$

$$y_t - \rho y_{t-1} = (x_t - \rho x_{t-1})'\beta + u_t$$

(Where did the first observation go?)
GLS: Unknown $\Omega$ (Green)

• Problem with GLS: $\Omega$ is unknown.

• Solution: Estimate $\Omega$. $\implies$ Feasible GLS (FGLS).

• For now, we will consider two methods of estimation:
  – Two-step, or Feasible estimation. Estimate $\Omega$ first, then do GLS. Emphasize same logic as White and Newey-West: We do not need to estimate $\Omega$. We need to find a matrix that behaves the same as $(1/T)X'\Omega X$.
  – Nice asymptotic properties of the FGLS estimator.

• ML estimation of $\beta$, $\sigma^2$, and $\Omega$ all at the same time.
  – We will examine two applications:
    - Harvey's model of heteroscedasticity
    - Beach-MacKinnon on the AR(1) model (see Lecture 13).

GLS: Specification of $\Omega$ (Green)

• $\Omega$ must be specified first.

• A full unrestricted $\Omega$ contains $T(T+1)/2 - 1$ parameters. (Why minus 1? Remember, tr($\Omega$) = $T$, so one element is determined.)

• $\Omega$ is generally specified (modeled) in terms of a few parameters. Thus, $\Omega = \Omega(\theta)$ for some small parameter vector $\theta$. It becomes a question of estimating $\theta$.

• Examples:
  (1) $\text{Var}[\varepsilon_i | X] = \sigma^2 \exp(\gamma' z_i)$. Variance a function of $\gamma$ and some variable $z_i$ (say, firm size or country).

  (2) $\varepsilon_i$ with AR(1) process. We have already derived $\sigma^2 \Omega$ as a function of $\rho$. 
Harvey’s Model of Heteroscedasticity (Green)

- The variance for observation $i$ is a function of $z_i$:
  \[
  \text{Var}[\varepsilon_i | \mathbf{X}] = \sigma^2 \exp(\gamma' z_i)
  \]
  But, errors are not auto/cross correlated:
  \[
  \text{Cov}[\varepsilon_i, \varepsilon_j | \mathbf{X}] = 0
  \]
  - The driving variable, $z$, can be firm size, a set of dummy variables - for example, for countries. This example is the one used for the estimation of the previous groupwise heteroscedasticity model.

- Then, we have a functional form for $\Sigma = \sigma^2 \Omega$
  \[
  \Sigma = \text{diagonal } [\exp(\theta + \gamma' z_i)],
  \]
  \[
  \theta = \log(\sigma^2)
  \]
  Once we specify $\Omega$ (and can be estimated), GLS is feasible.

GLS: AR(1) Model of Autocorrelation (Green)

- We have already derived $\Sigma = \sigma^2 \Omega$ for the AR(1) case..

\[
\sigma^2 \Omega = \left(\frac{\sigma_u^2}{1 - \rho^2}\right) \begin{bmatrix}
1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\
\rho & 1 & \rho & \ldots & \rho^{T-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1
\end{bmatrix}
\]

- Now, if we estimate $\sigma_u^2$ and $\rho$, we can do FGLS.
Estimated AR(1) Model (Greene)

AR(1) Model:  $e(t) = \rho_0 e(t-1) + u(t)$
Initial value of $\rho_0$ = .87566
Maximum iterations = 1
Method = Prais - Winsten
Iter= 1, SS= .022, Log-L= 127.593
Final value of $\rho_0$ = .959411
Std. Deviation: $e(t)$ = .076512
Std. Deviation: $u(t)$ = .021577
Autocorrelation: $u(t)$ = .253173

$N[0,1]$ used for significance levels

| Variable | Coefficient | Standard Error | t/St.Er | P[|Z|>z] |
|----------|-------------|----------------|---------|---------|
| Constant | -20.3373*** | .69623         | -29.211 | .0000   |
| LP       | -.11379***  | .03296         | -3.453  | .0006   |
| LY       | .87040***   | .08827         | 9.860   | .0000   |
| LPNC     | -.04028     | .06193         | -.650   | .5154   |
| LPUC     | .05426      | .12392         | .438    | .6615   |
| RHO      | .95941***   | .03949         | 24.295  | .0000   |

Standard OLS

| Variable | Coefficient | Standard Error | t/St.Er | P[|Z|>z] |
|----------|-------------|----------------|---------|---------|
| Constant | -21.2111*** | .75322         | -28.160 | .0000   |
| LP       | -.02121     | .04377         | -.485   | .6303   |
| LY       | 1.09587***  | .07771         | 14.102  | .0000   |
| LPNC     | -.37361**   | .15707         | -2.379  | .0215   |
| LPUC     | .02003      | .10330         | .194    | .8471   |

Two-Step Estimation (Green)

• The general result for estimation when $\Omega$ is estimated.

• GLS uses $[X'\Omega^{-1}X]^{-1} X' \Omega^{-1} y$ which converges in probability to $\beta$.

• We seek a vector which converges to the same thing that this does. Call it “Feasible GLS” or FGLS, based on $[X'\hat{\Omega}^{-1}X]^{-1} X' \hat{\Omega}^{-1} y$

• The object is to find a set of parameters such that $[X' \hat{\Omega}^{-1} X]^{-1} X' \hat{\Omega}^{-1} y - [X' \Omega^{-1} X]^{-1} X' \Omega^{-1} y \to 0$
Feasible GLS (Green)

For FGLS estimation, we do not seek an estimator of $\Omega$ such that

$$\hat{\Omega} - \Omega \rightarrow 0$$

This makes no sense, since $\hat{\Omega}$ is nxn and does not "converge" to anything. We seek a matrix $\Omega$ such that

$$(1/n)X'\hat{\Omega}^{-2}X - (1/n)X'\Omega^{-2}X \rightarrow 0$$

For the asymptotic properties, we will require that

$$(1/\sqrt{n})X'\hat{\Omega}^{-2}\epsilon - (1/n)X'\Omega^{-2}\epsilon \rightarrow 0$$

Note in this case, these are two random vectors, which we require to converge to the same random vector.

Two-Step FGLS (Green)

- **Theorem 8.5**: To achieve full efficiency, we do not need an *efficient* estimate of the parameters in $\Omega$, only a consistent one.

- Q: Why?
Harvey’s Model (Green)

• Examine Harvey’s model once again.

Estimation:
(1) Two-step FGLS: Use the OLS to estimate $\theta \Rightarrow \hat{\theta}$. Then, use 
\[ \{X'\Omega(\hat{\theta})^{-1}X\}^{-1}X'\Omega(\hat{\theta})^{-1}y \] to estimate $\beta$.

(2) Full ML estimation. Estimate all parameters simultaneously.
A handy result due to Oberhofer and Kmenta—the “zig-zag” approach.

Examine a model of groupwise heteroscedasticity.

---

Harvey’s Model: Groupwise Heteroscedasticity

• We have a sample, $y_{ig}$, $x_{ig}$, $x_{ig}$, with
N groups, each with $T_g$ observations.
Each group variance: $\text{Var}[\varepsilon_{ig}] = \sigma^2_g$

• Define a group dummy variable.
\[ d_{ig} = 1 \quad \text{if observation } ig \text{ is in group } j, \]
\[ = 0 \quad \text{otherwise.} \]

Then, model variances as:
\[ \text{Var}[\varepsilon_{ig}] = \sigma^2_g \exp(\theta_2 d_{i2} + \ldots + \theta_N d_{iN}) \]
\[ \text{Var}_1 = \sigma^2_g - \text{normalized variance (remember dummy trap!)} \]
\[ \text{Var}_2 = \sigma^2_g \exp(\theta_2) \]
... etc.
Harvey’s Model: Two-Step Procedure (Green)

• OLS is still consistent. Do OLS and keep e.

Step 1. Using e, calculate the group variances. That is,
- Est.Var_1 = e_1'e_1/T_1 estimates \( \sigma^2_g \)
- Est.Var_2 = e_2'e_2/T_2 estimates \( \sigma^2_g \exp(\theta_2) \)
- Estimator of \( \theta_2 \) is \( \ln[(e_2'e_2/T_2)/(e_1'e_1/T_1)] \)
- .... etc.


Step 3. Using WLS residuals, recompute variance estimators.

Iterate until convergence between steps 2 and 3.

GLS: General Remarks

• GLS is great (BLUE) if we know \( \Omega \). Very rare case.
• It needs the specification of \( \Omega \) –i.e., the functional form of autocorrelation and heteroscedasticity.
• If the specification is bad \( \Rightarrow \) estimates are biased.
• In general, GLS is used for larger samples, because more parameters need to be estimated.
• Feasible GLS is not BLUE (unlike GLS); but, it is consistent and asymptotically more efficient than OLS.
• We use GLS for inference and/or efficiency. OLS is still unbiased and consistent.
• OLS and GLS estimates will be different due to sampling error. But, if they are very different, then it is likely that some other CLM assumption is violated –likely, (A2').
Baltagi and Griffin’s Gasoline Data (Greene)

World Gasoline Demand Data, 18 OECD Countries, 19 years
Variables in the file are

COUNTRY = name of country
YEAR = year, 1960-1978
LGASPCAR = log of consumption per car
LINCOME = log of per capita income
LRPMG = log of real price of gasoline
LCARPCAP = log of per capita number of cars

See Baltagi (2001, p. 24) for analysis of these data. The article on which the analysis is based is Baltagi, B. and Griffin, J., "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures," European Economic Review, 22, 1983, pp. 117-137. The data were downloaded from the website for Baltagi’s text.
### White Estimator vs. Standard OLS (Greene)

#### BALTAGI & GRIFFIN DATA SET

**Standard OLS**

| Variable | Coefficient | Standard Error | t-ratio | P[|T|>t]| |
|----------|-------------|----------------|---------|--------|---|
| Constant | 2.39132562  | .11693429      | 20.450  | .0000  | |
| LINCOME | .88996166   | .03580581      | 24.855  | .0000  | |
| LRPMG   | -.89179791  | .03031474      | -29.418 | .0000  | |
| LCARPCAP| -.76337275  | .01860380      | -41.023 | .0000  | |

| White heteroscedasticity robust covariance matrix |

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|----------|-------------|----------------|----------|--------|-----------|
| Constant | 2.39132562  | .11794828      | 20.274   | .0000  |           |
| LINCOME | .88996166   | .04429158      | 20.093   | .0000  |           |
| LRPMG   | -.89179791  | .03890922      | -22.920  | .0000  |           |
| LCARPCAP| -.76337275  | .02152888      | -35.458  | .0000  |           |

### Baltagi and Griffin’s Gasoline Data (Greene) – Harvey’s Model

---

**Multiplicative Heteroskedastic Regression Model...**

Ordinary least squares regression ............

<table>
<thead>
<tr>
<th>LHS=LGASPCAR</th>
<th>Mean</th>
<th>4.29624</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>.54891</td>
<td></td>
</tr>
<tr>
<td>Number of observs.</td>
<td>342</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model size</th>
<th>Parameters</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degrees of freedom</td>
<td>338</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Residuals</th>
<th>Sum of squares</th>
<th>14.90436</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Wald statistic [17 d.f.]</th>
<th>699.43 (.0000) (Large)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B/P LM statistic [17 d.f.]</td>
<td>111.55 (.0000) (Large)</td>
</tr>
</tbody>
</table>

Cov matrix for b is 

\[ \sigma^2 \text{inv}(X'X)(X'WX)\text{inv}(X'X) \]

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|----------|-------------|----------------|----------|--------|-----------|
| Constant | 2.391133*** | .20010         | 11.951   | .0000  |           |
| LINCOME | .88996***   | .07358         | 12.094   | .0000  | -6.13943  |
| LRPMG   | -.89180***  | .06119         | -14.574  | .0000  | -9.04180  |
| LCARPCAP| -.76337***  | .03030         | -25.190  | .0000  | -9.04180  |
Baltagi and Griffin’s Gasoline Data (Greene) - Variance Estimates = log[\epsilon(i)’\epsilon(i)/T]

| Sigma | Coefficient   | Standard Error | b/St.Er. | P[|Z|>z] |
|-------|---------------|----------------|----------|--------|
| D1    | -2.60677***   | .72073         | -3.617   | .0003  |
| D2    | -1.52919**    | .72073         | -2.122   | .0339  |
| D3    | .47152        | .72073         | .654     | .5130  |
| D4    | -3.15102***   | .72073         | -4.372   | .0000  |
| D5    | -3.26236***   | .72073         | -4.526   | .0000  |
| D6    | -.09099       | .72073         | -.126    | .8995  |
| D7    | -1.88962***   | .72073         | -2.622   | .0087  |
| D8    | .60559        | .72073         | .840     | .4008  |
| D9    | -1.56624**    | .72073         | -2.173   | .0298  |
| D10   | -1.53284**    | .72073         | -2.127   | .0334  |
| D11   | -2.62835***   | .72073         | -3.647   | .0003  |
| D12   | -2.23638***   | .72073         | -3.103   | .0019  |
| D13   | -.77641       | .72073         | -1.077   | .2814  |
| D14   | -1.27341*     | .72073         | -1.767   | .0773  |
| D15   | -.57948       | .72073         | -.804    | .4214  |
| D16   | -1.81723**    | .72073         | -2.521   | .0117  |
| D17   | -2.93529***   | .72073         | -4.073   | .0000  |

Baltagi and Griffin’s Gasoline Data (Greene) - OLS vs. Iterative FGLS

<table>
<thead>
<tr>
<th>Ordinary Least Squares</th>
<th>Cov matrix for b is sigma^2*inv(X'X)(X'WX)inv(X'X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.39133***</td>
</tr>
<tr>
<td>LINCOME</td>
<td>.88996***</td>
</tr>
<tr>
<td>LRPMG</td>
<td>-.89180***</td>
</tr>
<tr>
<td>LCARPCAP</td>
<td>-.76337***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FGLS - Regression (mean) function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>LINCOME</td>
</tr>
<tr>
<td>LRPMG</td>
</tr>
<tr>
<td>LCARPCAP</td>
</tr>
</tbody>
</table>

• It looks like a substantial gain in reduced standard errors. OLS and GLS estimates a bit different => problems?