Lecture 11
GLS

1

CLM: Review

• Recall the CLM Assumptions

(**A1**) DGP: $\boldsymbol{y} = \mathbf{X}\,\beta + \boldsymbol{\varepsilon}$ is correctly specified.

(**A2**) $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$

(**A3**) $Var[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\,\mathbf{I}_T$

(**A4**) $\mathbf{X}$ has full column rank –rank($\mathbf{X}$) = k–, where T $\geq k$.

• OLS estimation: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{y}$

$Var[\mathbf{b}|\mathbf{X}] = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}$

$\Rightarrow$ **b** unbiased and efficient (MVUE)

• If (**A5**) $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$ $\Rightarrow \mathbf{b}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

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: $\{x_i, \varepsilon_i\}$ $i = 1, 2, ...,$ T is a sequence of independent observations. We require **X** to have finite means and variances. Similar requirement for $\boldsymbol{\varepsilon}$, but we also require E[$\boldsymbol{\varepsilon}$]=**0.** Two new assumptions:

(**A2'**) plim $(\mathbf{X}'\boldsymbol{\varepsilon}/T) = \mathbf{0}$.

(**A4'**) plim $(\mathbf{X}'\mathbf{X}/T)=\mathbf{Q}$.

⇒ We only get asymptotic results for **b** (consistency, asymptotic normality). Tests only have large sample distributions. Boostrapping or simulations may give us better finite sample behavior.

CLM: Review - Relaxing the Assumptions

(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$(\mathbf{Z}'\mathbf{X}/T) \neq \mathbf{0}$ (*relevant* condition)

(2) plim$(\mathbf{Z}'\boldsymbol{\varepsilon}/T) = \mathbf{0}$ (*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 $\mathbf{b}_{2SLS}$ (consistency, asymptotic normality). Tests only have asymptotic distributions. Small sample behavior may be bad. Problem: Finding **Z**.

(4) (**A1**) again! – Lecture 9. Any functional form is allowed. General estimation framework: M-estimation, with only asymptotic results. A special case: NLLS. Numerical optimization needed.

Generalized Regression Model (GRM)

• Now, we go back to the CLM Assumptions:

(**A1**) DGP: $\boldsymbol{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is correctly specified.

(**A2**) E[$\boldsymbol{\varepsilon}$|**X**] = 0

(**A3**) Var[$\boldsymbol{\varepsilon}$|**X**] = $\sigma^2\,\mathbf{I}_T$

(**A4**) **X** has full column rank – 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, σ^2. Instead, we will assume:

(**A3'**) Var[$\boldsymbol{\varepsilon}$|**X**] = $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Omega}$. where $\boldsymbol{\Omega} \neq \mathbf{I}_T$

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

Generalized Regression Model (GRM)

• Now, we relax (**A3**). The CLM assumes that errors are uncorrelated and all are drawn from a distribution with the same variance, σ^2.

(**A3**) Var[$\boldsymbol{\varepsilon}$|**X**] = $\sigma^2\mathbf{I}_T$

Instead, we will assume:

(**A3'**) Var[$\boldsymbol{\varepsilon}$|**X**] = $\boldsymbol{\Sigma}$ (sometimes written= $\sigma^2\boldsymbol{\Omega}$, where $\boldsymbol{\Omega} \neq \mathbf{I}_T$)

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1T} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{T1} & \sigma_{T2} & \cdots & \sigma_T^2 \end{bmatrix}$$

• Two Leading Cases:

– Pure heteroscedasticity: We model only the diagonal elements.

– Pure autocorrelation: We model only the off-diagonal elements.

11

GRM: Pure Heteroscedasticity

• Two Pure Cases:

– Pure heteroscedasticity: $E[\boldsymbol{\varepsilon}_i{'}\boldsymbol{\varepsilon}_j|\mathbf{X}] = \sigma_{ij} = \sigma_i^2$ if $i=j$

$= 0$ if $i \neq j$

$\Rightarrow \text{Var}[\boldsymbol{\varepsilon}_i|\mathbf{X}] = \sigma_i^2$

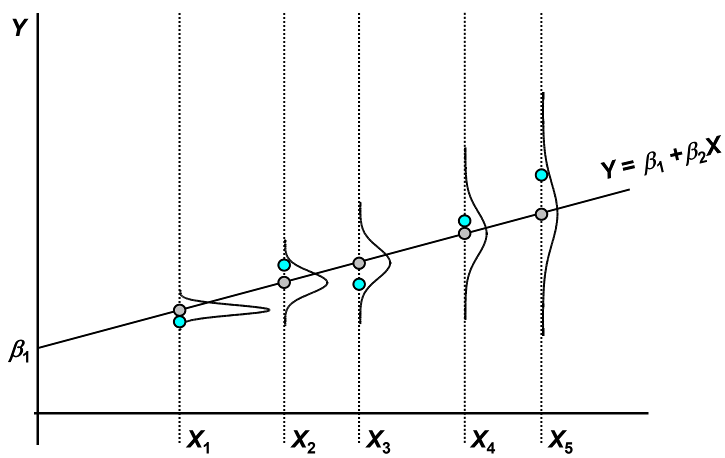
$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_T^2 \end{bmatrix}$$

• This type of variance-covariance structure is common in time series, where we observe the variance of the errors changing over time or subject to different regimes (say, bear and bull regimes).

7

GRM: Pure Heteroscedasticity

• 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.



8

GRM: Pure Cross-correlation

- Two Pure Cases:
  - Pure cross/auto-correlation: $E[\varepsilon_i'\varepsilon_j|\mathbf{X}] = \sigma_{ij}$ if $i \neq j$
    $= \sigma^2$ if $i = j$

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_{12} & \cdots & \sigma_{1T} \\ \sigma_{21} & \sigma^2 & \cdots & \sigma_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{T1} & \sigma_{T2} & \cdots & \sigma^2 \end{bmatrix}$$

- This type of variance-covariance structure is common in cross sections, where errors can show strong correlations, for example, when we model returns, the errors of two firms in the same industry can be subject to common (industry) shocks. Also common in time series, where we observe clustering of shocks over time.

9

GRM: Pure Cross-correlation

- Relative to pure cross/auto-correlation, 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.

10

GRM: Different Variance

• From (**A3**) $\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\, \mathbf{I}_T$ $\Rightarrow$ $\text{Var}[\mathbf{b}|\mathbf{X}] = \sigma^2\, (\mathbf{X}'\mathbf{X})^{-1}$.

• The true variance of **b** under (**A3'**) should be:

$$\begin{aligned}\text{Var}_T[\mathbf{b}|\mathbf{X}] &= E[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\, E[\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}|\mathbf{X}]\, (\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\, \mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\, (\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Example: We compute the true variance for the simplest case, a regression with only one explanatory variable and hetereoscedastic $\boldsymbol{\varepsilon}$:

$$\boldsymbol{y} = \mathbf{X}\,\beta + \boldsymbol{\varepsilon}, \qquad \varepsilon_i \sim D(0, \sigma_i^2)$$

$$\Rightarrow \qquad \text{Var}_T[\mathbf{b}|\mathbf{X}] = \left(\frac{1}{\sum_i^T (x_i-\bar{x})^2}\right)^2 \sum_{i=1}^T \sigma_i^2 (x_i-\bar{x})^2.$$

vs. $\quad \text{Var}[\mathbf{b}|\mathbf{X}] = \dfrac{\sigma^2}{\sum_i^T (x_i-\bar{x})^2} \neq \text{Var}_T[\mathbf{b}|\mathbf{X}].$

GRM: Different Variance

• Under (**A3'**), the OLS estimator of $\text{Var}[\mathbf{b}|\mathbf{X}]$ –i.e., $s^2\, (\mathbf{X}'\mathbf{X})^{-1}$– is biased.

• If we want to use OLS, we need to estimate $\text{Var}_T[\mathbf{b}|\mathbf{X}]$.

• To avoid the bias of inference based on OLS, we would like to estimate the unknown $\boldsymbol{\Sigma}$.

• But, $\boldsymbol{\Sigma}$ 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!

GRM: OLS Properties

• *Unbiased*

Given assumption (**A2**), the OLS estimator **b** is still *unbiased.* (Proof does not rely on $\boldsymbol{\Sigma}$):

$$\text{E}[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\, \text{E}[\mathbf{X'}\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}.$$

• *Consistency*

We relax (**A2**). Now, we assume use (**A2'**) instead. To get *consistency*, we need $\text{Var}_T[\mathbf{b}|\mathbf{X}] \to \infty$ as $T \to \infty$:

$$\begin{aligned}\text{Var}_T[\mathbf{b}|\mathbf{X}] &= (\mathbf{X'X})^{-1}\, \mathbf{X'\Sigma X}\, (\mathbf{X'X})^{-1} \\ &= (1/T)(\mathbf{X'X}/T)^{-1}\, (\mathbf{X'\Sigma X}/T)\, (\mathbf{X'X}/T)^{-1}\end{aligned}$$

Assumptions:

- plim $(\mathbf{X'X}/T) = \mathbf{Q_{XX}}$ a pd matrix of finite elements
- plim $(\mathbf{X'\Sigma X}/T) = \mathbf{Q_{X\Sigma X}}$ a finite matrix.

Under these assumptions, we get consistency for OLS.

GRM: OLS Properties

• *Asymptotic normality*?

$$\sqrt{T}\,(\mathbf{b} - \boldsymbol{\beta}) = (\mathbf{X'X}/T)^{-1}\, (\mathbf{X'}\boldsymbol{\varepsilon}/\sqrt{T})$$

Asymptotic normality for OLS followed from the application of the CLT to $\mathbf{X'}\boldsymbol{\varepsilon}/\sqrt{T}$:

$$b \xrightarrow{a} N(\beta, \frac{Q_{xx}^{-1} Q_{x\Sigma x} Q_{xx}^{-1}}{T})$$

where $\mathbf{Q_{X\Sigma X}} = \lim_{T\to\infty} \text{Var}[\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \varepsilon_t]$.

• 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 $\mathbf{X'}\boldsymbol{\varepsilon}/\sqrt{T}$ is not longer an independent sum. We need more assumptions to get asymptotic results.

GRM: Robust Covariance Matrix

• $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Omega}$ is unknown. It has Tx$(T$+1$)/2$ elements to estimate. Too many! A solution? Be explicit about (**A3'**): we model $\boldsymbol{\Sigma}$.

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

• We need to estimate $\text{Var}_T[\mathbf{b}|\mathbf{X}] = (\mathbf{X'X})^{-1}\,\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\,(\mathbf{X'X})^{-1}$

• It is important to notice a distinction between estimating

$\boldsymbol{\Sigma}$, a $(T$x$T)$ matrix ⇒ difficult with T observations.

& estimating

$\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X} = \sum_{j=1}^{T}\sum_{i=1}^{T}\sigma_{ij}\,\boldsymbol{x}_i\,\boldsymbol{x}_j{}'$, a $(k$x$k)$ matrix ⇒ easier!

GR Model: Robust Covariance Matrix

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

• We will estimate $\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X} = \sum_{j=1}^{T}\sum_{i=1}^{T}\sigma_{ij}\,\boldsymbol{x}_i\,\boldsymbol{x}_j{}'$, a $(k$x$k)$ matrix. That is, we are estimating [kx$(k+1)/2$ elements.

• This distinction is very important in modern applied econometrics:

– The White estimator

– The Newey-West estimator

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

GR Model: $\mathbf{X'\Sigma X}$

• Q: How does $\mathbf{X'\Sigma X}$ look like? Time series intuition.

We look at the simple linear model, with only one regressor (in this case, $\mathbf{x}_i\varepsilon_i$ is just a scalar). Assume $\mathbf{x}_i\varepsilon_i$ is *covariance stationary* (see Lecture 13) with autocovariances γ_j. Then, we derive $\mathbf{X'\Sigma X}$:

$$\mathbf{X'\Sigma X} = \text{Var}[\mathbf{X'}\boldsymbol{\varepsilon}/\sqrt{T}] = \text{Var}[(1/\sqrt{T})\,(x_1\varepsilon_1 + x_2\varepsilon_2 + \ldots + x_T\varepsilon_T)]$$

$$= (1/T)\,[T\gamma_0 + (T-1)(\gamma_1+\gamma_{-1}) + (T-2)(\gamma_2+\gamma_{-2}) + \ldots + 1\,(\gamma_{T-1}+\gamma_{1-T})]$$

$$= \gamma_0 + \frac{1}{T}\sum_{j=1}^{T-1}(T-j)(\gamma_j + \gamma_{-j})$$

$$= \sum_{j=-T+1}^{T-1}\gamma_j - \frac{1}{T}\sum_{j=1}^{T-1} j(\gamma_j + \gamma_{-j})$$

where γ_j is the autocovariance of $\mathbf{x}_i\varepsilon_i$ at lag j ($\gamma_0 = \sigma^2$ = variance of $\mathbf{x}_i\varepsilon_i$).

GR Model: $\mathbf{X'\Sigma X}$

Under some conditions (autocovariances are *"l-summable"*, so $\Sigma_j\, j|\gamma_j|<\infty$), then

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

<u>Note</u>: In the frequency domain, we define the spectrum of $\mathbf{x'}$e at frequency ω as:

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

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

Covariance Matrix: The White Estimator

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

$$\mathbf{Q}^* = (1/T)\, \mathbf{X}'\boldsymbol{\Sigma}\mathbf{X} = (1/T) \textstyle\sum_{i=1}^{T} \sigma_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$$

where

$$\mathbf{Q}^* = \mathbf{X}'\boldsymbol{\Sigma}\mathbf{X} = \begin{bmatrix} \sum_{i=1}^{T} \boldsymbol{x}_{1i}^2 \sigma_i^2 & \cdots & \sum_{i=1}^{T} \boldsymbol{x}_{1i}\boldsymbol{x}_{ki}\sigma_i^2 \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{T} \boldsymbol{x}_{ki}\boldsymbol{x}_{1i}\sigma_i^2 & \cdots & \sum_{i=1}^{T} \boldsymbol{x}_{ki}^2 \sigma_i^2 \end{bmatrix} = \textstyle\sum_{i=1}^{T} \sigma_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$$

• The OLS residuals, $\mathbf{e}$, are consistent estimators of $\boldsymbol{\varepsilon}$. This suggests using e_i^2 to estimate σ_i^2. That is,

we estimate $\quad \mathbf{Q}^* = (1/T) \sum_{i=1}^{T} \sigma_i^2\, \boldsymbol{x}_i \boldsymbol{x}_i'$

with $\quad \mathbf{S_0} = (1/T) \sum_{i=1}^{T} e_i^2\, \boldsymbol{x}_i \boldsymbol{x}_i'$

Covariance Matrix: The White Estimator

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

• Sketch of proof.

Suppose we observe ε_i. Then, each element of $\mathbf{Q}^*$ would be equal to

$$E[\varepsilon_i^2\, \boldsymbol{x}_i \boldsymbol{x}_i' | \boldsymbol{x}_i].$$

Then, by LLN $\quad \text{plim } (1/T) \sum_{i=1}^{T} \sigma_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' = \text{plim } (1/T) \sum_{i=1}^{T} \varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$

Q: Can we replace ε_i^2 by e_i^2? Yes, since the residuals $\mathbf{e}$ are consistent.

Then, the estimated HC variance is:

$$\text{Est. Var}_T[\mathbf{b}|\mathbf{X}] = (1/T)\, (\mathbf{X'X}/T)^{-1}\, [\textstyle\sum_{i=1}^{T} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'/T]\, (\mathbf{X'X}/T)^{-1}$$

Covariance Matrix: The White Estimator

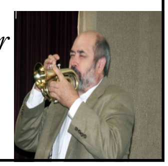
• Note that (**A3**) was not specified. That is, the White estimator is *robust* to a potential misspecifications of heteroscedasticity in (**A3**).

• 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"):

$$\text{HC0} = (\mathbf{X'X})^{-1}\, \mathbf{X'}\, \text{Diag}[e_i^2]\, \mathbf{X}\, (\mathbf{X'X})^{-1}$$

<u>Note</u>: The HC estimator is also called the *sandwich estimator* or the *White estimator* (also known as *Eiker-Huber-White estimator).*



Halbert White (1950-2012, USA)

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:

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

where $h_{ii} = \boldsymbol{x}_i\, (\mathbf{X'X})^{-1}\, \boldsymbol{x}_i'$.

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, *t*-tests and *F*-tests are asymptotically valid.

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

(5) The HC SE's can be larger or smaller than the OLS SE's (in general, HC SE's are larger when positively correlated to $\mathbf{x}_i$ or $\mathbf{x}_i^2$). 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:

```
# White SE in R
White_f <- function(y,X,b) {
T <- length(y); k <- length(b);
yhat <- X%*%b
e <- y-yhat
hhat <- t(X)*as.vector(t(e))
G <- matrix(0,k,k)
 za <- hhat[,1:k]%*%t(hhat[,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) }
```

The White Estimator: Application 1 – IBM

Example: We estimate t-values using OLS and White SE, for the 3 factor F-F model for IBM returns:

$$(r_{i=IBM,t} - r_f) = \beta_0 + \beta_1\,(r_{m,t} - r_f) + \beta_2\,SMB_t + \beta_3\,HML_t + \varepsilon_t$$

```
fit_ibm_ff3 <- lm(ibm_x ~ Mkt_RF + SMB + HML)      # OLS Regression with lm
b_i <-fit_ibm_ff3$coefficients                     # Extract OLS coefficients from fit_ibm
SE_OLS <- sqrt(diag(vcov(fit_ibm_ff3)))            # Extract OLS SE from fit_ibm
t_OLS <- b_i/SE_OLS                                # Calculate  OLS t-values

> b_i
 (Intercept)      Mkt_RF         SMB          HML
-0.005191356  0.910379487 -0.221385575 -0.139179020
> SE_OLS
(Intercept)      Mkt_RF         SMB          HML
0.002482305 0.056784474 0.084213761 0.084060299
> t_OLS
(Intercept)     Mkt_RF        SMB        HML
  -2.091345   16.032190   -2.628853   -1.655705
```

The White Estimator: Application 1 – IBM

Example (continuation):

```
> library(sandwich)
White <- vcovHC(fit_ibm_ff3, type = "HC0")
SE_White <- sqrt(diag(White))                          # White SE HC0
t_White <- b_i/SE_White

> SE_White
(Intercept)      Mkt_RF        SMB        HML
0.002505978 0.062481080 0.105645459 0.096087035
> t_White
(Intercept)      Mkt_RF        SMB        HML
  -2.071589   14.570482   -2.095552   -1.448468

> White <- vcovHC(fit_ibm_ff3, type = "HC3")            # White SE HC3 (refinement)
> SE_White <- sqrt(diag(White))# White SE HC0
> t_White <- b_i/SE_White
> SE_White
(Intercept)      Mkt_RF        SMB        HML
0.002533461 0.063818378 0.108316056 0.098800721
> t_White
(Intercept)      Mkt_RF        SMB        HML
  -2.049116   14.265162   -2.043885   -1.408684
```

The White Estimator: Application 2 – i_{MX}

Example: We estimate Mexican interest rates (i_{MX}) with a linear model including US interest rates, changes in exchange rates (MXN/USD), Mexican inflation and Mexican GDP growth, using quarterly data 1978:II – 2020:II (T=166):

$$i_{MX,t} = \beta_0 + \beta_1\, i_{US,t} + \beta_2\, e_t + \beta_3\, mx\_I_t + \beta_4\, mx\_y_t + \varepsilon_t$$

```
FMX_da <- read.csv("http://www.bauer.uh.edu/rsusmel/4397/FX_USA_MX.csv", head=TRUE,
sep=",")

us_i <- FMX_da$US_int                        # US short-term interest rates (i_US)
mx_CPI <- FMX_da$MX_CPI                      # Mexican CPI
mx_M1 <- FMX_da$MX_M1                        # Mexican Money Supply (M1)
mx_i <- FMX_da$MX_int                        # Mexican short-term interest rates (i_MX)
mx_GDP <- FMX_da$MX_GDP                      # Mexican GDP
S_mx <- FMX_da$MXN_USD                       # S_t = exchange rates (MXN/USD)
T <- length(mx_CPI)
mx_I <- log(mx_CPI[-1]/mx_CPI[-T])           # Mexican Inflation: Log  changes in CPI
mx_y <- log(mx_GDP[-1]/mx_GDP[-T])           # Mexican growth: Log  changes in GDP
```

The White Estimator: Application 2 – i_{MX}

Example (continuation):

```
mx_mg <- log(mx_M1[-1]/mx_M1[-T])        # Money growth: Log  changes in M1
e_mx <- log(S_mx[-1]/S_mx[-T])           # Log changes in S_t.
us_i_1 <- us_i[-1]/100                   # Adjust sample size.
mx_i_1 <- mx_i[-1]/100
mx_i_0 <- mx_i[-T]/100
fit_i <- lm(mx_i_1 ~ us_i_1 + e_mx + mx_I + mx_y)
> summary(fit_i)

Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept) 0.04022   0.01506    2.671   0.00834 **
us_i_1      0.85886   0.31211    2.752   0.00661 **
e_mx        -0.01064  0.02130   -0.499   0.61812
mx_I        3.34581   0.19439   17.212   < 2e-16 ***
mx_y        -0.49851  0.73717   -0.676   0.49985
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The White Estimator: Application 2 – i_{MX}

Example (continuation):

```
White <- vcovHC(fit_i, type = "HC0")
SE_White <- sqrt(diag(White))# White SE HC0
t_White <- b_i/SE_White

> SE_White
(Intercept)    us_i_1      e_mx       mx_I       mx_y
0.009665759 0.480130221 0.026362820 0.523925226 1.217901733
> t_White
(Intercept)    us_i_1      e_mx       mx_I       mx_y
  4.1613603  1.7888018  -0.4035554  6.3860367  -0.4093221
```

$\Rightarrow$ $i_{US,t}$ not longer significant at 5% level.

```
White3 <- vcovHC(fit_i, type = "HC3")    # Using popular refinement HC3
SE_White3 <- sqrt(diag(White3))          # White SE HC3
t_White <- b_i/SE_White3
> t_White3
(Intercept)    us_i_1      e_mx       mx_I       mx_y
  3.6338983  1.5589936  -0.2117600  5.4554986  -0.3519886
```

$\Rightarrow$ $i_{US,t}$ not longer significant at 10% level

Baltagi and Griffin's Gasoline Data (Greene)

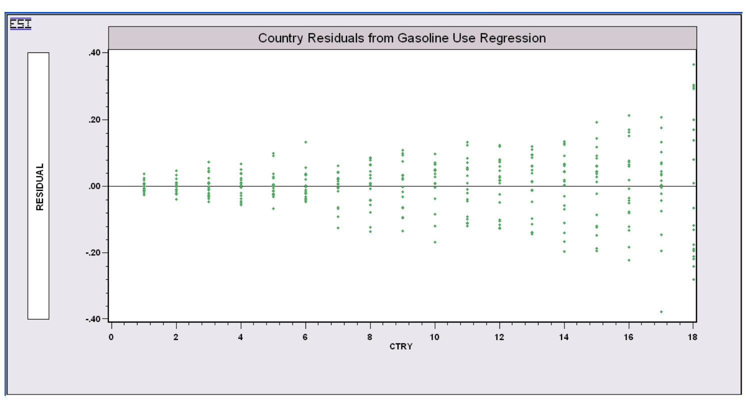
World Gasoline Demand Data, 18 OECD Countries, 19 years
Variables in the file are

COUNTRY = name of country
YEAR = year, 1960-1978
LGASPCAR = log of consumption per car
LINCOMEP = 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., "Gasolne 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

Standard OLS

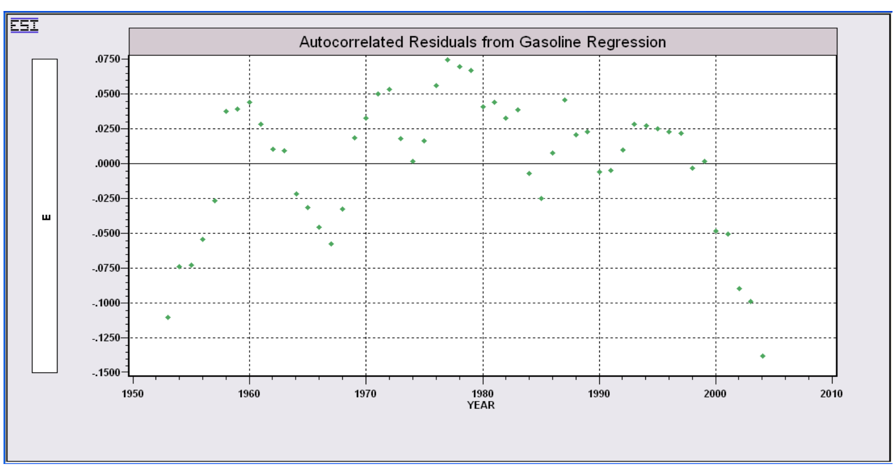
Variable	Coefficient	Standard Error	t-ratio	P[|T|>t]
Constant	2.39132562	.11693429	20.450	.0000
LINCOMEP	.88996166	.03580581	24.855	.0000
LRPMG	-.89179791	.03031474	-29.418	.0000
LCARPCAP	-.76337275	.01860830	-41.023	.0000

White heteroscedasticity robust covariance matrix

Constant	2.39132562	.11794828	20.274	.0000
LINCOMEP	.88996166	.04429158	20.093	.0000
LRPMG	-.89179791	.03890922	-22.920	.0000
LCARPCAP	-.76337275	.02152888	-35.458	.0000

Autocorrelated Residuals: Gasoline Demand

$$\log G = \beta_1 + \beta_2 \log Pg + \beta_3 \log Y + \beta_4 \log Pnc + \beta_5 \log Puc + \varepsilon$$



Newey-West Estimator

• Now, we also have autocorrelation. We need to estimate

$$\mathbf{Q}^* = (1/T)\ \mathbf{X'\Sigma X}\ = (1/T) \sum_{j=1}^{T} \sum_{i=1}^{T} \sigma_{ij}\, \boldsymbol{x}_i\, \boldsymbol{x}_j{}'$$

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

$$\Rightarrow \text{natural estimator of } \mathbf{Q}^*\text{: } (1/T) \sum_{j=1}^{T} \sum_{i=1}^{T} \boldsymbol{x}_i\, e_i\, e_j\, \boldsymbol{x}_j{}'$$

Using time series notation, estimator of $\mathbf{Q}^*$: $\sum_{t=1}^{T} \sum_{s=1}^{T} \boldsymbol{x}_t\, e_t\, e_s\, \boldsymbol{x}_s{}'$

• That is, we have a sum of the estimated autocovariances of $\mathbf{x}_t\varepsilon_t$, $\mathbf{\Gamma}_j$:

$$\Gamma_T(j) = \sum_{j=-(T-1)}^{T-1} E[x_t \varepsilon_t \varepsilon_{t-j} x_{t-j}{}']$$

Whitney Newey, USA

Kenneth D. West, USA

Newey-West Estimator

• Natural estimator of $\mathbf{Q}^*$: $\mathbf{S}_T = (1/T) \sum_{t=1}^{T} \sum_{s=1}^{T} \boldsymbol{x}_t e_t\ e_s \boldsymbol{x}_s{}'$

Note: If $\boldsymbol{x}_t \varepsilon_t$ are serially uncorrelated, the autocovariances vanish. We are left with the White estimator.

Under some conditions (autocovariances are *"l-summable"),* then

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

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

• We can estimate $\mathbf{Q}^*$ in two ways:

(1) parametrically, assuming a model to calculate γ_j.

(2) non-parametrically, using kernel estimation.

Note: (1) needs a specification of (**A3'**); while (2) does not.

Newey-West Estimator

• The parametric estimation uses an ARMA model – say, an AR(2) – to calculate γ_j.

• The non-parametric estimation uses:

$$S_T = \sum_{j=L}^{L} k(j)\hat{\Gamma}_T(j), \qquad \text{where } \hat{\Gamma}_T(j) = \frac{1}{T}\sum_{t=j+1}^{j} x_t e_t x_{t-j} e_{t-j} = \hat{\Gamma}_T(-j) \quad (j \geq 0).$$

• 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, $\mathbf{S}_T$, needs to be psd.

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

Newey-West Estimator

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

<u>**Issue 1**</u>: This sum has T^2 terms. It is difficult to get convergence.

<u>Solution</u>: 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 $\mathbf{\Gamma}_j$'s).

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

$$Q_T^* = \sum_{j=-L(T)}^{L(T)} \Gamma_T(j) \xrightarrow{p} Q^*$$

• Replacing $\mathbf{\Gamma}$(j) by its estimate, we get $\mathbf{S}_T$, which would be consistent for $\mathbf{Q}^*$ provided that $L(T)$ does not grow too fast with T.

Newey-West Estimator

• **Issue 2 (& 3)**: $\mathbf{S}_T$ 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 $\mathbf{Q}^*$:

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

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 t 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 $x \in$ R,

– $\int |k(x)|\, dx < \infty$;

– $k(.)$ is continuous at 0 and at all but a finite number of other points in R, and

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

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:

$\mathbf{S_0} = (1/T) \sum_{i=1}^{T} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$ – the White estimator.

(2) Add the Autocorrelation Component

$\mathbf{S}_T = \mathbf{S}_0 + (1/T)\sum_{l=1}^{L} k(l) \sum_{t=l+1}^{T} (\mathbf{x}_{t-l} e_{t-l} e_t \mathbf{x}_t' + \mathbf{x}_t e_t e_{t-l} \mathbf{x}_{t-l}')$

where

$k(\frac{j}{L(T)}) = \frac{L+1-|j|}{L+1}$ –The *Bartlett kernel*

$\Rightarrow$ linearly decaying weights.

Then,

Est. Var[**b**] $= (1/T)\, (\mathbf{X'X}/T)^{-1}\, \mathbf{S}_T\, (\mathbf{X'X}/T)^{-1}$ –NW's HAC Var.

• Under suitable conditions, as L, $T \rightarrow \infty$, and $L/T \rightarrow 0$, $\mathbf{S}_T \rightarrow \mathbf{Q}^*$. Asymptotic inferences on $\boldsymbol{\beta}$, based on OLS **b**, can be done with *t-test* and *Wald tests* using N(0,1) and χ^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.):

$$S_T^{WP} = 2\pi \sum_{j=1}^{T-1} K_T(2\pi j/T) I_{xeex}(2\pi j/T)$$

where $K_T = T^{-1} \sum_{u=0}^{T-1} K_T(u/L) e^{iwu}$ and I_{xe} is the periodogram of $x_t e_t$ at frequency ω:

$$I_{xeex}(\omega) = d_{xe}(\omega) \overline{d_{xe}(\omega)}', \quad \text{where } d_{xe}(\omega) = 2\pi \sum_{t=1}^{T} (x_t e_t) e^{-i\omega t}$$

• Under suitable conditions, as L & $T \rightarrow \infty$ and $L/T \rightarrow 0$,

$\mathbf{S}^{WP}{}_T \rightarrow \mathbf{Q}^*$.

NW Estimator: Kernel Choice

• Other kernels, $k_L(l)$, besides the Bartlett kernel, can be used:

- *Parzen kernel* –Gallant (1987).

$$
\begin{aligned}
k_L(l) &= 1 - 6l^2 + 6\,|l|^3 && \text{for} \quad 0 \le |l| \le 1/2 \\
&= 2(1 - |l|^3) && \text{for} \quad 0 \le |l| \le 1/2 \\
&= 0 && \text{otherwise}
\end{aligned}
$$

- *Quadratic spectral* (QS) kernel –Andrews (1991):

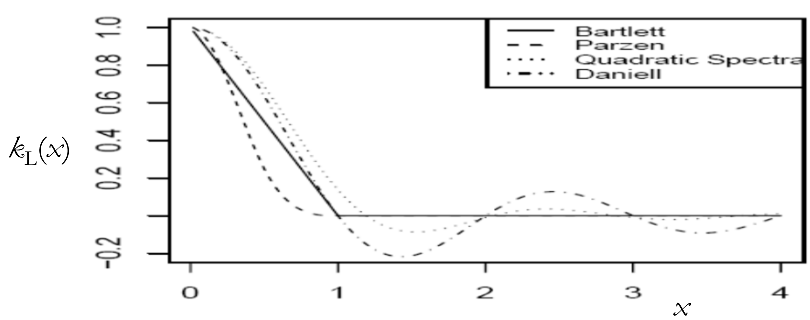
$$k_L(l) = 25/(12\pi^2 l^2)[\sin(6\pi\, l/5)/(6\pi\, l) - \cos(6\pi\, l/5)]$$

- *Daniell kernel* –Ng and Perron (1996):

$$k_L(l) = \sin(\pi\, l)/(\pi\, l)$$

• 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.)

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:

```
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):T] %*% 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: Application 1 – IBM

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

```
> t_OLS
(Intercept)      Mkt_RF        SMB        HML
  -2.091345   16.032190  -2.628853  -1.655705
```
⟹ with SE_OLS: SMB significant at 1% level

```
NW <- NeweyWest(fit_ibm_ff3, lag = 4, prewhite = FALSE)
SE_NW <- diag(sqrt(abs(NW)))
> t_NW <- b_i/SE_NW
> SE_NW
(Intercept)      Mkt_RF        SMB        HML
0.002527425 0.069918706 0.114355320 0.104112705
> t_NW
(Intercept)      Mkt_RF        SMB        HML
  -2.054010   13.020543  -1.935945  -1.336811
```
⟹ SMB close to significant at 5% level

- If we add more lags in the NW function(lag = 8)

```
NW <- NeweyWest(fit_ibm_ff3, lag = 8, prewhite = FALSE)
SE_NW <- diag(sqrt(abs(NW)))
t_NW <- b_i/SE_NW
> t_NW
(Intercept)      Mkt_RF        SMB        HML
  -2.033648   12.779060  -1.895993  -1.312649
```
⟹ not very different results.

NW Estimator: Application 2 – i_{MX}

Example: Mexican short-term interest rates

```
NW <- NeweyWest(fit_i, lag = 4, prewhite = FALSE)
SE_NW <- diag(sqrt(abs(NW)))
t_NW <- b_i/SE_NW
> SE_NW
(Intercept)     us_i_1       e_mx       mx_I       mx_y
 0.01107069 0.55810758 0.01472961 0.51675771 0.93960295
> t_NW
(Intercept)     us_i_1       e_mx       mx_I       mx_y
  3.6332593  1.5388750 -0.7222770  6.4746121 -0.5305582
```
⟹ $i_{US,t}$ not longer significant at 10% level

- If we add more lags in the text (lag = 8)

```
NW <- NeweyWest(fit_i, lag = 8, prewhite = FALSE)
SE_NW <- diag(sqrt(abs(NW)))
t_NW <- b_i/SE_NW
> t_NW
(Intercept)     us_i_1       e_mx       mx_I       mx_y
  3.0174983  1.4318654 -0.8279016  6.5897816 -0.5825521
```
⟹ similar results.

NW Estimator: Remarks

• Parametric estimators of $\mathbf{Q}^*$ 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 $\mathbf{Q}^*$ that are not consistent, but with better small sample properties. See Kiefer, Vogelsang and Bunzel (2000).

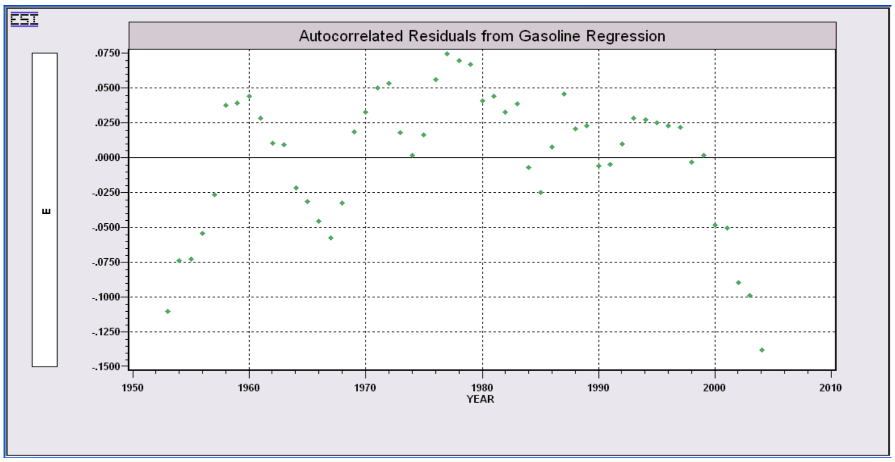
• The SE based on these inconsistent estimators of $\mathbf{Q}^*$ that are used for testing are referred as Heteroskedasticity-Autocorrelation Robust (HAR) SE.

• 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

$$\log G = \beta_1 + \beta_2 \log Pg + \beta_3 \log Y + \beta_4 \log Pnc + \beta_5 \log Puc + \varepsilon$$



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	.10330	.194	.8471

Variable	Coefficient	Standard Error	t-ratio	P[|T|>t]
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	.14176	.141	.8882

Generalized Least Squares (GLS)

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

Note: $\boldsymbol{\Omega}$ is symmetric $\Rightarrow$ exists $\mathbf{T} \ni\ \mathbf{T}'\mathbf{T} = \boldsymbol{\Omega}$ $\Rightarrow \mathbf{T}'^{-1}\,\boldsymbol{\Omega}\,\mathbf{T}^{-1} = \mathbf{I}$

• We transform the linear model in (**A1**) using $\mathbf{P} = \boldsymbol{\Omega}^{-1/2}$.

$$\mathbf{P} = \boldsymbol{\Omega}^{-1/2} \qquad \Rightarrow \mathbf{P}'\mathbf{P} = \boldsymbol{\Omega}^{-1}$$
$$\mathbf{P}\boldsymbol{y} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} \text{ or}$$
$$\boldsymbol{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*.$$
$$E[\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}|\mathbf{X}^*] = \mathbf{P}\,E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}^*]\,\mathbf{P}' = \mathbf{P}\,E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\,\mathbf{P}' = \sigma^2\,\mathbf{P}\,\boldsymbol{\Omega}\,\mathbf{P}'$$
$$= \sigma^2\,\boldsymbol{\Omega}^{-1/2}\,\boldsymbol{\Omega}\,\boldsymbol{\Omega}^{-1/2} = \sigma^2\,\mathbf{I}_T \qquad \Rightarrow \text{back to } (\mathbf{A3})$$

Generalized Least Squares (GLS)

• The transformed model is homoscedastic:

$$\text{Var}[\boldsymbol{\varepsilon}^*|\mathbf{X}^*] = E[\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}|\mathbf{X}^*] = \mathbf{P}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}^*]\mathbf{P}' = \sigma^2\,\mathbf{P\Omega P}' = \sigma^2\,\mathbf{I}_T$$

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

• Key assumption: $\boldsymbol{\Omega}$ is known, and, thus, $\mathbf{P}$ is also known; otherwise we cannot transformed the model.

• Q: Is $\boldsymbol{\Omega}$ known?

Alexander C. Aitken (1895 –1967, NZ)

Generalized Least Squares (GLS)

• **Aitken Theorem** (1935): The *Generalized Least Squares* estimator.

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} \quad \text{or}$$
$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*.$$
$$E[\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}|\mathbf{X}^*] = \sigma^2\mathbf{I}_T$$

We can use OLS in the transformed model. It satisfies G-M theorem.

Thus, the GLS estimator is:

$$\mathbf{b}_{GLS} = \mathbf{b}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\boldsymbol{y}^* = (\mathbf{X}'\mathbf{P}'\mathbf{PX})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\boldsymbol{y}$$
$$= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}$$

<u>Note I</u>: $\mathbf{b}_{GLS} \neq \mathbf{b}$. $\mathbf{b}_{GLS}$ is BLUE by construction, $\mathbf{b}$ is not.

<u>Note II</u>: 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:

$$\mathbf{b}_{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y} = \boldsymbol{\beta} + (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}$$
$$E[\mathbf{b}_{GLS}|\mathbf{X}] = \boldsymbol{\beta}$$

• Efficient Variance

$$\text{Var}[\mathbf{b}_{GLS}|\mathbf{X}] = E[(\mathbf{b}_{GLS} - \boldsymbol{\beta})(\mathbf{b}_{GLS} - \boldsymbol{\beta})'|\mathbf{X}]$$
$$= E[(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}'\boldsymbol{\Omega}^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}|\mathbf{X}]$$
$$= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\,\boldsymbol{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$
$$= \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$

<u>Note</u>: $\mathbf{b}_{GLS}$ is BLUE. This "best" variance can be derived from

$$\text{Var}[\mathbf{b}_{GLS}|\mathbf{X}] = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$

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 χ^2 distributions.

Consistency (Green)

Use Mean Square

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{\sigma^2}{n}\left(\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}}{n}\right)^{-1} \to \mathbf{0}\ ?$$

Requires to be $\left(\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}}{n}\right)$ "well behaved"

Either converge to a constant matrix or diverge.

Heteroscedasticity case:

$$\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega_{ii}}\mathbf{x}_i\mathbf{x}_i'$$

Autocorrelation case:

$$\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}}{n} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\omega_{ij}}\mathbf{x}_i\mathbf{x}_j'.\ n^2 \text{ terms. Convergence is unclear.}$$

Consistency – Autocorrelation case (Green)

$$\frac{X'\Omega^{-1}X}{T} = \frac{1}{T}\sum_{j=1}^{T}\sum_{i=1}^{T}\frac{1}{\omega_{ij}}x_i x_j' = \frac{\sigma_0^{\,2}}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\rho_{t-s}$$

- If the $\{X_t\}$ were uncorrelated –i.e., ρ_k=0–, then Var[$\mathbf{b}_{GLS}$|$\mathbf{X}$] → 0.

- We need to impose restrictions on the dependence among the X_t's. Usually, we require that the autocorrelation, ρ_k, gets weaker as *t-s* grows (and the double sum becomes finite).

Asymptotic Normality (Green)

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}\left(\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}}{n}\right)^{-1}\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}$$

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

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

$\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon} \rightarrow$ a mean to which we can apply the central limit theorem?

Heteroscedasticity case?

$\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i}{\sqrt{\omega_i}}\left(\frac{\varepsilon_i}{\sqrt{\omega_i}}\right)$. Var$\left(\frac{\varepsilon_i}{\sqrt{\omega_i}}\right) = \sigma^2$, $\frac{\mathbf{x}_i}{\sqrt{\omega_i}}$ is just data.

Apply Lindeberg-Feller. (Or assuming $\mathbf{x}_i / \sqrt{\omega_i}$ 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

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\Omega}^{ij}\,\mathbf{x}_i\mathbf{x}_j\varepsilon_i\varepsilon_i$$

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

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

• *Intuition*: $\{Z_t\}$ 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*."

Time Series: Stationarity

• Consider the joint probability distribution of the collection of RVs:

$$F(y_{t_1}, y_{t_2}, \ldots, y_{t_T}) = F(Y_{t_1} \le y_{t_1}, Y_{t_2} \le y_{t_2}, \ldots, Y_{t_T} \le y_{t_T})$$

To do statistical analysis with dependent observations, we need some extra assumptions. We need some form of invariance on the structure of the time series.

If the distribution F is changing with every observation, estimation and inference become very difficult.

• Stationarity is an invariant property: the statistical characteristics of the time series do not change over time.

• There different definitions of stationarity, they differ in how strong is the invariance of the distribution over time.

Time Series: Stationarity

• We say that a process is stationary of

1st order if $F(y_{t_1}) = F(y_{t_{1+k}})$ for any t_1, k

2nd order if $F(y_{t_1}, y_{t_2}) = F(y_{t_{1+k}}, y_{t_{2+k}})$ for any t_1, t_2, k

Nth-order if $F(y_{t_1}, \ldots, y_{t_T}) = F(y_{t_{1+k}}, \ldots, y_{t_{T+k}})$ for any t_1, ..., t_T, k

• *Nth-order* stationarity is a strong assumption (& difficult to verify in practice). *2nd order* stationarity is weaker: only consider mean and covariance (easier to verify in practice).

• Moments describe a distribution. We calculate moments as usual:

$$\mathrm{E}[Y_t] = \mu$$

$$\mathrm{Var}(Y_t) = \sigma^2 = E[(Y_t - \mu)^2]$$

$$\mathrm{Cov}(Y_{t_1}, Y_{t_2}) = E[(Y_{t_1} - \mu)(Y_{t_2} - \mu)] = \gamma(t_1 - t_2)$$

Time Series: Stationarity & Autocovariances

• Stationarity requires the moments to be constant.

Terminology:
$\text{Cov}(Y_{t_1}, Y_{t_2}) = \gamma(t_1 - t_2)$ is called the *auto-covariance function.*

Notes: $\gamma(t_1 - t_2)$ is a function of $k = t_1 - t_2$.
$\gamma(0)$ is the variance.

• The autocovariance function is symmetric. That is,

$$\gamma(t_1 - t_2) = \text{Cov}(Y_{t_1}, Y_{t_2}) = \text{Cov}(Y_{t_2}, Y_{t_1}) = \gamma(t_2 - t_1)$$

Remark: The autocovariance measures the (linear) dependence between the Y_t's separated by k periods.

Time Series: Stationarity & Autocorrelations

• From the autocovariances, we derive the autocorrelations:

$$\text{Corr}(Y_{t_1}, Y_{t_2}) = \rho(Y_{t_1}, Y_{t_2}) = \frac{\gamma(t_1 - t_2)}{\sigma_{t_1}\sigma_{t_2}} = \frac{\gamma(t_1 - t_2)}{\gamma(0)}$$

the last step takes assumes: $\sigma_{t_1} = \sigma_{t_2} = \sqrt{\gamma(0)}$

• $\text{Corr}(Y_{t_1}, Y_{t_2}) = \rho(Y_{t_1}, Y_{t_2})$ is called the *auto-correlation function* (ACF), – think of it as a function of $k = t_1 - t_2$. The ACF is also symmetric.

• Unlike autocovoriances, autocorrelations are not unit dependent. It is easier to compare dependencies across different time series.

• Stationarity requires all these moments to be independent of time. If the moments are time dependent, we say the series is *non-stationary.*

Time Series: Stationarity & Constant Moments

• For strictly stationary process (constant moments), we need:

$$\mu_t = \mu$$
$$\sigma_t = \sigma$$

because $F(y_{t_1}) = F(y_{t_{1+k}}) \quad \Rightarrow \quad \mu_{t_1} = \mu_{t_{1+k}} = \mu$

$$\sigma_{t_1} = \sigma_{t_{1+k}} = \sigma$$

Then,

$$F(y_{t_1}, y_{t_2}) = F(y_{t_{1+k}}, y_{t_{2+k}}) \Rightarrow \text{Cov}(y_{t_1}, y_{t_2}) = \text{Cov}(y_{t_{1+k}}, y_{t_{2+k}})$$

$$\Rightarrow \rho(t_1, t_2) = \rho(t_1 + k, t_2 + k)$$

Let $t_1 = t - k$ & $t_2 = t$

$$\Rightarrow \rho(t_1, t_2) = \rho(t - k, t) = \rho(t, t - k) = \rho(k) = \rho_k$$

The correlation between any two RVs depends on the time difference. Given the symmetry, we have $\rho(k) = \rho(-k)$.

Time Series: Weak Stationary

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

That is, Z_t is covariance stationary if:

$E(Z_t) =$ constant
$Var(Z_t) =$ constant
$Cov(Z_{t_1}, Z_{t_2}) = E[(Z_{t_1} - \mu_{t_1})(Z_{t_2} - \mu_{t_2})] = \gamma(t_1 - t_2) = f(t_1 - t_2)$

Remark: Covariance stationarity is only concerned with the covariance of a process, only the mean, variance and covariance are time-invariant. *N^{th}-order* stationarity is stronger and assumes that the whole distribution is invariant over time.

Time Series: Stationarity – Example

Example: Assume $\varepsilon_t \sim WN(0, \sigma^2)$.

$$y_t = \phi\, y_{t-1} + \varepsilon_t. \qquad \text{(AR(1) process)}$$

- **Mean**

Taking expectations on both side:

$$E[y_t] = \phi\, E[y_{t-1}] + E[\varepsilon_t]$$
$$\mu = \phi\, \mu + 0$$
$$E[y_t] = \mu = 0 \qquad \text{(assuming } \phi \neq 1\text{)}$$

- **Variance**

Applying the variance on both side:

$$Var[y_t] = \gamma(0) = \phi^2\, Var[y_{t-1}] + Var[\varepsilon_t]$$
$$\gamma(0) = \sigma^2/(1 - \phi^2) \qquad \text{(assuming } |\phi| < 1\text{)}$$

Time Series: Stationarity – Example

Example (continuation): $y_t = \phi\, y_{t-1} + \varepsilon_t$ (AR(1) process)

- **Covariance**

$$\gamma(1) = Cov[y_t, y_{t-1}] = E[y_t\, y_{t-1}] = E[(\phi\, y_{t-1} + \varepsilon_t)\, y_{t-1}]$$
$$= \phi\, E[y_{t-1}^2] = \phi\, Var[y_{t-1}^2] = \phi\, \gamma(0)$$
$$= \phi\, [\sigma^2/(1 - \phi^2)]$$

$$\gamma(2) = Cov[y_t, y_{t-2}] = E[y_t\, y_{t-2}] = E[(\phi\, y_{t-1} + \varepsilon_t)\, y_{t-2}]$$
$$= \phi\, E[y_t\, y_{t-1}] = \phi\, Cov[y_t, y_{t-1}] = \phi^2\, \gamma(0)$$
$$= \phi^2\, [\sigma^2/(1 - \phi^2)]$$

$\vdots$

$$\gamma(k) = Cov[y_t, y_{t-k}] = \phi^k\, \gamma(0)$$

$\Rightarrow$ If $|\phi| < 1$, the process is covariance stationary: mean, variance and covariance are constant.

Note: If $|\phi| < 1$, the dependence gets weaker as k increases.

Time Series: Stationarity – Example

Example (continuation): $y_t = \phi\, y_{t-1} + \varepsilon_t$ (AR(1) process)

- **Covariance**

  $\gamma(k) = \text{Cov}[y_t, y_{t-k}] = \phi^k\, \gamma(0)$

Note: From the autocovariance function, we can derive the auto-correlation function:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\phi^k\, \gamma(0)}{\gamma(0)} = \phi^k$$

Time Series: Non-Stationarity – Example

Example: Assume $\varepsilon_t \sim \text{WN}(0, \sigma^2)$.

$y_t = \mu + y_{t-1} + \varepsilon_t$ (Random Walk with drift process)

Doing backward substitution:

$$\begin{aligned} y_t &= \mu + (\mu + y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= 2 * \mu + y_{t-2} + \varepsilon_t + \varepsilon_{t-1} \\ &= 2 * \mu + (\mu + y_{t-3} + \varepsilon_{t-2}) + \varepsilon_t + \varepsilon_{t-1} \\ &= 3 * \mu + y_{t-3} + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} \end{aligned}$$

$$\Rightarrow y_t = \mu\ t + \sum_{j=0}^{t-1} \varepsilon_{t-j} + y_0$$

- **Mean** & **Variance**

  $E[y_t] = \mu\ t + y_0$

  $\text{Var}[y_t] = \gamma(0) = \sum_{j=0}^{t-1} \sigma^2 = \sigma^2\, t$

$\Rightarrow$ the process is non-stationary; that is, moments are time dependent.

Time Series: Stationarity – Remarks

• The main characteristic of time series is that observations are dependent.

• To analyze time series, however, we need to assume that some features of the series are not changing. If we have non-stationary series (say, mean or variance are changing with each observation), it is not possible to make inferences.

• Stationarity is an invariant property: the statistical characteristics of the time series do not vary over time.

• If IBM is weak stationary, then, the returns of IBM may change month to month or year to year, but the average return and the variance in two equal lengths time intervals will be more or less the same.

Time Series: Stationarity – Remarks

• In the long run, say 100-200 years, the stationarity assumption may not be realistic. After all, technological change has affected the return of IBM over the long run. But, in the short-run, stationarity seems likely to hold.

• In general, time series analysis is done under the stationarity assumption.

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 ~ N(0, 1) independent of $\{U_t\}$.

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

$$\bar{Y}_n = \frac{1}{n}\sum_{t=1}^{n} Y_t \xrightarrow{no} E(Y_t) = \frac{1}{2}$$

$$\bar{Y}_n - Z \xrightarrow{p} \frac{1}{2}$$

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*.

Time Series: Ergodicity of mean

• Intuition behind Ergodicity:
We go to a casino to play a game with 20% return, but on average, one gambler out of 100 goes bankrupt. If 100 gamblers play the game, there is a 99% chance of winning and getting a 20% return. This is the *ensemble scenario*. Suppose that gambler 35 is the one that goes bankrupt. Gambler 36 is not affected by the bankruptcy of gamble 35.

Suppose now that instead of 100 gamblers you play the game 100 times. This is the *time series* scenario. You keep winning 20% every day until day 35 where you go bankrupt. There is no day 36 for you.

Result: The probability of success from the group (ensemble scenario) does not apply to one person (time series scenario).

Ergodicity describes a situation where the ensemble scenario outcome applies to the time series scenario.

Time Series: Ergodicity of mean

• With dependent observations, we cannot use the LLN used before. The *ergodicity theorem* plays the role of the LLN with dependent observations.

The formal definition of ergodicity is complex and is seldom used in time series analysis. One consequence of ergodicity is the ergodic theorem, which is extremely useful in time series.

It states conditions under which if Z_t is an ergodic stochastic process then

$$\frac{1}{T}\sum_{t=1} g(Z_t) \overset{a.s}{\rightarrow} \mathrm{E}[g(Z)]$$

for any function $g(.)$. And, for any time shift k

$$\frac{1}{T}\sum_{t=1} g(Z_{t_1+k}, Z_{t_2+k}, \ldots, Z_{t_\tau+k}) \overset{a.s}{\rightarrow} \mathrm{E}[g(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_\tau}))]$$

where a.s. means *almost sure convergence*, a strong form of convergence.

Time Series: Ergodicity of mean

• We want to estimate the mean of the process $\{Z_t\}$, $\mu(Z_t)$. But, we need to distinguishing between *ensemble average* (with m observations) and *time average* (with $T = n$ observations):

- Ensemble Average: $\bar{\bar{z}} = \frac{\sum_{i=1}^{m} z_i}{m}$

- Time Series Average: $\bar{z} = \frac{\sum_{t=1}^{n} z_t}{n}$

Q: Which estimator is the most appropriate?
A: Ensemble Average. But, it is impossible to calculate. We only observe one Z_t, with dependent observations.

• Q: Under which circumstances we can use the time average (with 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.

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.

$$\mathrm{E}[\bar{z}] = \frac{\sum_{t=1}^{T} E[z_t]}{T} = \mu$$

2. Q: Is the variance going to zero as T grows? It depends.

$$\begin{aligned}
\mathrm{var}[\,\bar{z}\,] &= \mathrm{var}[\,(z_1 + z_2 + \cdots + z_T)/T] \\
&= \{\mathrm{var}[\,z_1] + \mathrm{var}[z_2] + \cdots + \mathrm{var}[z_T] \\
&\quad + 2\,\mathrm{cov}[z_1, z_2] + 2\,\mathrm{cov}[z_1, z_3] + \cdots + 2\,\mathrm{cov}[z_1, z_T] \\
&\quad + 2\,\mathrm{cov}[z_2, z_3] + 2\,\mathrm{cov}[z_2, z_4] + \ \ldots + 2\,\mathrm{cov}[z_2, z_T] \\
&\quad + 2\,\mathrm{cov}[z_3, z_4] \ + 2\,\mathrm{cov}[z_3, z_5] + \ldots + 2\,\mathrm{cov}[z_3, z_T] + \\
&\quad \ldots \\
&\quad +2\,\mathrm{cov}[z_{T-1}, z_T]\}/T^2
\end{aligned}$$

Time Series: Ergodicity of mean

• Dividing the RHS by γ_0, and recalling that $\rho_k = \rho_{-k}$, we get:

$$\begin{aligned}
\mathrm{var}[\,\bar{z}\,] &= \frac{\gamma_0}{T^2}\,\{T\rho_0 + 2(T-1)\rho_1 + 2(T-2)\rho_2 + \ \ldots + 2\,\rho_{T-1}\} \\
&= \frac{\gamma_0}{T^2}\{T\rho_0 + 2\sum_{k=1}^{T-1}(T-k)\rho_k\} \\
&= \frac{\gamma_0}{T^2}\sum_{k=T-1}^{T-1}(T-|k|)\rho_k \\
&= \frac{\gamma_0}{T}\sum_{k=T-1}^{T-1}(1-\frac{|k|}{T})\rho_k
\end{aligned}$$

Then,

$$\lim_{T\to\infty} \mathrm{var}[\,\bar{z}\,] = \lim_{T\to\infty} \frac{\gamma_0}{T}\sum_{k=0}(1-\frac{|k|}{T})\rho_k \xrightarrow{?} 0$$

• 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.

Time Series: Ergodicity of mean

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 ρ_k. Conditions weaker than "they are all zero;" but, strong enough to exclude the sequence of identical copies.

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

$$\sum_{t=1}^{T-1} \sum_{k=1}^{T-t} \rho_k \leq \sum_{t=1}^{T-1} \sum_{k=1}^{T-t} |\rho_k| \leq \sum_{t=1}^{T-1} \sum_{k=1}^{\infty} |\rho_k|$$

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).

Time Series: Ergodicity of mean

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

$$\bar{z} \xrightarrow{p} \mathrm{E}[Z_t] = \mu$$

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

$$\rho_k \to 0, \text{ as } k \to \infty$$

• A sufficient condition to ensure ergodicity for second moments is:

$$\sum_{k=T-1}^{T-1} |\rho_k| < \infty$$

A process which is ergodic in the first and second moments is usually referred as *ergodic in the wide sense*.

Time Series: Ergodicity of 2nd moments

• *Ergodicity under Gaussian Distribution*

If $\{Z_t\}$is a stationary Gaussian process,

$$\sum_{k=T-1}^{T-1} |\rho_k| < \infty$$

is sufficient to ensure ergodicity for all moments.

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

Review: GLS

• GRM: Assumptions (**A1**), (**A2**), (**A3'**) & (**A4**) hold. That is,

(**A1**) DGP: $\boldsymbol{y} = \mathbf{X}\,\beta + \boldsymbol{\varepsilon}$ is correctly specified.

(**A2**) $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$

(**A3'**) $Var[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{\Sigma} = \sigma^2\,\boldsymbol{\Omega}$ ($\boldsymbol{\Omega}$ is symmetric $\Rightarrow \mathbf{P}'\mathbf{P} = \boldsymbol{\Omega}^{-1}$)

(**A4**) $\mathbf{X}$ has full column rank –i.e., rank($\mathbf{X}$)=k–, where T ≥ k.

• We transform the linear model in (**A1**) using $\mathbf{P} = \boldsymbol{\Omega}^{-1/2}$.

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} \quad \text{or}$$

$$\boldsymbol{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*.$$

$$E[\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}|\mathbf{X}^*] = \mathbf{P}\,E[\boldsymbol{\varepsilon\varepsilon}'|\mathbf{X}^*]\mathbf{P}' = \mathbf{P}\,E[\boldsymbol{\varepsilon\varepsilon}'|\mathbf{X}]\,\mathbf{P}' = \sigma^2\mathbf{P}\boldsymbol{\Omega}\mathbf{P}'$$

$$= \sigma^2\,\boldsymbol{\Omega}^{-1/2}\,\boldsymbol{\Omega}\,\boldsymbol{\Omega}^{-1/2} = \sigma^2\mathbf{I}_T \qquad \Rightarrow \text{back to } (\mathbf{A3})$$

• The transformed model is homoscedastic: We have do OLS:

$$\mathbf{b}_{GLS} = \mathbf{b}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\,\mathbf{X}^{*\prime}\boldsymbol{y}^*$$

$$= (\mathbf{X}'\mathbf{P}'\mathbf{PX})^{-1}\,\mathbf{X}'\mathbf{P}'\mathbf{P}\boldsymbol{y} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}$$

Review: GLS – Properties

- The GLS estimator is:

$$\mathbf{b}_{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}$$

Note I: $\mathbf{b}_{GLS} \neq \mathbf{b}$. $\mathbf{b}_{GLS}$ is BLUE by construction, $\mathbf{b}$ is not.

- Properties: Unbiased & Consistent

- Efficient Variance

$\mathbf{b}_{GLS}$ is BLUE. The "best" variance can be derived from

$$\text{Var}[\mathbf{b}_{GLS}|\mathbf{X}] = \sigma^2(\mathbf{X}^{*\prime}\,\mathbf{X}^*)^{-1} = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$

Then, the usual OLS variance for $\mathbf{b}$ is biased and inefficient!

Note II: Both unbiased and consistent. In practice, both estimators will be different, but not that different.

GLS: Steps

- Steps for GLS:

Step 1. Find transformation matrix $\mathbf{P} = \boldsymbol{\Omega}^{-1/2}$ (in the case of heteroscedasticity, $\mathbf{P}$ is a diagonal matrix).

Step 2. Transform the model: $\mathbf{X}^* = \mathbf{PX}$ & $\mathbf{y}^* = \mathbf{Py}$.

Step 3. Do GLS; that is, OLS with the transformed variables.

- Key step to do GLS: **Step 1**, getting the transformation matrix:

$$\mathbf{P} = \boldsymbol{\Omega}^{-1/2}.$$

(Weighted) GLS: Pure Heteroscedasticity

- Find the transformation matrix $\mathbf{P} = \mathbf{\Omega}^{-1/2}$.

$$(\mathbf{A3'})\ \ Var[\varepsilon] = \mathbf{\Sigma} = \sigma^2\mathbf{\Omega} = \sigma^2\begin{bmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & 0 \\ 0 & 0 & & 0 \\ 0 & 0 & \dots & \omega_T \end{bmatrix}$$

$$\mathbf{\Omega}^{-1/2} = \mathbf{P} = \begin{bmatrix} 1/\sqrt{\omega_1} & 0 & \dots & 0 \\ 0 & 1/\sqrt{\omega_2} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 1/\sqrt{\omega_T} \end{bmatrix}$$

- Now, transform **y** & **X**:

$$\boldsymbol{y}^* = \mathbf{Py} = \begin{bmatrix} 1/\sqrt{\omega_1} & 0 & \dots & 0 \\ 0 & 1/\sqrt{\omega_2} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 1/\sqrt{\omega_T} \end{bmatrix} * \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} y_1/\sqrt{\omega_1} \\ y_2/\sqrt{\omega_2} \\ \vdots \\ y_T/\sqrt{\omega_T} \end{bmatrix}$$

(Weighted) GLS: Pure Heteroscedasticity

- Each observation of y, y_i, is divided by $\sqrt{\omega_i}$. Similar transformation occurs with **X**:

$$\boldsymbol{X}^* = \mathbf{PX} = \begin{bmatrix} 1/\sqrt{\omega_1} & 0 & \dots & 0 \\ 0 & 1/\sqrt{\omega_2} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 1/\sqrt{\omega_T} \end{bmatrix} * \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ 1 & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{2T} & \cdots & x_{kT} \end{bmatrix} =$$

$$= \begin{bmatrix} 1/\sqrt{\omega_1} & x_{21}/\sqrt{\omega_1} & \dots & x_{k1}/\sqrt{\omega_1} \\ 1/\sqrt{\omega_2} & x_{22}/\sqrt{\omega_2} & \dots & x_{k2}/\sqrt{\omega_2} \\ \vdots & \vdots & \dots & \vdots \\ 1/\sqrt{\omega_T} & x_{2T}/\sqrt{\omega_T} & \dots & x_{kT}/\sqrt{\omega_n} \end{bmatrix}$$

- Now, we can do OLS with the transformed variables:

$$\mathbf{b}_{GLS} = \mathbf{b}^* = (\mathbf{X}^{*\prime}\,\mathbf{X}^*)^{-1}\,\mathbf{X}^{*\prime}\,\mathbf{y}^* = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{\Omega}^{-1}\boldsymbol{y}$$

(Weighted) GLS: Pure Heteroscedasticity

• In the case of heteroscedasticity, GLS is also called *Weighted Least Squares* (WLS): Think of $[\omega_i]^{-1/2}$ as weights. The GLS estimator is:

$$\mathbf{b}_{\text{GLS}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y} = \left(\sum_{i=1}^{T}\frac{1}{\omega_i}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{T}\frac{1}{\omega_i}\boldsymbol{x}_i y_i\right)$$

Observations with lower (bigger) variances –i.e., lower (bigger) ω_i– are given higher (lower) weights in the sums: More precise observations, more weight!

• The GLS variance is given by:

$$\hat{\sigma}^2_{GLS} = \frac{\sum_{i=1}^{T}\left(\frac{y_i - \boldsymbol{x}_i'\,\mathbf{b}_{\text{GLS}}}{\omega_i}\right)^2}{T-K}$$

(Weighted) GLS: Pure Heteroscedasticity

Example: It is common to find that squared market returns (Mkt_RF^2) influence the heteroscedasticity in stock returns. We use DIS returns. Suppose we assume: (**A3'**) σ_i^2 = $(\text{Mkt\_RT}_i)^2$.

Steps for GLS:

1. Find transformation matrix, **P**, with i^{th} diagonal element: $1/\sqrt{\sigma_i^2}$
2. Transform model: Each y_i and x_i is divided ("weighted") by
   σ_i = sqrt[$(\text{Mkt\_RT}_i)^2$].
3. Do GLS, that is, OLS with transformed variables.

```
dis_x <- lr_dis – RF                                      # Disney's excess returns
T <- length(dis_x)
Mkt_RF2 <- Mkt_RF^2                                       # (A3')
y_w <- dis_x/sqrt(Mkt_RF2)                                # transformed y = y*
x0 <- matrix(1,T,1)
xx_w <- cbind(x0, Mkt_RF, SMB, HML)/sqrt(Mkt_RF2)         # transformed X = X*
fit_dis_wls <- lm(y_w ~ xx_w)                             # GLS
```

(Weighted) GLS: Pure Heteroscedasticity

Example (continue):

```
> summary(fit_dis_wls)

Call:
lm(formula = y_w ~ xx_w)

Residuals:
    Min      1Q  Median      3Q     Max
-59.399  -0.891   0.316   1.503  77.434

Coefficients:
                Estimate   Std. Error t value Pr(>|t|)
xx_w           -0.006607   0.001586  -4.165 3.59e-05 ***
xx_wMkt_RF      1.588057   0.334771   4.744 2.66e-06 ***     ⇒ OLS b: 1.26056
xx_wSMB        -0.200423   0.067498  -2.969  0.00311 **      ⇒ OLS b: -0.028993
xx_wHML        -0.042032   0.072821  -0.577  0.56404         ⇒ OLS b: 0.174545
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.984 on 566 degrees of freedom
Multiple R-squared:  0.09078,   Adjusted R-squared:  0.08435
F-statistic: 14.13 on 4 and 566 DF,  p-value: 5.366e-11
```

GLS: First-order Autocorrelation Case

- We assume an AR(1) process for the ε_t:

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + u_t, \qquad u_t\text{: non-autocorrelated error} \sim D(0, \sigma_u^2)$$

Steps for GLS:

1. To find the transformation matrix **P**, we need to derive the implied (**A3'**) based on the AR(1) process for ε_t:

(1) Find diagonal elements of $\mathbf{\Omega}$: $\text{Var}[\varepsilon_t] = \sigma_{ii} = \sigma_\varepsilon^2$

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + u_t \qquad \text{(the autoregressive form)}$$

$$\Rightarrow \quad \text{Var}[\varepsilon_t] = \rho^2\, \text{Var}[\varepsilon_{t-1}] + \text{Var}[u_t] \qquad (\text{Var}[\varepsilon_t] = \text{Var}[\varepsilon_{t-1}] = \sigma_\varepsilon^2)$$

$$\Rightarrow \quad \sigma_\varepsilon^2 = \sigma_u^2/(1-\rho^2) \qquad \text{–we need to assume } |\rho| < 1.$$

(2) Find the off-diagonal elements of $\mathbf{\Omega}$: $\sigma_{ij} = \gamma_{l=j-i}$:

$$\Rightarrow \quad \sigma_{ij} = \gamma_l = \text{Cov}[\varepsilon_i, \varepsilon_j] = E[\varepsilon_i\, \varepsilon_j] \qquad l = j - i$$

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

• Let $\varepsilon_t = \rho\, \varepsilon_{t-1} + u_t$ -AR(1) with u_t = WN error ~ D(0, σ_u^2)

• Then,

$$\begin{aligned} \varepsilon_t &= \rho\, \varepsilon_{t-1} + u_t && \text{(the autoregressive form)} \\ &= \rho\, (\rho\, \varepsilon_{t-2} + u_{t-1}) + u_t && = \ldots \\ &= u_t + \rho\, u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + \ldots \\ &= \textstyle\sum_{j=0}^{t} \rho^j u_{t-j} && \text{(a moving average)} \end{aligned}$$

• $\text{Var}[\varepsilon_t] = \sum_{j=0}^{t} \rho^{2j} Var[u_{t-j}] = \sum_{j=0}^{t} \rho^{2j} \sigma_u^2$

$$= \frac{\sigma_u^2}{(1-\rho^2)} \qquad \text{–we need to assume } |\rho| < 1.$$

• Easier:

$\text{Var}[\varepsilon_t] = \rho^2\, \text{Var}[\varepsilon_{t-1}] + \text{Var}[u_t] \quad \Rightarrow \text{Var}[\varepsilon_t] = \dfrac{\sigma_u^2}{(1-\rho^2)}$

GLS: AR(1) Case – Autocovariances

(2) Find $\sigma_{ij} = \gamma_l$. We call γ_l the *autocovariance* at lag $l = j - i$. It is the autocorrelation between two errors separated in time by *l periods:*

$$\sigma_{ij} = \gamma_l = \text{Cov}[\varepsilon_i, \varepsilon_j] = \text{E}[\varepsilon_i\, \varepsilon_j] \qquad l = j - i$$

$$\begin{aligned} \gamma_1 &= Cov[\varepsilon_t, \varepsilon_{t-1}] = E[(\rho\varepsilon_{t-1} + u_t)\, \varepsilon_{t-1}] \\ &= \rho\, E[\varepsilon_{t-1}\, \varepsilon_{t-1}] + E[u_t\, \varepsilon_{t-1}] \\ &= \rho\, \text{Var}[\varepsilon_{t-1}] = \rho\, \sigma_\varepsilon^2 \\ &= \frac{\rho\sigma_u^2}{(1-\rho^2)} \end{aligned}$$

$$\begin{aligned} \gamma_2 &= Cov[\varepsilon_t, \varepsilon_{t-2}] = E[(\rho\varepsilon_{t-1} + u_t)\, \varepsilon_{t-2}] \\ &= \rho\, E[\varepsilon_{t-1}\, \varepsilon_{t-2}] + E[u_t\, \varepsilon_{t-2}] \\ &= \rho\, Cov[\varepsilon_t, \varepsilon_{t-1}] = \rho\, \gamma_1 \end{aligned}$$

GLS: AR(1) Case – Autocovariances

$$\gamma_2 = \rho\,\gamma_1 = \frac{\rho^2\sigma_u^2}{(1-\rho^2)}$$

$$\begin{aligned}\gamma_3 = Cov[\varepsilon_t, \varepsilon_{t-3}] &= E[(\rho\varepsilon_{t-1} + u_t)\,\varepsilon_{t-3}] \\ &= \rho\, E[\varepsilon_{t-1}\,\varepsilon_{t-3}] + E[u_t\,\varepsilon_{t-3}] \\ &= \rho\, Cov[\varepsilon_t, \varepsilon_{t-2}] = \rho\,\gamma_2 \\ &= \rho^2\gamma_1 = \frac{\rho^3\sigma_u^2}{(1-\rho^2)}\end{aligned}$$

$\vdots$

$$\gamma_l = Cov[\varepsilon_t, \varepsilon_{t-l}] = \rho^{l-1}\,\gamma_1$$

- If we define $\gamma_0 = \sigma_\varepsilon^2 = \sigma_u^2/(1 - \rho^2)$, then

$$\gamma_l = \rho^l\,\gamma_0$$

GLS: AR(1) Case – Autocorrelation Matrix Σ

- Now, we get (**A3'**) $\boldsymbol{\Sigma} = \sigma^2\,\boldsymbol{\Omega}$.

$$\textbf{(A3')}\quad \sigma^2\Omega = \left(\frac{\sigma_u^2}{1-\rho^2}\right)\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix}$$

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

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

GLS: AR(1) Case – Transformed y & X: y* & X*

2. With $\mathbf{P} = \mathbf{\Omega}^{-1/2}$, we transform the data to do GLS.

$$\mathbf{P} = \mathbf{\Omega}^{-1/2} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 \\ -\rho & 1 & 0 & \dots & 0 \\ 0 & -\rho & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & -\rho & 0 \end{bmatrix}$$

:

$$\mathbf{y}^* = \mathbf{P}\,\boldsymbol{y} = \begin{pmatrix} \left(\sqrt{1-\rho^2}\right) y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \dots \\ y_T - \rho_{T-1} \end{pmatrix}$$

$\Rightarrow$ GLS: Transformed $\mathbf{y}^*$.

GLS: AR(1) Case – Transformed y & X: y* & X*

2. Transformed $\mathbf{x}_k$ column (independent variable k) of matrix $\mathbf{X}$ is:

$$\mathbf{P} = \mathbf{\Omega}^{-1/2} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 \\ -\rho & 1 & 0 & \dots & 0 \\ 0 & -\rho & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & -\rho & 0 \end{bmatrix}$$

:

$$\boldsymbol{x}_k^* = \mathbf{P}\,\boldsymbol{x}_k = \begin{pmatrix} \left(\sqrt{1-\rho^2}\right) x_{k1} \\ x_{k2} - \rho\, x_{k1} \\ x_{k3} - \rho\, x_{k2} \\ \dots \\ x_T - \rho\, x_{T-1} \end{pmatrix}$$

$\Rightarrow$ GLS: Transformed $\mathbf{X}^*$.

3. GLS is done with transformed data. In (**A3'**) we assume ρ known.

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 multiply the DGP by ρ and subtract it from it. That is,

$$
\begin{aligned}
y_t &= \boldsymbol{x}_t{}'\boldsymbol{\beta} + \varepsilon_t, \qquad\qquad \varepsilon_t = \rho\varepsilon_{t-1} + u_t \\
\rho y_{t-1} &= \rho \boldsymbol{x}_{t-1}{}'\boldsymbol{\beta} + \rho\varepsilon_{t-1} \\
&\text{-------------------} \\
y_t - \rho y_{t-1} &= (\boldsymbol{x}_t - \rho\boldsymbol{x}_{t-1})'\boldsymbol{\beta} + (\varepsilon_t - \rho\varepsilon_{t-1}) \\
y_t^* &= \boldsymbol{x}_t^*{}'\boldsymbol{\beta} + u_t
\end{aligned}
$$

Now, we have the errors, u_t, which are uncorrelated. We can do OLS with the pseudo differences.

Note: $y_t^* = y_t - \rho y_{t-1}$ & $\boldsymbol{x}_t^* = \boldsymbol{x}_t - \rho\boldsymbol{x}_{t-1}$ are *pseudo differences*.

FGLS: Unknown Ω

• The problem with GLS is that $\boldsymbol{\Omega}$ is unknown. For example, in the AR(1) case, ρ is unknown.

• Solution: Estimate $\boldsymbol{\Omega}$. ⇒ *Feasible GLS* (FGLS).

• In general, there are two approaches for GLS

(1) Two-step, or *Feasible estimation*: - First, estimate $\boldsymbol{\Omega}$ first.
- Second, do GLS.

Similar logic to HAC procedures: We do not need to estimate $\boldsymbol{\Omega}$, difficult with T observations. We estimate $(1/T)\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}$.

– Nice asymptotic properties for FGLS estimator. Not longer BLUE

(2) ML estimation of $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\Omega}$ at the same time (joint estimation of all parameters). With some exceptions, rare in practice.

FGLS: Two-Step Estimation (Green)

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

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

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

• The object is to find a set of parameters such that

$$[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y} - [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} \rightarrow \mathbf{0}$$

FGLS: Asymptotic Details (Green)

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

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

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

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

For the asymptotic properties, we will require that

$$(1/\sqrt{n})\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\varepsilon - (1/n)\mathbf{X}'\boldsymbol{\Omega}^{-1}\varepsilon \rightarrow \mathbf{0}$$

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

FGLS: Specification of Ω

- Ω must be specified first.
- Ω is generally specified (modeled) in terms of a few parameters. Thus, $\Omega = \Omega(\theta)$ for some small parameter vector θ. Then, we need to estimate θ.

Examples:

(1) $\text{Var}[\varepsilon_i|\mathbf{X}] = \sigma^2 f(\boldsymbol{\gamma}'\mathbf{z}_i)$. Variance a function of $\boldsymbol{\gamma}$ and some variable $\mathbf{z}_i$ (say, market volatility, firm size, country dummy, etc). In general, f is an exponential to make sure the variance is positive.

(2) ε_i with AR(1) process. We have already derived $\sigma^2\,\Omega$ as a function of ρ.

Technical note: To achieve full efficiency, we do not need an *efficient* estimate of the parameters in Ω, only a consistent one.

FGLS: Estimation – Steps

- Steps for FGLS:

1. Estimate the model proposed in (**A3'**). Get $\hat{\sigma}_i^2$ & $\hat{\sigma}_{ij}$
2. Find transformation matrix, **P**, using the estimated $\hat{\sigma}_i^2$ & $\hat{\sigma}_{ij}$.
3. Using **P** from Step 2, transform model: $\mathbf{X}^*= \mathbf{PX}$ and $\mathbf{y}^*= \mathbf{Py}$.
4. Do FGLS, that is, OLS with $\mathbf{X}^*$ & $\mathbf{y}^*$.

Example: In the pure heteroscedasticity case (**P** is diagonal):

1. Estimate the model proposed in (**A3'**). Get $\hat{\sigma}_i^2$.
2. Find transformation matrix, **P**, with i^{th} diagonal element: $1/\hat{\sigma}_i$
3. Transform model: Each y_i and x_i is divided ("weighted") by $\hat{\sigma}_i$.
4. Do FGLS, that is, OLS with transformed variables.

FGLS: Estimation – Heteroscedasticity

Example: Last lecture, we found that Mkt_RF^2 and SMB^2 are drivers of the heteroscedasticity in DIS returns: Suppose we assume:
(**A3'**) $\sigma_i^2 = \gamma_0 + \gamma_1 (\text{Mkt\_RT}_i)^2 + \gamma_3 (\text{SMB}_i)^2$

- Steps for FGLS:

1. Use OLS squared residuals to estimate (**A3'**):

```
fit_dis_ff3 <- lm(dis_x ~ Mkt_RF + SMB + HML)
e_dis <- fit_dis_ff3$residuals
e_dis2 <- e_dis^2
fit_dis2 <- lm(e_dis2 ~ Mkt_RF2 + SMB2)
summary(fit_dis2)
var_dis2 <- fit_dis2$fitted          # Estimated variance vector, with elements σ̂i².
```

2. Find transformation matrix, **P**, with i^{th} diagonal element: $1/\hat{\sigma}_i$

```
w_fgls <- sqrt(var_dis2)             # 1/σ̂i
```

3. Transform model: Each y_i and x_i is "weighted" by $1/\hat{\sigma}_i$.

```
y_fw <- dis_x/w_fgls                           # transformed y
xx_fw <- cbind(x0, Mkt_RF, SMB, HML)/w_fgls    # transformed X
```

FGLS: Estimation – Heteroscedasticity

Example (continuation):

4. Do GLS, that is, OLS with transformed variables.

```
fit_dis_fgls <- lm(y_fw ~ xx_fw - 1)
> summary(fit_dis_fgls)

Coefficients:
              Estimate   Std. Error t value Pr(>|t|)
xx_fw         -0.003097  0.002696  -1.149    0.251
xx_fwMkt_RF    1.208067  0.073344  16.471   <2e-16 ***
xx_fwSMB      -0.043761  0.105280  -0.416    0.678
xx_fwHML       0.125125  0.100853   1.241    0.215     ⇒ not longer significant at 10%.
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9998 on 566 degrees of freedom
Multiple R-squared:  0.3413,    Adjusted R-squared:  0.3366
F-statistic: 73.31 on 4 and 566 DF,  p-value: < 2.2e-16
```

FGLS: Estimation – Heteroscedasticity

Example (continuation): Compare OLS and GLS and FGLS results

	b_{OLS}	SE	b_{GLS}	SE	b_{FGLS}	SE
Intercept	0.00417	0.00279	-0.00661	0.00159	-0.00310	0.00270
Mkt_RF	**1.26056**	0.06380	**1.58806**	0.33477	**1.20807**	0.07334
SMB	-0.02899	0.09461	-0.20042	0.06750	-0.04376	0.10528
HML	0.17455	0.09444	-0.04203	0.07282	0.12513	0.10085

• Comments:
- The GLS estimates are quite different than OLS estimates (remember OLS is unbiased and consistent). Very likely the assumed functional form in (**A3'**) was not a good one.
- The FGLS results are similar to the OLS, as expected, if model is OK. FGLS is likely a more precise estimator (HML is not longer significant at 10%.

Harvey's Model of Heteroscedasticity (Green)

• The variance for observation *i* is a function of $\mathbf{z}_i$:

$$\text{Var}[\varepsilon_i|\mathbf{X}] = \sigma^2 \exp(\boldsymbol{\gamma}'\mathbf{z}_i)$$

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 $\boldsymbol{\Sigma} = \sigma^2\,\boldsymbol{\Omega}$

$$\boldsymbol{\Sigma} = \text{diagonal }[\exp(\theta + \boldsymbol{\gamma}'\mathbf{z}_i)],$$

$$\theta = \log(\sigma^2)$$

Once we specify $\boldsymbol{\Omega}$ (and can be estimated), GLS is feasible.

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

• We have already derived $\boldsymbol{\Sigma} = \sigma^2\,\boldsymbol{\Omega}$ for the AR(1) case..

$$\sigma^2\boldsymbol{\Omega} = \left(\frac{\sigma_u^2}{1-\rho^2}\right)\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix}$$

• Now, if we estimate σ_u^2 and ρ, we can do FGLS.

.

Estimated AR(1) Model (Greene)

```
AR(1) Model:      e(t) = rho * e(t-1) + u(t)
Initial value of rho        =         .87566
Maximum iterations          =              1
Method = Prais - Winsten
Iter=  1, SS=         .022, Log-L=    127.593
Final value of Rho     =             .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  b/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|     .05426           .12392         .438   .6615
    LPUC|    -.04028           .06193        -.650   .5154
     RHO|     .95941***        .03949       24.295   .0000
--------+--------------------------------------------------
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           .10330         .194   .8471
```

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

$\{\mathbf{X}' [\boldsymbol{\Omega}(\hat{\theta})]^{-1} \mathbf{X}\}^{-1} \mathbf{X}' [\boldsymbol{\Omega}(\hat{\theta})]^{-1} \mathbf{y}$ to estimate $\boldsymbol{\beta}$.

(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.

Andrew C. Harvey, England

Harvey's Model: Groupwise Heteroscedasticity

• We have a sample, y_{ig}, x_{ig},…, with N groups, each with T_g observations.

Each group variance: $\text{Var}[\varepsilon_{ig}] = \sigma_g^2$

• 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:

$\text{Var}[\varepsilon_{ig}] = \sigma_g^2 \exp(\theta_2 d_2 + \ldots + \theta_N d_N)$

$\text{Var}_1 = \sigma_g^2$ –normalized variance (remember dummy trap!)

$\text{Var}_2 = \sigma_g^2 \exp(\theta_2)$

... 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.$Var_1 = \mathbf{e}_1'\mathbf{e}_1/T_1$ estimates σ_g^2
- Est.$Var_2 = \mathbf{e}_2'\mathbf{e}_2/T_2$ estimates $\sigma_g^2 \exp(\theta_2)$
- Estimator of θ_2 is $\ln[(e_2'e_2/T_2)/(e_1'e_1/T_1)]$
- .... etc.

Step 2. Now, use FGLS –weighted least squares. Keep WLS residuals

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 **Ω**. Very rare case.

• It needs the specification of **Ω** –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'**).

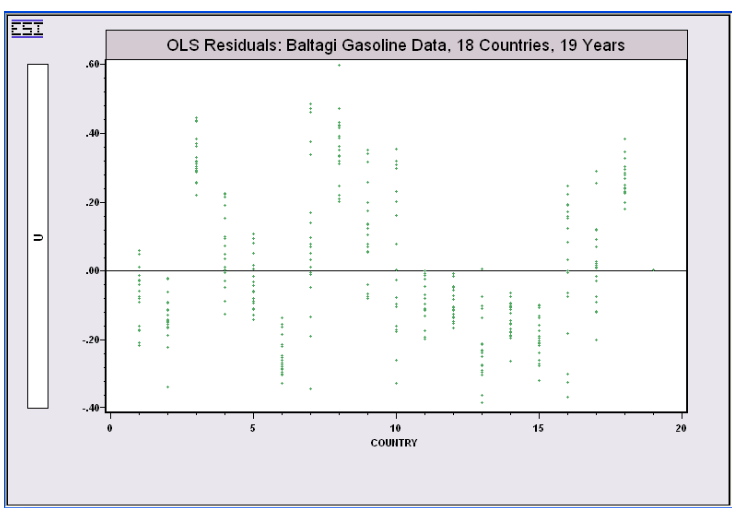
Baltagi and Griffin's Gasoline Data (Greene)

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
LINCOMEP = 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., "Gasolne 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.

Baltagi and Griffin's Gasoline Data (Greene) - ANOVA



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
 LINCOMEP|     .88996166        .03580581    24.855    .0000
 LRPMG   |   -.89179791        .03031474   -29.418    .0000
 LCARPCAP|   -.76337275        .01860830   -41.023    .0000

| White heteroscedasticity robust covariance matrix  |
+----------------------------------------------------+
 Constant|    2.39132562        .11794828    20.274    .0000
 LINCOMEP|     .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 ............
LHS=LGASPCAR Mean                  =        4.29624
             Standard deviation    =         .54891
             Number of observs.    =            342
Model size   Parameters            =              4
             Degrees of freedom    =            338
Residuals    Sum of squares        =       14.90436
Wald   statistic [17 d.f.] =        699.43 (.0000)  (Large)
B/P LM statistic [17 d.f.] =        111.55 (.0000)  (Large)
Cov matrix for b is sigma^2*inv(X'X)(X'WX)inv(X'X)
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
Constant|    2.39133***       .20010        11.951   .0000
LINCOMEP|     .88996***       .07358        12.094   .0000     -6.13943
   LRPMG|    -.89180***       .06119       -14.574   .0000      -.52310
LCARPCAP|    -.76337***       .03030       -25.190   .0000     -9.04180
--------+-------------------------------------------------------------
```

Baltagi and Griffin's Gasoline Data (Greene) - Variance Estimates = log[e(i)'e(i)/T]

Sigma	.48196***	.12281	3.924	.0001	
D1	-2.60677***	.72073	-3.617	.0003	.05556
D2	-1.52919**	.72073	-2.122	.0339	.05556
D3	.47152	.72073	.654	.5130	.05556
D4	-3.15102***	.72073	-4.372	.0000	.05556
D5	-3.26236***	.72073	-4.526	.0000	.05556
D6	-.09099	.72073	-.126	.8995	.05556
D7	-1.88962***	.72073	-2.622	.0087	.05556
D8	.60559	.72073	.840	.4008	.05556
D9	-1.56624**	.72073	-2.173	.0298	.05556
D10	-1.53284**	.72073	-2.127	.0334	.05556
D11	-2.62835***	.72073	-3.647	.0003	.05556
D12	-2.23638***	.72073	-3.103	.0019	.05556
D13	-.77641	.72073	-1.077	.2814	.05556
D14	-1.27341*	.72073	-1.767	.0773	.05556
D15	-.57948	.72073	-.804	.4214	.05556
D16	-1.81723**	.72073	-2.521	.0117	.05556
D17	-2.93529***	.72073	-4.073	.0000	.05556

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

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]
	Ordinary Least Squares			
	Cov matrix for b is sigma^2*inv(X'X)(X'WX)inv(X'X)			
Constant	2.39133***	.20010	11.951	.0000
LINCOMEP	.88996***	.07358	12.094	.0000
LRPMG	-.89180***	.06119	-14.574	.0000
LCARPCAP	-.76337***	.03030	-25.190	.0000
	FGLS - Regression (mean) function			
Constant	1.56909***	**.06744**	23.267	.0000
LINCOMEP	.60853***	**.02097**	29.019	.0000
LRPMG	-.61698***	**.01902**	-32.441	.0000
LCARPCAP	-.66938***	**.01116**	-59.994	.0000

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