The Classical Model

Heteroskedasticity and
Correlations Across Errors

Heteroskedasticity

- Recall Assumption 5 of the CLRM: that all errors have the same variance. That is,

$$\text{Var}(\varepsilon_i) = \sigma^2 \text{ for all } i = 1,2,\ldots,n$$

- **Heteroskedasticity** is a violation of this assumption. It occurs if different observations' errors have different variances. For example,

$$\text{Var}(\varepsilon_i) = \sigma_i^2$$

  - In this case, we say the errors are **heteroskedastic.**
- Because heteroskedasticity violates an assumption of the CLRM, we know that least squares is not BLUE when the errors are heteroskedastic.
- Heteroskedasticity occurs most often in **cross-sectional** data. These are data where observations are all for the same time period (e.g., a particular month, day, or year) but are from different entities (e.g., people, firms, provinces, countries, etc.)

Pure Heteroskedasticity

- There are two basic types of heteroskedasticity (pure & impure)
- **Pure Heteroskedasticity** arises if the model is *correctly specified*, but the errors are heteroskedastic, e.g., the DGP is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

  where $\text{Var}(\varepsilon_i) = \sigma_i^2$
- There are many ways to specify the heteroskedastic variance σ_i^2.
- A very simple specification is **discrete heteroskedasticity**, where the errors are drawn from one of two distributions, a "wide" distribution (with Large variance σ_L^2) or a "narrow" distribution (with Small variance σ_S^2) … **draw a picture.**
- A common specification is to assume that the error variance is **proportional** to a the square of variable *Z* (that may or may not be one of the independent variables). In this case,
- $$\text{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 Z_i^2$$
- and each observation's error is drawn from its own distribution with mean zero and variance $\sigma^2 Z_i^2$. (**draw some pictures**)
  - An example: suppose we're modeling household expenditures on leisure activities (movies, skiing, vacations, etc.). At low levels of household income, there will be less household-to-household variation in leisure spending (in $ terms) than at high levels of household income. That is, the error variance will be proportional to household income (*Z*). This is because poor people have less room in their budget for such variance.

Inefficiency

- Why is OLS inefficient when we have pure heteroskedasticity?
- It is because there is another linear estimator that uses the data better, and can deliver a lower-variance estimated coefficient
- Eg, what if some observations had zero-variance on their errors, but others had positive variance
  - A linear estimator that delivers a lower-variance coefficient is to run OLS on *only* those observations with zero-variance. Trash all the rest of the data

Impure Heteroskedasticity

- **impure heteroskedasticity** can arise if the model is *mis-specified* (e.g., due to an omitted variable) and the specification error induces heteroskedasticity. For example, suppose the DGP is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

but we estimate $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i^*$

where $\varepsilon^*_i = \beta_2 X_{2i} + \varepsilon_i$

and where ε_i is a classical error term.

- Then if X_{2i} itself has a heteroskedastic component (e.g., the value of X_{2i} comes from either a "wide" or "narrow" distribution), then omitting it from the model makes the mis-specified error term ε^*_i behave heteroskedastically.
- Of course the solution here is simple: don't omit X_2 from the model!

Consequences of Heteroskedasticity

- We know that heteroskedasticity violates Assumption 5 of the CLRM, and hence OLS is not BLUE. What more can we say?

OLS estimates remain unbiased

We only need Assumptions 1-3 to show that the OLS estimator is unbiased, hence a violation of Assumption 5 has no effect on this property

Variance of the OLS estimator is inflated

Consider the case when the variance of the error rises with X: the variance of the estimated slope coefficient is bigger (draw picture).

Standard formulae for standard errors of OLS estimates are wrong.

They don't take account of the extra sampling variation (#2 above)

OLS Variance is Wrong

Let $Y_i = \alpha + \beta X_i + \varepsilon_i$, and $E\left[\left(X_i - \overline{X}\right)\varepsilon_i\right] = 0$ and let ε be heteroskedastic: $E\left[\left(\varepsilon_i\right)^2\right] = \sigma_i^2$

- $\hat{\beta} = \beta + \left(\dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)$

$E\left[\hat{\beta}\right] = \beta$

$$V\left[\hat{\beta}\right] = E\left[\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)^2\right] = E\left[\left(\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)^2\right] = \frac{1}{\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^2} E\left[\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \varepsilon_i\right)^2\right]$$

$$= \frac{1}{\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^2} E\left[\left(X_1 - \overline{X}\right)\left(X_1 - \overline{X}\right)\varepsilon_1\varepsilon_1 + \left(X_1 - \overline{X}\right)\left(X_2 - \overline{X}\right)\varepsilon_1\varepsilon_2 + \ldots + \left(X_{n-1} - \overline{X}\right)\left(X_n - \overline{X}\right)\varepsilon_{n-1}\varepsilon_n + \left(X_n - \overline{X}\right)\left(X_n - \overline{X}\right)\varepsilon_n\varepsilon_n\right]$$

$$= \frac{1}{\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^2} E\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \left(\varepsilon_i\right)^2\right] = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 E\left[\left(\varepsilon_i\right)^2\right]}{\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\right)^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sigma_i^2}{\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\right)^2}$$

Testing for Heteroskedasticity

- There are many formal tests available
- In addition to formal tests **always look at residual plots!!**
- Look for specification errors too, since apparent heteroskedasticity may just be due to an omitted variable, for instance.
- Of the many formal tests available, the most useful is the *White Test*. It is quite general, and designed to test for heteroskedasticity of an unknown form (e.g., when we don't know that the error variance is proportional to some Z)
- The White test is quite common, and you can do it EViews with a couple of clicks (we'll see this in a moment).
- The next slide discusses how the test works …

Three Steps of the White test

1. Estimate the regression model of interest (call this equation 1) & collect the residuals, e_i.

   Suppose equation 1 is: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

2. Square the residuals. Regress e_i^2 on all the independent variables from equation 1, their squares, and cross products (call this equation 2).

   So our equation 2 is:

   $$e_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{1i}^2 + \alpha_5 X_{2i}^2 + \alpha_6 X_{3i}^2 + \alpha_7 X_{1i} X_{2i} + \alpha_8 X_{1i} X_{3i} + \alpha_9 X_{2i} X_{3i} + u_i$$

3. Test the overall significance of equation 2.

   The test statistic is nR^2, where n is the sample size and R^2 is the proportion of variation explained in equation 2. Under the null of no heteroskedasticity, this test statistic has a Chi-square(k^*) distribution asymptotically, where k^* is the number of slope coefficients in equation 2.

   Critical values of the Chi-square distribution are in the text (table B-8).

   if the test statistic exceeds the critical value, reject the null.

   In Eviews, you first run the regression, then, under View, select “Residual Diagnostics”, select “Heteroskedasticity Tests”, select “White”

What to do if errors are heteroskedastic …

- If you find evidence of heteroskedasticity – whether through a formal test by looking at residual plots – you have several options

1. Use OLS to estimate the regression and "fix" the standard errors
   A. We know OLS is unbiased, it's just that the usual formula for the standard errors is wrong (and hence tests can be misleading)
   B. We can get **consistent** estimates of the standard errors (as the sample size goes to infinity, a consistent estimator gets arbitrarily close to the true value in a probabilistic sense) called **White's Heteroskedasticity-Consistent** standard errors
   C. When specifying the regression in EViews, click the OPTIONS tab, check the "Coefficient Covariance Matrix" box, and the "White" button
   D. Most of the time, this approach is sufficient
2. Try Weighted Least Squares (WLS) – if you know the source of the heteroskedasticity and want a more efficient estimator
3. Try re-defining the variables – again, if you think you understand the source of the problem (taking log of dependent variable often helps)

The Park Test

- If we suspect we know the source of the heteroskedasticity, e.g., if we suspect

$$\text{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 Z_i^2$$

then we can do better.

- There are three steps:

1. Estimate the equation of interest:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

and collect the residuals e_i

2. Estimate the auxiliary regression:

$$ln(e_i^2) = \alpha_0 + \alpha_1 lnZ_i + u_i$$

where the error term u_i satisfies the assumptions of the CLRM. (*why logs?*).

3. Test the statistical significance of lnZ_i in the auxiliary regression with a t-test, i.e., test the null

$$H_0: \alpha_1 = 0 \text{ , } H_1: \alpha_1 \neq 0$$

- Problem with this approach: we need to know *Z*!

Weighted Least Squares

- We know that OLS is not BLUE when errors are heteroskedastic
- Suppose we want to estimate the regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

and we know (or suspect) that

$$\text{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 Z_i^2.$$

- We could rewrite the model as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + Z_i u_i$$

where $\text{Var}(u_i) = \sigma^2$.

- The BLUE of this model is called Weighted Least Squares (WLS). It is just least squares on the transformed model:

$$Y_i/Z_i = \beta_0/Z_i + \beta_1 X_{1i}/Z_i + \beta_2 X_{2i}/Z_i + u_i$$

where we divide everything through by Z_i.

- Notice that the transformed model has a homoskedastic error, and hence OLS is BLUE in the transformed model.
- You can do all this in EViews using the "Weight" option (but note EViews defines the weight as $1/Z_i$)

Redefining Variables

- Sometimes the best alternative is to go back to the drawing board – and redefine the variables in a way that is consistent with economic theory & common sense **and** that makes the errors homoskedastic.
- Using a logarithmic dependent variable may help
  - homoskedasticity in the semi-log model means the error variance is a constant **proportion** of the dependent variable.
- Other transformations may help, too, e.g., deflating by a scale variable

Redefining Variables – An Example

- Another example: suppose we're estimating the regression:

$$EXP_i = \beta_0 + \beta_1 GDP_i + \beta_2 POP_i + \beta_3 CIG_i + \varepsilon_i$$

where EXP_i is medical expenditure in province i
GDP_i is GDP in province i
POP_i is population in province i
CIG_i is the number of cigarettes sold in province i

Then "large" provinces (Quebec and Ontario) will have much larger error variance than "small" provinces (e.g., PEI), just because the scale of their medical expenditures (and everything else) will be so much bigger. We could take logs, but even better, estimate:

$$EXP_i/POP_i = \alpha_0 + \alpha_1 GDP_i/POP_i + \alpha_3 CIG_i/POP_i + u_i$$

which puts expenditures, GDP, and cigarette sales in "per-capita" terms. This way, each province contributes about the same amount of information, and should stabilize the error variances too.

Serial Correlation

- **Serial correlation** occurs when one observation's error term (ε_i) is correlated with another observation's error term (ε_j): $Corr(\varepsilon_i, \varepsilon_j) \neq 0$
- We say the errors are **serially correlated**
- This usually happens because there is an important relationship (economic or otherwise) between the observations. Examples:
  - **Time series data** (when observations are measurements of the same variables at different points in time)
  - **Cluster sampling** (when observations are measurements of the same variables on related *subjects*, e.g., more than one member of the same family, more than one firm operating in the same market, etc.)
    - Example: Suppose you are modeling calorie consumption with data on a random sample of families, one observation for each family member. Because families eat together, random shocks to calorie consumption (i.e., errors) are likely to be correlated within families.
- Serial correlation violates Assumption 4 of the CLRM. So we know that least squares is not BLUE when errors are serially correlated.

Pure Serial Correlation

- There are two basic types of serial correlation (pure & impure)
- **Pure Serial Correlation** arises if the model is *correctly specified*, but the errors are serially correlated, e.g., the DGP is:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t$$

where $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$

and u_t is a "classical" error term (i.e., it satisfies the assumptions of the CLRM)

- **Note:** we used the subscript *t* (instead of *i*) to denote the observation number. This is standard for models of time series data (*t* refers to a time period), which is where serial correlation arises most frequently.
- **Note also:** this kind of serial correlation is called **first-order autocorrelation** (or first-order autoregression, or AR(1) for short), and ρ is called the autocorrelation coefficient
  - this kind of serial correlation is very common in time-series settings

Impure Serial Correlation

- **Impure serial correlation** arises if the model is *mis-specified* (e.g., due to an omitted variable) and the specification error induces serial correlation. For example, suppose the DGP is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$$

but we estimate $Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon^*_t$
where $\varepsilon^*_t = \beta_2 X_{2t} + \varepsilon_t$
and suppose $X_{2t} = \gamma X_{2t-1} + u_t$
and where ε_t and u_t are classical error terms.

- Because of the specification error (omitting X_2 from the model), the error term in the mis-specified model is:

$$\begin{aligned} \varepsilon^*_t &= \beta_2 X_{2t} + \varepsilon_t \\ &= \beta_2(\gamma X_{2t-1} + u_t) + \varepsilon_t \\ &= \gamma \varepsilon^*_{t-1} + \beta_2 u_t + \varepsilon_t - \gamma \varepsilon_{t-1} \end{aligned}$$

and is therefore correlated with the error term of observation *t-1* (that is, ε^*_{t-1})

- This omitted variables problem does not cause **bias** because the omitted variable is **not correlated** with the included regressor.
- But, it does cause **inefficiency** because the omitted variable is correlated with itself over time.

Some examples

- We saw the example of first-order autocorrelation already: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$
- this requires $-1 < \rho < 1$ (**why? what if** $\rho = 0$?)
- $\rho < 0$ is an example of **negative serial correlation**, where ε_t and ε_{t-1} tend to have opposite signs. This case is difficult to interpret & pretty rare in economic data. (**Draw some pictures**)
- $\rho > 0$ is an example of **positive serial correlation**. In this case, ε_t and ε_{t-1} tend to have the same sign. This is very common in economic data **(draw some pictures).**
  - Happens frequently in time series data if macroeconomic "shocks" take time to work their way through the economy
  - Example: modeling the price of oranges. Oranges can only be grown in warm climates. They are transported by truck before being sold to consumers. Thus their price is influenced by the price of gasoline (and hence of oil). An unexpected shock to the supply of oil (say, due to the invasion of an oil producing country …) leads to an increase in the price of oil that may last several years, and shows up as a series of positive "shocks" to the price of oranges.
- We can also have higher order autocorrelation, e.g.,

  $\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + u_t$ (second-order autocorrelation)
- Autocorrelation need not be between adjacent periods, e.g., with quarterly data we might have

  $\varepsilon_t = \rho\varepsilon_{t-4} + u_t$ (seasonal autocorrelation, where today's error is correlated with the error 1 year ago, say due to seasonal demand)

Consequences of Serial Correlation

- We know that serial correlation violates Assumption 4 of the CLRM, and hence OLS is not BLUE. What more can we say?

1. OLS estimates remain unbiased

   We only need Assumptions 1-3 to show that the OLS estimator is unbiased, hence a violation of Assumption 4 has no effect on this property

2. The OLS estimator is no longer the best (minimum variance) linear unbiased estimator

   Serial correlation implies that errors are partly predictable. For example, with positive serial correlation, then a positive error today implies tomorrow's error is likely to be positive also. The OLS estimator ignores this information; more efficient estimators are available that do not.

3. Standard formulae for standard errors of OLS estimates are wrong.

   Standard formulae for OLS standard errors assume that errors are not serially correlated – have a look at how we derived these in lecture 12 (we needed to use assumption 4 of the CLRM). Since our t-test statistic depends on these standard errors, we should be careful about doing t-tests in the presence of serial correlation.

Testing for Serial Correlation

- There are a number of formal tests available
- In addition to formal tests **always look at residual plots!!**
- The most common test is the *Durbin-Watson* (DW) *d Test*
- This is a test for first-order autocorrelation **only**
- Some caveats: the model needs to have an intercept term, and can't have a *lagged dependent variable* (i.e., Y_{t-1} as one of the independent variables)
- If we write the error term as

  $$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

  DW tests the null hypothesis

  $$H_0 : \rho \leq 0 \quad \text{(no positive autocorrelation)}$$

  or

  $$H_0 : \rho = 0 \quad \text{(no autocorrelation)}$$

  against the appropriate alternative.
- This test is so common that almost every software package automatically calculates the value of the *d* statistic whenever you estimate a regression
  - but they almost never report p-values … for reasons we'll see in a moment

The Durbin-Watson *d* test

- The test statistic is based on the least squares residuals $e_1, e_2, \ldots, e_T$ (where T is the sample size). The test statistic is:

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

- Consider the extremes:
  - if $\rho = 1$, then we expect $e_t = e_{t-1}$. When this is true, $d \approx 0$.
  - if $\rho = -1$, then we expect $e_t = -e_{t-1}$ and $e_t - e_{t-1} = -2e_{t-1}$. When this is true, $d \approx 4$.
  - if $\rho = 0$, then we expect $d \approx 2$.
- Hence values of the test statistic "far" from 2 indicate that serial correlation is likely present.
- Unfortunately, the distribution theory for d is a little funny – for some values of d, the test is **inconclusive**
  - For a given significance level, there are two critical values: $0 < d_L < d_U < 2$
  - For a one-sided test, $H_0 : \rho \leq 0$, $H_1 : \rho > 0$
    - Reject H_0 if $d < d_L$
    - Do not reject H_0 if $d > d_U$
    - Test is inconclusive if $d_L \leq d \leq d_U$
  - (see the text for decision rules for a two-sided test)
- Do an example … and show how to plot the residuals

What to do if errors are serially correlated …

- If you find evidence of serial correlation – whether through a formal test or just by looking at residual plots – you have several options available to you

1. Use OLS to estimate the regression and "fix" the standard errors
   A. We know OLS is unbiased, it's just that the usual formula for the standard errors is wrong (and hence tests can be misleading)
   B. We can get **consistent** estimates of the standard errors (as the sample size goes to infinity, a consistent estimator gets arbitrarily close to the true value in a probabilistic sense) called **Newey-West** standard errors
   C. When specifying the regression in EViews, click the OPTIONS tab, check the "coefficient covariance matrix" box, and the "HAC Newey-West" button
   D. Most of the time, this approach is sufficient
2. Try Generalized Least Squares (GLS) – if you want a more efficient estimator

Generalized Least Squares

- We know that OLS is not BLUE when errors are serially correlated
- Rather, the BLUE is a generalization of OLS called Generalized Least Squares (GLS)
- Suppose we want to estimate the regression:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t$$

  but we suspect that ε_t is serially correlated. In particular, suppose we think (say, based on a DW test) that

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

- Then we could write the model as:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \rho \varepsilon_{t-1} + u_t$$

- GLS is a method of estimating β_0 , β_1 and ρ (and it's the BLUE of β_0 β_1)
- There are several different ways of calculating the GLS estimator (the text discusses two) -- the mechanics are beyond the scope of this course (we have computers for that!)
- The simplest: in EViews, just add AR(1) as an independent variable! (or AR(4) for the seasonal model we saw before, or AR(1) AR(2) for the second-order autoregressive model, etc.).

Non-Spherical Errors

- If errors are uncorrelated across observations and have identical variances, we say they are *spherical*.
- If errors are correlated across observations or have variances that differ across observations, we say that they are *non-spherical*.
- When you have nonspherical errors, OLS gives the wrong variances, but you can get correct ones: White or HAC-Newey-West (or clustered)
- When you have nonspherical errors, OLS is not efficient. Its variance is not the smallest possible among linear estimators. But, WLS is efficient if you know the form of the heteroskedasticity, and GLS is efficient if you know the form of the correlation across observations.