# 1 The Multiple Regression Model: Freeing Up the Classical Assumptions

- Some or all of classical assumptions were crucial for many of the derivations of the previous chapters.

- Derivation of the OLS estimator itself only required the assumption of a linear relationship between $Y$ and $X$

- But to show that OLS estimator had desirable properties did require assumptions

- Proof of the Gauss-Markov theorem required all classical assumptions except for the assumption of Normal errors

- Derivation of confidence intervals and hypothesis testing procedures required all classical assumptions.

- But what if some or all of the classical assumptions are false?

- Chapter begins with discussing some general theory, before considering some special cases.

- Two general categories: problems which call for use of *Generalized Least Squares (or GLS) Estimator*

- *Heteroskedasticity* and *autocorrelated errors* will be discussed in this category.

- Second category relates to the use of the so-called *Instrumental Variables (IV) Estimator*.

## 1.1 Basic Theoretical Results

In previous lectures derived theoretical results using multiple regression model with classical assumptions

$$Y_i = \alpha + \beta_1 X_{1i} + .. + \beta_k X_{ki} + \varepsilon_i.$$

1. $E(\varepsilon_i) = 0$ – mean zero errors.

2. $var(\varepsilon_i) = E\left(\varepsilon_i^2\right) = \sigma^2$ – constant variance errors (homoskedasticity).

3. $cov\left(\varepsilon_i \varepsilon_j\right) = 0$ for $i \neq j$ (errors uncorrelated with one another)

4. $\varepsilon_i$ is Normally distributed

5. $X_i$ is fixed. It is not a random variable.

Remember: Assumption 1 is innocuous (if the error had a non-zero mean we could include it as part of the intercept — it would have no effect on estimation of slope coefficients in the model).

Assumption 4 can be relaxed (approximately) by using asymptotic theory (not discussed in this course, but see Appendix to Chapter 3 if you are interested)

Assumption 5 we will still maintain (we will discuss this more later on in the context of "instrumental variables" estimation).

For now Assumptions 2 and 3.

Heteroskedasticity relates to Assumption 2.

Autocorrelation (also called serial correlation) relates to Assumption 3.

Basic ideas:

- Under classical assumptions, Gauss Markov theorem says "OLS is BLUE". But if Assumptions 2 and 3 are violated OLS this no longer holds (OLS is still unbiased, but is no longer "best". i.e. no longer minimum variance).

- Concepts/proofs/derivations often use following strategy. The model can be transformed to create a new model which does satisfy classical assumptions. We know OLS (on the transformed model) will be BLUE. (And all the theory we worked out for the OLS estimator will hold — except it will hold for the transformed model).

- The OLS estimator using such a transformed model is called the Generalized Least Squares (GLS) estimator.

## 1.2 Heteroskedasticity

Heteroskedasticity occurs when the error variance differs across observations.

Assumption 2 replaced by $var(\varepsilon_i) = \sigma^2\omega_i^2$ for $i = 1, .., N$.

### 1.2.1 Some theoretical results assuming $\omega_i^2$ is known

**What are the properties of the OLS estimator if heteroskedasticity is present?** To make derivations easier, let us go back to the simple regression model:

$$Y_i = \beta X_i + \varepsilon_i$$

where all the classical assumptions hold, except for Assumption 2. We now have heteroskedasticity.

Remember that OLS estimator can be written in various ways:

$$\widehat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} = \beta + \frac{\sum X_i \varepsilon_i}{\sum X_i^2}$$

Before, under classical assumptions, we proved:

$$\widehat{\beta} \text{ is } N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right),$$

which we used to derive confidence intervals and hypothesis testing procedures.

Under heteroskedasticity, most of our previous derivations still work. The error variance did not appear in our proofs for unbiasedness of OLS nor showing it was Normal.

Hence, we will not repeat the derivations here but simply state the following results:

- Under the present assumptions (i.e. allowing for heteroskedasticity), OLS is still unbiased (i.e. $E\left(\widehat{\beta}\right) = \beta$) and it is Normally distribution.

New result:

Under the present assumptions,

$$var\left(\widehat{\beta}\right) = \frac{\sigma^2 \sum X_i^2 \omega_i^2}{\left(\sum X_i^2\right)^2}.$$

Proof (using various properties of variance operator)

$$
\begin{aligned}
var\left(\widehat{\beta}\right) &= var\left(\beta + \frac{\sum X_i \varepsilon_i}{\sum X_i^2}\right) \\
&= var\left(\frac{\sum X_i \varepsilon_i}{\sum X_i^2}\right) \\
&= \frac{1}{\left(\sum X_i^2\right)^2} var\left(\sum X_i \varepsilon_i\right) \\
&= \frac{1}{\left(\sum X_i^2\right)^2} \sum X_i^2 var\left(\varepsilon_i\right) \\
&= \frac{\sigma^2}{\left(\sum X_i^2\right)^2} \sum X_i^2 \omega_i^2
\end{aligned}
$$

Key Theoretical Point: If heteroskedasticity is present, the variance of the OLS estimator is different than what it was under the classical assumptions.

Key Point for Empirical Practice:

- If heteroskedasticity is present and you ignore it, simply using the OLS estimator in a software package, the software package will use the incorrect formula for $var\left(\widehat{\beta}\right)$.

- Software package will use the formula which obtains under the classical assumptions, where it should be using $var\left(\widehat{\beta}\right) = \frac{\sigma^2 \sum X_i^2 \omega_i^2}{\left(\sum X_i^2\right)^2}$.

- Since $var\left(\widehat{\beta}\right)$ enters the formula for confidence intervals and test statistics, THESE WILL BE INCORRECT.

- In summary: OLS is still unbiased if heteroskedasticity is present (so as an estimate it may be okay), but everything else (confidence intervals, hypothesis tests, etc.) will be incorrect (unless you make sure the computer is using the correct $var\left(\widehat{\beta}\right) = \frac{\sigma^2 \sum X_i^2 \omega_i^2}{\left(\sum X_i^2\right)^2}$ formula).

- The only case where using OLS is acceptable is if you make sure the computer is using the correct $var\left(\widehat{\beta}\right) = \frac{\sigma^2 \sum X_i^2 \omega_i^2}{\left(\sum X_i^2\right)^2}$ formula. This is a point we will return to later in our discussion of something called a heteroskedasticity consistent estimator (to be defined later).

### 1.2.2 The Generalized Least Squares Estimator under Heteroskedasticity

Idea: Transform the model to create a new model which does obey classical assumptions.

The original regression model is:

$$Y_i = \beta X_i + \varepsilon_i \tag{1}$$

Consider a transformed model where we divide both sides by $\omega_i$:

$$\frac{Y_i}{\omega_i} = \beta \frac{X_i}{\omega_i} + \frac{\varepsilon_i}{\omega_i}$$

or (to make the notation compact):

$$Y_i^* = \beta X_i^* + \varepsilon_i^* \tag{2}$$

Transformed model given in (2) satisfies the classical assumptions.Key thing to verify:

$$\begin{aligned} var\left(\varepsilon_i^*\right) &= var\left(\frac{\varepsilon_i}{\omega_i}\right) \\ &= \frac{1}{\omega_i^2} var\left(\varepsilon_i\right) \\ &= \frac{\sigma^2 \omega_i^2}{\omega_i^2} = \sigma^2. \end{aligned}$$

So error variances in (2) are constant.

Important point: The transformed model in (2) satisfies classical assumptions. Hence, all our OLS results (using transformed model) can be used to say OLS (on transformed model) is BLUE, OLS confidence intervals (using transformed data) are correct, etc. etc.

**The Generalized Least Squares Estimator** The previous reasoning suggests OLS using tranformed data provides a good estimator:

$$\widehat{\beta}_{GLS} = \frac{\sum X_i^* Y_i^*}{\sum X_i^{*2}}$$

In terms of the original data this is:

$$\widehat{\beta}_{GLS} = \frac{\sum \frac{X_i Y_i}{\omega_i^2}}{\sum \frac{X_i^2}{\omega_i^2}}$$

This is called the Generalized Least Squares (GLS) estimator (and I have written "GLS" as a subscript on it to make explicit it is not the sames as OLS).

Intuition: This is sometimes called the "weighted least squares" estimator. Each observation is "weighted" with weights inversely proportional to its error variance.

Note: I am still working with the simple regression model, but the extension to multiple regression is immediate. Simply divide every explanatory variable (and the dependent variable) by $\omega_i$ and then do OLS on the transformed model.

**Properties of the GLS estimator (under heteroskedasticity)** Since GLS is equivalent to OLS on transformed model, we can use all our OLS results from Chapter 3 (and apply them to the transformed model).

That is, plug in $X_i^*$ and $Y_i^*$ instead of $X_i$ and $Y_i$ in all our old formulae.

So, since the transformed model satisfies the classical assumptions, we can immediately draw on our old results to say:

$$\widehat{\beta}_{GLS} \text{ is } N\left(\beta, \frac{\sigma^2}{\sum X_i^{*2}}\right).$$

Thus, (under the current assumptions) GLS is unbiased with

$$
\begin{aligned}
var\left(\widehat{\beta}_{GLS}\right) &= \frac{\sigma^2}{\sum X_i^{*2}} \\
&= \frac{\sigma^2}{\sum\left(\frac{X_i^2}{\omega_i^2}\right)}
\end{aligned}
$$

Note: This is not the same as the OLS formula.

Important point:

Gauss-Markov theorem tells us that, under the classical assumptions, OLS is BLUE.

Here $\widehat{\beta}_{GLS}$ is equivalent to OLS estimation of a transformed model which does satisfy the classical assumptions. Hence, under heteroskedasticity, it follows immediately that $\widehat{\beta}_{GLS}$ is BLUE.

An implication of this is that:

$$var\left(\widehat{\beta}_{GLS}\right) \leq var\left(\widehat{\beta}_{OLS}\right)$$

where $\widehat{\beta}_{OLS}$ is OLS using the original (not transformed) data.

Thus, it follows that GLS is a better estimator than OLS. Both are unbiased, but GLS has a smaller variance (it is more efficient).

The fact that

$$\widehat{\beta}_{GLS} \text{ is } N\left(\beta, \frac{\sigma^2}{\sum X_i^{*2}}\right).$$

can be used to derive confidence intervals and hypothesis tests exactly as before. We will not repeat this material (formulae are same as before except with $X_i^*$ and $y_i^*$ instead of $X_i$ and $y_i$).

### 1.2.3 Heteroskedasticity: Estimation if Error variances are unknown

The derivations above assumed that $\omega_i^2$ is known. In practice, it will usually be the case that $\omega_i^2$ is unknown.

How to proceed? Either figure out what $\omega_i^2$ or replace $\omega_i^2$ by an estimate (it can be show that, if the estimate is consistent, then GLS is a consistent estimator).

Alternatively, a *heteroskedasticity consistent estimator* (HCE) can be used.

Digression: *consistency* is an asymptotic concept (asymptotic derivations not done in this course)

Intuition 1: Consistency has some similarities to unbiasedness.

Intuition 2: A consistent estimator is one which, as sample size goes to infinity, go to true value.

## Fixing up a Heteroskedasticity Problem by Logging

- In some cases, log linear regressions will be homoskedastic even if linear regression is heteroskedastic

- Note: if variables have values which are zero or negative you cannot log them.

- But even if you log some of your variables (or even only log the dependent variable) it is sometimes enough to fix up a heteroskedasticity problem

- Remember: be careful with interpretation of coefficients when you log variables (see Chapter 4)

- Heteroskedasticity tests (see below) can be used to see whether logging fixes us a heteroskedasticity problem

- Note: solving a heteroskedasticity problem by logging is not called GLS

**Doing GLS by Transforming the Model** In many cases, the heteroskedasticity can be related to an explanatory variable. Hence it is common to use the multiple regression model:

$$Y_i = \alpha + \beta_1 X_{1i} + .. + \beta_k X_{ki} + \varepsilon_i,$$

under the classical assumptions except that

$$var\left(\varepsilon_i\right) = \sigma^2 \omega_i = \sigma^2 Z_i^2$$

where $Z_i$ is an explanatory variable (usually $Z_i$ will be one of $X_{2i}, .., X_{ki}$).

This captures the idea "the error variances vary directly with an explanatory variable".

If you suspect "the error variances vary inversely with an explanatory variable" you could use:

$$var\left(\varepsilon_i\right) = \sigma^2 \frac{1}{Z_i^2}$$

Note: variances must be positive which is why I have used $Z_i^2$. An alternative choice is to use the exponential function (e.g. $var\left(\varepsilon_i\right) = \sigma^2 \exp\left(Z_i\right)$).

Remember: under heteroskedasticity, GLS says we should tranform our data as:

$$\frac{Y_i}{\omega_i} = \beta \frac{X_i}{\omega_i} + \frac{\varepsilon_i}{\omega_i}.$$

and then use OLS on transformed model. But here we have $\omega_i = Z_i$. So can divide all your variables by $Z_i$ and then do OLS.

Empirical tip: Experiment with different choices for $Z$ (usually, it will be one of $X_1, .., X_k$)

Note: cannot divide by zero and, hence, you cannot use this transformation for a variable which has $Z_i = 0$ for any observation

Cannot use of this transformation with dummy variables.

If the heteroskedasticity is characterized by $f(Z_i) = \exp(Z_i)$ then zero values of $Z_i$ are acceptable.

Above has "error variances vary directly with $Z$"

If error variances vary inversely with $Z$ (e.g. $f(Z_i) = \frac{1}{Z_i^2}$), transformed model becomes:

$$Y_i Z_i = \alpha Z_i + \beta_1 X_{1i} Z_i + .. + \beta_k X_{ki} Z_i + \varepsilon_i Z_i$$

GLS estimator obtained by multiplying all your variables by $Z_i$ and then doing OLS with these new variables.

**What if heteroskedasticity is present, but you cannot relate it to a single variable, Z?** It is desirable to do GLS (as describe above if you can). If you cannot, remember that OLS is still unbiased so is an adequate second best estimator. But the variance formula we derived under the classical assumptions not longer holds. The correct formula is:

$$var\left(\widehat{\beta}\right) = \frac{\sigma^2 \sum X_i^2 \omega_i^2}{\left(\sum X_i^2\right)^2}.$$

So one thing you can do is use OLS with this correct formula to calculate the variance.

Problem: we do not know $\sigma^2\omega_i^2$.

Solution: Replace it with an estimate.

Since

$$var\left(\varepsilon_i\right) = E\left(\varepsilon_i^2\right) = \sigma^2\omega_i^2$$

this suggests that we can use the OLS residuals:

$$\widehat{\varepsilon}_i^2$$

as estimates of $\sigma^2\omega_i^2$.

Thus, an estimate of $var\left(\widehat{\beta}\right)$ is

$$\widehat{var\left(\widehat{\beta}\right)} = \frac{\sum X_i^2\widehat{\varepsilon}_i^2}{\left(\sum X_i^2\right)^2}.$$

It can be shown that this estimate is consistent.

Summary: Use OLS to estimate $\widehat{\beta}$, then use $\widehat{var\left(\widehat{\beta}\right)}$ in formulae for confidence intervals, etc.

This is an example of a heteroskedasticity consistent estimator (HCE). There are others and they can be automatically calculated in more sophisticated computer packages such as Stata or PC Give (but not in Excel).

Advantages: HCEs are easy to calculate and you do not need to know the form that the heteroskedasticity takes.

Disadvantages: HCEs are not as efficient as the GLS estimator (i.e. they will have larger variance).

### 1.2.4 Testing for Heteroskedasticity

If heteroskedasticity is NOT present, then OLS is fine (it is BLUE). But if it is present, you should use GLS (or a HCE). Thus, it is important to know if heteroskedasticity is present. There are many tests, here I will describe some of the most common ones.

**Goldfeld Quandt test** This is good for the case where you suspect heteroskedasticity depends on an explanatory variable, $Z$ (which will often be one of $X_{2i}, .., X_{ki}$).

Basic idea: if you divide up your data into high $Z$ and low $Z$ parts, and run two separate regressions then they should have different error variances (if heteroskedasticity is present).

Details:

1. Order the data by the magnitude of $Z$.

2. Omit the middle $d$ observations (no hard and fast rule to choose $d$, common choice $d = .2N$)

3. Run two separate regressions, one using the observations with low values for $Z$, the other using observations with high $Z$.

4. Calculate the sum of squares residuals (SSR) for each of the two regressions (call them $\text{SSR}_{LOW}$ and $\text{SSR}_{HIGH}$).

5. Calculate the Goldfeld-Quandt test statistic which is:

$$GQ = \frac{\text{SSR}_{HIGH}}{\text{SSR}_{LOW}}.$$

Under the hypothesis of homoskedasticity, $GQ$ has an $F_{.5(N-d-4),.5(N-d-4)}$ distribution (and can use F statistical tables to get critical value). Reject homoskedasticity (and, thus, conclude heteroskedasticity is present) if $GQ$ is greater than the critical value.

Note: Test above assumes error variances vary directly with $Z$. If you suspect that the error variances vary inversely with $Z$, then reverse the ordering of the data in Step 1

Empirical tip: Try various choices for $Z$

**The White Test for Heteroskedasticity** Goldfeld-Quandt test is good if a logical choice of a single $Z$ suggests itself (or if heteroskedasticity is related to a single variable and you are patient enough experiment with different choices for $Z$). White test is good if there are several possible explanatory variables which might influence the error variance.

That is:

$$var\left(\varepsilon_i\right) = \sigma^2 f\left(\gamma_0 + \gamma_1 Z_{1i} + .. + \gamma_p Z_{pi}\right),$$

Where $f\left(\right)$ is a positive function.

Loosely speaking, this captures the idea: error variance might depend on any or all of the variables $Z_1, .., Z_p$ (which may be the same as the explanatory variables in the regression itself).

White test involves the following steps:

- Run OLS on the original regression (ignoring heteroskedasticity) and obtain the residuals, $\widehat{\varepsilon}_i$.

- Run a second regression of the equation:

$$\widehat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 Z_{1i} + .. + \gamma_p Z_{pi} + v_i$$

and obtain the $R^2$ from this regression.

- Calculate the White test statistic:

$$W = NR^2$$

- This test statistic has a $\chi^2(p)$ distribution which can be used to get a critical value from.

- An advantage of the White test is that it need only be done once.

- Just need to choose $Z_1, .., Z_p$ (usually the explanatory variables in the original regression).

- A disadvantage is that, if the test indicates that heteroskedasticity is present, it does not offer much guidance on how you should try and transform the model to do GLS.

- All you know is that heteroskedasticity is present and is related to one (or several) of the variables $Z_1, .., Z_p$.

- Note these advantages/disadvantages are the exact opposite of the Goldfeld-Quandt test.

- Goldfeld-Quandt test requires selection of a single $Z$ (or doing many tests with many choices of $Z$). But if you can find one $Z$ which is related to the heteroskedasticity, this suggests how to transform model to do GLS.

### 1.2.5 Recommendations for Empirical Practice

- If you think you might have a heteroskedasticity problem, begin by doing White heteroskedasticity test.

- If your tests indicate heteroskedasticity is present, then do some Goldfeld-Quandt tests to see if you can associate the heteroskedasticity with a particular explanatory variable.

- Sometimes simple things (e.g. logging some or all of your variables) will be enough to fix the problem. (Although the resulting estimator is NOT called a GLS estimator)

- Sometimes multiply/dividing all your explanatory variables by some variable ($Z$) is enough to fix the problem.

- Note: Every time you try such a transformation you must do heteroskedasticity test (White test will be simplest) to check if it has fixed the problem.

- If you cannot find a transformation which fixes the heteroskedasticity problem, then use a HCE. (But you cannot easily do this in Excel).

- Remember: if heteroskedasticity is present, then hypothesis tests involving $\beta$'s will be incorrect. So wait until after you have corrected the problem (or are using an HCE) before doing hypothesis testing (e.g. to find out which of your explanatory variables are insignificant).

- Textbook contains two examples (one of which forms basis for Computer Problem Sheet 2)

## 1.3 Autocorrelation

We will continue our discussion of problems which call for the use of the Generalized Least Squares Estimator by considering an important topic called *autocorrelation*.

This is used with time series data, so we will use $t = 1, .., T$ to denote observations (rather than $i = 1, .., N$)

## 1.4 Reminder of Basic Theoretical Results

In previous lectures derived theoretical results using multiple regression model with classical assumptions

$$Y_t = \alpha + \beta_1 X_{1t} + .. + \beta_k X_{kt} + \varepsilon_t$$

1. $E(\varepsilon_t) = 0$ – mean zero errors.

2. $var(\varepsilon_t) = E\left(\varepsilon_t^2\right) = \sigma^2$ – constant variance errors (homoskedasticity).

3. $E(\varepsilon_t \varepsilon_s) = 0$ for $t \neq s$ — $\varepsilon_t$ and $\varepsilon_s$ are uncorrelated with one another.

4. $\varepsilon_t$ is Normally distributed

5. $X_{2t}, .., X_{kt}$ are fixed. They are not a random variable.

Remember: Assumption 1 is innocuous (if the error had a non-zero mean we could include it as part of the intercept — it would have no effect on estimation of slope coefficients in the model).

Assumption 4 can be relaxed (approximately) by using asymptotic theory.

Assumption 5 we will still maintain.

Autocorrelation (also called serial correlation) relates to Assumption 3.

Basic ideas:

- Under classical assumptions, Gauss Markov theorem says "OLS is BLUE". But if Assumptions 2 and 3 are violated OLS this no longer holds (OLS is still unbiased, but is no longer "best". i.e. no longer minimum variance).

- Concepts/proofs/derivations use following strategy. The model can be transformed to create a new model which does satisfy classical assumptions. We know OLS (on the transformed model) will be BLUE. (And

all the theory we worked out for the OLS estimator will hold — except it will hold for the transformed model).

- The OLS estimator using such a transformed model is called the Generalized Least Squares (GLS) estimator.

## 1.5 Autocorrelated Errors

- We will work with the multiple regression model under the classical assumptions, with the exception that the errors follow and *autoregressive process of order 1 (AR(1))*:

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

where it is $u_t$ which satisfies classical assumptions. So $E(u_t) = 0$, $var(u_t) = \sigma^2$ and $cov(u_t, u_s) = 0$ (for $t \neq s$).

- We also assume $-1 < \rho < 1$. To preview later material, this restriction ensure *stationarity* and means you do not have to worry about problems relating to *unit roots* and *cointegration* (definitions will be provided to you later on).

- We will focus on the AR(1) cases, but note that the AR(p) errors case is a simple extension:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + .. + \rho_p \varepsilon_{t-p} + u_t$$

### 1.5.1 Variances and Covariances of $\varepsilon_t$

- The assumptions above specified properties of $u_t$, but we need to know properties of $\varepsilon_t$.

- Notation:

$$\sigma_\varepsilon^2 = var\left(\varepsilon_t\right) = E\left(\varepsilon_t^2\right)$$

where last equal sign follows since errors have mean zero.

- Derivation of variance of regression errors (textbook does derivation in different way):

$$\begin{aligned}
\sigma_{\varepsilon}^{2} &= var\left(\rho\varepsilon_{t-1} + u_t\right) \\
&= \rho^2 var\left(\varepsilon_{t-1}\right) + var\left(u_t\right) \\
&= \rho^2\sigma_{\varepsilon}^{2} + \sigma^2 \\
&= \frac{\sigma^2}{1-\rho^2}
\end{aligned}$$

- In the previous derivations we have used properties of variance operator, the fact that $\varepsilon_{t-1}$ and $u_t$ are independent of one another and that $\varepsilon_t$ is homoskedastic.

- The derivation of covariance between different regression errors is done in Problem Sheet 3:

$$cov\left(\varepsilon_t, \varepsilon_{t-1}\right) = \rho\sigma_{\varepsilon}^{2}$$

- For errors more than one period apart, we can show:

$$cov\left(\varepsilon_t, \varepsilon_{t-s}\right) = \rho^s \sigma_\varepsilon^2$$

- Thus, we have established that the regression model with autocorrelated errors violates assumption 3. That is, the regression errors are NOT uncorrelated with one another.

- Hence, we need to work with a GLS estimator.

### 1.5.2 The GLS Estimator for the Autocorrelated Errors Case

- Remember: GLS can be interpreted as OLS on a suitably transformed model.

- In this case, the appropriate transformation is refered to as "quasi-differencing".

- To explain what this is, consider the regression model:

$$Y_t = \alpha + \beta_1 X_{1t} + .. + \beta_k X_{kt} + \varepsilon_t$$

- This model will hold for every time period so we can take it at period $t-1$ and multiply both sides of the equation by $\rho$:

$$\rho Y_{t-1} = \rho\alpha + \rho\beta_1 X_{1t-1} + .. + \rho\beta_k X_{kt-1} + \rho\varepsilon_{t-1}$$

- Subtract this equation from the original regression equation:

$$Y_t - \rho Y_{t-1} = \alpha - \rho\alpha + \beta_1 (X_{1t} - \rho X_{1t-1}) \\ +.. + \beta_k (X_{kt} - \rho X_{kt-1}) + \varepsilon_t - \rho\varepsilon_{t-1}$$

or

$$Y_t^* = \alpha^* + \beta_1 X_{1t}^* + .. + \beta_k X_{kt}^* + u_t$$

- But $u_t$ satisfies the classical assumptions so OLS on this transformed model will be GLS (which will be BLUE).

- Note that the transformed variables are "quasi-differenced"

$$Y_t^* = Y_t - \rho Y_{t-1}$$
$$X_{1t}^* = (X_{1t} - \rho X_{1t-1})$$
etc.

The case with $\rho = 1$ (which we do not consider) is called "differenced" – this is not quite the same so we say "quasi" differenced.

- One (relatively minor) issue: if our original data is from $t = 1, .., T$ then $Y_1^* = Y_1 - \rho Y_0$ will involve $Y_0$ (and same issue for explanatory variables). But we do not observe such "initial conditions". There are many ways of treating initial conditions.

- What we do (simplest, most common thing) is work with data from $t = 2, .., T$ (and use $t = 1$ values for variables as initial conditions).

- Summary: If we knew $\rho$, then we could quasi-difference the data and do OLS using the transformed data (which is equivalent to GLS).

- In practice, we rarely (if ever) know $\rho$. Hence, we replace $\rho$ by an estimate: $\widehat{\rho}$. There are several ways of getting a $\widehat{\rho}$, we now turn to one, called the Cochrane-Orcutt procedure.

### 1.5.3 The Cochrane-Orcutt Procedure

- Remember: with autocorrelated errors, GLS is BLUE. However, OLS (on original data) is still unbiased.

- Cochrane-Orcutt procedure begins with OLS and then uses OLS residuals to estimate $\rho$.

- Cochrane-Orcutt procedure goes through following steps:

1. Do OLS regression of $Y_t$ on intercept, $X_{1t}, .., X_{kt}$ and produce the OLS residuals, $\widehat{\varepsilon}_t$.

2. Do OLS regression of $\widehat{\varepsilon}_t$ on $\widehat{\varepsilon}_{t-1}$ which will provide a $\widehat{\rho}$.

3. Quasi-difference all variables to produce

$$Y_t^* = Y_t - \widehat{\rho} Y_{t-1}$$
$$X_{1t}^* = (X_{1t} - \widehat{\rho} X_{1t-1})$$
etc.

4. Do OLS regression of $Y_t^*$ on intercept, $X_{1t}^*, .., X_{kt}^*$, thus producing GLS estimates of the coefficients.

### 1.5.4 Autocorrelation Consistent Estimators

- Remember: with heteroskedasticity we discussed heteroskedasticity consistent estimator (HCE).

- Less efficient than GLS, but is a correct second-best solution when GLS difficult to implement.

- Similar issues hold autocorrelated errors.

- There exist *autocorrelation consistent estimators* which allow for the correct use of OLS methods when you have autocorrelated errors.

- We will not explain these, but many popular econometrics software packages include them. The most popular is the *Newey-West estimator*.

### 1.5.5 Testing for Autocorrelated Errors

- If $\rho = 0$ then doing OLS on the original data is fine (OLS is BLUE). However, if $\rho \neq 0$, then a GLS estimator such as the Cochrane-Orcutt estimator is better.

- This motivates testing $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$.

- There are several such tests, here we describe some of the most popular.

**Breusch-Godfrey Test** AR(p) errors:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + .. + \rho_p \varepsilon_{t-p} + u_t.$$

$H_0 : \rho_1 = 0, \rho_2 = 0, .., \rho_p = 0$

Breusch-Godfrey test involves the following steps:

1. Run a regression of $Y_t$ on an intercept, $X_1, .., X_k$ using OLS and produce the residuals, $\widehat{\varepsilon}_t$.

2. Run second regression of $\widehat{\varepsilon}_t$ on intercept, $X_1, .., X_k, \widehat{\varepsilon}_{t-1}, .., \widehat{\varepsilon}_{t-p}$ using OLS and produce the $R^2$.

3. Calculate the test statistic:

$$LM = TR^2.$$

If $H_0$ is true, then $LM$ has an (approximate) $\chi^2(p)$ distribution.

Thus, critical value taken from statistical tables for the Chi-square distribution.

### 1.5.6 The Box-Pierce and Ljung Tests

- These test $H_0 : \rho_1 = 0, \rho_2 = 0, .., \rho_p = 0$

- Both based on idea that, if the errors are not autocorrelated, then the correlations between different errors should be zero.

- Replace errors by residuals.

- $\widehat{\varepsilon}_t$ are residuals from OLS regression of $Y$ on an intercept, $X_1, .., X_k$,

- Correlations between $\widehat{\varepsilon}_t$ and $\widehat{\varepsilon}_{t-s}$ are:

$$r_s = \frac{\sum_{t=s+1}^{T} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-s}}{\sum_{t=s+1}^{T} \widehat{\varepsilon}_t^2}.$$

- *Box-Pierce test statistic* (sometimes called the *Q test statistic*) is:

$$Q = T \sum_{j=1}^{p} r_j^2,$$

- The $p$ means that AR(p) errors are being tested for.

- The *Ljung test statistic* is:

$$Q^* = T\,(T+2) \sum_{j=1}^{p} \frac{r_j^2}{T-j}.$$

- Critical values for both taken from $\chi^2\,(p)$ tables.

- Many econometrics software packages present these test statitics

- Warning: in some cases, one of the explanatory variables will be the dependent variable from a previous period ("lagged dependent variable"). For instance:

$$Y_t = \alpha + \delta Y_{t-1} + \beta X_t + \varepsilon_t.$$

- The Box-Pierce and Ljung tests are not appropriate in this case. The Breusch-Godfrey test, however, is still appropriate.

- The textbook discusses two other approaches: the Durbin-Watson statistic and Durbin's h-test.

## 1.6 Instrumental Variable Methods

- Overview: Under the classical assumptions, OLS is BLUE.

- When we relax some of the assumptions (e.g. to allow for heteroskedasticity or autocorrelated errors), then OLS is no longer BLUE but it is still unbiased and (if a consistent estimator is used to give a good estimate for $var\left(\widehat{\beta}\right)$) then OLS will be correct (although it will be less efficient than GLS).

- However, in the case we are about to consider, OLS will be biased and an entirely different estimator will be called for – the instrumental variables (IV) estimator.

- This set of notes will consider relaxing the assumption that the explanatory variables are not random variables.

- For simplicity, we will work with the simple regression model, but results generalize to the case of multiple regression.

# 2 Theory Motivating the IV Estimator

In previous lectures derived theoretical results using regression model with classical assumptions

$$Y_i = \beta X_i + \varepsilon_i$$

1. $E(\varepsilon_i) = 0$ – mean zero errors.

2. $var(\varepsilon_i) = E\left(\varepsilon_i^2\right) = \sigma^2$ – constant variance errors (homoskedasticity).

3. $E\left(\varepsilon_i \varepsilon_j\right) = 0$ for $i \neq j$ — $\varepsilon_i$ and $\varepsilon_j$ are uncorrelated with each another.

4. $\varepsilon_i$ is Normally distributed

5. $X_i$ is fixed. It is not a random variable.

Remember: Assumption 1 is innocuous.

Assumption 4 can be relaxed (approximately) by using asymptotic theory.

Assumptions 2 and 3 were discussed in lectures on heteroskedasticity and autocorrelated errors.

Now we will focus on relaxing Assumption 5.

Note: When explanatory variables are be random, many derivations we did before with expected value and variance operators become much more difficult/impossible. For this reason, most relevant results are asymptotic.

But asymptotic methods not covered in course (see appendix to Chapter 5 if you are interested)

This section provides some intuition, hints at derivations and discussion of things relevant for empirical practice.

## 2.1 Case 1: Explanatory Variable is Random But is Uncorrelated with Error

- If $X_i$ is now a random variable, we have to make some assumptions about its distribution.

- Assume $X_i$ are i.i.d. (independent and identically distributed) random variables with:

$$E(X_i) = \mu_X$$

$$var(X_i) = \sigma_X^2$$

- In Case 1 we will assume explanatory variable and errors are uncorrelated with one another:

$$cov(X_i, \varepsilon_i) = E(X_i \varepsilon_i) = 0$$

- Remember, under classical assumptions:

$$\widehat{\beta} \text{ is } N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right).$$

- This result can still be shown to hold approximately in this case (we will not provide details, some given in textbook)

- Bottom line: If we relax the assumptions of Normality and fixed explanatory variables we get exactly the same results as for OLS under the classical assumptions (but here they hold approximately), *provided explanatory variables are uncorrelated with the error term*.

## 2.2 Case 2: Explanatory Variable is Correlated with the Error Term

- We will work with the simple regression model under classical assumptions except for Assumption 5.

- Assume $X_i$ are i.i.d. (independent and identically distributed) random variables with:

$$E(X_i) = \mu_X$$

$$var(X_i) = \sigma_X^2$$

- In Case 2 we will assume explanatory variable and errors are correlated with one another:

$$cov(X_i, \varepsilon_i) = E(X_i \varepsilon_i) \neq 0$$

- It turns out that, in this case, OLS is biased and a new estimator is called for. That estimator is the instrumental variables (IV) estimator.

- Why is this? We will not provide proof, but offer a hint.

- The proof that OLS is biased begins in the same manner as the proof of Chapter 3. We can get up to the following stage in the proof:

$$E\left(\widehat{\beta}\right) = \beta + E\left(\frac{\sum X_i \varepsilon_i}{\sum X_i^2}\right)$$

- But at this stage we can go no farther other than to note that there is no reason to think that $E\left(\frac{\sum X_i \varepsilon_i}{\sum X_i^2}\right) = 0$ and, in fact, it is not.

- Intuition: (ignore $\sum X_i^2$ in the denominator), we could write the numerator as $E\left(\sum X_i \varepsilon_i\right) = \sum E\left(X_i \varepsilon_i\right) = \sum cov\left(X_i \varepsilon_i\right) \neq 0$ .

- Important point: if the error and explanatory variable are correlated, then OLS is biased and should be avoided.

- Soon will offer some explanation for why this might occur, but first introduce new estimator to handle this case.

## 2.3 The Instrumental Variables Estimator

- An instrumental variable, $Z_i$, is a random variable which is uncorrelated with the error but is correlated with the explanatory variable.

- Formally, an instrumental variable is assumed to satisfy the following assumptions:

$$E(Z_i) = \mu_Z$$

$$var(Z_i) = \sigma_Z^2$$

$$cov(Z_i, \varepsilon_i) = E(Z_i \varepsilon_i) = 0$$

$$cov\left(X_i, Z_i\right) = E\left(X_i Z_i\right) - \mu_Z \mu_X = \sigma_{XZ} \neq 0$$

- Assuming an instrumental variable exists (something we will return to later), we can introduce the instrumental variables estimator:

$$\widehat{\beta}_{IV} = \frac{\sum_{i=1}^{N} Z_i Y_i}{\sum_{i=1}^{N} X_i Z_i}$$

- The asymptotic derivations in the appendix (not covered in this course) imply (approximately):

$$\widehat{\beta}_{IV} \text{ is } N\left(\beta, \frac{\left(\sigma_Z^2 + \mu_Z^2\right)\sigma^2}{N\left(\sigma_{XZ} + \mu_X \mu_Z\right)^2}\right).$$

- This formula can be used to calculate confidence intervals, hypothesis tests, etc. (comparable to Chapter 3 derivations)

- In practice, the unknown means and variances can be replaced by their sample counterparts. Thus, $\mu_X$ can be replaced by $\overline{X}$, $\sigma_Z^2$ by the sample variance of $\frac{(Z_i - \overline{Z})^2}{N-1}$, etc.

- No additional details of how this is done, but note that econometrics software packages do IV estimation.

### 2.3.1 Using the IV Estimator in Practice

- what if you have a multiple regression model involving more than one explanatory variable?

- Answer: you need at least one instrumental variable for each explanatory variable that is correlated with the error.

- what if you have more instrumental variables than you need?

- Use the *generalized instrumental variables estimator* (GIVE).

- Explanation of GIVE given in textbook (I will not cover this in course).

- Most econometrics software packages will calculate GIVEs for you

- Testing is discussed in textbook. Hausman test and Sargan test (not responsible for in this course)

## 2.4 Why Might the Explanatory Variable Be Correlated with Error?

- There are many different reasons why the explanatory variables might be correlated with the errors.

- "Errors in Variables" problem (discussed below).

- Simultaneous equations model covered in textbook (but will not cover in this course)

- There are also other models which imply $X$ and $\varepsilon$ correlated

### 2.4.1 Errors in Variables

- What if you want to run the regression:

$$Y_i = \beta X_i + \varepsilon_i.$$

This regression satisfies the classical assumptions, but you do not observe $x_i$, but instead observe:

$$X_i^* = X_i + v_i,$$

where $v_i$ is i.i.d. with mean zero, variance $\sigma_\nu^2$ and is independent of $\varepsilon_i$.

- In other words, $X$ is observed with error.

- Replacing $X_i$ in the original regression yields a new regression:

$$\begin{aligned} Y_i &= \beta\left(X_i^* - v_i\right) + \varepsilon_i \\ &= \beta X_i^* + \varepsilon_i^* \end{aligned}$$

where $\varepsilon_i^* = \varepsilon_i - \beta v_i$

- What is the covariance between the explanatory variable, $X_i^*$, and the error, $\varepsilon_i^*$, in this new regression?

$$\begin{aligned} cov\left(X_i^*, \varepsilon_i^*\right) &= E\left[\left(X_i + v_i\right)\left(\varepsilon_i - \beta v_i\right)\right] \\ &= -\beta\sigma_\nu^2 \neq 0 \end{aligned}$$

- Hence measurement error in explanatory variables (but not dependent variable), causes them to be correlated with the regression error.

### 2.4.2 An example where the explanatory variable could be correlated with the error

- Suppose we are interested in estimating the returns to schooling and have data from a survey of many individuals on:

The dependent variable: $Y =$ income

The explanatory variable: $X =$ years of schooling

And other explanatory variables like experience, age, occupation, etc.. which we will ignore here to simplify the exposition.

- My contention is that, in such a regression it probably is the case that $X$ is correlated with the error and, thus, OLS will be inconsistent.

- To understand why, first think of how errors are interpreted in this regression.

- An individual with a positive error is earning an unusually high level of income. That is, his/her income is more than his/her education would suggest.

- An individual with a negative error is earning an unusually low level of income. That is, his/her income is less than his/her education would suggest.

- What might be correlated with this error? Perhaps each individual has some underlying quality (e.g. intelligence, ambition, drive, talent, luck – or even family encouragement). This quality would like be associated with the error (e.g. individuals with more drive tend to achieve unusually high incomes).

- But this quality would also effect the schooling choice of the individual. For instance, ambitious students would be more likely to go to university.

- Summary: Ambitious, intelligent, driven individuals would both tend to have more schooling and more income (i.e. positive errors).

- So both the error and the explanatory variable would be influenced by this quality. Error and explanatory variable probably would be correlated with one another.

### 2.4.3 How do you choose instrumental variables?

- There is a lot of discussion in the literature how to do this. But this is too extensive and complicated for this course, so we offer a few practical thoughts.

- An instrumental variable should be correlated with explanatory variable, but not with error.

- Sometimes economic theory (or common sense) suggests variables with this property.

- In our example, we want a variable which is correlated with the schooling decision, but is unrelated to error (i.e. factors which might explain why individuals have unusually high/low incomes)

- An alternative way of saying this: we want to find a variable which affects schooling choice, but has no direct effect on income.

- Characteristics of parents or older siblings have been used as instruments.

- Justification: if either of your parents had a university degree, then you probably come from a family where education is valued (increase the chances you go to university). However, your employer will not care that your parents went to university (so no direct effect on your income).

- Other researchers have used geographical location variables as instruments.

- Justification: if you live in a community where a university/college is you are more likely to go to university. However, your employer will not care where you lived so location variable will have no direct effect on your income.

# 3 Chapter Summary

Chapter discusses violations of classical assumptions and breaks into a "GLS" part and an "IV" part.

1. If errors either have different variances (heteroskedasticity) or are correlated with one another, then OLS is unbiased, but is no longer the best estimator. The best estimator is GLS.

2. If heteroskedasticity is present, then the GLS estimator can be calculated using OLS on a transformed model. If suitable transformation cannot be found, then heteroskedasticity consistent estimator should be used.

3. There are many tests for heteroskedasticity, including the Goldfeld-Quandt test and the White test.

4. If errors are autocorrelated, GLS estimator is OLS on a transformed model. The required transformation involves quasi-differencing each variable. The Cochrane-Orcutt procedure is a popular way of implementing the GLS estimator.

5. There are many tests for autocorrelated errors, including the Breusch-Godfrey test, the Box-Pierce test and the Ljung test.

6. In many applications, it is implausible to treat the explanatory variables and fixed. Hence, it is important to allow for them to be random variables.

7. If explanatory variables are random and all of them are uncorrelated with the regression error, then standard methods associated with OLS (as developed in Chapters 2 and 3) still work.

8. If explanatory variables are random and some of them are correlated with the regression error, then OLS is biased. The instrumental variables estimator is not.

9. In multiple regression, at least one instrument is required for every explanatory variable which is correlated with the error.

10. If you have valid instruments, then the Hausman test can be used to test if the explanatory variables are correlated with the error.

11. In general, cannot test whether an instrumental variable is a valid one. However, if you have more instruments than the minimum required, the Sargan test can be used.

12. Explanatory variables can be correlated with error they are measured with error.