

UNIVERSITY OF ESSEX
DEPARTMENT OF ECONOMICS

EC501: Econometric Methods and Applications

Lecture Notes 2011/12
(Preliminary and incomplete)

João Santos Silva

jmcSS@essex.ac.uk

Room: 5B.204

Tel.: (01206) 872769

Office hours: please check my webpage regularly.

Course preliminaries

- **Classes**

- Begin next week.
- Problem set each week (on CMR)
- Solutions available (on CMR) week following class.

- **Computing classes**

- Begin University week 7
- Computing package: Stata

- **Assessment**

- Mid term exam: week 8, Friday 25th November, 5pm (**one hour**).
- Final exam: next May/June (two hours).
- Overall mark: whichever is the greater, EITHER 50% Mid term mark, 50% Exam mark OR 100% Exam mark.

- **Textbooks**

W.H. Greene (2012) *Econometric Analysis* (7th edition), Pearson.

J.H. Stock and M.M. Watson (2012) *Introduction to Econometrics* (3rd edition), Pearson.

J.M. Wooldridge (2009) *Introductory Econometrics* (4th. Edition), South Western College.

- **Further reading**

A.S. Goldberger (1991) *A Course in Econometrics*, Harvard University Press.

University week: 2

The classical linear regression model (CLRM)

Outline

1. Introduction (what is linear regression);
2. Notation and model specification;
3. Model assumptions
4. Estimation by ordinary least squares (OLS)

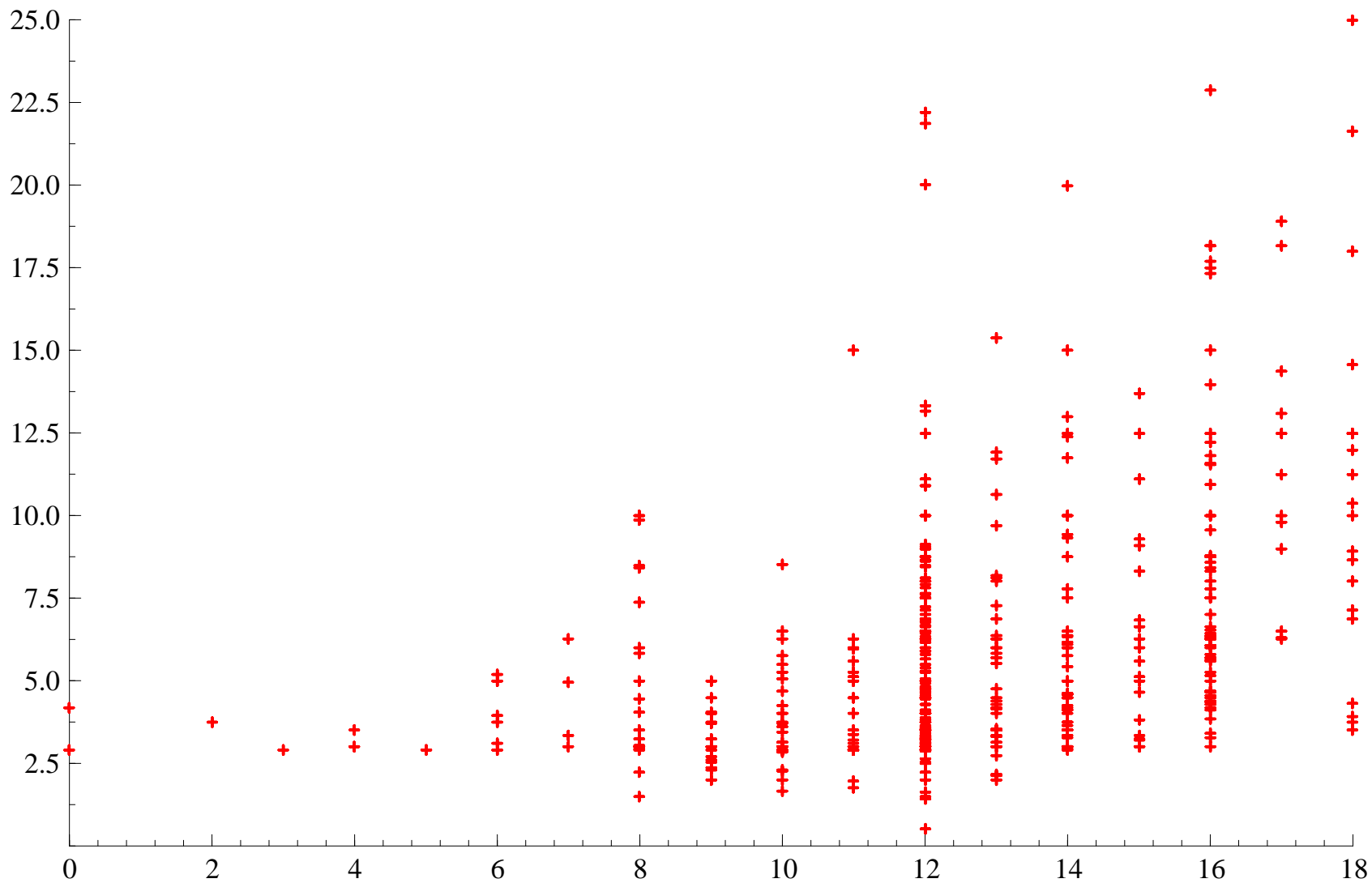
Reading: Greene: chapters 1 to 3, Appendices A (A.1–A.4, A.8) and B (B.1–B.8)

1) Introduction (what is linear regression)

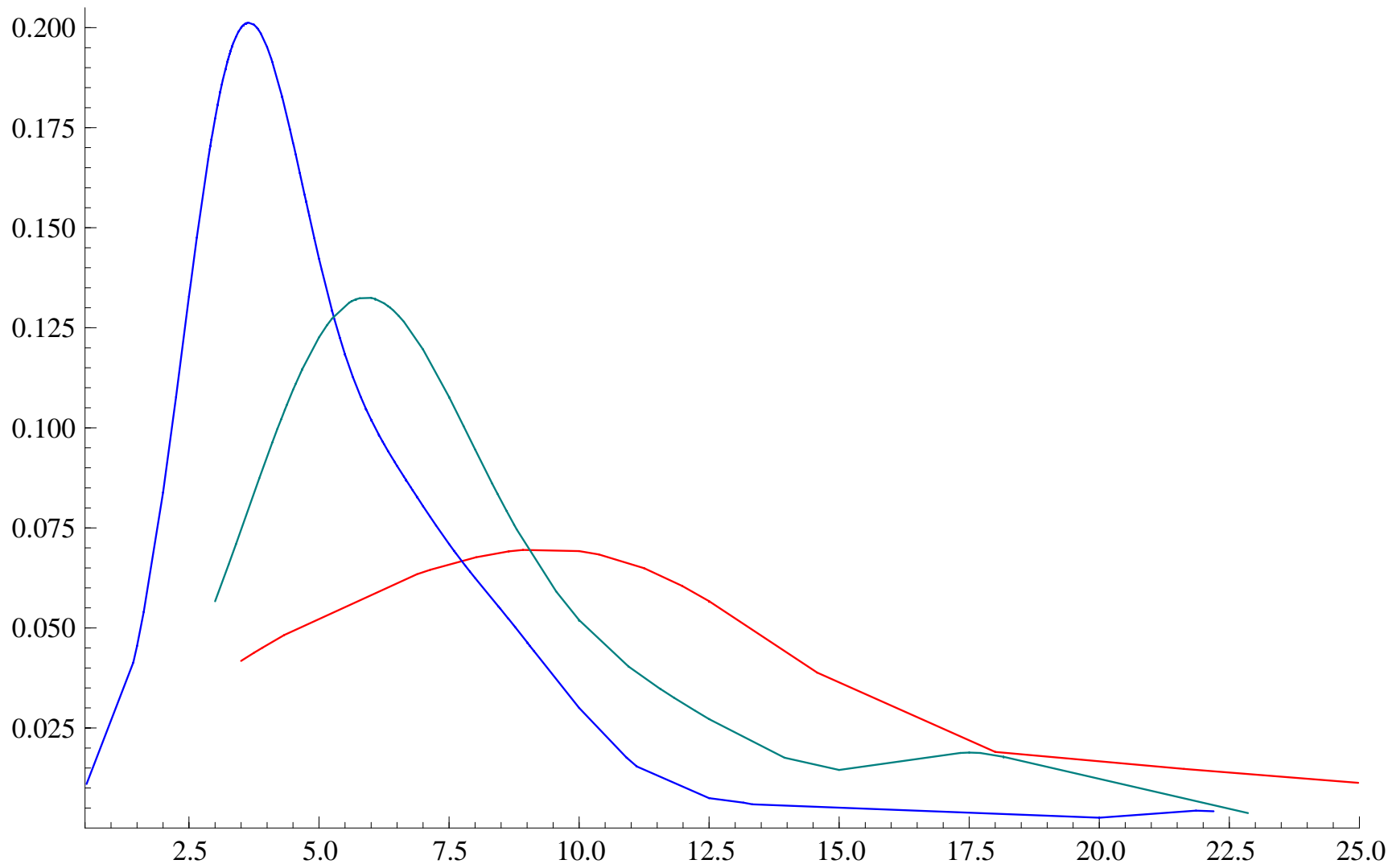
- Economic theory is interested in the relation between certain economic variables.
- Typically, economic theory leads to deterministic functions relating the variables of interest, e.g., $Q_i = AK_i^\alpha L_i^\beta$, $Q_i = f(P_i)$, $C_i = \alpha + \beta Y_i$.
- However, in reality, economic variables are random and are not related by deterministic functions.
- If we want to estimate the unknown parameters of these functions, we need to use data which are obtained by sampling from the joint distribution of the relevant variables.
- We observe data from the joint distribution

$$f(y, x) = f(y|x) g(x)$$

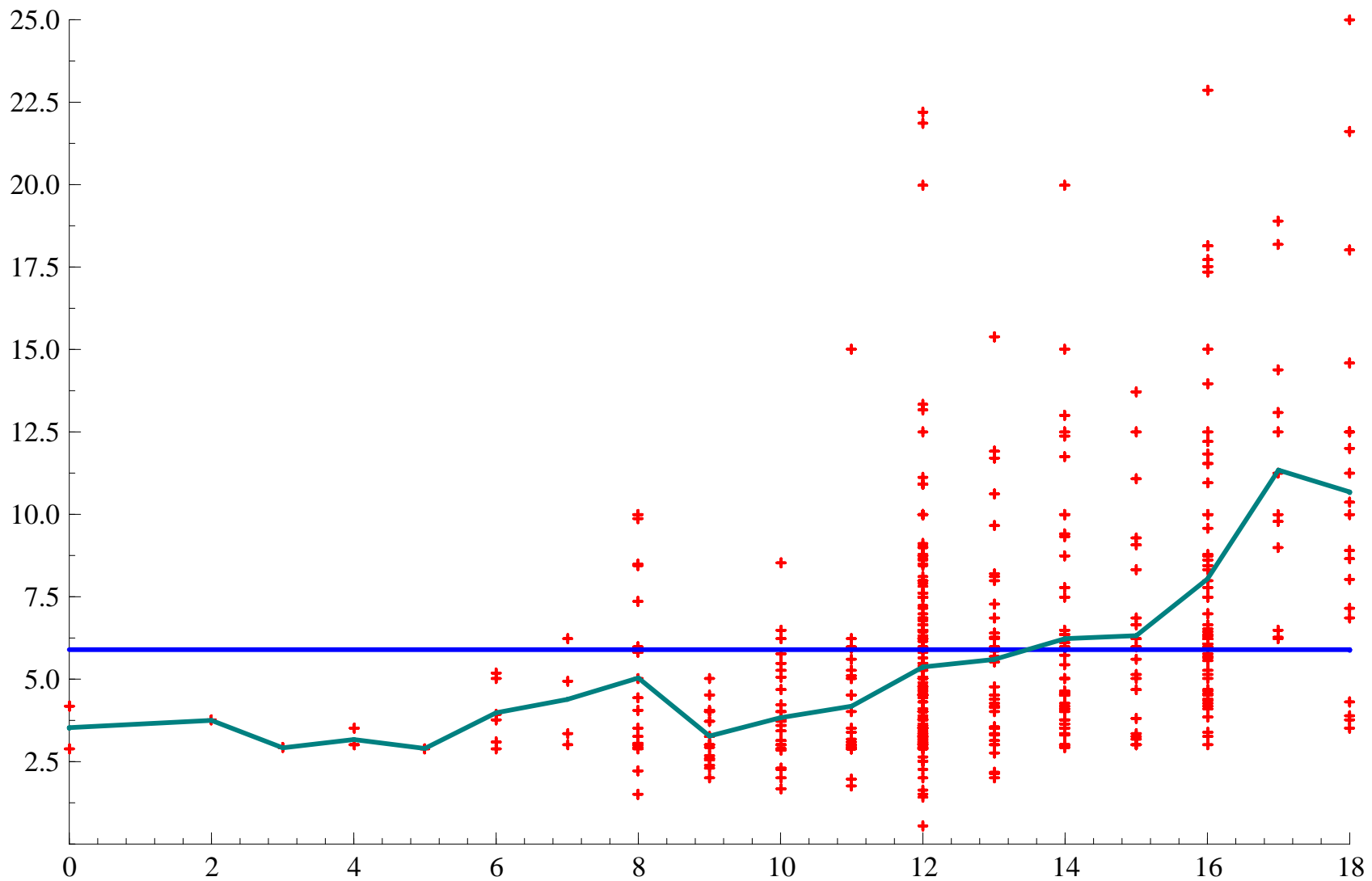
- We will be especially interested in the mean of y given x : $E(y|x)$.
- Moreover, we will generally assume that $E(y|x)$ is a linear function of x .



Hourly wage vs years of education (US data)



Estimated wage densities for 12, 16, and 18 years of education



Hourly wage vs years of education (US data)

- So, essentially, we will:
 - specify a **linear** relationship between 2 (or more) variables;
 - estimate the parameters (coefficients) of the model from a sample of data;
 - conduct hypothesis tests using the estimated model;
 - compare it with alternatives;
 - produce forecasts into the future.

2) Notation and model specification

- We shall define the following variables and parameters:

y	dependent variable;
x_1, \dots, x_K	independent/explanatory variables, or regressors;
β_1, \dots, β_K	parameters/coefficients of interest;
ϵ	random disturbance.

In the above:

y, x_1, \dots, x_K	observable;
β_1, \dots, β_K	unknown (need to be estimated);
ϵ	unobservable.

- We will be concerned with **single equation** models of the form:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where i indexes observations and n denotes sample size.

- Notice that the model has to be linear on the parameters, but not on the regressors.
- **Example:** set $K = 3$ and let $x_{i1} = 1$ and $x_{i3} = x_{i2}^2$ for all $i = 1, \dots, n$, then

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i2}^2 + \epsilon_i, \quad i = 1, \dots, n.$$

This regression model includes an **intercept/constant** β_1 , and a quadratic term.

- It is convenient to express the model more concisely in **matrix notation**. For that, we begin by writing the model for each observation:

$$\begin{aligned}
 y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_K x_{1K} + \epsilon_1, \\
 y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_K x_{2K} + \epsilon_2, \\
 &\vdots \\
 y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_K x_{nK} + \epsilon_n.
 \end{aligned}$$

- Now group (stack) observations on each variable into $n \times 1$ vectors:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}, \quad (j = 1, \dots, K), \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

- The model can then be written as

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \dots + \beta_K \mathbf{x}_K + \boldsymbol{\epsilon}.$$

- Now place the \mathbf{x}_j vectors next to each other to form an $n \times K$ matrix \mathbf{X} and place the β_j into a $K \times 1$ column vector β :

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_K] = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}.$$

- Then the model can be written as

$$\begin{array}{ccccccc} \mathbf{y} & = & \mathbf{X} & \beta & + & \epsilon. & \\ (n \times 1) & & (n \times K) & (K \times 1) & & (n \times 1) & \end{array} \quad (2)$$

- Another useful and frequently used expression for a single observation is

$$y_i = \mathbf{x}'_i \beta + \epsilon_i, \quad i = 1, \dots, n, \quad (3)$$

where $\mathbf{x}'_i = [x_{i1}, \dots, x_{ij}, \dots, x_{iK}]$ denotes the i 'th **row** of \mathbf{X} .

3) Model assumptions

In order to complete the specification of the (Neo)CLRM we need some **assumptions**.

1. Linearity

$E(y_i|\mathbf{X}) = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK}$. This implies that $E(\epsilon_i|\mathbf{X}) = 0$ for all $i = 1, \dots, n$ so that $E(\epsilon) = \mathbf{0}$ (an $n \times 1$ vector of zeros).

2. Full rank

The matrix \mathbf{X} has rank K . This ensures the columns of \mathbf{X} are linearly independent i.e., the probability that $\mathbf{x}'_i \lambda = \mathbf{0}$ is smaller than 1 for any fixed $\lambda \neq \mathbf{0}$.

3. Data Generation

We can use $\{(y_i, \mathbf{x}_i) : i = 1 \dots n\}$, a random sample of size n of the population of interest.

- Actually, the sample of \mathbf{x}_i does not have to be representative of the population of interest, all we need is that the observations of y_i are representative, conditionally on the observed \mathbf{x}_i .

4. Homoskedasticity and nonautocorrelation

$E(\epsilon\epsilon'|\mathbf{X}) = \sigma^2 I_n$ where I_n is the $n \times n$ identity matrix. Written more fully this means

$$E(\epsilon\epsilon'|\mathbf{X}) = \begin{bmatrix} \text{var}(\epsilon_1|\mathbf{X}) & \text{cov}(\epsilon_1, \epsilon_2|\mathbf{X}) & \dots & \text{cov}(\epsilon_1, \epsilon_n|\mathbf{X}) \\ \text{cov}(\epsilon_2, \epsilon_1|\mathbf{X}) & \text{var}(\epsilon_2|\mathbf{X}) & \dots & \text{cov}(\epsilon_2, \epsilon_n|\mathbf{X}) \\ \vdots & \vdots & & \vdots \\ \text{cov}(\epsilon_n, \epsilon_1|\mathbf{X}) & \text{cov}(\epsilon_n, \epsilon_2|\mathbf{X}) & \dots & \text{var}(\epsilon_n|\mathbf{X}) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}.$$

Hence $\text{var}(\epsilon_i|\mathbf{X}) = \sigma^2$ for all i (constant variance, or **homoskedasticity**) and $\text{cov}(\epsilon_i, \epsilon_j|\mathbf{X}) = 0$ for $i \neq j$ (**nonautocorrelation**).

5. Normality

Conditionally on \mathbf{X} , the disturbances are jointly normally distributed, with mean and variance as above: $\epsilon|\mathbf{X} \sim N(0, \sigma^2 I_n)$. **We will soon drop this assumption.**

4) Estimation by ordinary least squares (OLS)

- Since β is unknown, how can we estimate it?
- A clue is given by an important property of the conditional expectation.
- $E(y|x)$ is the function of x that minimizes

$$E \left[(y - h(x))^2 \right]. \quad (4)$$

- So, the **Analogy principle** suggests that β should be estimated by minimizing the sample analog of equation (4).
- Let $\mathbf{b}_0 = [b_{01}, b_{02}, \dots, b_{0K}]'$ be an arbitrary set of estimates of the elements of $\beta = [\beta_1, \dots, \beta_K]'$. Inserting these into (1) yields a set of **residuals**:

$$e_{0i} = y_i - b_{01}x_{i1} - b_{02}x_{i2} - \dots - b_{0K}x_{iK} = y_i - \mathbf{x}'_i \mathbf{b}_0, \quad i = 1, \dots, n. \quad (5)$$

- The least squares method chooses the estimate of β so as to minimise the sum of squared residuals over all observations i.e.,

$$\mathbf{b} = \arg \min_{\mathbf{b}_0} S(\mathbf{b}_0)$$

where $S(\mathbf{b}_0)$ is the sample analog of (4) and is defined by

$$S(\mathbf{b}_0) = \sum_{i=1}^n e_{0i}^2 = \sum_{i=1}^n (y_i - b_{01}x_{i1} - b_{02}x_{i2} - \dots - b_{0K}x_{iK})^2. \quad (6)$$

- This is a simple unconstrained optimisation problem, so that \mathbf{b} is obtained by:
 - (a) differentiating (6) with respect to each b_{0j} ;
 - (b) solving the K equations (first-order conditions) simultaneously.

- Let $\mathbf{e}_0 = (e_{01}, e_{02}, \dots, e_{0n})'$, so that $\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{e}_0$. Then

$$\sum_{i=1}^n e_{0i}^2 = \mathbf{e}_0' \mathbf{e}_0 = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)' (\mathbf{y} - \mathbf{X}\mathbf{b}_0)$$

and so the least squares criterion function is:

$$S(\mathbf{b}_0) = \mathbf{e}_0' \mathbf{e}_0 = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)' (\mathbf{y} - \mathbf{X}\mathbf{b}_0). \quad (7)$$

- To do the minimisation, expand $S(\mathbf{b}_0)$ to obtain:

$$S(\mathbf{b}_0) = \mathbf{y}'\mathbf{y} - \mathbf{b}_0' \mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0' \mathbf{X}'\mathbf{X}\mathbf{b}_0.$$

- But $\mathbf{b}_0' \mathbf{X}'\mathbf{y}$ is a scalar and hence $\mathbf{b}_0' \mathbf{X}'\mathbf{y} = (\mathbf{y}'\mathbf{X}\mathbf{b}_0)'$ so that

$$S(\mathbf{b}_0) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}_0' \mathbf{X}'\mathbf{y} + \mathbf{b}_0' \mathbf{X}'\mathbf{X}\mathbf{b}_0. \quad (8)$$

- Differentiating (8) with respect to \mathbf{b}_0 gives a $K \times 1$ vector:

$$\frac{\partial S(\mathbf{b}_0)}{\partial \mathbf{b}_0} = \begin{pmatrix} \partial S(\mathbf{b}_0)/\partial b_{01} \\ \partial S(\mathbf{b}_0)/\partial b_{02} \\ \vdots \\ \partial S(\mathbf{b}_0)/\partial b_{0K} \end{pmatrix}$$

which can be written as:

$$\frac{\partial S(\mathbf{b}_0)}{\partial \mathbf{b}_0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_0 = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) = -2\mathbf{X}'\mathbf{e}_0. \quad (9)$$

- The vector \mathbf{b} minimizes $S(\mathbf{b}_0)$ and therefore sets this derivative equal to a $K \times 1$ vector of zeros. That is

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0$$

- Thus we require

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0 \Rightarrow \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b},$$

and hence

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (10)$$

- Assumption 2 (full rank) ensures that the matrix $\mathbf{X}'\mathbf{X}$ is invertible and, hence, that the OLS estimator \mathbf{b} can be obtained.
- To check that (10) is the minimum of $S(\mathbf{b}_0)$, differentiate (9) again:

$$\frac{\partial^2 S(\mathbf{b}_0)}{\partial \mathbf{b}_0 \partial \mathbf{b}_0'} = 2\mathbf{X}'\mathbf{X},$$

which is **positive definite** under Assumption 2 and hence \mathbf{b} corresponds to a minimum.

University week: 3
The CLRM (continued)

Outline

1. Review;
2. Algebraic results;
3. Goodness-of-fit;
4. Statistical properties of OLS;
5. Estimation of the error variance.

Reading: Greene: chapters 3 and 4.

1) Review

- **Assessment:** the overall mark is EITHER 50% Mid term mark, 50% Exam mark OR 100% Exam mark, whichever is greater.
- The (Neo)CLRM: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

Assumptions: A1. Linearity

$$E(y_i|\mathbf{X}) = \mathbf{x}'_i\boldsymbol{\beta}$$

A2. Full rank

$$\text{rank}(\mathbf{X}) = K$$

A3. Random Sample

$\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ is a random sample of the population of interest.

A4. Spherical disturbances $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}) = \sigma^2 I_n$

A5. Normality $\boldsymbol{\epsilon}|\mathbf{X} \sim N(0, \sigma^2 I_n)$

- Ordinary least squares (OLS): $\mathbf{b} = \arg \min_{\mathbf{b}_0} S(\mathbf{b}_0)$ where

$$S(\mathbf{b}_0) = \mathbf{e}'_0 \mathbf{e}_0 = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0) \rightarrow \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

2) Algebraic results

- What are the elements of the $K \times K$ matrix $\mathbf{X}'\mathbf{X}$ and $K \times 1$ vector $\mathbf{X}'\mathbf{y}$?
- Recall that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K]$. Then:

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_K \end{bmatrix} \mathbf{y} = \begin{bmatrix} \mathbf{x}'_1\mathbf{y} \\ \mathbf{x}'_2\mathbf{y} \\ \vdots \\ \mathbf{x}'_K\mathbf{y} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{iK}y_i \end{bmatrix};$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_K \end{bmatrix} [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K] = \begin{bmatrix} \mathbf{x}'_1\mathbf{x}_1 & \mathbf{x}'_1\mathbf{x}_2 & \dots & \mathbf{x}'_1\mathbf{x}_K \\ \mathbf{x}'_2\mathbf{x}_1 & \mathbf{x}'_2\mathbf{x}_2 & \dots & \mathbf{x}'_2\mathbf{x}_K \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{x}'_K\mathbf{x}_1 & \mathbf{x}'_K\mathbf{x}_2 & \dots & \mathbf{x}'_K\mathbf{x}_K \end{bmatrix};$$

- That is:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{iK} \\ \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{iK} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^n x_{iK}x_{i1} & \sum_{i=1}^n x_{iK}x_{i2} & \cdots & \sum_{i=1}^n x_{iK}^2 \end{bmatrix}.$$

- Hence \mathbf{b} can be computed from knowledge of the sample sums of squares and cross-products – the raw data are not needed.
- NB: if $x_{i1} = 1$ for all $i = 1, \dots, n$ then $\mathbf{x}'_1 \mathbf{y} = \sum y_i$, $\mathbf{x}'_1 \mathbf{x}_1 = n$ and $\mathbf{x}'_1 \mathbf{x}_k = \sum x_{ik}$ for $k = 2, \dots, K$, where \sum denotes $\sum_{i=1}^n$

- Recall that the model is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

- Now, define the $n \times 1$ vector of **predicted/fitted values** of \mathbf{y} as:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$$

- Also, define the $n \times 1$ vector of OLS **residuals** as:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}.$$

- Then, the estimated model is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}. \tag{11}$$

- Therefore: $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$ (actual = fitted + residual).

- Note that

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}, \quad (12)$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the **projection matrix** such that $\mathbf{P}\mathbf{X} = \mathbf{X}$.

- Also,

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{y} - \mathbf{P}\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{y} = \mathbf{M}\mathbf{y}. \quad (13)$$

where $\mathbf{M} = \mathbf{I} - \mathbf{P}$.

- The matrices \mathbf{P} and \mathbf{M} have special properties:

(a) **symmetric**: $\mathbf{M} = \mathbf{M}'$, $\mathbf{P} = \mathbf{P}'$;

(b) **idempotent**: $\mathbf{M}^2 = \mathbf{M}$, $\mathbf{P}^2 = \mathbf{P}$;

(c) **orthogonal**: $\mathbf{P}\mathbf{M} = \mathbf{P}'\mathbf{M} = \mathbf{M}'\mathbf{P} = \mathbf{M}\mathbf{P} = \mathbf{0}$.

3) Goodness-of-fit

- How well does the estimated model fit the data?
- Attempt to measure this in terms of the proportion of the variation in y explained by the model.
- We use

$$\begin{aligned} R^2 &= 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \\ &= 1 - \frac{RSS}{TSS} \quad \begin{array}{l} \leftarrow \text{residual sum of squares} \\ \leftarrow \text{total sum of squares} \end{array} \end{aligned} \tag{14}$$

where $\bar{y} = \sum y_i/n$ is the sample mean.

- Note that $0 \leq R^2 \leq 1$.
- $R^2 = 0 \Rightarrow \sum e_i^2 = \sum (y_i - \bar{y})^2$ so the model is as good as the mean!
- $R^2 = 1 \Rightarrow RSS = 0$ or $e_i = 0$ for all $i = 1, \dots, n$, and so there is a perfect fit.
- NB: do not use R^2 computed in this way for models that do not contain an intercept!
- The R^2 can also be defined as: $[\text{corr}(y_i, \hat{y}_i)]^2$, and this is well defined even if the model has no constant, but in these models it cannot be interpreted as the explained proportion of the variance.

- Because RSS will never rise (and will typically fall) by adding more regressors, it is possible to inflate the R^2 just by using more irrelevant regressors.
- It has been suggested that goodness-of-fit should be measured by the **adjusted** R^2 , or \bar{R}^2 :

$$\bar{R}^2 = 1 - \frac{RSS/(n - K)}{TSS/(n - 1)} = 1 - (1 - R^2) \frac{(n - 1)}{(n - K)}. \quad (15)$$

- \bar{R}^2 incurs a penalty if adding more regressors (increasing K) does not significantly reduce the RSS .
- Too much emphasis is often placed on R^2 and \bar{R}^2 – other aspects of the model are much more important.
- Notice that none of the assumptions requires the model to have a good fit!

4) Statistical properties of OLS

- **Linearity (A2):** The OLS vector \mathbf{b} is a **linear** function of \mathbf{y}

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

– Furthermore, using **(A1)**, it is also a linear function of the unobservable random vector ϵ

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon) \quad (\text{because } \mathbf{y} = \mathbf{X}\beta + \epsilon) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon \\ &= \underline{\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon} \quad (\text{because } (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = I)\end{aligned}\tag{16}$$

- **Expectation (A1–A3):** Recall that $E(\mathbf{b}) = E_{\mathbf{X}} [E[\mathbf{b}|\mathbf{X}]]$ and notice that

$$E[\mathbf{b}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}[\mathbf{y}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta.$$

Then, $E(\mathbf{b}) = E_{\mathbf{X}} [\beta] = \beta$, i.e., the OLS estimator \mathbf{b} is **unbiased**.

- Recall from (16) that $\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$.
- **Variance (A1–A4):** The conditional covariance matrix of \mathbf{b} is:

$$\begin{aligned}
\text{var}(\mathbf{b}|\mathbf{X}) &= E[(\mathbf{b} - \beta)(\mathbf{b} - \beta)'|\mathbf{X}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\epsilon\epsilon'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \text{ since } E(\epsilon\epsilon'|\mathbf{X}) = \sigma^2 I_n \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

– Although it is not very important, we note that $\text{var}(\mathbf{b}) = \sigma^2 E_{\mathbf{X}} [(\mathbf{X}'\mathbf{X})^{-1}]$.

- **Normality (A1–A5):** From (16), it follows that

$$\mathbf{b}|\mathbf{X} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

- Clearly OLS is a **linear unbiased estimator** (LUE) of β .
- But how does OLS compare to other LUEs?

Gauss-Markov theorem: Under Assumptions 1–4 (note that Assumption 5 is not needed for this), the OLS estimator \mathbf{b} of β is the **best linear unbiased estimator** (BLUE) in the sense that it has minimum variance within the class of LUEs.

- What does this mean? Take any other LUE, call it \mathbf{b}_1 . Then

$$\text{var}(\mathbf{b}_1|\mathbf{X}) \geq \text{var}(\mathbf{b}|\mathbf{X})$$

in the sense that the matrix $\text{var}(\mathbf{b}_1|\mathbf{X}) - \text{var}(\mathbf{b}|\mathbf{X})$ is positive semi-definite.

5) Estimation of error variance

- In order to use the distribution of \mathbf{b} to conduct hypothesis tests, we need an estimate of σ^2 , which is the variance of the unobservable disturbances ϵ_i .
- If the disturbances ϵ_i were observed, we could estimate their variance from the sum of their squares.
- The sample analogs of the ϵ_i are the residuals e_i , and we shall base our estimator on the sum of their squares.
- Recall from (13) that

$$\mathbf{e} = \mathbf{M}\mathbf{y}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ has the property that $\mathbf{M}\mathbf{X} = \mathbf{0}$.

- Substitute $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for \mathbf{y} :

$$\mathbf{e} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{M}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\epsilon} = \mathbf{M}\boldsymbol{\epsilon}. \quad (17)$$

- Then, defining $\mathbf{M} = [m_{ij}]$, the *RSS* can be written as

$$RSS = \mathbf{e}'\mathbf{e} = \boldsymbol{\epsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon} = \sum_{i=1}^n \sum_{j=1}^n \varepsilon_i \varepsilon_j m_{ij},$$

because $\mathbf{M}' = \mathbf{M}$ and $\mathbf{M}^2 = \mathbf{M}$.

- The expectation of the *RSS* is:

$$E(\mathbf{e}'\mathbf{e}|\mathbf{X}) = \sum_{i=1}^n \sum_{j=1}^n m_{ij} E(\varepsilon_i \varepsilon_j | \mathbf{X})$$

and because by **(A4)** $E(\varepsilon_i \varepsilon_j | \mathbf{X}) = 0$ for $i \neq j$, we have that

$$E(\mathbf{e}'\mathbf{e}|\mathbf{X}) = \sum_{i=1}^n m_{ii} E(\varepsilon_i \varepsilon_i | \mathbf{X}) = E(\varepsilon_i \varepsilon_i | \mathbf{X}) \sum_{i=1}^n m_{ii} = \sigma^2 \text{tr}(\mathbf{M})$$

- But what is $\text{tr}(\mathbf{M})$?

- Using the properties of $\text{tr}(\cdot)$, we have that:

$$\begin{aligned}\text{tr}(\mathbf{M}) &= \text{tr}(I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \text{tr}(I_n) - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= \text{tr}(I_n) - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] \\ &= \text{tr}(I_n) - \text{tr}[I_K] \\ &= n - K.\end{aligned}$$

- Therefore $E(\mathbf{e}'\mathbf{e}) = \sigma^2(n - K)$ and so:

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K} = \frac{\sum e_i^2}{n - K} \quad (18)$$

is an unbiased estimator of σ^2 because

$$E(s^2) = \frac{E(\mathbf{e}'\mathbf{e})}{n - K} = \sigma^2.$$

University week: 4

Inference in the (Neo)CLRM

Outline

1. Review;
2. Standard errors and t-tests;
3. Tests of multiple linear restrictions (F-tests).

Reading: Greene: chapter 5.

NOTE: Matrices and vectors will no longer appear in bold typeface. Also, often we will not make explicit that the distribution is conditional on X .

1) Review

- Model: $y = X\beta + \epsilon$.
- OLS estimator: $b = (X'X)^{-1} X'y$.
- $E(b) = \beta$ and hence b is unbiased.
- $\text{var}(b|X) = \sigma^2 (X'X)^{-1}$.
- b is minimum variance (Gauss-Markov Theorem).
- OLS is BLUE (Best Linear Unbiased Estimator).

2) Standard errors and t-tests

- Recall that, under normality (Assumption 5), we have

$$b|X \sim N[\beta, \sigma^2 (X'X)^{-1}]$$

- For each element of b , b_k , we have

$$b_k|X \sim N[\beta_k, \sigma^2 S_{kk}], \quad k = 1, \dots, K, \quad (19)$$

where S_{kk} is the k 'th diagonal element of $(X'X)^{-1}$.

- If σ^2 were known then

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S_{kk}}} \sim N(0, 1),$$

which could be used to test hypotheses.

- Testing hypotheses is important because the omission of relevant variables, or the inclusion of irrelevant ones, affects the properties of the estimator. **See Problem Set 3, Question 4.**

- So to test

$$H_0 : \beta_k = \beta_k^* \quad \text{NULL HYPOTHESIS}$$

$$\text{against } H_1 : \beta_k \neq \beta_k^* \quad \text{ALTERNATIVE HYPOTHESIS}$$

we would use

$$z_k^* = \frac{b_k - \beta_k^*}{\sqrt{\sigma^2 S_{kk}}} \sim N(0, 1) \quad \text{under } H_0.$$

- Let \bar{z} denote the critical value (more on this shortly) for the $N(0, 1)$.

Decision rule: if $|z_k^*| \geq \bar{z}$ reject H_0 ;

if $|z_k^*| < \bar{z}$ do not reject H_0 .

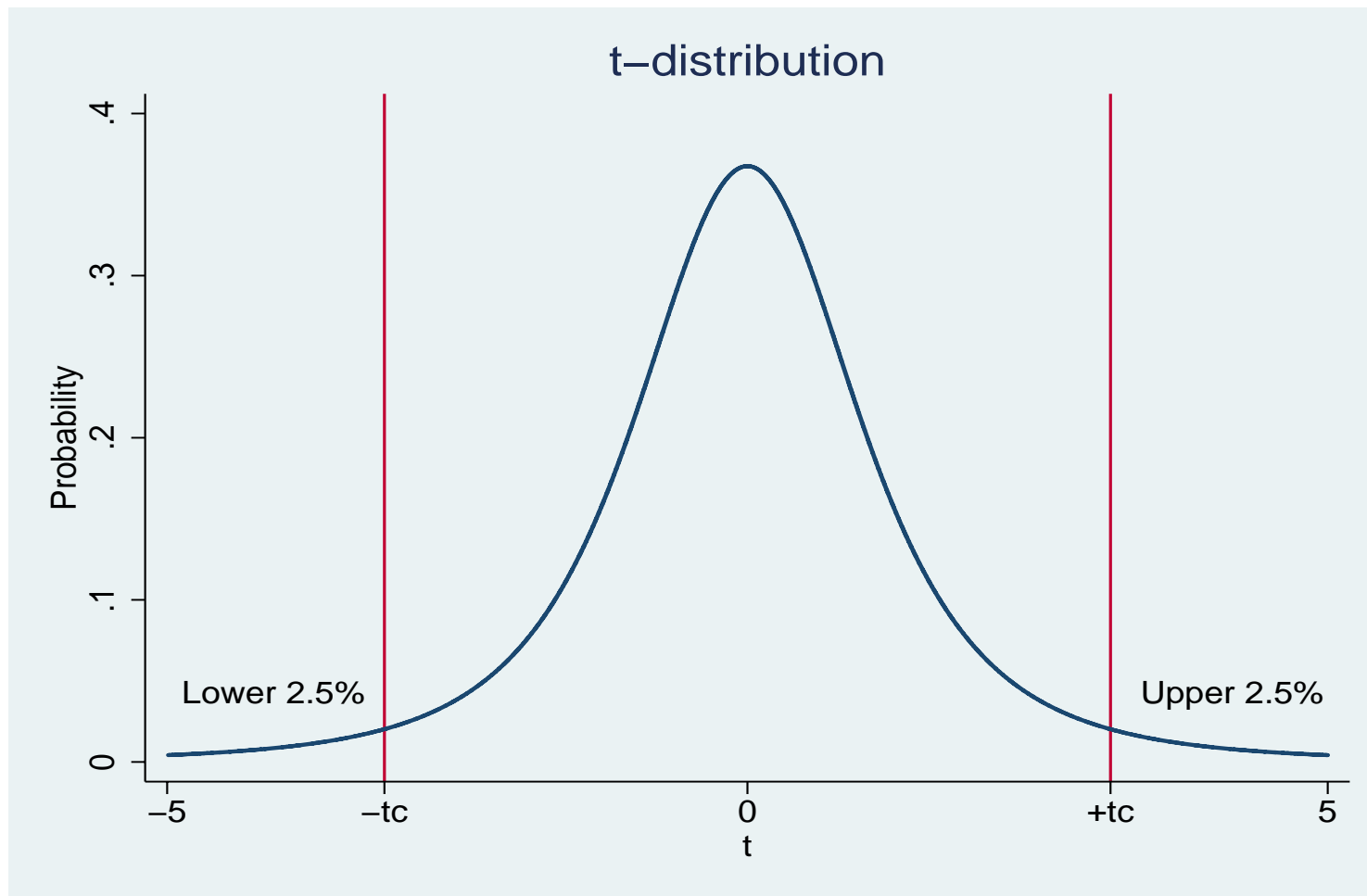
- **But** σ^2 is not known!

- Hence we use s^2 in its place and the new test statistic has a different distribution:

$$t_k = \frac{b_k - \beta_k^*}{\sqrt{s^2 S_{kk}}} \sim t_{n-K} \text{ under } H_0.$$

- The t-distribution with $n - K$ degrees of freedom has fatter tails than the $N(0, 1)$ due to the imprecision associated with using the estimator s^2 rather than σ^2 .
- The t-distribution approaches the standard normal as $n - K \rightarrow \infty$.

- We need to choose a **significance level** for the test: let's conduct the test at the 5% level of significance.
- We can find a number $t_{n-K}^{0.025}$ ($= tc$) such that 2.5% of the t_{n-K} distribution lies above it.
- Then, 5% of the t_{n-K} distribution lies outside the interval $[-tc, tc]$.



- If H_0 is really true, there is only a 5% chance of obtaining a t_k value outside $[-tc, tc]$.
- If this occurs, we regard it as evidence **against** H_0 .
- The **decision rule** is:
 - if $|t_k| \geq tc$, reject H_0 ;
 - if $|t_k| < tc$, do not reject H_0 .
- The the interval $[-tc, tc]$ is, therefore, called the **acceptance region**.
- The area outside this interval is the **rejection region**.

- A common hypothesis to test is

$$H_0 : \beta_k = 0 \text{ against } H_1 : \beta_k \neq 0.$$

- Under H_0 , **given the other regressors**, the conditional mean of y does not depend on x_k .
- That is to say, given the other regressors, x_k does not help to predict y .
- The test statistic is

$$t_k = \frac{b_k}{\sqrt{s^2 S_{kk}}} = \frac{\text{estimate}}{\text{standard error}},$$

which is routinely computed in regression software (e.g., Stata).

3) Tests of multiple linear restrictions (F-tests)

- What if we want to jointly test more than one restriction?
- As an example, consider the two models

$$y = X_1\beta_1 + \epsilon, \quad (20a)$$

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon, \quad (20b)$$

$$\begin{array}{lll} y : n \times 1 & X_1 : n \times K_1 & \beta_1 : K_1 \times 1 \\ \epsilon : n \times 1 & X_2 : n \times K_2 & \beta_2 : K_2 \times 1 \end{array}$$

- Model (20a) is obtained from (20b) by setting the K_2 elements of the vector β_2 equal to zero.

- In this case, the hypothesis of interest is:

$$H_0 : \beta_2 = 0 \text{ (} K_2 \text{ restrictions) against } H_1 : \beta_2 \neq 0.$$

- This involves a test of more than one restriction so we cannot use the simple t-test (we can test the restrictions individually but this says nothing about their **joint** significance).
- We can write (20b) as

$$y = [X_1 : X_2] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon = X\beta + \epsilon \quad (21)$$

and so the restrictions in H_0 are:

$$[0 : I_{K_2}] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = 0 \text{ or } R\beta = 0.$$

- This framework can be easily generalized.
- Consider Model (21) and suppose we want to test the set of J linear restrictions:

$$H_0 : R\beta - q = 0 \text{ against } H_1 : R\beta - q \neq 0$$

where R is $J \times K$ and q is $J \times 1$.

- For example, suppose $K = 3$ and we wish to test $\beta_1 + \beta_2 = 0$ and $\beta_1 - 2\beta_2 + \beta_3 = 1$.
- Here there are 2 restrictions ($J = 2$) and we have

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

- We will now study **two methods** of testing H_0 .

- **Method 1** only needs the unrestricted estimator $b = (X'X)^{-1}X'y$.
- We know that (under normality)

$$b \sim N[\beta, \sigma^2(X'X)^{-1}]$$

$$Rb \sim N[R\beta, \sigma^2R(X'X)^{-1}R']$$

$$Rb - q \sim N[R\beta - q, \sigma^2R(X'X)^{-1}R'].$$

- Under H_0 we know that $R\beta - q = 0$ and hence

$$Rb - q \sim N[0, \sigma^2R(X'X)^{-1}R'] \text{ under } H_0.$$

Theorem: If $z \sim N(\mu, V)$ ($q \times 1$) then $(z - \mu)'V^{-1}(z - \mu) \sim \chi_q^2$.

- Hence

$$W_1 = (Rb - q)' [\sigma^2 R(X'X)^{-1}R']^{-1} (Rb - q) \sim \chi_J^2.$$

- **But** σ^2 is unknown and so this distribution cannot be used.
- However, consider $W_2 = e'e/\sigma^2$.
- We know that $e = M\epsilon$ where $M = I - X(X'X)^{-1}X'$ is symmetric idempotent.

Theorem: If $x \sim N(0, \sigma^2 I)$ and A is idempotent of rank r , then $\frac{x'Ax}{\sigma^2} \sim \chi_r^2$.

- Here, $\epsilon \sim N(0, \sigma^2 I)$ under Assumption 5, and $e'e = \epsilon'M\epsilon$, so

$$W_2 = \frac{e'e}{\sigma^2} \sim \chi_{n-K}^2$$

because, for a symmetric idempotent matrix, $\text{rank}(M) = \text{tr}(M) = n - K$.

- We have therefore found two (independent) random variables, W_1 and W_2 , each with χ^2 distributions.
- To test the hypothesis of interest we need yet another theorem.

Theorem: If $c_1 \sim \chi_{n_1}^2$, $c_2 \sim \chi_{n_2}^2$ and c_1 and c_2 are independent, then $\frac{c_1}{c_2} \times \frac{n_2}{n_1} \sim F_{n_1, n_2}$.

- Using this theorem we find that

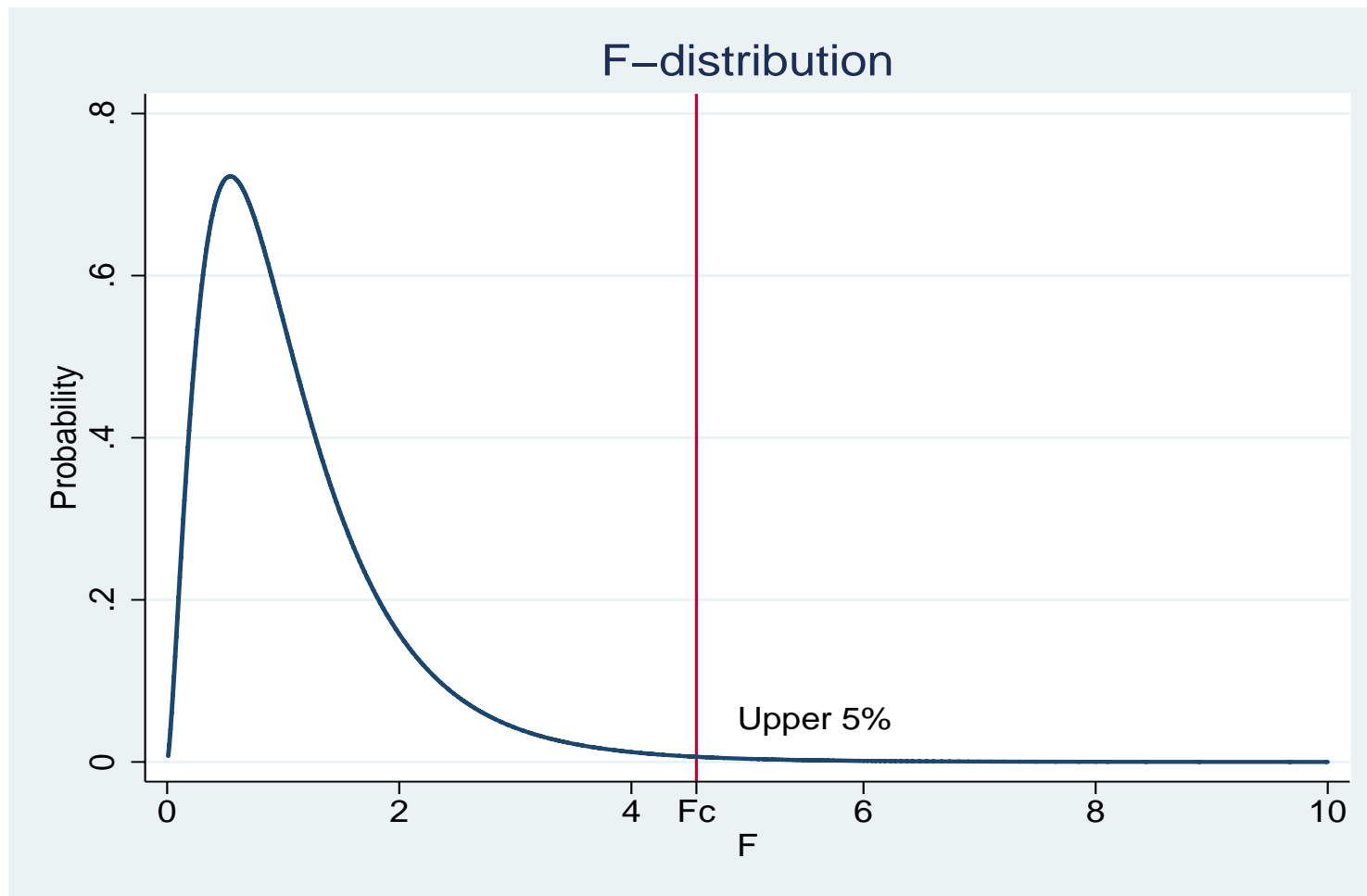
$$F = \frac{W_1}{W_2} \times \frac{(n - K)}{J} \sim F_{J, n-K} \text{ under } H_0.$$

- Written more fully,

$$F = \frac{(Rb - q)' [\sigma^2 R(X'X)^{-1} R']^{-1} (Rb - q)}{e'e/\sigma^2} \cdot \frac{(n - K)}{J}$$

$$= (Rb - q)' [s^2 R(X'X)^{-1} R']^{-1} (Rb - q) / J.$$

- Let $F_{J,n-K}^{0.05}$ ($= Fc$) denote the 5% critical value from the $F_{J,n-K}$ distribution.
- The decision rule is:
 - if $F \geq Fc$, reject H_0 ;
 - if $F < Fc$, do not reject H_0 .



- **Method 2** proceeds in 4 steps and involves estimation with and without the restrictions imposed:

1. Estimate $y = X\beta + \epsilon$ and obtain $S = e'e$.
2. Impose the restrictions, estimate the restricted model, and obtain $S_R = e'_R e_R$, where e_R is the $n \times 1$ vector of residuals from the restricted model.
3. Compute

$$F = \left(\frac{S_R - S}{S} \right) \left(\frac{n - K}{J} \right).$$

4. Under H_0 , $F \sim F_{J, n-K}$ and therefore the decision rule is as in Method 1.

● Points to note:

1. $F > 0$.
2. J = number of restrictions: degrees of freedom for numerator;
 $n - K$ = degrees of freedom for denominator;
 K = number of regressors in **unrestricted** model.
3. Example of imposing restrictions: consider the model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad H_0 : \beta_1 + \beta_2 = 1$$

$$\Rightarrow \beta_2 = 1 - \beta_1$$

$$\Rightarrow y_i = \beta_1 x_{i1} + (1 - \beta_1) x_{i2} + \epsilon_i$$

$$\Rightarrow (y_i - x_{i2}) = \beta_1 (x_{i1} - x_{i2}) + \epsilon_i$$

i.e., simply regress $y_i^* = y_i - x_{i2}$ on $x_i^* = x_{i1} - x_{i2}$.

Example of STATA output

Number of obs = 35
F(2, 32) = 30523.24
Prob > F = 0.0000
R-squared = 0.9995
Adj R-squared = 0.9994
Root MSE = .02102

ly	Coef.	Std. Err.	t	P> t
lx2	.9816755	.0464554	21.13	0.000
lx3	.0313708	.0402388	0.78	0.441
_cons	.0020919	.0058119	0.36	0.721

University week: 5

Large sample methods

Outline

1. Review;
2. Large sample concepts;
3. The method of maximum likelihood.
4. Large sample hypothesis tests;

Reading: Greene: chapter 14 and Appendix D.

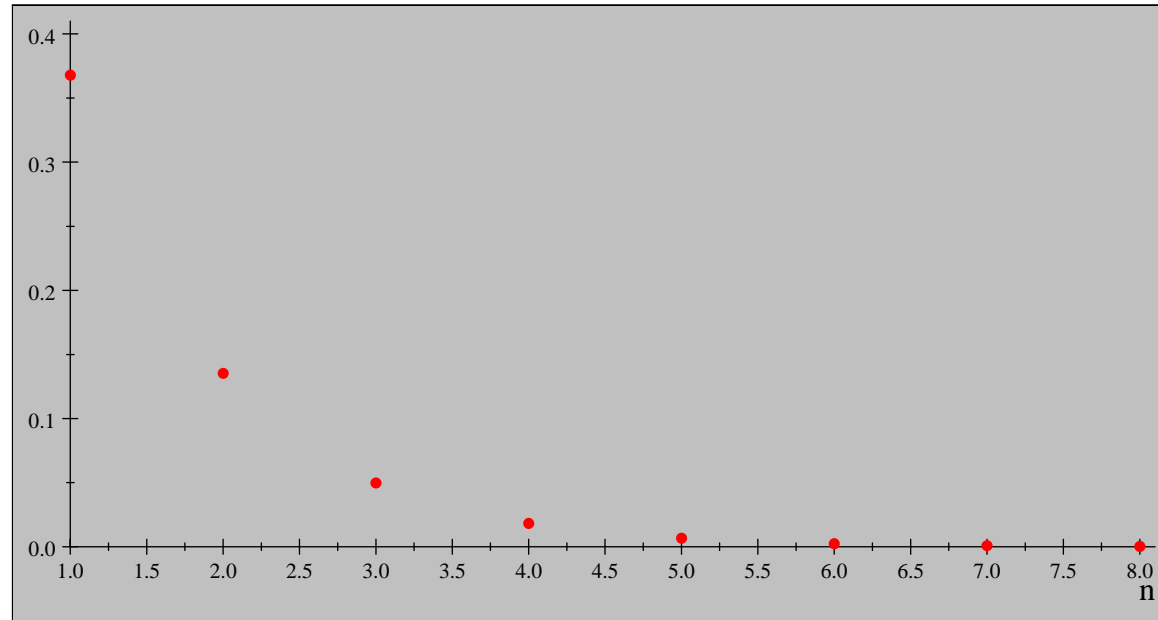
1) Review

- (Neo)CLRM: $y = X\beta + \epsilon$.
- The OLS estimator, $b = (X'X)^{-1} X'y$, is BLUE.
- Hypothesis tests based on exact distributions: t- and F-statistics.
- Exact distributions rely on normality assumption.
- Next week we will see what happens when we relax the assumption of normality.
- To do this we shall use large sample (asymptotic) methods.
- Today we will introduce important large sample concepts and explore even further the implications of the normality assumption.

2) Large sample concepts

- If we wish to relax some of the assumptions of the (Neo)CLRM, then exact finite sample results are typically not available.
- For example, if we relax the normality assumption, t-statistics no longer have t-distributions, F-statistics no longer have F-distributions, etc.
- Hence the critical values from these distributions are not correct, and incorrect inferences may be drawn from the tests.
- However, for large enough n these results hold **approximately**.
- We therefore use large sample methods to find out the properties of estimators and test statistics as $n \rightarrow \infty$.
- But if large samples are available and the normality assumption holds, then we can also use additional results.

- Consider a sequence of numbers indexed by n , e.g., $\{x_n = e^{-n}\} = \left\{\frac{1}{e}, \frac{1}{e^2}, \frac{1}{e^3}, \dots, \frac{1}{e^n}, \dots\right\}$.



- We can define the **limit** of this sequence as $n \rightarrow \infty$:

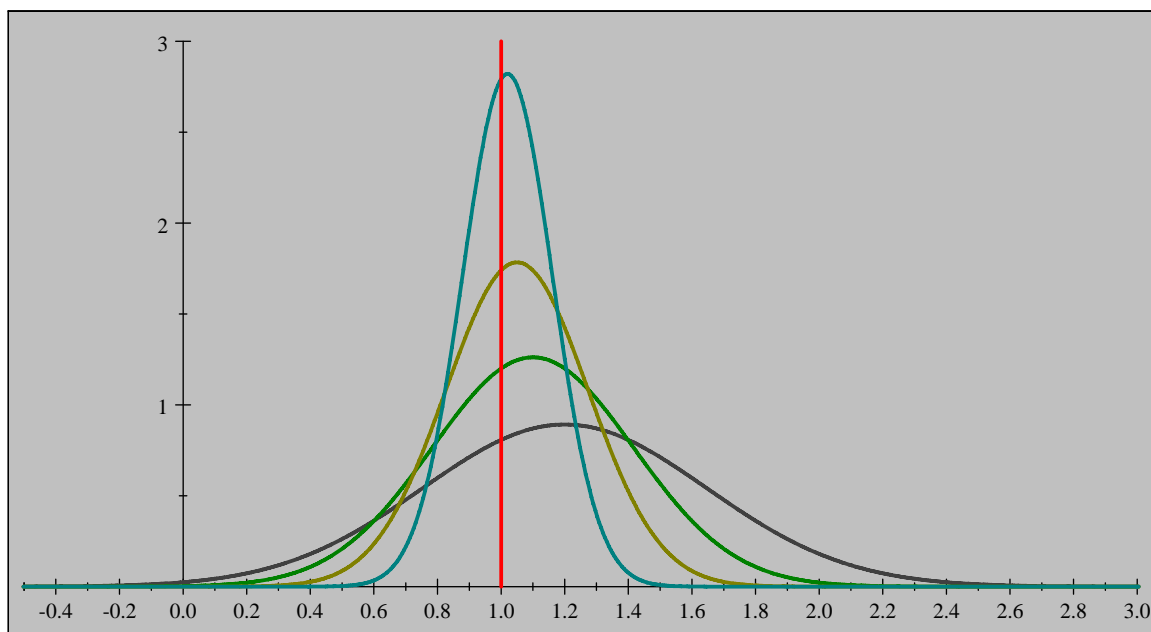
$$\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} e^{-n} = 0.$$

- The sequence $\{x_n\}$ is said to **converge** to zero.

- What happens if the elements are random variables?
- The sequence of random variables $\{x_n\}$ **converges in probability** to a constant c if

$$\lim_{n \rightarrow \infty} \Pr(|x_n - c| > \epsilon) = 0 \text{ for any } \epsilon > 0.$$

- This is written as $x_n \xrightarrow{p} c$ or $\text{plim } x_n = c$.
- In words: there exists a positive number ϵ such that, as n gets larger and larger, the probability that the distance between x_n and c is larger than ϵ converges to zero.



- If $\text{plim } b = \beta$ then b is said to be a **consistent** estimator of β .
- A useful property of the plim operator is:

Slutsky's Theorem: If $h(\cdot)$ is a continuous function and $\text{plim } x_n = c$, then

$$\text{plim } h(x_n) = h(\text{plim } x_n) = h(c).$$

- This property is **not** shared by the expectations operator: in general, $E[h(x)] \neq h[E(x)]$ for a random variable x (this is called **Jensen's inequality**).
- Another useful result is:

If x_n is a random sequence such that $\lim_{n \rightarrow \infty} E(x_n) = c$ and $\lim_{n \rightarrow \infty} \text{var}(x_n) = 0$, then $\text{plim } x_n = c$ (see Theorem D-1 on p. 1107 of Greene).

- When they exist, this enables us to establish consistency by examining the limiting properties of the expectation and variance.

- We are also interested in the distribution of random variables.
- Suppose $F_n(\cdot)$, the distribution function of x_n , converges to a distribution function $F(\cdot)$ as $n \rightarrow \infty$; then $F(\cdot)$ is the **limiting distribution** of x_n .
- If x is a random variable having distribution function $F(\cdot)$, then x_n is said to **converge in distribution** to x .
 - For example, if $x \sim N(0, \sigma^2)$ and x_n converges in distribution to x , then we write

$$x_n \xrightarrow{d} x \sim N(0, \sigma^2) \quad \text{or} \quad x_n \xrightarrow{d} N(0, \sigma^2).$$

- A useful result concerning convergence in distribution is:

Cramer's Theorem: If A_n is a matrix sequence such that $\text{plim } A_n = A$, and b_n is a vector sequence such that $b_n \xrightarrow{d} b \sim N(0, Q)$, then

$$A_n b_n \xrightarrow{d} Ab \sim N(0, AQA').$$

- This is will be useful in studying the limiting distribution of the OLS estimator.

3) The method of maximum likelihood

- Consider the classical model

$$y_i = x_i' \beta + \epsilon_i, \quad i = 1, \dots, n; \quad \epsilon_i | X \sim NID(0, \sigma^2).$$

- **NB:** ‘ $NID(0, \sigma^2)$ ’ means ‘normally and independently distributed with mean zero and variance σ^2 ’.
- The normality assumption means that the conditional probability density function (pdf) for ϵ_i is

$$f(\epsilon_i | X, \beta, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{\epsilon_i^2}{2\sigma^2} \right\}, \quad -\infty < \epsilon_i < \infty.$$

- The independence of the ϵ_i means the joint pdf for the $n \times 1$ vector ϵ is:

$$f(\epsilon | X, \beta, \sigma^2) = \prod_{i=1}^n f(\epsilon_i | X) = \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left\{ -\frac{\sum_{i=1}^n \epsilon_i^2}{2\sigma^2} \right\}.$$

- Because $y = X\beta + \epsilon$ and $\sum \epsilon_i^2 = \epsilon'\epsilon = (y - X\beta)'(y - X\beta)$ we obtain the joint pdf for y :

$$f(y|X, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} \right\}. \quad (22)$$

- This is a function of y for **given** X , β and σ^2 .
- **But** in econometrics we need to estimate β and σ^2 for **given** y and X .
- The method of **maximum likelihood** takes the probability density in (22) and chooses the values of β and σ^2 which are most likely to have generated the observed y .
- When regarded as a function of β and σ^2 for given y , the function in (22) is called the **likelihood function** $L(\beta, \sigma^2; y, X)$:

$$L(\beta, \sigma^2; y, X) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} \right\} \quad (23)$$

- We need to maximise (23) with respect to β and σ^2 .
- It is easiest to take logs and maximize the log-likelihood function:

$$\ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{S(\beta)}{2\sigma^2}, \quad (24)$$

where $S(\beta) = (y - X\beta)'(y - X\beta)$ is the familiar sum of squares function.

- The first-derivatives of the log-likelihood function are:

$$\frac{\partial \ln L}{\partial \beta} = -\frac{\partial S(\beta)}{\partial \beta} \cdot \frac{1}{2\sigma^2}; \quad (25a)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{S(\beta)}{2\sigma^4}. \quad (25b)$$

- Clearly, from (25a) and (25b):

$$\hat{\beta}_{ML} = b = (X'X)^{-1}X'y,$$

$$\hat{\sigma}_{ML}^2 = \frac{(y - X\hat{\beta}_{ML})'(y - X\hat{\beta}_{ML})}{n} = \frac{\sum e_i^2}{n} \neq s^2.$$

- For “well behaved” (regular) problems, the MLE $\hat{\theta}$ has the following properties:

1. **Consistency:** $\text{plim } \hat{\theta} = \theta_0$.
2. **Asymptotic normality:** $\hat{\theta} \stackrel{a}{\sim} N(\theta_0, I(\theta_0)^{-1})$, where $I(\theta) = -E \left[\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right]$ is the **information matrix**.
3. **Asymptotic efficiency:** $\hat{\theta}$ achieves the Cramer-Rao lower bound for consistent estimators, meaning that in general, for a consistent estimator $\tilde{\theta}$, $\text{var}(\tilde{\theta}) \geq I(\theta_0)^{-1}$.

- Although we have concentrated on the (Neo)CLRM, maximum likelihood can be applied in a wide variety of problems, provided we can write down the density function!
- Consider now a general estimation problem where we want to estimate an $m \times 1$ parameter vector θ whose true value is θ_0 (in the previous case, $\theta = (\beta, \sigma^2)$).
- Let

$$g(\theta)_{(m \times 1)} = \frac{\partial \ln L(\theta)}{\partial \theta},$$

$$H(\theta)_{(m \times m)} = \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'}.$$

- For regular problems, the maximum likelihood estimator (MLE) satisfies $g(\hat{\theta}) = 0$.

4) Large sample hypothesis tests

- Sometimes we want to test hypotheses involving **nonlinear** restrictions of the form

$$H_0 : c(\theta) = 0 \text{ against } H_1 : c(\theta) \neq 0 \quad (26)$$

where θ is an $m \times 1$ vector of parameters and the function $c : \mathbb{R}^m \rightarrow \mathbb{R}^J$ ($J \leq m$) i.e., $c(\theta)$ is $J \times 1$.

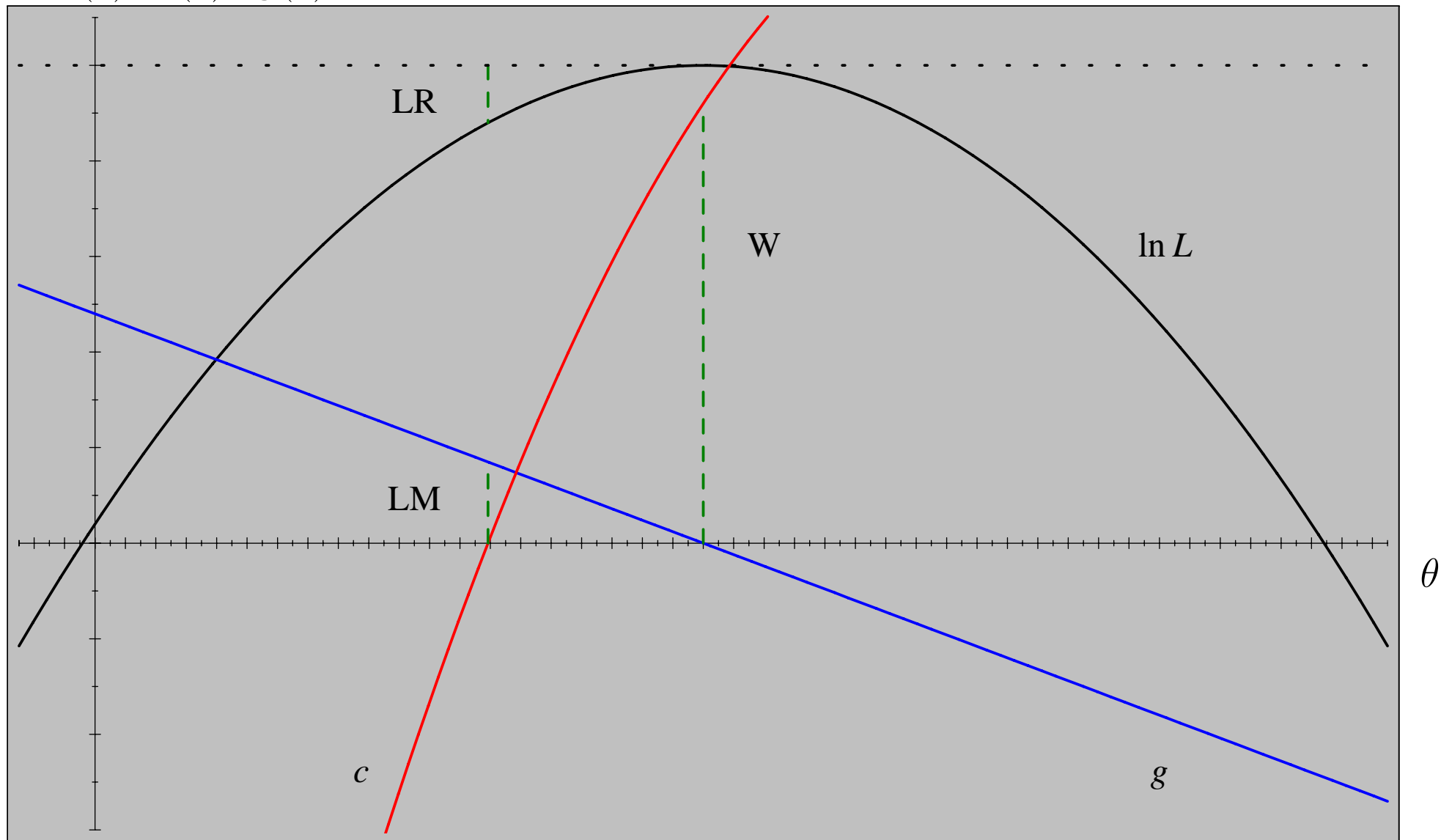
- Linear restrictions are a special case: $c(\theta) = R\theta - q$.
- Let $\hat{\theta}$ be the unrestricted MLE and $\hat{\theta}_R$ be the restricted MLE:

$$\hat{\theta} = \arg \max_{\theta} L(\theta); \quad \hat{\theta}_R = \arg \max_{\theta} L(\theta) \text{ s.t. } c(\theta) = 0,$$

where $L(\theta)$ is the likelihood function.

- There are three large sample tests based on $L(\theta)$.

$\ln L(\theta), c(\theta), g(\theta)$



- **Wald Test:** Based on unrestricted estimator $\hat{\theta}$:

$$W = c(\hat{\theta})' \left[C(\hat{\theta}) I(\hat{\theta})^{-1} C(\hat{\theta})' \right]^{-1} c(\hat{\theta}) \xrightarrow{d} \chi_J^2 \quad (27)$$

under H_0 as $n \rightarrow \infty$, where

$$C(\theta)_{(J \times m)} = \frac{\partial c(\theta)}{\partial \theta'}$$

and $I(\theta)$, as before, denotes the information matrix

$$I(\theta)_{(m \times m)} = -E [H(\theta)]$$

- Both t- and F-tests are special cases of Wald tests.
- Easy to use when restrictions are difficult to impose on the model.

- **Likelihood Ratio Test:** Based on both $\hat{\theta}$ and $\hat{\theta}_R$:

$$LR = 2 \left[\ln L(\hat{\theta}) - \ln L(\hat{\theta}_R) \right] \xrightarrow{d} \chi_J^2$$

under H_0 as $n \rightarrow \infty$.

- Sometimes written as $LR = -2 \ln \lambda$, where $\lambda = L(\hat{\theta}_R)/L(\hat{\theta})$.
- Very easy to compute when $\hat{\theta}$ and $\hat{\theta}_R$ are available.
- **Lagrange Multiplier (score) Test:** Based on $\hat{\theta}_R$:

$$LM = g(\hat{\theta}_R)' \left[I(\hat{\theta}_R) \right]^{-1} g(\hat{\theta}_R) \xrightarrow{d} \chi_J^2$$

under H_0 as $n \rightarrow \infty$, where (as before)

$$g(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta}.$$

- Easy to use when it is difficult to estimate unrestricted model.

- Which test should be used, and when?
- These tests are asymptotically equivalent under the null and are only asymptotically valid.
- The LM test is often interesting when we want to check for departures from the maintained assumptions as we do not have to fully specify the alternative (more on this later).
- The Wald test is easy to modify so that it is valid even if some assumptions of the model do not hold (more on this later).
- Note that different inferences can be drawn from testing the same hypothesis with the different tests. In the CLRM, for example, it can be shown that

$$LM \leq LR \leq W.$$

- In such circumstances, if LM rejects H_0 , then so will LR and W , while if W does not reject H_0 , then neither will LM or LR .

University week: 6

Large sample methods (continued)

Outline

1. Review;
2. OLS in large samples: Relaxing normality;
3. Instrumental variables estimation.

Reading: Greene: chapters 4 and 8.

1) Review

- Model: $y = X\beta + \epsilon$.
- OLS estimator is BLUE under classical assumptions.
- Large sample concepts: consistency and limiting distribution.
- MLE under normality: equivalent to OLS in classical model.

2) OLS in large samples: Relaxing normality

- Consider the model

$$y_i = x_i' \beta + \epsilon_i; \quad \epsilon_i | X \sim IID(0, \sigma^2).$$

- **NB:** ‘ $IID(0, \sigma^2)$ ’ means ‘independently and identically distributed with mean zero and variance σ^2 ’.
- In this model $E(b) = \beta$ and $var(b) = \sigma^2(X'X)^{-1}$ as usual.
- It is typically assumed that (this is questionable for time series data)

$$\text{plim} \frac{X'X}{n} = Q_{xx} \quad (K \times K \text{ nonsingular}), \quad (28)$$

i.e. $\text{plim} \sum_{i=1}^n x_{ij}x_{ik}/n = q_{jk}$.

- Is b a consistent estimator of β ?
- Recall that from (16) we have

$$b = \beta + (X'X)^{-1} X'\epsilon.$$

- Therefore, by Slutsky's theorem:

$$\begin{aligned} \text{plim}(b) &= \beta + \text{plim} \left(\frac{1}{n} X'X \right)^{-1} \text{plim} \left(\frac{1}{n} X'\epsilon \right) \\ &= \beta + Q_{xx}^{-1} \text{plim} \left(\frac{1}{n} X'\epsilon \right). \end{aligned}$$

- Therefore, for b to be consistent for β , we need $\text{plim} \left(\frac{1}{n} X'\epsilon \right) = 0$.
- Because $E \left(\frac{1}{n} X'\epsilon \right) = 0$, we just need to show that

$$\lim_{n \rightarrow \infty} \text{var} \left(\frac{1}{n} X'\epsilon \right) = 0.$$

- By definition

$$\begin{aligned} \text{var} \left(\frac{1}{n} X' \epsilon | X \right) &= \frac{1}{n^2} E(X' \epsilon \epsilon' X | X) \\ &= \frac{1}{n^2} X' E(\epsilon \epsilon' | X) X \\ &= \frac{\sigma^2}{n^2} X' X = \frac{\sigma^2}{n} \frac{X' X}{n} \end{aligned}$$

- Hence

$$\text{var} \left(\frac{1}{n} X' \epsilon \right) = \frac{\sigma^2}{n} E \left(\frac{X' X}{n} \right).$$

- For “well behaved” data, $E \left(\frac{X' X}{n} \right)$ is a finite constant and therefore

$$\lim_{n \rightarrow \infty} \text{var} \left(\frac{1}{n} X' \epsilon \right) = 0 \times \lim_{n \rightarrow \infty} E \left(\frac{X' X}{n} \right) = 0.$$

- This shows that b , the OLS estimator of β , is consistent.

- What about the large sample distribution of b ?
- Recall the definition of Q_{xx} in (28) and note that under our assumptions a Central Limit Theorem leads to

$$\frac{1}{\sqrt{n}}X'\epsilon \xrightarrow{d} N(0, \sigma^2 Q_{xx}). \quad (29)$$

- Using again $b = \beta + (X'X)^{-1} X'\epsilon$, we obtain

$$\sqrt{n}(b - \beta) = \left(\frac{X'X}{n} \right)^{-1} \frac{1}{\sqrt{n}}X'\epsilon.$$

- Then, from Cramer's Theorem

$$\sqrt{n}(b - \beta) \xrightarrow{d} N(0, \sigma^2 Q_{xx}^{-1} Q_{xx} Q_{xx}^{-1}) = N(0, \sigma^2 Q_{xx}^{-1}). \quad (30)$$

- We can use this result to justify using the normal distribution in large (but finite) samples:

$$(b - \beta) \stackrel{a}{\sim} N(0, \sigma^2 n^{-1} Q_{xx}^{-1}) \quad \Rightarrow \quad b \stackrel{a}{\sim} N(\beta, \sigma^2 (X'X)^{-1}).$$

3) Instrumental variables estimation

- Recall that $E(y_i|\mathbf{X}) = x_i'\beta$ implies that $E(\epsilon_i|\mathbf{X}) = 0$, with $\epsilon_i = y_i - x_i'\beta$.
- However, often economists are interested in models that are not conditional expectations.
- A simple example involves the consumption function in a closed economy:

$$C_i = \beta_1 + \beta_2 Y_i + \epsilon_i,$$

$$Y_i = C_i + I_i + G_i$$

$$= \beta_1 + \beta_2 Y_i + \epsilon_i + I_i + G_i = \frac{1}{1 - \beta_2} (\beta_1 + I_i + G_i + \epsilon_i).$$

- So, Y_i , the regressor in the consumption function, is correlated with (in fact, depends on) ϵ_i , the disturbance.
- Because $E(\epsilon_i|Y_i) \neq 0$ (Y_i is “endogenous”), β_1 and β_2 cannot be parameters of a conditional expectation.

- Consider the model

$$y = X\beta + \epsilon, \quad \epsilon_i = IID(0, \sigma^2) \quad i = 1, \dots, n. \quad (31)$$

- We have examined the properties of the OLS estimator in this model assuming

$$\text{plim} \left(\frac{1}{n} X'X \right) = Q_{xx} \text{ (nonsingular);}$$

$$\text{plim} \left(\frac{1}{n} X'\epsilon \right) = 0.$$

- Because $b = \beta + (X'X/n)^{-1}(X'\epsilon/n)$ we find that $\text{plim } b = \beta$.

- When the elements of X are correlated with ϵ in the limit, we have

$$\text{plim} \left(\frac{1}{n} X'\epsilon \right) = \gamma \neq 0.$$

- Therefore, $\text{plim}(b) = \beta + Q_{xx}^{-1}\gamma \neq \beta$, and so b is no longer consistent.

- How can we construct a consistent estimator of β ?
- Suppose we can find a set of L **instrumental variables**, the observations being contained in the $n \times L$ matrix Z .
- We require Z to satisfy the following properties:
 - (a) $\text{plim} (n^{-1} Z' Z) = Q_{zz}$ (positive definite);
 - (b) $\text{plim} (n^{-1} Z' X) = Q_{zx}$ ($L \times K$, rank K): “ Z correlated with X ;”
 - (c) $\text{plim} (n^{-1} Z' \epsilon) = 0$: “ Z uncorrelated with ϵ .”
- Where do instruments come from?
- Recall that we require Z to be uncorrelated with ϵ and correlated with X .
- Reconsider the consumption function example:

$$C_i = \beta_1 + \beta_2 Y_i + \epsilon_i,$$

$$Y_i = C_i + I_i + G_i.$$

- If I_i and G_i are exogenous then they would be suitable instruments because they are uncorrelated with ϵ_i but determine Y_i (and hence are correlated with Y_i).
- In the above example, the complete model is known and the choice of instruments is obvious.
- But often it is infeasible and/or impractical to specify a large set of equations when we are really only interested in one.
- The **general rule** is that instruments must be variables that affect y_i only through their effect on the endogenous regressors they instrument.
- The IV estimator may be obtained in two steps:
 1. Regress the variables in X on those in Z and obtain the fitted values, \hat{X} ;
 2. Regress y on \hat{X} to obtain b_{IV} , the IV estimator of β .

Let's look at these two steps in more detail.

● **STEP 1:**

– We have $X = [x_1, \dots, x_j, \dots, x_K]$.

– Regressing each x_k ($k = 1, \dots, K$) on Z is equivalent to estimating the equation

$$x_k = Za_k + u_k,$$

the estimator for which is $\hat{a}_k = (Z'Z)^{-1}Z'x_k$ ($L \times 1$).

– The fitted values are

$$\hat{x}_k = Z\hat{a}_k = Z(Z'Z)^{-1}Z'x_k = P_Zx_k \quad (n \times 1),$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix for Z .

– Hence

$$\hat{X} = [\hat{x}_1, \dots, \hat{x}_j, \dots, \hat{x}_K] = [P_Zx_1, \dots, P_Zx_j, \dots, P_Zx_K] = P_ZX.$$

• **STEP 2:**

– Write $X = \hat{X} + u$ and substitute into (31) to get

$$y = (\hat{X} + u)\beta + \epsilon = \hat{X}\beta + (\epsilon + u\beta).$$

– We now estimate $y = \hat{X}\beta + \epsilon$:

$$\begin{aligned} b_{IV} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y \\ &= (X'P_Z'P_ZX)^{-1}X'P_Z'y \\ &= (X'P_ZX)^{-1}X'P_Zy \\ &= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y. \end{aligned} \tag{32}$$

- To show that b_{IV} is consistent, substitute (31) into (32) to obtain

$$\begin{aligned}
 b_{IV} &= \beta + [X'Z(Z'Z)^{-1}Z'X]^{-1} X'Z(Z'Z)^{-1}Z'\epsilon \\
 &= \beta + \left[\frac{X'Z}{n} \left(\frac{Z'Z}{n} \right)^{-1} \frac{Z'X}{n} \right]^{-1} \frac{X'Z}{n} \left(\frac{Z'Z}{n} \right)^{-1} \frac{Z'\epsilon}{n}.
 \end{aligned}$$

- Taking the probability limit, we obtain

$$\text{plim}(b_{IV}) = \beta + [Q'_{zx}Q^{-1}_{zz}Q_{zx}]^{-1}Q'_{zx}Q^{-1}_{zz} \times 0,$$

so that $\text{plim}(b_{IV}) = \beta$ and b_{IV} is a consistent estimator of β .

- Under appropriate conditions ensuring that

$$\frac{1}{\sqrt{n}}Z'\epsilon \xrightarrow{d} N(0, \sigma^2 Q_{zz})$$

we find that, as $n \rightarrow \infty$:

$$\begin{aligned} \sqrt{n}(b_{IV} - \beta) &= \left[\frac{X'Z}{n} \left(\frac{Z'Z}{n} \right)^{-1} \frac{Z'X}{n} \right]^{-1} \frac{X'Z}{n} \left(\frac{Z'Z}{n} \right)^{-1} \frac{Z'\epsilon}{\sqrt{n}} \\ &\xrightarrow{d} N(0, \sigma^2 (Q'_{zx} Q_{zz}^{-1} Q_{zx})^{-1}) \end{aligned}$$

- Hence for large (but finite) sample size n

$$b_{IV} \overset{a}{\sim} N\left(\beta, \sigma^2 (\hat{X}'\hat{X})^{-1}\right). \quad (33)$$

- The distribution in (33) cannot be used in practice because σ^2 is unknown.

- We can estimate σ^2 using

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n - K} \text{ where } \hat{\epsilon} = y - Xb_{IV}.$$

- Note that the appropriate residuals are $y - Xb_{IV}$ and **not** $y - \hat{X}b_{IV}$.
- The IVE can be obtained from the minimisation problem:

$$b_{IV} = \arg \min_{\beta} S_{IV}(\beta) \text{ where } S_{IV}(\beta) = (y - X\beta)'P_Z(y - X\beta).$$

University week: 7

The generalised linear regression model; heteroskedasticity

Outline

1. Review;
2. The generalised linear regression model;
3. Properties of OLS in this context;
4. The generalized least squares estimator;
5. Heteroskedasticity: testing and modelling.

Reading: Greene: chapter 9.

1) Review

- Model: $y = X\beta + \epsilon$.
- OLS estimator: BLUE under classical assumptions.
- MLE: equivalent to OLS in classical model, efficient.
- Relax assumptions: OLS consistent and asymptotically normal (need other assumptions).
- IV: consistent under regressor-error correlation.
- What happens if $E(\epsilon\epsilon'|X) \neq \sigma^2 I$?

2) The generalised linear regression model

- Consider the (generalized) linear regression model

$$y = X\beta + \epsilon, \quad E(\epsilon|X) = 0, \quad (34)$$

but where we now assume that

$$E(\epsilon\epsilon'|X) = \sigma^2\Omega,$$

where Ω is an $n \times n$ symmetric positive definite matrix.

- The diagonal elements of Ω need not be the same, and the off-diagonal elements need not be zero.
- The (Neo)CLRM results when $\Omega = I_n$.
- This model captures the two main cases of **non-spherical disturbances**:

- **Heteroskedasticity:** here $var(\epsilon_i|X) = \sigma_i^2$ while $cov(\epsilon_i, \epsilon_j|X) = E(\epsilon_i\epsilon_j|X) = 0$ for $i \neq j$, so that

$$E(\epsilon\epsilon'|X) = \sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_{11} & 0 & \dots & 0 \\ 0 & \omega_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} = \Sigma,$$

where $\sigma_i^2 = \sigma^2\omega_{ii}$ ($i = 1, \dots, n$).

- **Autocorrelation (serial correlation):** In this case $var(\epsilon_i|X) = \sigma^2$ while $cov(\epsilon_i, \epsilon_j|X) = E(\epsilon_i\epsilon_j|X) = \sigma^2\rho_{|i-j|}$ for $i \neq j$, so that

$$E(\epsilon\epsilon'|X) = \sigma^2\Omega = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-2} & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-3} & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \dots & \rho_{n-4} & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{n-2} & \rho_{n-3} & \rho_{n-4} & \dots & 1 & \rho_1 \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & \rho_1 & 1 \end{bmatrix}.$$

3) Properties of OLS in this context

- Recall that $b = \beta + (X'X)^{-1}X'\epsilon$ so that $E(b) = \beta$ if $E(\epsilon|X) = 0$, i.e., b is still unbiased.
- Under very general conditions, b is also consistent.
- However, $\text{var}(b|X) = E[(b - \beta)(b - \beta)'|X]$ has the form

$$\begin{aligned}\text{var}(b|X) &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] \\ &= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1},\end{aligned}$$

which is typically not equal to $\sigma^2(X'X)^{-1}$.

- Hence inferences based on the usual t- and F-tests are invalid and may be misleading.
- Furthermore, OLS may not be BLUE, i.e., we may be able to find a LUE with a smaller variance than OLS.

4) The generalized least squares estimator

- Because Ω is positive definite we can find a matrix P ($n \times n$) such that

$$P'P = \Omega^{-1} \text{ and therefore } P^{-1}(P')^{-1} = \Omega.$$

- Note that $P\Omega P' = I_n$, and that P is **not** the projection matrix we have seen before.
- Now we pre-multiply the model (34) by P :

$$Py = PX\beta + P\epsilon \text{ or } y_* = X_*\beta + \epsilon_*. \quad (35)$$

- Then, if $E(\epsilon_*|X_*) = 0$, the transformed model satisfies the (Neo)CLRM assumptions:

$$E(\epsilon_*\epsilon_*'|X_*) = E(P\epsilon\epsilon'P'|X_*) = PE(\epsilon\epsilon'|X_*)P' = \sigma^2 P\Omega P' = \sigma^2 I_n.$$

- Hence OLS applied to (35) is BLUE.

- This is the generalised least squares (GLS) estimator:

$$\begin{aligned}\hat{\beta} &= (X_*' X_*)^{-1} X_*' y_* \\ &= (X' P' P X)^{-1} X' P' P y \\ &= (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y.\end{aligned}$$

- Alternatively, we can define the GLS estimator as $\hat{\beta} = \arg \min_{\beta} S_*(\beta)$, with

$$S_*(\beta) = (y - X\beta)' \Omega^{-1} (y - X\beta) = \epsilon_*' \epsilon_*$$

- What are the properties of this estimator?

- Using $y = X\beta + \epsilon$ we obtain $\hat{\beta} = \beta + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\epsilon$.

$$E(\hat{\beta}) = \beta \text{ if } E(\epsilon_*|X_*) = 0$$

$$\text{var}(\hat{\beta}|X) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X\right]$$

$$= E\left[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\epsilon\epsilon'\Omega^{-1}X(X'\Omega^{-1}X)^{-1}|X\right]$$

$$= \sigma^2(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1}X)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1} = \sigma^2(X_*'X_*)^{-1}.$$

- The GLS estimator is the BLUE for this model, and hence

$$\text{var}(b) - \text{var}(\hat{\beta}) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} - \sigma^2(X'\Omega^{-1}X)^{-1}$$

is a positive semi-definite matrix.

- Finally, $\hat{\beta}$ is a consistent estimator of β if

$$\text{plim} \frac{X' \Omega^{-1} X}{n} = Q_* \quad \text{and} \quad \text{plim} \frac{X' \Omega^{-1} \epsilon}{n} = 0.$$

- For t- and F-tests we need an estimate of σ^2 , and can use

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}' \hat{\epsilon}}{n - K},$$

where $\hat{\epsilon} = y_* - X_* \hat{\beta}$.

- The estimated covariance matrix of $\hat{\beta}$ is

$$\hat{\sigma}^2 (X' \Omega^{-1} X)^{-1} = \hat{\sigma}^2 (X_*' X_*)^{-1}.$$

- To test the J linear restrictions $H_0: R\beta = q$, we can use an F-test based on the transformed model. Let $\hat{\beta}_c$ denote the constrained estimator of β under H_0 . Then we can use

$$F = \frac{S_*(\hat{\beta}_c) - S_*(\hat{\beta})}{S_*(\hat{\beta})} \cdot \frac{n - K}{J} \sim F_{J, n-K} \quad \text{under } H_0.$$

- Despite being BLUE, the GLS estimator assumes that Ω is known, but typically this is not true in practice.
- The **feasible GLS** (FGLS) estimator is based on an estimator of Ω , denoted $\hat{\Omega}$, and is given by (Greene uses $\hat{\beta}$ for $\hat{\beta}_F$)

$$\hat{\beta}_F = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y.$$

- $\hat{\beta}_F$ and $\hat{\beta}$ are asymptotically equivalent if

$$\text{plim} \left[\left(\frac{X'\hat{\Omega}^{-1}X}{n} \right) - \left(\frac{X'\Omega^{-1}X}{n} \right) \right] = 0,$$

$$\text{plim} \left[\left(\frac{X'\hat{\Omega}^{-1}\epsilon}{\sqrt{n}} \right) - \left(\frac{X'\Omega^{-1}\epsilon}{\sqrt{n}} \right) \right] = 0.$$

- Notice that we only need consistency of $\hat{\Omega}$ for $\hat{\beta}_F$ to be equivalent to GLS.

5) Heteroskedasticity: testing and modelling

- Consider the heteroskedastic linear regression model

$$y_i = x_i' \beta + \epsilon_i, \quad E(\epsilon_i | X) = 0, \quad E(\epsilon_i \epsilon_j | X) = \begin{cases} \sigma_i^2, & i = j, \\ 0, & i \neq j. \end{cases} \quad (36)$$

- This is a special case of the generalised linear regression model in which:

$$\sigma^2 \Omega = \sigma^2 \begin{bmatrix} \omega_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \omega_{nn} \end{bmatrix}, \quad P = \begin{bmatrix} 1/\sqrt{\omega_{11}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\sqrt{\omega_{nn}} \end{bmatrix}.$$

- Therefore, the GLS estimator is obtained as the OLS estimator in

$$Py = PX\beta + P\epsilon \quad \text{or} \quad y_* = X_*\beta + \epsilon_*.$$

- That is, $\hat{\beta}$ is the Weighted Least Squares (WLS) estimator

$$\hat{\beta} = \left(\sum_{i=1}^n \frac{x_i x_i'}{\omega_{ii}} \right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{\omega_{ii}} = \left(\sum_{i=1}^n \frac{x_i x_i'}{\sigma_i^2} \right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2}. \quad (37)$$

- **But** σ_i^2 is typically unknown, so we use the FGLS estimator

$$\hat{\beta}_F = \left(\sum_{i=1}^n \frac{x_i x_i'}{\hat{\sigma}_i^2} \right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{\hat{\sigma}_i^2}. \quad (38)$$

- The question in practice is: how do we model σ_i^2 , i.e., how do we obtain $\hat{\sigma}_i^2$?
- We can try to relate the variance to a set of observable variables, contained in a $p \times 1$ vector of functions of x_i , but it is generally difficult to know how to specify this relation.
- A possible approach is to be less ambitious and just try to approximate σ_i^2 . This can be done following these steps:
 - a) Obtain b , the OLS estimate of β , and the corresponding fitted values $\hat{y}_i = x_i' b$ and residuals $e_i = y_i - x_i' b$;
 - b) Use OLS to estimate

$$\ln(e_i^2) = \alpha_0 + \alpha_1 \hat{y}_i + \alpha_2 \hat{y}_i^2 + u_i,$$

approximate σ_i^2 with $\exp(\hat{\alpha}_0 + \hat{\alpha}_1 \hat{y}_i + \hat{\alpha}_2 \hat{y}_i^2) = \exp(\widehat{\ln(e_i^2)})$.

- c) Obtain (38) using this approximation to σ_i^2 .

- Because this method is unlikely to lead to a consistent estimator of σ_i^2 , the WLS estimator obtained in this way will not be fully efficient, but if the heteroskedasticity is strong it may be much more efficient than simple OLS.
- A major problem is that if the estimator of σ_i^2 is not consistent, the FGLS standard errors are also invalid.
- So, in order to have a practical procedure, we need to do two things:
 - (a) Find a way of computing standard errors that are valid even if there is heteroskedasticity;
 - (b) Find a way to check for the presence of heteroskedasticity and to gauge its strength.

- Consistent standard errors for OLS

- Recall that $var(b|X) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$.

- Under heteroskedasticity this is

$$var(b|X) = (X'X)^{-1} \left(\sum_{i=1}^n \sigma_i^2 x_i x_i' \right) (X'X)^{-1}.$$

- It can be shown that

$$plim \frac{1}{n} \sum_{i=1}^n \sigma_i^2 x_i x_i' = plim \frac{1}{n} \sum_{i=1}^n e_i^2 x_i x_i'.$$

- This leads to the **Eicker-White heteroskedasticity consistent estimator** of $var(b|X)$:

$$\widehat{var}(b|X) = (X'X)^{-1} \left(\sum_{i=1}^n e_i^2 x_i x_i' \right) (X'X)^{-1}.$$

- Although this can be used for inference using b , it does not make b BLUE!

- The same approach can be used to obtain valid standard errors after FGLS estimation when the estimate of σ_i^2 is not consistent.

- Testing for heteroskedasticity

- Usually, a linear regression model is estimated assuming the classical assumptions hold, and then tested to see whether these assumptions appear to be satisfied.
- Therefore, after OLS estimation, we want to test

$$H_0 : \sigma_i^2 = \sigma^2, \text{ against } H_1 : \sigma_i^2 = f(z_i, \alpha).$$

where z_i is a $p \times 1$ vector of squares and cross products of the regressors and $f(z_i, \alpha)$ is such that $f(z_i, 0) = \sigma^2$.

- So, the test for $E(\epsilon_i^2 | x_i) = \sigma^2$ can be recast as

$$H_0 : \alpha = 0 \text{ against } H_1 : \alpha \neq 0.$$

- Modifying a Lagrange Multiplier test proposed by Breusch and Pagan we can check the validity of the null without specifying $f(z_i, \alpha)$ (only mild regularity conditions are required).

- To perform the test, we regress the squared residuals e^2 on a constant and on z_i and compute the R^2 from this auxiliary regression.
- Under H_0 , we have that as $n \rightarrow \infty$

$$nR^2 \sim \chi_p^2,$$

where p is the dimension of z_i .

- Alternatively, just use an F-statistic to test that in the regression of e^2 on a constant and on z_i all slope parameters are equal to zero.
- In cases where candidate elements for z_i are not obvious, a convenient approach is to use fitted values from the initial OLS regression and its squares.
- That is, set $z_i = \begin{bmatrix} \hat{y}_i & \hat{y}_i^2 \end{bmatrix}$.
- This particular procedure is proposed by Wooldridge and uses results from the work of Breusch, Pagan, White, and Koenker.
- Finally, note that other tests for heteroskedasticity are also available.

University week: 8

Serial correlation and dynamic models

Outline

1. Review;
2. Serial correlation (autocorrelation);
3. Examples of autocorrelation processes;
4. Testing for serial correlation;
5. Estimation of dynamic models (lagged dependent variables).

Reading: Greene: chapter 20.

1) Review;

- Generalised Model: $y = X\beta + \epsilon$, $E(\epsilon\epsilon'|X) = \sigma^2\Omega$.
- GLS estimator: BLUE if Ω known.
- FGLS estimator: asymptotically equivalent to GLS if Ω estimated consistently.
- Heteroskedasticity: Ω diagonal with different elements on diagonal.
- He have seen how to test for heteroskedasticity.
- If heteroskedasticity is a problem, we can use a robust covariance matrix and WLS.
- What if $E(\epsilon\epsilon'|X)$ is not diagonal (serial correlation)?

2) Serial correlation (autocorrelation);

- We shall consider the linear regression model

$$y_t = x_t' \beta + \epsilon_t, \quad t = 1, \dots, T, \quad \text{or} \quad y = X\beta + \epsilon, \quad (39)$$

where we are using a t subscript to index observations in time, and T denotes sample size.

- We shall continue to assume that $E(\epsilon|X) = 0$ with $E(\epsilon\epsilon'|X) = \sigma^2\Omega$, where Ω is no longer diagonal.
- Notice that autocorrelation and heteroskedasticity have very different natures:
 - Heteroskedasticity is a characteristic of the population;
 - Autocorrelation is a characteristic of the sampling scheme.
- What are the elements of Ω ?

- Recall that

$$E(\epsilon\epsilon'|X) = \sigma^2\Omega = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{T-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \cdots & \gamma_0 \end{bmatrix}.$$

where the **autocovariances** are defined as $cov(\epsilon_t, \epsilon_{t-s}|X) = cov(\epsilon_{t+s}, \epsilon_t|X) = \gamma_s$.

- Note that $\gamma_0 = cov(\epsilon_t, \epsilon_t|X) = var(\epsilon_t|X) = \sigma^2$ which we assume as constant i.e., there is no heteroskedasticity.
- The autocovariances depend on the units of measurement, so we cannot tell whether a value of 0.1 is large or small without further information.
- The **autocorrelations** do not depend on the units of measurement, and are defined as

$$\rho_s = \frac{\gamma_s}{\gamma_0}.$$

- Note that $\rho_0 = 1$ and that $-1 \leq \rho_s \leq 1$ for all s .

3) Examples of autocorrelation processes;

- It is obviously impractical to try to estimate all T autocovariances $(\gamma_0, \dots, \gamma_{T-1})$ or autocorrelations from a sample of T observations.
- We therefore try to model the autocovariances (or autocorrelations) in terms of a small number of parameters.
- Autoregressive disturbances: AR(1)
 - The first-order autoregressive process, or AR(1) process, is given by

$$\epsilon_t = \rho\epsilon_{t-1} + u_t, \quad |\rho| < 1, \quad (40)$$

where

$$E(u_t|X) = 0, \quad E(u_t^2|X) = \sigma_u^2, \quad E(u_t u_s|X) = 0 \text{ for } t \neq s.$$

- What are the properties of ϵ_t ?

– By repeated backward substitution:

$$\epsilon_t = \rho(\rho\epsilon_{t-2} + u_{t-1}) + u_t$$

$$= \rho^2\epsilon_{t-2} + u_t + \rho u_{t-1}$$

$$= \rho^2(\rho\epsilon_{t-3} + u_{t-2}) + u_t + \rho u_{t-1}$$

$$= \rho^3\epsilon_{t-3} + u_t + \rho u_{t-1} + \rho^2 u_{t-2}$$

⋮

$$= \rho^s\epsilon_{t-s} + u_t + \rho u_{t-1} + \dots + \rho^{s-2}u_{t-s+2} + \rho^{s-1}u_{t-s+1}$$

$$\text{i.e., } \epsilon_t = \rho^s\epsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i} \quad (s > 0).$$

– It is sometimes convenient to let $s \rightarrow \infty$ to obtain

$$\epsilon_t = \sum_{i=0}^{\infty} \rho^i u_{t-i},$$

because $\lim_{s \rightarrow \infty} \rho^s = 0$ due to $|\rho| < 1$.

– Note that ϵ_t depends on current and all lagged values of u_t .

– It is straightforward (!) to show that:

$$E(\epsilon_t | X) = 0,$$

$$\gamma_0 = \text{var}(\epsilon_t | X) = \frac{\sigma_u^2}{1 - \rho^2},$$

$$\gamma_s = \text{cov}(\epsilon_t, \epsilon_{t-s} | X) = \frac{\rho^s \sigma_u^2}{1 - \rho^2},$$

and hence $\rho_s = \gamma_s / \gamma_0 = \rho^s$ (this notation is due to Greene!).

- Using the AR(1) process, the entire covariance matrix of ϵ is determined by only two parameters: σ_u^2 and ρ .
- It is therefore a very **parsimonious** method of capturing autocorrelation.
- The AR(1) process can be extended to the AR(p) process which includes p lags of ϵ_t :

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \dots + \rho_p \epsilon_{t-p} + u_t. \quad (41)$$

- The formulae for the autocovariances are more complicated and depend on the AR parameters ρ_1, \dots, ρ_p and σ_u^2 .
- An alternative process is the first-order moving average, or MA(1), defined by

$$\epsilon_t = u_t + \lambda u_{t-1}, \quad (42)$$

where u_t has the same properties as in the AR(1) case.

- The properties of ϵ_t are easily derived:

$$E(\epsilon_t|X) = 0,$$

$$\gamma_0 = \text{var}(\epsilon_t|X) = \sigma_u^2(1 + \lambda^2),$$

$$\gamma_1 = \text{cov}(\epsilon_t, \epsilon_{t-1}|X) = \lambda\sigma_u^2,$$

$$\gamma_s = \text{cov}(\epsilon_t, \epsilon_{t-s}|X) = 0, \quad s > 1.$$

- The MA(1) process can be generalised to the MA(q):

$$\epsilon_t = u_t + \lambda_1 u_{t-1} + \dots + \lambda_q u_{t-q}. \quad (43)$$

with $\gamma_s = 0$ for $s > q$.

- An important difference between an MA and an AR process is that for the latter the autocorrelations tail off smoothly, whereas for the MA process they abruptly become zero after q lags.

4) Testing for serial correlation;

- Usually we estimate a linear model by OLS assuming that the classical assumptions hold, and then attempt to test whether those assumptions appear to be satisfied for the estimated model.
- We therefore need to be able to test whether the assumption of zero covariance of the disturbances appears to hold, based on the estimated residuals.
- Until the 1990s, the most commonly used test for serial correlation was the **Durbin-Watson test** for first-order autocorrelation.
- Nowadays, the Lagrange Multiplier test is more popular because it can be applied in a wider set of circumstances and can test for higher-order serial correlation.

- The null and alternative hypotheses for the **Breusch-Godfrey** LM test are as follows:

H_0 : no serial correlation in ϵ_t ;

H_1 : ϵ_t is AR(p) or MA(p).

- Notice that because an LM test is used, we do not need to be specific about the nature of the serial correlation process under the alternative.
- The test statistic is most easily calculated by estimating the following auxiliary regression:

$$e_t = x_t' \gamma + \rho_1 e_{t-1} + \dots + \rho_p e_{t-p} + u_t. \quad (44)$$

- The test statistic is simply $T \times R^2$ and, under H_0 , we have as $T \rightarrow \infty$,

$$LM = TR^2 \sim \chi_p^2.$$

- Note that the R^2 is from the auxiliary regression in (44).

5) Estimation of dynamic models (lagged dependent variables).

- A dynamic model is one containing lagged values of the dependent variable y_t and/or the regressors x_t , e.g.:

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \beta_4 y_{t-1} + \epsilon_t.$$

- Typically, we reserve the use of the label "dynamic model" for models including the lagged dependent variable.
- We will now focus on

$$E(y_t | x_t, x_{t-1}, y_{t-1}) = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \beta_4 y_{t-1},$$

with $E(\epsilon_t | x_t, x_{t-1}, y_{t-1}) = 0$.

- Notice that this is different from $E(y_t | X)$, where X includes the lagged values of y_t .
- Indeed, for $t < T$, we have that $E(y_t | X) = y_t$, which is obviously not the model we want.

- The change in the conditioning set changes the properties of the OLS estimator.
- Consider the AR(1) model for y_t :

$$y_t = \beta y_{t-1} + \epsilon_t,$$

where $E(\epsilon_t|y_{t-1}) = 0$, $E(\epsilon_t^2|y_{t-1}) = \sigma^2$ and $E(\epsilon_t\epsilon_s|y_{t-1}, y_{s-1}) = 0$ for $t \neq s$.

- Here,

$$b = \frac{\sum_{t=1}^T y_{t-1}y_t}{\sum_{t=1}^T y_{t-1}^2} = \beta + \frac{\sum_{t=1}^T y_{t-1}\epsilon_t}{\sum_{t=1}^T y_{t-1}^2}.$$

- Typically, b is not unbiased, because

$$E\left(\frac{\sum_{t=1}^T y_{t-1}\epsilon_t}{\sum_{t=1}^T y_{t-1}^2}\right) = E\left(\frac{\sum_{t=1}^T \frac{y_{t-1}}{\sum_{t=1}^T y_{t-1}^2} \epsilon_t}{1}\right) \neq \left(\frac{E\sum_{t=1}^T y_{t-1}\epsilon_t}{E\sum_{t=1}^T y_{t-1}^2}\right) = 0.$$

- However, provided $\text{plim } T^{-1} \sum_{t=1}^T y_{t-1} \epsilon_t = 0$, then b is consistent because

$$\text{plim } b = \beta + \left(\frac{\text{plim } T^{-1} \sum_{t=1}^T y_{t-1} \epsilon_t}{\text{plim } T^{-1} \sum_{t=1}^T y_{t-1}^2} \right) = \beta.$$

- The important feature here is that y_{t-1} and ϵ_t are **uncorrelated**.
- The OLS estimator would not be consistent if, for example, ϵ_t was an MA(1) process, because y_{t-1} and ϵ_t would then be correlated.
- If OLS is used in models with lagged dependent variables, then it is important to test for serial correlation in the disturbances.

- How should we proceed if, having estimated the model

$$y_t = x_t' \beta + \epsilon_t,$$

(where the vector x_t may include lagged values of y_t) we find evidence of first-order serial correlation of the form

$$\epsilon_t = \rho \epsilon_{t-1} + u_t?$$

- We could attempt to estimate the model allowing for the serially correlated disturbance, and there are ways of doing this.
- However, we need to ask **why** is there autocorrelation in the disturbances? This can often arise because of:
 - (a) **misspecification** (fitting a linear model when y and x are related non-linearly);
 - (b) **neglected dynamics** (not including enough lags of y_t and x_t).
- Neglected dynamics is generally both serious and reasonably easy to deal with.

- In most cases, the problem can be eliminated by modelling the dynamics by including lagged variables as regressors, rather than simply confining dynamics to the error term.
- For example, notice that a model of the form

$$y_t = x_t' \beta + \epsilon_t; \quad \epsilon_t = \rho \epsilon_{t-1} + u_t$$

can be rewritten as

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

- More generally, this approach leads to models of the form

$$y_t = \rho_1 y_{t-1} + \dots + \rho_p y_{t-p} + x_t' \beta + \epsilon_t,$$

where x_t can contain current and lagged values of the other regressors.

- In general, the ultimate purpose is to specify a model that incorporates all the relevant information available at period t .
- That is, a model such that $E[\epsilon_t | x_t, y_{t-1}, x_{t-1}, y_{t-2}, \dots] = 0$.

University week: 9

Unit roots and cointegration

Outline

1. Review;
2. Stationarity;
3. Unit roots in the AR(1) model;
4. Testing for unit roots;
5. Cointegration.

Reading: Greene: chapters 21 and 22.

1) Review

- Implicit (so far) has been some notion of stationarity e.g., $E(\epsilon_t|x_t) = 0$, $E(\epsilon_t^2|x_t) = \sigma^2$, $E(\epsilon_t\epsilon_s|x_t) = 0$ ($t \neq s$).
- Finite-sample and asymptotic results have rested on variables being “well behaved”.
- Many time series are not stationary but contain deterministic/stochastic trends
- For example, in Problem Set 1, we considered a model where Y fluctuates around a linear trend $\beta_1 + \beta_2 i$ and found that:

$$X'X = \begin{bmatrix} n & \sum_{i=1}^n i \\ \sum_{i=1}^n i & \sum_{i=1}^n i^2 \end{bmatrix} = \begin{bmatrix} n & \frac{n(n+1)}{2} \\ \frac{n(n+1)}{2} & \frac{n(n+1)(2n+1)}{6} \end{bmatrix}$$

- In this case, $n^{-1}X'X$ does not converge to a finite limit.
- For cases like this, new tools are needed.

2) Stationarity

Definition: A time series $\{y_t\}$ ($t = 1, \dots, T$) is said to be (covariance) **stationary** if:

(i) $E(y_t) = \mu$ (constant – does not depend on t);

(ii) $var(y_t) = \sigma^2 < \infty$ (constant – does not depend on t);

(iii) $cov(y_t, y_s) = \lambda_{|t-s|}$ (only depends on $|t - s|$ and not on t or s alone).

- The most fundamental stationary process is **white noise**, which we shall denote ϵ_t , and which has the properties:

$$E(\epsilon_t) = 0, \quad E(\epsilon_t^2) = \sigma^2, \quad E(\epsilon_t \epsilon_s) = 0 \quad (t \neq s).$$

- We shall write $\epsilon_t \sim WN(0, \sigma^2)$.
- The white noise ϵ_t was used to characterise both the $AR(p)$ and the $MA(q)$ processes.

3) Unit roots in the AR(1) model

- AR processes are used extensively in econometrics – we shall focus on the AR(1)
- Consider the AR(1) process

$$y_t = \gamma y_{t-1} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2). \quad (45)$$

- For y_t to be stationary we need $-1 < \gamma < 1$ (commonly written $|\gamma| < 1$).
- If $|\gamma| > 1$ the series is explosive, while if $|\gamma| = 1$ we have a unit root.
- More generally, for AR(p) processes of the form $y_t = \gamma_1 y_{t-1} + \dots + \gamma_p y_{t-p} + \epsilon_t$, we need to consider the roots of the equation $\gamma(z) = 1 - \gamma_1 z - \dots - \gamma_p z^p = 0$. If the roots lie outside the unit circle (modulus greater than one) the process is stationary; if their modulus is equal to one there is a unit root; while if they are inside the unit circle (modulus less than one) the process is nonstationary, or explosive.
- Explosive series are not frequent in economics and consequently we focus on the case $\gamma = 1$, which leads to $y_t = y_{t-1} + \epsilon_t$ being a **random walk**.

- What are the properties of a random walk (RW)?
- By repeated substitution, we find that

$$\begin{aligned}
 y_t &= y_{t-1} + \epsilon_t \\
 &= (y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\
 &= (y_{t-3} + \epsilon_{t-2}) + \epsilon_{t-1} + \epsilon_t
 \end{aligned}$$

⋮

$$\Rightarrow y_t = y_0 + \sum_{i=1}^t \epsilon_i.$$

- This implies that $E(y_t) = y_0$ (which we shall assume to be fixed) and $var(y_t) = t\sigma^2 \rightarrow \infty$ as $t \rightarrow \infty$.
- Notice that the first difference of a RW is stationary: $y_t - y_{t-1} = \Delta y_t = \epsilon_t$.

- Often it is appropriate to include an intercept in the model, so that (45) becomes

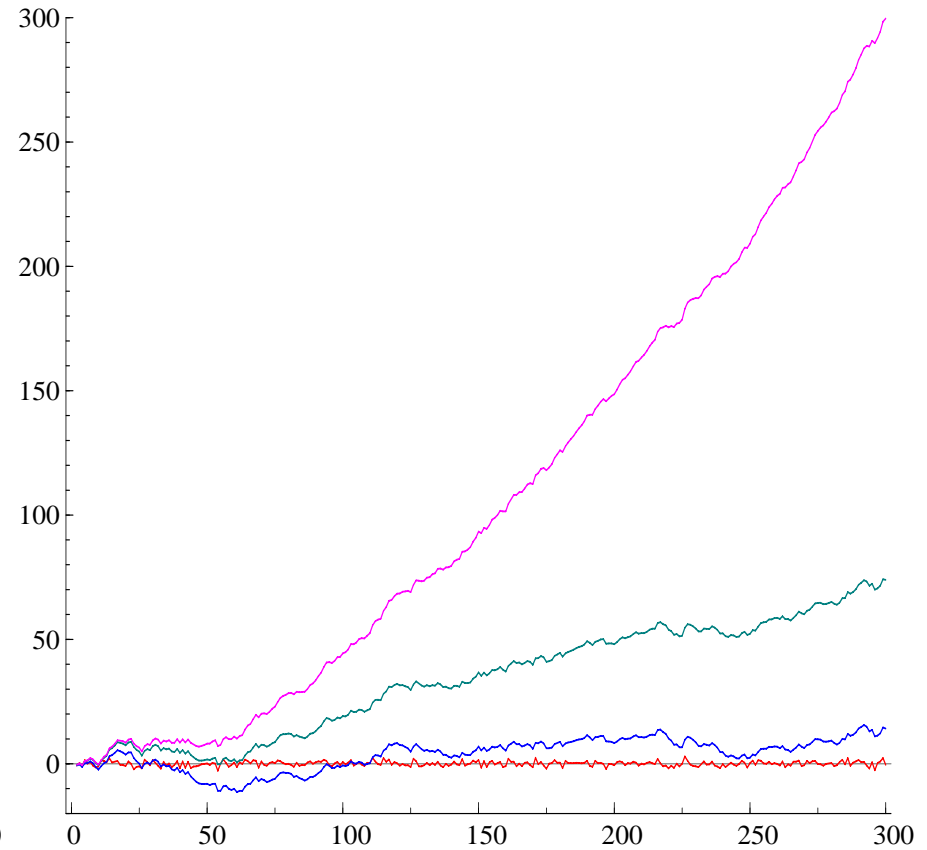
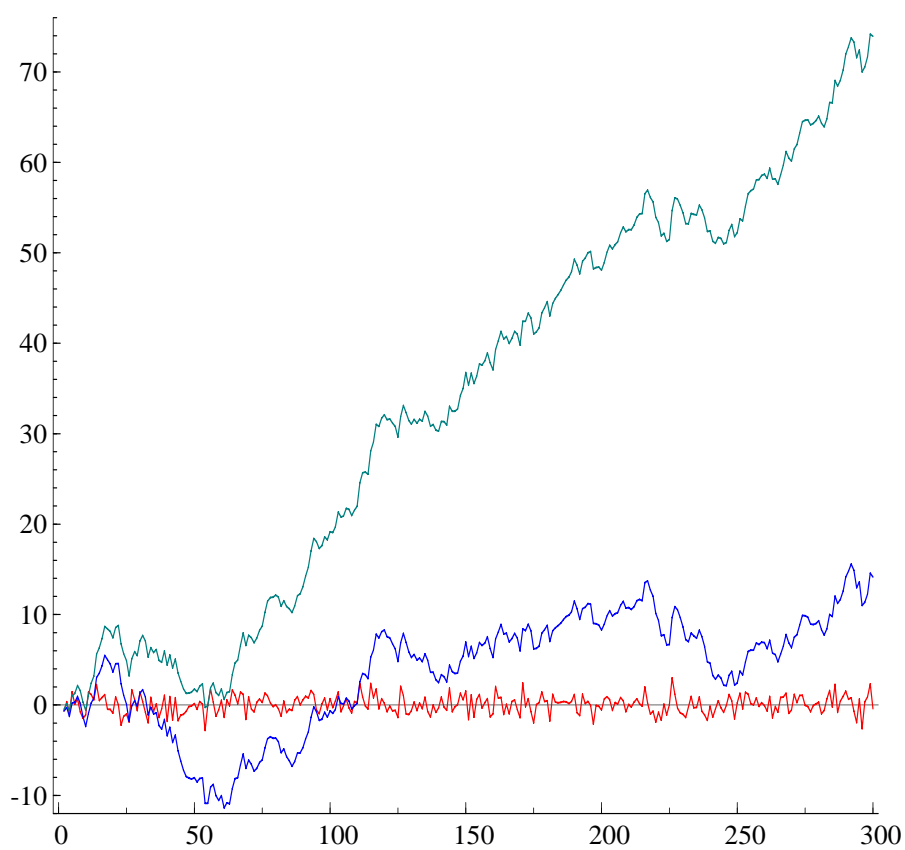
$$y_t = \mu + \gamma y_{t-1} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2).$$

- If $\gamma = 1$ then $\Delta y_t = \mu + \epsilon_t$ (**RW with drift**) and changes in y_t are equal to a constant μ plus a random component ϵ_t (and they are stationary).
- Sometimes a linear trend is also appropriate, in which case (45) becomes

$$y_t = \mu + \beta t + \gamma y_{t-1} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2).$$

- If $\gamma = 1$ then $\Delta y_t = \mu + \beta t + \epsilon_t$ (**RW with drift and trend**) and changes in y_t are equal to a linear trend $\mu + \beta t$ plus a random component ϵ_t (stationary around the trend).
- A process with a unit root is often called **integrated of order one**, or I(1), because it requires differencing once to become stationary.
- In this terminology, a stationary series is I(0), because it does not need differencing to become stationary.

- Examples of random walks: the pictures below graph a white noise (in red), a RW (in blue) a RW with drift (in green), and a RW with drift and trend (in magenta).



4) Testing for unit roots

- For the AR(1) defined by (45), subtracting y_{t-1} from both sides leads to

$$\Delta y_t \equiv y_t - y_{t-1} = \gamma^* y_{t-1} + \epsilon_t, \quad \gamma^* = \gamma - 1. \quad (46)$$

- Note that $-1 < \gamma < 1 \Rightarrow -1 < 1 + \gamma^* < 1 \Rightarrow -2 < \gamma^* < 0$.
- Testing for a unit root is therefore equivalent to testing

$$H_0: \gamma^* = 0 \quad \text{unit root}$$

$$\text{against } H_1: \gamma^* < 0 \quad \text{stationary AR(1),}$$

which is most easily carried out using a t-test in (46).

- **But** y_t is nonstationary under H_0 , and so the usual limiting (normal) distribution does not apply for the t-ratio.

- The appropriate distribution is the **Dickey-Fuller** distribution, for which critical values are tabulated.
- The test statistic is

$$DF = \frac{\hat{\gamma}^*}{se(\hat{\gamma}^*)},$$

where $se(\cdot)$ denotes the estimated standard error.

- Note that this is a **one-sided** test – we are looking for significantly **negative** values of DF in order to reject H_0 .
- Let \overline{DF} denote the critical value (where $\overline{DF} < 0$).
- The decision rule for the test is:

if $DF \leq \overline{DF}$ reject H_0 (y_t is stationary);

if $DF > \overline{DF}$ do not reject H_0 (y_t is nonstationary).

- This test procedure can be adapted for the case in which the model has a drift or a trend.
- For the case with a drift, the test regression becomes

$$\Delta y_t = \mu + \gamma^* y_{t-1} + \epsilon_t,$$

and the same testing procedure applies (but with **different critical values**).

- With a drift and a trend, the test regression becomes

$$\Delta y_t = \mu + \beta t + \gamma^* y_{t-1} + \epsilon_t,$$

and again the same testing procedure applies (but with **different critical values**).

- How do we choose the model used to test stationarity? Look at the plot of the series!
- If the series tends to increase or decrease over time, include the trend, otherwise just include the drift (the model without the drift is rarely used).

- If an $AR(p)$ model is more appropriate than an $AR(1)$, we can augment the test regression with $p - 1$ lags of Δy_t :

$$\Delta y_t = \gamma^* y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta y_{t-i} + \epsilon_t,$$

$$\Delta y_t = \mu + \gamma^* y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta y_{t-i} + \epsilon_t,$$

$$\Delta y_t = \mu + \beta t + \gamma^* y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta y_{t-i} + \epsilon_t.$$

- This is the **augmented Dickey-Fuller**, or **ADF**, test, and can also have the effect of removing serial correlation from the residuals (desirable because the critical values depend on ϵ_t being white noise).
- Augmenting the test does not change the critical values to use!

5) Cointegration

- Suppose we have 2 independent random walks:

$$y_t = y_{t-1} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2),$$

$$x_t = x_{t-1} + u_t, \quad u_t \sim WN(0, \omega^2),$$

where $E(\epsilon_t u_s) = 0$ for all t and s .

- Suppose we regress y on x , obtaining

$$y_t = b_1 + b_2 x_t + e_t.$$

- Common sense suggests that we would have:
 - (a) $t_2 = b_2 / se(b_2)$ to be ‘small,’ thereby not rejecting $H_0 : \beta_2 = 0$;
 - (b) R^2 to be low (y and x are independent).

- However, because both y_t and x_t are I(1) (random walks), the results we have seen so far are invalid.
- This leads to what is known as a **spurious regression**, and we find that:
 - (a) $t_2 \rightarrow \infty$ as $T \rightarrow \infty$, i.e., the larger the sample, the higher the probability of **rejecting** $H_0 : \beta_2 = 0$;
 - (b) R^2 will be ‘acceptable,’ suggesting a reasonable fit.
- Therefore, we have to be especially careful when estimating regressions with I(1) data.
- A characteristic of a spurious regression is that e_t will also be I(1).

- Suppose, however, that we run a regression involving only I(1) regressors and find e_t to be stationary, i.e., e_t is I(0).
- What does this imply about the relationship between y_t and x_t ?
- It suggests that the linear combination $y_t - b_1 - b_2x_t$ of the two I(1) series is stationary.
- In this case y_t and x_t are said to be **cointegrated**.
- Many time series in economics have been found to be I(1), and so cointegration is often important for economic theory.
- In particular, the fact that two series are cointegrated suggests that there is a long-run relation between them.

- To test for the presence of cointegration between a pair of variables y_t and x_t , the following procedure can be used:

(a) Test for the orders of integration of y_t and x_t .

(b) If both variables are $I(1)$, regress y on x to obtain

$$y_t = b_1 + b_2x_t + e_t. \quad (47)$$

(c) Apply the DF/ADF test (**without drift or trend**) to e_t . If $e_t \sim I(1)$ there is no cointegration, but if $e_t \sim I(0)$, there is evidence of cointegration.

- **Note:** a different set of critical values is required for the tests of cointegration (different from the ones for the DF/ADF test without drift or trend).
- This test is called the Engle-Granger test (EG), and it can also be “augmented” if there are signs of serial correlation in the residuals of the auxiliary regression used to perform the test.

- Things to note:
 - in (47), standard inference is **not valid** because the series are I(1);
 - e_t is required only to be stationary, or I(0), but it can still be highly autocorrelated;
 - it is possible to consider cointegration between a **set** of I(1) variables x_1, \dots, x_K ;
 - with more than two variables the methods become more complicated as it is possible to have between 1 and $K - 1$ linear combinations of the variables that are cointegrated.
- To estimate and perform standard inference with models involving I(1) variables, we will first make them stationary by taking first differences.
- Recall also that cointegration describes a long-run, or equilibrium, relationship between the variables, and that the error from this relation is stationary.
- The error from this long-run relationship contains information about the short-run behaviour of the series.

- Therefore, the dynamic behaviour of these variables can be described by an **error correction model**, or ECM, of the form:

$$\Delta y_t = \lambda e_{t-1} + \sum_{i=1}^p \alpha_i \Delta y_{t-i} + \sum_{i=0}^q \delta_i \Delta x_{t-i} + \epsilon_t,$$

where e_t is defined by (47) and $\epsilon_t \sim WN(0, \sigma^2)$.

- Here y_t is responding not only to lagged changes in y and x , but also to the disequilibrium in the previous period.
- We expect $\lambda < 0$:
 - if $e_{t-1} < 0 \Rightarrow y_{t-1} < (b_1 + b_2 x_{t-1}) \Rightarrow \Delta y_t$ should be positive;
 - if $e_{t-1} > 0 \Rightarrow y_{t-1} > (b_1 + b_2 x_{t-1}) \Rightarrow \Delta y_t$ should be negative.

University week: 10

Simultaneous equations

Outline

1. Review;
2. Simultaneity;
3. Identification;
4. Two stages least squares (2SLS) estimation.

Reading: Greene: chapter 10.

1) Review;

- We have examined (in some detail) the CLRM and have relaxed most of the assumptions.
- The CLRM is a single equation, but ...
- Sometimes we are interested in complete systems of two or more equations e.g. demand-supply models, models of (sectors of) the economy.
- New problems arise in the estimation of simultaneous equations models.

2) Simultaneity;

- Consider the following simple model of demand and supply:

$$\text{Demand equation: } q_{d,t} = \alpha_1 p_t + \alpha_2 x_t + \epsilon_{d,t}, \quad (48)$$

$$\text{Supply equation: } q_{s,t} = \beta_1 p_t + \epsilon_{s,t}, \quad (49)$$

$$\text{Equilibrium condition: } q_{d,t} = q_{s,t} = q_t, \quad (50)$$

where: q_d, q_s : demand and supply (unobserved);

q : equilibrium quantity (observed);

p : price (observed);

x : income (observed);

ϵ_d, ϵ_s : random disturbances (unobserved).

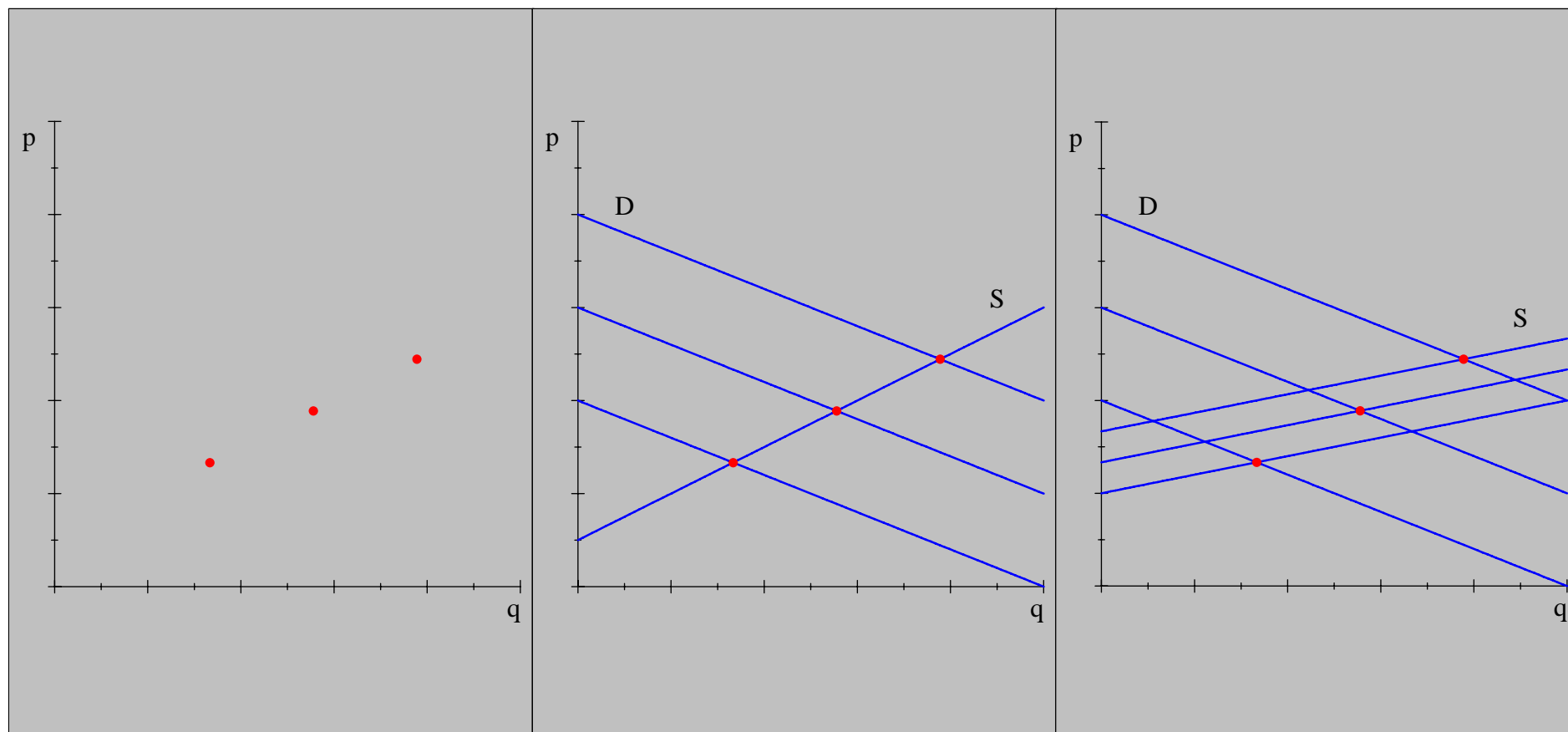
- Eliminating q_d and q_s using (50) we obtain:

$$\text{Demand equation: } q_t = \alpha_1 p_t + \alpha_2 x_t + \epsilon_{d,t}, \quad (51)$$

$$\text{Supply equation: } q_t = \beta_1 p_t + \epsilon_{s,t}. \quad (52)$$

- Equations (48)–(50) are **structural equations** and form the **structural model**.
- A structural model is a model suggested by economic theory.
- In this particular model, p and q are jointly determined, or **endogenous**, while x is determined outside the model, and is therefore **exogenous**.
- Note that lagged values of endogenous variables, called **predetermined** variables, can often be treated as exogenous.
- Given observations on q and p , how can we know whether we are estimating the demand, the supply, or any other line?

- In other words: can the parameters of the structural model be **identified**?
- For us to be able to identify one equation, we must obtain observations when that equation is fixed and the other is moved by some exogenous reason, as in (2).



(1)

(2)

(3)

- In order to proceed, we need to solve (51) and (52) for q and p , leading to the **reduced form** of the model.

$$q_t = \pi_1 x_t + v_{1,t},$$

$$p_t = \pi_2 x_t + v_{2,t},$$

$$\pi_1 = \frac{\beta_1 \alpha_2}{\beta_1 - \alpha_1}, \quad v_{1,t} = \frac{\beta_1 \epsilon_{d,t} - \alpha_1 \epsilon_{s,t}}{\beta_1 - \alpha_1},$$

$$\pi_2 = \frac{\alpha_2}{\beta_1 - \alpha_1}, \quad v_{2,t} = \frac{\epsilon_{d,t} - \epsilon_{s,t}}{\beta_1 - \alpha_1}.$$

- Assuming that $E(v_{i,t}|x_t) = 0$, $i = 1, 2$, it is clear that π_1 and π_2 can be consistently estimated by OLS.
- However, equations (51) and (52) cannot be estimated by OLS because the endogenous regressors are correlated with the errors.
- Can estimates of α_1 , α_2 and β_1 be derived **uniquely** from the estimates of π_1 and π_2 ?

- We have seen that we require shifts in the demand and supply equations (excluding the random disturbances) to trace out points **along** the other equation.
- In our system

$$\text{Demand equation: } q_t = \alpha_1 p_t + \alpha_2 x_t + \epsilon_{d,t},$$

$$\text{Supply equation: } q_t = \beta_1 p_t + \epsilon_{s,t},$$

only the supply equation (β_1) can be identified because changes in x shift the demand curve and thereby trace out points on the supply curve.

- There is no independent shift in supply that is capable of tracing out points on the demand curve.
- We shall now look at identification more generally.

3) Identification;

- A general notation for the structural model containing M equations is:

$$\gamma_{11}y_{t1} + \gamma_{21}y_{t2} + \dots + \gamma_{M1}y_{tM} + \beta_{11}x_{t1} + \dots + \beta_{K1}x_{tK} = \epsilon_{t1}$$

$$\gamma_{12}y_{t1} + \gamma_{22}y_{t2} + \dots + \gamma_{M2}y_{tM} + \beta_{12}x_{t1} + \dots + \beta_{K2}x_{tK} = \epsilon_{t2}$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$\gamma_{1M}y_{t1} + \gamma_{2M}y_{t2} + \dots + \gamma_{MM}y_{tM} + \beta_{1M}x_{t1} + \dots + \beta_{KM}x_{tK} = \epsilon_{tM}$$

where $t = 1, \dots, T$.

- Writing the model in matrix form will help to manipulate it.

- In matrix form (placing each equation side-by-side):

$$\begin{aligned}
 & [y_{t1}, y_{t2}, \dots, y_{tM}] \begin{bmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1M} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2M} \\ \vdots & \vdots & & \vdots \\ \gamma_{M1} & \gamma_{M2} & \dots & \gamma_{MM} \end{bmatrix} \\
 & + [x_{t1}, x_{t2}, \dots, x_{tK}] \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2M} \\ \vdots & \vdots & & \vdots \\ \beta_{K1} & \beta_{K2} & \dots & \beta_{KM} \end{bmatrix} = [\epsilon_{t1}, \epsilon_{t2}, \dots, \epsilon_{tM}] \\
 & \Rightarrow y'_t \Gamma + x'_t B = \epsilon'_t. \tag{53}
 \end{aligned}$$

- From (53) it is easy to solve for y'_t to obtain the reduced form:

$$y'_t = -x'_t B \Gamma^{-1} + \epsilon'_t \Gamma^{-1} = x'_t \Pi + v'_t \quad (54)$$

where $\Pi = -B \Gamma^{-1}$ ($K \times M$) and $v'_t = \epsilon'_t \Gamma^{-1}$ ($1 \times M$).

- NB: Γ needs to be nonsingular (completeness condition).
- We shall assume that

$$E(\epsilon_t | x_t) = 0, \quad E(\epsilon_t \epsilon'_t | x_t) = \Sigma, \quad E(\epsilon_t \epsilon'_s | x_t, x_s) = 0 \quad (t \neq s).$$

and hence (because $v_t = (\Gamma^{-1})' \epsilon_t$)

$$E(v_t | x_t) = 0, \quad E(v_t v'_t | x_t) = (\Gamma^{-1}) \Sigma \Gamma^{-1} = \Omega, \quad E(v_t v'_s | x_t, x_s) = 0 \quad (t \neq s).$$

- We can consistently estimate Π by OLS, but what about Γ and B ?

- Notice that Π has $K \times M$ parameters, whereas Γ and B have $M \times M + K \times M$ parameters (Σ and Ω are both $M \times M$).
- Thus, **unless** further restrictions are imposed, it is not possible to identify Γ and B .
- What types of restrictions can be imposed?
 - **Normalisation restrictions:** there will be a coefficient of 1 on one y_{tj} in each equation.
 - **Zero/exclusion restrictions:** some variables are excluded from certain equations and hence the corresponding coefficient is zero.
 - Other types of restrictions can be imposed, but we will not talk about them.
- As we have seen in the example of the demand system, it is possible that some equations of the model are identified while others are not.
- Therefore, identification proceeds on an **equation-by-equation** basis.

- Consider equation j , $y_t' \Gamma_j + x_t' B_j = \epsilon_{tj}$, and define:

Y_{tj} : included endogenous ($M_j \times 1$);

Y_{tj}^* : excluded endogenous ($M_j^* \times 1$);

x_{tj} : included exogenous ($K_j \times 1$);

x_{tj}^* : excluded exogenous ($K_j^* \times 1$),

where $M_j + M_j^* + 1 = M$ and $K_j + K_j^* = K$.

- Equation j is then (noting that $\gamma_j^* = 0$ and $\beta_j^* = 0$)

$$y_{tj} = Y_{tj}' \gamma_j + Y_{tj}^{*'} \gamma_j^* + x_{tj}' \beta_j + x_{tj}^{*'} \beta_j^* + \epsilon_{tj}$$

$$= Y_{tj}' \gamma_j + x_{tj}' \beta_j + \epsilon_{tj}.$$

- Implicitly, $\Gamma_j' = [1, -\gamma_j', 0']$, $B_j' = [-\beta_j', 0']$.

- We can also partition the reduced form:

$$[y_{tj}, Y'_{tj}, Y_{tj}^{*'}] = [x'_{tj}, x_{tj}^{*'}] \begin{bmatrix} \pi_j & \underline{\Pi}_j & \bar{\Pi}_j \\ \pi_j^* & \underline{\Pi}_j^* & \bar{\Pi}_j^* \end{bmatrix} + [v_{tj}, V'_{tj}, V_{tj}^{*'}].$$

- We know that $\Pi = -B\Gamma^{-1}$ implying:

$$\begin{matrix} \Pi & \Gamma \\ (K \times M) & (M \times M) \end{matrix} = \begin{matrix} -B \\ (K \times M) \end{matrix}.$$

- So, taking column j of this expression gives $\Pi\Gamma_j = -B_j$ or

$$\begin{bmatrix} \pi_j & \underline{\Pi}_j & \bar{\Pi}_j \\ \pi_j^* & \underline{\Pi}_j^* & \bar{\Pi}_j^* \end{bmatrix} \begin{bmatrix} 1 \\ -\gamma_j \\ 0 \end{bmatrix} = \begin{bmatrix} \beta_j \\ 0 \end{bmatrix}.$$

- Now, we need to solve this system for γ_j and β_j .

- The two relevant equations are:

$$\begin{matrix} \pi_j & - & \underline{\Pi}_j & \gamma_j & = & \beta_j & ; \\ (K_j \times 1) & & (K_j \times M_j) & (M_j \times 1) & & (K_j \times 1) & \end{matrix}$$

$$\begin{matrix} \pi_j^* & - & \underline{\Pi}_j^* & \gamma_j & = & 0 & . \\ (K_j^* \times 1) & & (K_j^* \times M_j) & (M_j \times 1) & & (K_j^* \times 1) & \end{matrix}$$

- The second set of K_j^* equations needs to be solved for M_j unknowns (γ_j).

- Hence, it is necessary (but not sufficient) that the model verifies the **order condition**:

$K_j^* \geq M_j$ — The number of excluded exogenous (and predetermined) variables must be at least as large as the number of included endogenous variables.

- The sufficient condition is the **rank condition**: $\text{rank} [\pi_j^*, \underline{\Pi}_j^*] = \text{rank} [\underline{\Pi}_j^*] = M_j$.

- If we can do this, we then solve the first set of K_j equations for K_j unknowns (β_j).

4) Two stages least squares (2SLS) estimation.

- Recall that equation j can be written

$$y_{tj} = Y'_{tj}\gamma_j + x'_{tj}\beta_j + \epsilon_{tj}, \quad t = 1, \dots, T.$$

- OLS is not a consistent estimator of γ_j and β_j because of the endogeneity of Y'_{tj} i.e.,

$$\text{plim} \frac{1}{T} \sum_{t=1}^T Y_{tj} \epsilon_{tj} \neq 0.$$

(We know that Y_{tj} depends on ϵ_{tj} from the reduced form).

- We can therefore use an IV (see slides 77 to 85) estimator of which 2SLS is a special case.

- The exogenous variables in the model form perfect instruments, because:
 - (a) they are uncorrelated with ϵ_{tj} (they are exogenous);
 - (b) they are correlated with Y_{tj} (they help determine Y_{tj} via the reduced form).
- Stacking the observations over $t = 1, \dots, T$:

$$\begin{array}{ccccccc}
 y_j & = & Y_j & \gamma_j & + & X_j & \beta_j & + & \epsilon_j \\
 T \times 1 & & T \times M_j & M_j \times 1 & & T \times K_j & K_j \times 1 & & T \times 1
 \end{array}$$

or equivalently

$$y_j = [Y_j, X_j] \begin{bmatrix} \gamma_j \\ \beta_j \end{bmatrix} + \epsilon_j = Z_j \delta_j + \epsilon_j.$$

- As the name implies, the 2SLS estimator of δ_j can be obtained in two steps:

(a) Regress the variables in Y_j on X and obtain the fitted values \hat{Y}_j .

Effectively this estimates the reduced form for Y_j :

$$\begin{array}{ccccccc} Y_j & = & X & \Pi_j & + & V_j & \\ T \times M_j & & T \times K & K \times M_j & & T \times M_j & \end{array}$$

$$\Rightarrow \hat{\Pi}_j = (X'X)^{-1}X'Y_j$$

$$\Rightarrow \hat{Y}_j = X\hat{\Pi}_j = X(X'X)^{-1}X'Y_j = PY_j,$$

where $P = X(X'X)^{-1}X'$ is the projection matrix for X .

(b) Regress y_j on \hat{Y}_j and X_j to obtain $\hat{\gamma}_{j,2SLS}$ and $\hat{\beta}_{j,2SLS}$.

Let $\hat{Z}_j = \begin{bmatrix} \hat{Y}_j \\ X_j \end{bmatrix}$ so that the model becomes

$$y_j = \hat{Z}_j \delta_j + \epsilon_j,$$

leading to the estimator

$$\hat{\delta}_{j,2SLS} = (\hat{Z}_j' \hat{Z}_j)^{-1} \hat{Z}_j' y_j.$$

- The resulting estimator is consistent, i.e.,

$$\text{plim } \hat{\delta}_{j,2SLS} = \delta_j.$$

- The (asymptotic) covariance matrix of $\hat{\delta}_{j,2SLS}$ is

$$\text{Asy. Var. } \left(\hat{\delta}_{j,2SLS} \right) = \sigma_j^2 \left(\hat{Z}_j' \hat{Z}_j \right)^{-1}.$$

- In order to conduct inference, we therefore need to estimate $\sigma_j^2 = E(\epsilon_{tj}^2|x_t)$.
- Let $e_j = y_j - Z_j\hat{\delta}_{j,2SLS}$ denote the $T \times 1$ vector of residuals (note that we use Z_j and not \hat{Z}_j).
- Then we can estimate σ_j^2 using

$$\hat{\sigma}_j^2 = \frac{e_j'e_j}{T - M_j - K_j}.$$

University week: 11

Panel Data

Outline

1. Review;
2. General panel data model;
3. Pooled OLS;
4. Random effects;
5. Fixed effects.

Reading: Greene: chapter 11.

1) Review;

- We have examined single equation and simultaneous equations models in which variables are indexed by observation number e.g. y_t .
- This includes both cross-section and time-series data.
- Sometimes, we are interested in data that relate to a cross-section over time.
- In this case we need to index both the position in the cross-section as well as in time e.g. y_{it} is the observation on variable y for the i 'th cross-sectional unit in time period t .
- Special models are needed to explore the features of this kind of data.

2) General panel data model

- General panel data features

- Combine observations on cross-section over time → panel/longitudinal data.
- Usually cross-section dimension $>$ time series dimension.
- Focus on cross-sectional heterogeneity.
- As panels evolve, the time series dimension is growing as well → more interest recently in dynamic panel models.
- Observations have two subscripts, i and t .

- The general form of the panel data model we shall consider is

$$y_{it} = x'_{it}\beta + z'_i\alpha + \epsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T, \quad (55)$$

where

y_{it} observable scalar dependent variable,

x_{it} $K \times 1$ vector of **observable** regressors (no intercept),

β $K \times 1$ vector of unobservable parameters,

$z'_i\alpha$ (scalar) **heterogeneity** or **individual effect**,

z_i $H \times 1$ vector with **unobservable** individual specific variables,

α $H \times 1$ vector of unobservable parameters,

ϵ_{it} scalar random disturbance satisfying $E(\epsilon_{it}|x_{i1}, x_{i2}, \dots, z_i) = 0$,

$E(\epsilon_{it}^2|x_{i1}, x_{i2}, \dots, z_i) = \sigma^2$ and $E(\epsilon_{it}\epsilon_{js}|x_{i1}, x_{i2}, \dots, z_i) = 0$ if $i \neq j$ and/or $t \neq s$,

n number of cross-sectional units (e.g., individuals),

T number of time periods (e.g., years).

- We will consider three different approaches to the estimation of this model.

3) Pooled OLS

- If the unobservable individual heterogeneity is uncorrelated with x_{it} then we can write

$$\begin{aligned}y_{it} &= x'_{it}\beta + E(z'_i\alpha) + [z'_i\alpha - E(z'_i\alpha)] + \epsilon_{it} \\ &= x'_{it}\beta + \mu + u_i + \epsilon_{it}.\end{aligned}$$

- It is convenient to define $w'_{it} = [1, x'_{it}]$, $\delta = [\mu, \beta']'$ and $\eta_{it} = u_i + \epsilon_{it}$, so that the model can be written as

$$y_{it} = w'_{it}\delta + \eta_{it}. \tag{56}$$

- Stacking the observations for each i over $t = 1, \dots, T$ gives

$$y_i = W_i\delta + \eta_i, \quad i = 1, \dots, n,$$

where y_i is $T \times 1$ and W_i is $T \times (K + 1)$.

- It will be assumed that

$$\begin{aligned}
 E(\epsilon_{it}|W_i) &= 0, & E(u_i|W_i) &= 0, & E(\epsilon_{it}\epsilon_{js}|W_i, W_j) &= 0 \quad (t \neq s, i \neq j), \\
 E(\epsilon_{it}^2|W_i) &= \sigma_\epsilon^2, & E(u_i^2|W_i) &= \sigma_u^2, & E(u_i u_j|W_i, W_j) &= 0 \quad (i \neq j), \\
 E(\eta_{it}|W_i) &= 0 & E(\eta_{it}^2|W_i) &= \sigma_u^2 + \sigma_\epsilon^2, & E(\epsilon_{it} u_j|W_i, W_j) &= 0 \text{ for all } i, t \text{ and } j, \\
 E(\eta_{it}\eta_{is}|W_i, W_j) &= \sigma_u^2 \quad (t \neq s), & E(\eta_{it}\eta_{js}|W_i, W_j) &= 0 \quad (i \neq j).
 \end{aligned}$$

- Estimating (56) by OLS leads to the so-called **pooled OLS** estimator, which is consistent for δ (β and μ).
- However, the presence of the individual effect u_i causes serial correlation in the error terms.
- Therefore, pooled OLS is not efficient and the usual covariance matrix is invalid.
- However, an asymptotically valid covariance matrix can be obtained (option `cluster` in Stata).
- An efficient estimator can be obtained using GLS.

4) Random effects

- As before, the model to be estimated is

$$\begin{aligned}y_{it} &= x'_{it}\beta + \mu + u_i + \epsilon_{it}. \\ &= w'_{it}\delta + \eta_{it}\end{aligned}$$

- We maintain the same assumptions about u_i and ϵ_{it} .
- In particular, we continue to assume that the unobservable individual heterogeneity is uncorrelated with x_{it} .
- Note that $E(\eta_i\eta'_i|W_i) = \Sigma$, where

$$\Sigma = \begin{bmatrix} \sigma_\epsilon^2 + \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\epsilon^2 + \sigma_u^2 & \dots & \sigma_u^2 \\ & & \vdots & \\ \sigma_u^2 & \sigma_u^2 & \dots & \sigma_\epsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\epsilon^2 I_T + \sigma_u^2 1_T 1'_T,$$

while $E(\eta_i\eta'_j|W_i, W_j) = 0$ for $i \neq j$.

- Stacking over i gives:

$$y = W\delta + \eta, \quad (57)$$

where y is $nT \times 1$, W is $nT \times (K + 1)$, and

$$E(\eta\eta') = \Omega = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \Sigma \end{bmatrix} = I_n \otimes \Sigma,$$

- Estimation of (57) by OLS leads to the pooled OLS estimator we have seen before.
- However, (57) is a generalised linear regression model and can therefore be estimated efficiently by GLS if Ω is known:

$$\hat{\delta} = (W'\Omega^{-1}W)^{-1}W'\Omega^{-1}y = \left(\sum_{i=1}^n W_i'\Sigma^{-1}W_i \right)^{-1} \sum_{i=1}^n W_i'\Sigma^{-1}y_i.$$

- If Ω is unknown an FGLS estimator can be used, provided consistent estimates of σ_ϵ^2 and σ_u^2 can be found.

- It can be shown that s_{Pooled}^2 is a consistent estimator of $\sigma_\epsilon^2 + \sigma_u^2$.
- A consistent estimator of σ_ϵ^2 is provided by s_{LSDV}^2 , which is an estimate of the variance of the error term in the model we will consider next.
- We can therefore estimate σ_u^2 consistently using $\hat{\sigma}_u^2 = s_{Pooled}^2 - s_{LSDV}^2$.
- It is possible to test for random effects based on the residuals from pooled OLS: e_{it} .
- The test, due to Breusch and Pagan (1980), is of

$$H_0 : \sigma_u^2 = 0 \text{ against } H_1 : \sigma_u^2 \neq 0.$$

- The LM statistic, which essentially checks for serial correlation, is

$$LM = \frac{nT}{2(T-1)} \left[\frac{\sum_i (\sum_t e_{it})^2}{\sum_i \sum_t e_{it}^2} - 1 \right]^2 \sim \chi_1^2 \text{ under } H_0.$$

- Please note that this IS NOT the Breusch and Pagan test for heteroskedasticity!
- **If z_i is correlated with x_{it} , then both pooled OLS and the Random Effects estimator are biased and inconsistent due to the omitted variables.**

5) Fixed effects

- If the unobservable individual heterogeneity is correlated with x_{it} , we can write $z_i'\alpha = \alpha_i$ and the model to be estimated becomes

$$y_{it} = x_{it}'\beta + \alpha_i + \epsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T. \quad (58)$$

- Stacking the observations for each i over $t = 1, \dots, T$ gives

$$y_i = X_i\beta + 1_T\alpha_i + \epsilon_i, \quad i = 1, \dots, n,$$

where

$$y_i \text{ and } \epsilon_i \text{ are } T \times 1 \text{ vectors,} \quad X_i \text{ is } T \times K, \quad E(\epsilon_i\epsilon_i'|X_i, z_i) = \sigma^2 I_T,$$

$$1_T \text{ denotes a } T \times 1 \text{ vector of ones,} \quad \alpha_i \text{ is a scalar,} \quad E(\epsilon_i\epsilon_j'|X_i, z_i) = 0 \text{ for } i \neq j.$$

- We can now stack these n equations.

- This leads to

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \beta + \begin{pmatrix} 1_T & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1_T \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$y = X\beta + D\alpha + \epsilon, \tag{59}$$

where

y and ϵ are $nT \times 1$ vectors, X is $nT \times K$,

$D = I_n \otimes 1_T$ is $nT \times n$, α is $n \times 1$.

- This is often called the **least squares dummy variable (LSDV) model**.
- It is essentially a classical regression model which captures the unobservable effects by taking α_i to be a group-specific constant term, and can be estimated by OLS.
- **This estimator is consistent for β even if z_i is correlated with x_{it} .**

- If n is large, there are some computational tricks that can be employed to exploit the sparsity of D , as follows.
- Let $M_D = I - D(D'D)^{-1}D'$ and note that $M_DD = 0$.
- Then the OLS estimator of β can be obtained by pre-multiplying the model (59) by M_D :

$$\begin{aligned}
 M_D y &= M_D X \beta + M_D \epsilon \quad (\text{because } M_D D = 0) \\
 \Rightarrow y_* &= X_* \beta + \epsilon_* \\
 \Rightarrow b &= (X_*' X_*)^{-1} X_*' y_* = (X' M_D X)^{-1} X' M_D y.
 \end{aligned}$$

- Notice that pre-multiplying by M_D is the same as obtaining deviations from group means, e.g., $y_* = y_{it} - \bar{y}_i$ with $\bar{y}_i = T^{-1} \sum_{t=1}^T y_{it}$.
- Therefore, β can be estimated by regressing $y_{it} - \bar{y}_i$ on $x_{it} - \bar{x}_i$.
- Notice also that with this method it is not possible to estimate the parameters of any regressors that do not vary with t because these regressors are eliminated when we compute the differences to the group means.

- The estimator of α can be obtained from the group means of the residuals $y = Xb$.
- In particular (note that a is only consistent if $T \rightarrow \infty$),

$$a = (D'D)^{-1}D'(y - Xb).$$

- The covariance matrices of a and b can be shown to be

$$\text{var}(b|X_i, z_i) = \sigma^2(X'M_D X)^{-1}, \quad \text{var}(a|X_i, z_i) = \frac{\sigma^2}{T}I_n + \sigma^2\bar{X}(X'M_D X)^{-1}\bar{X}',$$

where $\bar{X}' = [T^{-1}\sum_t x'_{1t} \cdots T^{-1}\sum_t x'_{nt}]$.

- We can estimate σ^2 using

$$s_{LSDV}^2 = \frac{e'e}{nT - n - K} = \frac{(y - Xb)'M_D(y - Xb)}{nT - n - K}.$$

- A **test for the significance of the group effects** can be carried out with an F -statistic based on the sums of squared residuals in the restricted and unrestricted regressions.
- The test is of the $n - 1$ restrictions:

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_n (= \alpha) \text{ against } H_1 : H_0 \text{ is false.}$$

- Define:

S_{LSDV} : Sum of squared residuals in unrestricted regression (58);

S_{Pooled} : Sum of squared residuals in restricted (pooled) regression ($\alpha_i = \alpha$).

- Then the F -statistic (which is closely related to the Breusch-Pagan test seen earlier) is:

$$F = \frac{S_{Pooled} - S_{LSDV}}{S_{LSDV}} \times \frac{nT - n - K}{n - 1} \sim F_{n-1, nT-n-K} \text{ under } H_0.$$