Causal Inference, Time Series and Economic History

1. Introduction to Time Series Analysis

Jason Lennard

Overview

- An introduction to the course
  - Aims, format, outline and assessment
- An introduction to time series analysis
  - Primer on time series data
  - Derivation of Ordinary Least Squares (OLS) estimates
  - Assumptions under which OLS is the Best Linear Unbiased Estimator (BLUE)

Part I: Introduction to the Course

Aims

- To understand key quantitative methodologies in economic history
- To understand the principles of research design
  - This is a fundamental skill for research (whether it's a PhD dissertation or an academic paper)

Format

- 6 classes
- We will study a mix of theory, simulations, paper replications and quantitative historiography
- Discussion papers are to be read *before* the seminar

Course Outline

Week	Topic	Date	Time	Room
1	Introduction to time series analysis	Wednesday 11 May	9:30-12:30	Alfa1:1104
2	Stationarity, filtering and seasonal adjustment	Wednesday 11 May	14:30-17:30	Alfa1:1104
3	Single-equation models	Thursday 12 May	9:30-12:30	Alfa1:1104
4	Vector autoregressions	Thursday 12 May	14:30-17:30	Alfa1:1104
5	Narrative methods	Friday 13 May	9:30-12:30	Alfa1:1104
6	Instrumental variables and natural experiments	Friday 13 May	14:30-17:30	Alfa1:1104

Assessment

- 2,500 word essay due Monday 5 September 2022 (100%)

Part II: Introduction to Time Series Analysis

Time Series Data (y_t)

- One unit (an individual, firm, industry, country, etc.) observed at multiple points in time, i.e., $i = 1, t > 1$
- Example: GDP per capita of Austria between 1870 and 1913, i.e., $i = 1, t = 44$
- Covered in this course

Table: GDP per Capita (1990 Int. GK$)

	Austria
1870	1,863
1871	1,979
⋮	⋮
1913	3,465

Cross Sectional Data (y_i)

- Multiple units (individuals, firms, industries, countries, etc.) observed at one point in time, i.e., $i > 1, t = 1$
- Example: GDP per capita of Western European countries in 1870, i.e., $i = 12, t = 1$
- Not covered in this course

Table: GDP per Capita (1990 Int. GK$)

	Austria	Belgium	...	United Kingdom
1870	1,863	2,692	...	3,190

Panel Data (y_{it})

- Multiple units (individuals, firms, industries, countries, etc.) observed at multiple points in time, i.e., $i > 1$, $t > 1$
- Example: GDP per capita of Western European countries between 1870 and 1913, i.e., $i = 12$, $t = 44$
- Not covered in this course

Table: GDP per Capita (1990 Int. GK$)

	Austria	Belgium	...	United Kingdom
1870	1,863	2,692	...	3,190
1871	1,979	2,682	...	3,332
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1913	3,465	4,220	...	4,921

The Time Series Regression Model

$$y_t = \alpha + \beta x_t + u_t \tag{1}$$

where y_t is the dependent variable
x_t is the independent variable
α is the intercept
β is the coefficient
u_t is the error term

Deriving OLS Estimates of α and β

The problem is to find values of $\hat{\alpha}$ and $\hat{\beta}$ to minimize the sum of squared residuals (SSR):

$$SSR = \sum_{t=1}^{n} u_t^2 \tag{2}$$

The first step is to write equation (2) in terms of $\hat{\alpha}$ and $\hat{\beta}$
The time series regression model, $y_t = \hat{\alpha} + \hat{\beta} x_t + u_t$, shows that:

$$u_t = y_t - \hat{\alpha} - \hat{\beta} x_t \tag{3}$$

Therefore equation (2) can be re-written as:

$$SSR = \sum_{t=1}^{n} (y_t - \hat{\alpha} - \hat{\beta} x_t)^2 \tag{4}$$

Deriving OLS Estimates of α and β

We can solve this minimisation problem using calculus by taking the partial derivative of SSR, $\sum_{t=1}^{n}(y_t - \hat{\alpha} - \hat{\beta}x_t)^2$, with respect to $\hat{\alpha}$ and $\hat{\beta}$:

$$\frac{\partial SSR}{\partial \hat{\alpha}} = -2\sum_{t=1}^{n}(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \tag{5}$$

$$\frac{\partial SSR}{\partial \hat{\beta}} = -2\sum_{t=1}^{n}x_t(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \tag{6}$$

Deriving OLS Estimates of α and β

The goal now is to find an expression for $\hat{\alpha}$. Divide both sides of equation (5), $-2\sum_{t=1}^{n}(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0$, by -2:

$$\sum_{t=1}^{n}(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \tag{7}$$

$$\sum_{t=1}^{n} y_t - \hat{\alpha}\sum_{t=1}^{n} 1 - \hat{\beta}\sum_{t=1}^{n} x_t = 0 \tag{8}$$

$$\sum_{t=1}^{n} y_t = \hat{\alpha}\sum_{t=1}^{n} 1 + \hat{\beta}\sum_{t=1}^{n} x_t \tag{9}$$

$$n\bar{y}_t = \hat{\alpha}n + \hat{\beta}n\bar{x}_t \tag{10}$$

As $\sum_{t=1}^{n} y_t = n\bar{y}$, $\sum_{t=1}^{n} x_t = n\bar{x}$

Deriving OLS Estimates of α and β

Divide equation (10), $n\bar{y}_t = \hat{\alpha}n + \hat{\beta}n\bar{x}_t$, by n:

$$\bar{y}_t = \hat{\alpha} + \hat{\beta}\bar{x}_t \tag{11}$$

$$\hat{\alpha} = \bar{y}_t - \hat{\beta}\bar{x}_t \tag{12}$$

Deriving OLS Estimates of α and β

The goal now is to find an expression for $\hat{\beta}$. Divide both sides of equation (6), $-2\sum_{t=1}^{n} x_t(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0$, by -2:

$$\sum_{t=1}^{n} x_t(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \tag{13}$$

Replacing $\hat{\alpha}$ in equation (13) with equation (12), $\hat{\alpha} = \bar{y}_t - \hat{\beta}\bar{x}_t$, gives:

$$\sum_{t=1}^{n} x_t[y_t - (\bar{y} - \hat{\beta}\bar{x}) - \hat{\beta}x_t] = 0 \tag{14}$$

Expanding out of the round brackets:

$$\sum_{t=1}^{n} x_t(y_t - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_t) = 0 \tag{15}$$

Deriving OLS Estimates of α and β

Expanding out of the brackets again, $\sum_{t=1}^{n} x_t(y_t - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_t) = 0$:

$$\sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} x_t \bar{y} + \sum_{t=1}^{n} x_t \hat{\beta}\bar{x} - \sum_{t=1}^{n} x_t \hat{\beta} x_t = 0 \tag{16}$$

Bring the last two terms over to the right-hand side:

$$\sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} x_t \bar{y} = \sum_{t=1}^{n} x_t \hat{\beta} x_t - \sum_{t=1}^{n} x_t \hat{\beta}\bar{x} \tag{17}$$

$$\hat{\beta} \sum_{t=1}^{n} (x_t x_t - x_t \bar{x}) = \sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} x_t \bar{y} \tag{18}$$

$$\hat{\beta} = \frac{\sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} x_t \bar{y}}{\sum_{t=1}^{n} (x_t x_t - x_t \bar{x})} \tag{19}$$

Deriving OLS Estimates of α and β

Collecting terms in equation (19), $\hat{\beta} = \frac{\sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} x_t \bar{y}}{\sum_{t=1}^{n} (x_t x_t - x_t \bar{x})}$:

$$\hat{\beta} = \frac{\sum_{t=1}^{n} x_t (y_t - \bar{y})}{\sum_{t=1}^{n} x_t (x_t - \bar{x})} \tag{20}$$

See the appendix for the details of this step:

$$\hat{\beta} = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})^2} \tag{21}$$

$$\hat{\beta} = \frac{Cov(x_t, y_t)}{Var(x_t)} \tag{22}$$

Example: Okun's Law

- Download "Okun's Law.xlsx" from Moodle
- Okun's law is an association between changes in real GDP and unemployment
- Let's try to calculate the intercept and slope in Excel and compare the results with the output in Stata

Gauss-Markov Assumptions

- The Gauss-Markov assumptions are a set of criteria that if met mean that OLS is BLUE (Best Linear Unbiased Estimator)
- These assumptions are a crucial way to critically evaluate research design
- Violations of these assumptions has consequences for $\hat{\beta}$ and $se(\hat{\beta})$
- The five assumptions are listed in order of (my perceived) importance

Gauss-Markov Assumption 1

Linear in parameters

- The time series process follows a model that is linear in parameters, such as:

$$y_t = \alpha + \beta x_t + u_t$$

$$y_t = \alpha + \beta \ln(x_t) + u_t$$

$$y_t = \alpha + \beta x_t + \gamma x_t^2 + u_t$$

- An example of a time series process that is non-linear in parameters:

$$y_t = \alpha + \beta^2 x_t + u_t$$

Gauss-Markov Assumption 2

No perfect collinearity

- No independent variable is constant or a perfect linear combination of the others
- In other words, the independent variables must not be perfectly correlated
- For example, if we wanted to estimate:

$$y_t = \alpha + \beta x_t + \gamma z_t + u_t$$

- But:

$$z_t = \delta + \theta x_t$$

- Then we would have perfect collinearity
- Note that the absence of an error term makes the relationship exact

Gauss-Markov Assumption 3

Homoscedasticity

- Conditional on $\mathbf{X}$, the variance of u_t is the same for all t:
  $Var(u_t \mid \mathbf{X}) = Var(u_t) = \sigma^2, t = 1, 2, \ldots, n$
- *Consequence:* Heteroscedasticity means that OLS standard errors are biased
- *Test:* Breusch-Pagan or White test for heterscoedasticity (see Wooldridge, *Introductory Econometrics*, chapters 8 and 10)
- *Potential solution:* Use an estimator that is robust to potential heteroscedasticity such as the Newey-West (1987) estimator

Gauss-Markov Assumption 4

No serial correlation

- Conditional on $\mathbf{X}$, the errors in two different time periods are uncorrelated: $Corr(u_t, u_s \mid \mathbf{X}) = 0$, for all $t \neq s$
- *Consequence:* Serial correlation means that OLS standard errors are biased
- *Test:* Breusch-Godfrey test (see Wooldridge, *Introductory Econometrics*, chapter 12)
- *Potential solution:* Use an estimator that is robust to potential serial correlation such as the Newey-West (1987) estimator

Gauss-Markov Assumption 5

Zero conditional mean

- For each t, the expected value of the error u_t, given the explanatory variables for *all* time periods, is zero: $E(u_t \mid \mathbf{X}) = 0, t = 1, 2, \ldots, n$
- Expressed differently, $Cov(u_t, \mathbf{X}) = 0$
- In other words, the error term at time t, u_t, is uncorrelated with each explanatory variable in *every* period
- *Consequence:* Non-zero conditional mean means that OLS coefficients are biased
- *Test:* Difficult to test
- *Potential solution:* Many! Some of which will be covered in this course

Gauss-Markov Assumption 5

There are a number of reasons why the zero conditional mean assumption might fail:

1. Measurement error
2. Omitted variable bias
3. Reverse causality

Measurement Error

- Measurement error is the difference between the observed variable and the true variable: $e_t = x_t - x_t^*$
- For example, historical estimates of GDP are measured with error. In the United Kingdom in the late 19th century, measurement error is ± 20 per cent (Solomou and Weale, 1991)
- *Consequence:* Can lead to attenuation bias ($|\hat{\beta}| < |\beta|$)
- *Potential solution:* Collect more accurate data or instrumental variables

Measurement Error

Proof

We want to estimate the following equation:

$$y_t = \alpha + \beta x_t^* + u_t$$

But x_t^* is unobserved. As we only observe x_t, we actually estimate (as $e_t = x_t - x_t^*$, therefore $x_t^* = x_t - e_t$):

$$y_t = \alpha + \beta(x_t - e_t) + u_t$$

$$y_t = \alpha + \beta x_t + (u_t - \beta e_t)$$

Measurement Error

Proof

Replacing u_t with the new residual term in our expression for the zero conditional mean assumption, $Cov(u_t, x_t) = 0$:

$$Cov(u_t - \beta e_t, x_t) = 0$$

Assuming u_t and x_t are uncorrelated:

$$-\beta Cov(e_t, x_t) = 0$$

Substituting x_t for $x_t^* + e_t$:

$$-\beta Cov(e_t, x_t^* + e_t) = 0$$

And assuming that the measurement error and the true variable are uncorrelated:

$$-\beta Cov(e_t, e_t) = 0$$

$$-\beta Var(e_t) \neq 0$$

Measurement Error

Example

- Download “Measurement Error.xlsx” from Moodle
- The “e_multiplier” parameter controls the degree of time-varying measurement error
- The “e_shifter” parameter controls the degree of time-invariant measurement error
- Let’s vary the degree of measurement error and see how $\hat{\alpha}$ and $\hat{\beta}$ differ from α and β

Omitted Variable Bias

- Omitted variable bias arises when a relevant variable is omitted from the regression
- In other words, when a variable that is correlated with the dependent and independent variable is not included in the model
- *Consequence:* Omitted variable bias means that OLS coefficients are biased
- *Potential solution:* Include the omitted variable

Omitted Variable Bias

Proof

We want to estimate the following equation:

$$y_t = \alpha + \beta x_t + \gamma z_t + u_t$$

But instead we estimate:

$$y_t = \alpha + \beta x_t + e_t$$

where $e_t = \gamma z_t + u_t$
Plugging e_t into our expression for the zero conditional mean assumption, $Cov(u_t, x_t) = 0$:

$$Cov(\gamma z_t + u_t, x_t) = 0$$

Omitted Variable Bias

Proof

Assuming that the population error and the included independent variable are uncorrelated:

$$\gamma Cov(z_t, x_t) = 0$$

Therefore, OLS is only unbiased if $\gamma = 0$ (the omitted variable is not correlated with the dependent variable) or $Cov(z_t, x_t) = 0$ (the included and omitted independent variables are not correlated with each other)

Reverse Causality

- Reverse causality occurs when x_t not only affects but is affected by y_t
- *Consequence:* Reverse causality means that OLS coefficients are biased
- *Potential solution:* Many

Reverse Causality

Proof

Suppose the population process is a system of equations:

$$y_t = \alpha + \beta x_t + u_t \tag{23}$$

$$x_t = \delta + \theta y_t + e_t \tag{24}$$

Consider this simple thought experiment:

1. Shock the error term in equation (23), u_t
2. y_t changes in equations (23) and (24)
3. x_t changes in equations (23) and (24)

Therefore, there is a correlation between x_t and u_t that violates the zero conditional mean assumption, $Cov(u_t, x_t) = 0$

Reverse Causality

Direction of the bias in $\hat{\beta}$

$$\hat{\beta} = \beta + \frac{Cov(u_t, x_t)}{Var(x_t)}$$

- If β is positive (negative) and the covariance term is positive (negative), $\hat{\beta}$ will *overstate* the true absolute magnitude of the effect
- If β is positive (negative) and the covariance term is negative (positive), $\hat{\beta}$ will *understate* the true absolute magnitude of the effect

Next Class

- *Class discussion paper:* Edvinsson, R., 'New annual estimates of Swedish GDP, 1800-2010', *Economic History Review*, 66 (2013), pp. 1101-26.
- Stationarity, filtering and seasonal adjustment

Further Reading

- Wooldridge, *Introductory Econometrics*, chapters 2 and 10
- Stock, J. H., and Watson, M. W., *Introduction to econometrics*, chapter 4

Appendix: Equations (20)-(21)

Starting with the numerator in equation (21):

$$\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y}) = \sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} x_t \bar{y} - \sum_{t=1}^{n} \bar{x} y_t + \sum_{t=1}^{n} \bar{x}\bar{y} \tag{25}$$

$$\sum_{t=1}^{n} x_t y_t - n\bar{x}\bar{y} - \bar{x}n\bar{y} + n\bar{x}\bar{y} \tag{26}$$

$$\sum_{t=1}^{n} x_t y_t - n\bar{x}\bar{y} \tag{27}$$

$$\sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} x_t \bar{y} \tag{28}$$

$$\sum_{t=1}^{n} x_t (y_t - \bar{y}) \tag{29}$$

which is the numerator of equation (20)

Appendix: Equations (20)-(21)

Moving on to the denominator in equation (21):

$$\sum_{t=1}^{n}(x_t - \bar{x})^2 = \sum_{t=1}^{n}(x_t - \bar{x})(x_t - \bar{x}) \tag{30}$$

$$\sum_{t=1}^{n} x_t x_t - \sum_{t=1}^{n} x_t \bar{x} - \sum_{t=1}^{n} \bar{x} x_t + \sum_{t=1}^{n} \bar{x}\bar{x} \tag{31}$$

$$\sum_{t=1}^{n} x_t x_t - n\bar{x}\bar{x} - n\bar{x}\bar{x} + n\bar{x}\bar{x} \tag{32}$$

$$\sum_{t=1}^{n} x_t x_t - n\bar{x}\bar{x} \tag{33}$$

$$\sum_{t=1}^{n} x_t x_t - \sum_{t=1}^{n} x_t \bar{x} \tag{34}$$

$$\sum_{t=1}^{n} x_t (x_t - \bar{x}) \tag{35}$$

which is the denominator of equation (20)