# Table of contents

1. Introduction
2. Recap on OLS
3. Interpretation of regression model
4. Group discussions
5. Moving from theory to empirical work
6. Case studies
7. References
Housekeeping I

- Lectures on Sat 9.30-12.30 every two weeks (with breaks).
- My email address is mcdonaj5@tcd.ie.
- My office hour is Fri 1.30-2.30pm. Room 3022, Dept of Economics, Arts block.
- Am around most of Friday.
About the module

- Not a technical econometrics course.
- I assume knowledge up to the EC8006 *Data Analysis* module.
- Concepts will be illustrated using examples from published papers.
- Some pointers on using econometrics for dissertations.
- More confident when reading published papers.
- More confident using stata.
- Interactive course. You are expected to have read the papers.
Some suggested textbooks

Data description v economic modelling

- Start by drawing a distinction between describing data and the more ambitious task of empirically modelling the economic processes involved.
  - Data indicates graduates earn more than non-graduates.
  - Want to explain why graduates earn more than non-graduates. Understanding the causal relationship.
- Most academic and professional econometric work is concerned with identification of causal relationships.
Education and earnings

![Education pays graph]

- **Unemployment rate in 2009**
  - Doctoral degree: 2.5%
  - Professional degree: 2.3%
  - Master's degree: 3.9%
  - Bachelor's degree: 6.8%
  - Associate degree: 8.6%
  - Some college, no degree: 9.7%
  - High school graduate: 14.6%
  - Less than a high school diploma: 19.0%
  - Average, all workers: 7.9%

- **Median weekly earnings in 2009**
  - Doctoral degree: $1,532
  - Professional degree: $1,529
  - Master's degree: $1,257
  - Bachelor's degree: $1,025
  - Associate degree: $761
  - Some college, no degree: $699
  - High school graduate: $626
  - Less than a high school diploma: $454
  - Average, all workers: $774

How economists think

Parameters and exogenous variables → MODEL → Endogenous variables

Inputs → Outputs
Economic theory suggests relationships but rarely specifies magnitude.

We want to test economic relationships by investigating their compatibility with data.

We want to estimate the empirical magnitude of effects.

We may want to forecast.

Example: Consumer theory suggests quantities purchased should depend on prices and income.

- We wish to test whether such relationships exist.
- We want to know whether demands are income elastic or inelastic, whether different goods are substitutes or complements etc..
Regression

Regression is the most common tool in applied economics.
Helps us to understand the causal relationships between variables.

A simple regression is a relationship between two variables: \( Y \) and \( X \); or Dependent and Independent / explanatory variables.

Multiple regression - there are two or more regressors/independent variables (\( X_1, X_2, X_3..., X_n \)).

Regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values of the explanatory variable(s).
What variables to include?

- Explanatory variable(s) should cause/influence the dependent variable.
- Sometimes **structural** equations come direct from the economic theory and have direct casual interpretation.
- Other times theory will give you a **guide** to variables.
- Avoid the “kitchen sink” regression.
- But “general to specific” approach is attractive.
Recap on the OLS regression model

Assume a **linear** relationship exists between Y and X, which shows how Y changes with X:

\[ Y = \alpha + \beta X + u \]  

- **Y** = dependent variable.
- **X** = independent or explanatory variables.
- **\( \alpha \)** = vertical intercept of straight line.
- **\( \beta \)** = how a change in X affects Y. Slope of line or \( \frac{\Delta Y}{\Delta X} \).
- **u** is a random error term.
Recap on OLS

OLS assumptions: the non-technical version

1. The model must be linear in the parameters.
2. The data are a random sample of the population i.e. residuals are statistically independent/uncorrelated from each other.
3. The independent variables are not too strongly collinear.
4. The independent variables are measured precisely such that measurement error is negligible.
5. The expected value of the residuals is always zero.
6. The residuals have constant variance (homogeneous variance)
7. The residuals are Normally distributed.
This could be a model relating a person’s wage to their education:

\[ wage = \alpha + \beta \text{education} + u \]  

- wage is dollars per hour.
- \( \alpha \) wages when no education. Sometimes not relevant.
- educ is years of education.
- \( \beta \) measures the change in hourly wage given another year of education, holding all other factors fixed or \( \frac{\Delta \text{wage}}{\Delta \text{educ}} \)
- u captures random unobserved factors such as ability, experience, motivation.
Mechanics of OLS

- What we have: actual data on $X$ and $Y$.
- What we do not know: $\alpha$, $\beta$ or $u$.
- Regression analysis uses data ($X$ and $Y$) to estimate $\alpha$, $\beta$ and $u$.
- These estimates of $\alpha$ and $\beta$ are $\hat{\alpha}$ and $\hat{\beta}$.
- We can use statistical theory to satisfy ourselves that the OLS estimates are the “best” available.
**BLUE estimates**

- Under certain circumstances, OLS is **Best**, **Linear**, **Unbiased**, and **Estimator**.
- The proof rests on central limit theorem, normal distributions, sampling theory etc..
- Most textbooks provide a formal proof.
- If you do the exercise over and over again with different parts of the population, the sampling distribution of $\hat{\beta}$ is centred on $\beta$. 
BLUE estimates

- A **unbiased** estimator will yield a mean that is the value of the true parameter of the population.
- A **consistent** estimator is one which approaches the real value of the parameter in the population as the size of the sample, n, increases.
- An **efficient** estimator has the lowest variance.
Sampling theory
Mechanics the OLS estimates

- The error term $u$ plays an important role in OLS.
- $u$ is defined as the difference between particular data point and the true regression line: $u = Y - \alpha - \beta \times X$
- However if we replace $\alpha$ and $\beta$ by their estimates $\hat{\alpha}$ and $\hat{\beta}$, we get a residual (e). $e = Y - \hat{\alpha} - \hat{\beta} \times X$
- The $\hat{\alpha}$ and $\hat{\beta}$ are obtained by minimizing the sum of squared residuals and are called the “Ordinary Least Squared estimators” (OLS).
Mechanics the OLS estimates

- Criteria: Minimise the sum of the squared residuals over the sample data i.e.

\[
\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 = \sum_{i=1}^{n} e_i^2
\]  

- Why is the estimate of the error term squared? To give positive and negative values equal treatment;
Geometry of OLS

$Y = \hat{\alpha} + \hat{\beta}X$
A geometry approach to OLS

- We want to find the **best fitting line** through the XY plot which contains the sample observations.
- Choose the ”best fitting” line which makes the residuals ($e$) as small as possible (formally the one that minimizes the sum of squared residuals).
- The $\hat{\alpha}$ and $\hat{\beta}$ obtained in this way (i.e.obtained by minimizing the sum of squared residuals) are called the “Ordinary Least Squared estimators” of $\alpha$ and $\beta$. 
Mechanics the OLS estimates

- We never worry about doing the maths of the minimisation problem that gives us $\hat{\alpha}$ and $\hat{\beta}$.
- This is done for us by the software (Excel/Stata/eviews/R).
- We’ll have lots of practice on this in two weeks.
- We get something like this.
- $u$ is a random disturbance term.
Recap on OLS

```
.tsset y
   time variable: year, 1964 to 1982
.regress rinu rgnp rinrate

Source | SS    | df | MS
-----|-------|----|----
Model | 20746.3449 | 2  | 10373.1724
Residual | 4738.62733  | 16 | 296.164208
Total  | 25484.9722  | 18 | 1415.83179

Number of obs = 19
F( 2, 16)   = 35.03
Prob > F    = 0.0000
R-squared   = 0.8141
Adj R-squared = 0.7908
Root MSE    = 17.209

.rinp | Coef.  | Std. Err. | t   | P>|t| | [95% Conf. Interval]
-----|--------|-----------|-----|-----|-------------------|-------------------|
    |        |           |     |     |                   |                   |
rgnp | .1691365 | .0205665  | 8.22 | 0.000 | .1255375  | .2127354 |
rinrate | -1.001439 | 2.368749  | -0.42 | 0.678 | -6.022963 | 4.020085 |
_cons | -12.5336 | 24.91527  | -0.50 | 0.622 | -65.35161 | 40.28441 |
```

.predict resid01, residuals
Multiple regression is just an extension

Assume a linear relationship exists between $Y$ and $X_1$ and $X_2$, which shows how $Y$ change with $X_1$ and $X_2$:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u \tag{4}$$

- $Y = \text{dependent variable}$.
- $X_1, X_2 = \text{independent or explanatory variable}$.
- $\alpha = \text{Constant. Relevant}$?
- $\beta_1 = \frac{\Delta Y}{\Delta X_1}; \beta_2 = \frac{\Delta Y}{\Delta X_2}$
Multiple regression is just an extension

- This could be a model relating a person’s wage to their education and experience.

\[ wage = \alpha + \beta_1 \text{education} + \beta_2 \text{experience} + u \]  

- \(\text{wage}\) - dollars per hour.
- \(\text{educ}\) - years of education; \(\text{exp}\) - years of experience
- \(\alpha\) = Wages when education and experience are zero. Relevant?
- \(\beta_1\) measures the change in hourly wage given another year of education, holding all other factors fixed; \(\beta_2\) measures the change in hourly wage given another year of experience, holding all other factors fixed.
- \(u\) captures unobserved factors such as ability.
Additions to the “vanilla” OLS model

- Can add **time trends** if variables are growing.
- Can add a **dummy variable** which is an artificial 0-1 variable created to represent an attribute with two or more distinct categories/levels. Men/women, OECD/non-OECD member.
- **Interactive** terms if variables reinforce each other eg Male*White.
- **Squared terms** of X if the relationship appears to be non-linear. Diminishing returns to experience?
- If we want Y to be a **probability** between 0 and 1 eg, being employed, alternatives to OLS.
Non-linear relationships

Copyright 2014. Laerd Statistics.
Recap on OLS

Trends

S&P Composite Index: Regression to Trend
Real (inflation-adjusted) Price since 1871 with Regression
Variance measured below

This log-scale chart illustrates regression to the trend across 140 years of market history. The peak in 2000 was an unprecedented 152% above trend — nearly double the peak in 1929.
The latest monthly average of daily closes is 70% above trend.

Dr John McDonagh
MSc Economic Policy Methods Seminar
October 10, 2015
Different types of dataset

- **Cross section data** are different units observed at same point in time. Good for investigating relationships between variables which vary widely across countries / individuals but may lack variation in other variables.

- **Time series data** are same unit observed over time and good for investigating where these vary over time.

- **Panel data** combines cross sectional and time series. This is a powerful econometric techniques in capturing unobserved variables.

- **Financial data** are often high frequency and volatility. Good for event studies.
The coefficients

- An OLS estimate gives
  \[ Y = 2 + 0.5X \quad (6) \]

  - What does the intercept coefficient \( \alpha \) mean?
    \( \alpha = 2 \) tells us that if \( X = 0 \), then \( Y \) will be 2 units.

  - What does the slope coefficient \( \beta \) mean?
    \( \beta = 0.5 \) tells us that if \( X \) increases by 1 unit, \( Y \) will increase by 0.5 units.
Cholesterol and TV example

```
. regress cholesterol time_tv

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5.04902329</td>
<td>1</td>
<td>5.04902329</td>
</tr>
<tr>
<td>Residual</td>
<td>28.3220135</td>
<td>98</td>
<td>.289000137</td>
</tr>
<tr>
<td>Total</td>
<td>33.3710367</td>
<td>99</td>
<td>.337081179</td>
</tr>
</tbody>
</table>

Number of obs = 100
F( 1, 98) = 17.47
Prob > F = 0.0001
R-squared = 0.1513
Adj R-squared = 0.1426
Root MSE = .53759

| cholesterol | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-------------|--------|-----------|-------|------|----------------------|
| time_tv     | .0440691| .0105434  | 4.18  | 0.000| .0231461 - .0649921  |
| _cons       | -2.134777| 1.813099  | -1.18 | 0.242| -5.732812 - 1.463259 |
```
The coefficients

- **OLS regression of average daily time spent watching TV in minutes as a function of cholesterol concentration in mmol/L.**

- The regression equation

\[
\text{cholesterol} = -2.135 + 0.044 \times (\text{watchingtv})
\]  

- So an extra minute of watching TV increases cholesterol concentration by 0.044 mmol/L.
You’ll often see regressions with **logs** of variables.

Unit changes in log-X translate into percentage changes in X. Coefficients are **unit-free** so doesn’t matter how TV is measured (minutes) or cholesterol concentration mmol/L.

\[
\log(\text{cholesterol}) = -0.344 + 0.032 \times (\log(\text{watchingtv}))
\] (8)

So a 1% increase in watching TV increases cholesterol concentration by 3%. This is an **elasticity**.
Interpretation of regression model

**t stats**

- Standard errors and t-stats automatically printed from the software.
- These test the null hypothesis that the coefficient is equal to zero (no effect) $H_0: \beta = 0$.
- A low p-value (<0.05) / high t-stat indicates that you can reject the null hypothesis that the coefficient is zero (effect).
- High p-value (>0.05) / low t-stat indicates that you cannot reject the null hypothesis that the coefficient is zero (no effect).
- P-value on TV nearly zero / t stat 4 in cholesterol example. Highly significant.
- Also use f-tests for more complex hypothesis.
Interpretation of regression model: Measuring fit: \( R^2 \)

- \( R^2 \) measures how “well” the sample regression fits the data.
- \( R^2 = \) square of the correlation coefficient between the actual and the predicted values of \( Y \).
- Total variability in \( Y = \) Variability “explained” by \( X + \) Variability that cannot be “explained” and is left as an error.
- Returning to our cholesterol example, the \( R^2 \) was 0.15.
- What does this mean?
Interpretation of regression model

Measuring fit: $R^2$

- $R^2 = 1$ means perfect fit. All data points lie exactly on regression line.
- $R^2 = 0$ means $X$ does not have any explanatory power for $Y$ (i.e. $X$ has no influence on $Y$.) Bigger values of $R^2$ imply $X$ has more explanatory power for $Y$.
- A low value of $R^2$ need not mean that the model is wrong but does imply that it fails to capture the main influences.
- Where might a low $R^2$ still be useful?
The OLS regression model requires several assumptions to be **BLUE**.

- Primarily concerned with the residuals of the model.
- Eye balling the data does help before formal tests.
- **Hososcedasticity** - the probability distribution of the errors has constant variance.
- Independence of errors - the error values are statistically independent of each other. **No autocorrelation.** An issue in time series data.
- **Normality** of error - error error values are normally distributed for any given value of x.
Assessing the model 4. Checking residuals

- Figure (a) shows what a good plot of the residuals should look like. The points are scattered along the x axis fairly evenly with a higher concentration at the axis.
- Figure (b) shows residuals that are not homoscedastic: the variance of the residuals increases as x increases.
- Figure (c) shows residuals are not independent: they follow a non-linear trend line along x. Not correctly specified model (this plot comes from trying to fit a linear regression model to data that follow a quadratic trend line).
Plotting residuals

(A) What you want to see
(B) Not Homoscedastic
(C) Not Independent
Moving from the textbook OLS model to empirical work

- The OLS regression model requires a few assumptions to work to be **BLUE**.
- Primarily concerned with the residuals of the model.
- Can seem overly technical but ignoring them can lead to **bias** even in large samples (inconsistency) and **inefficient** estimates.
- Issues when presenting and submitting articles to journals.
- With practice you can recognise some of the main issues at an early stage and test.
- Lets look at some examples in groups.
Group discussions

Reminder on the reasons for correlation

There can be three reasons for correlation which are not mutually exclusive:

- **Cause**: Changes in X drive changes in Y.
- **Reverse Cause**: Changes in Y drive changes in X.
- **Correlated variable**: Changes in Z drives X and Y.
- Z might be time.
An economist believes there is a link between the quality of institutions (civil service, law and order, etc.) and GDP per capita.

The economist has data on both (assume that there is good, reliable data on institutional quality.)

An OLS regression of GDP per capita on institutions shows a positive, significant and large coefficient on institutions.

Do you have any concerns on such a regression?

Taken from Hall and Jones (1999).
Group discussion 1

GDP per capita (Y)

Good institutions (X)
An economist suggests a new model of inflation that outperforms anything used by central banks around the world.

An OLS model of inflation as a function of cumulative rainfall shows an $R^2$ of 0.99, a large coefficient with a p-value of 0.

Do you have any concerns on such a regression?

This example was used by Hendry (1980) to illustrate issue with many macroeconomic datasets.
Group discussion 2

Cumulative Inflation (Y)

Rainfall (X)
An analyst with an interest in the economics of education has final exam marks for 400 students in a course. They have data on a number of variables including the mark obtained by a student’s room-mate in the same exam. They use OLS to regress student’s final exam mark on the room-mate’s mark. Do you have any concerns on such a regression? Taken from Stinebrickner (2006).
How Rodrik et al tackle case I

Figure 1. The “deep” determinants of income.
1. Data/sample issues

- Need to assume our data are a **random** sample.
- If data are **not representative** or **unreliable** this can threaten results. Not easy to test.
- Common issue with historical data. But does not prevent analysis. Are results robust to different proxy variables?
- Larger sample size more reliable. Standard errors etc depend inversely on the sample size.
- Sometimes may end up observing lots of zeros for a variable. People not working, zero consumption of a good. Cannot ignore them and alternatives to OLS.
2,3 & 4. Endogeneity

- **Simultaneity, omitted variable** and **systematic** error are classic examples of Endogeneity.
- Regressing $Y$ on $X$ will lead to an OLS estimator that is **inconsistent** and **biased**.
- The reason is that the residual is correlated to the $X$ variable: formally $\text{Cov}(X, U)$ is non-zero.
- Endogeneity is perhaps the biggest challenge in econometrics.
Simultaneity

- **Definition.** If changes in the $Y$ variable also cause changes in the $X$ variable.
- **Consequences.** The OLS estimator will be biased and will not disappear in large samples (is inconsistent).
- **Examples** in economics:
  - Prices and quantities.
  - Taxes and Government spending.
  - Growth and institutions.
Omitted variables I

- **Definition.** Omitted variables where $X_2$ is omitted.

  $$ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u $$  

- **Consequences.** A bias depending on $\beta_2$ and relationship between $X_1$ and $X_2$. Bias will be $\beta_1 \left( \frac{Cov(X_1, X_2)}{\text{var}(X_2)} \right)$

- **Examples** in economics: ability in wage equations.

- **Solution:** add the omitted variable (or proxy).
Omitted variables II

**Definition.** Omitted variables where explanatory variables are often unobserved or unobservable and affect both $Y$ and $X$.

**Consequences.** The OLS estimator is biased if the omitted variable is relevant and does not disappear in large samples (OLS is inconsistent).

**Examples** in economics:

- Effect of smoking on wages (ignoring the level of education).
- Effect of smoking on Cancer (ignoring the physical health state).
- Effect of good corporate governance on wages (ignoring management culture).
Measurement error/errors-in-variables

- **Definition.** The relevant *explanatory* variables is estimated with systematic error.

- **Consequences.** The OLS estimator is biased does not disappear in large samples (OLS is inconsistent).

- **Examples in economics:**
  - Recall bias: how much time did you spend unemployed last year?
  - Rounding bias: how much money did you spend on food last week?
  - Measures of permanent rather than actual income.

- **Less severe consequences when measurement error affects the dependent variable.**
Solutions to Endogeneity

- **Instrumental variable** approach has become increasingly popular in econometrics.
- Introduce new variable $Z$.
- Replace the actual realized values of $x$ (which are correlated with $u$) by predicted values of $x$ that are related to the actual $x$ - but uncorrelated with $u$ - then we can obtain a consistent estimator of $Z$. $Z$ must be able to explain $X$; but not be linked to $Y$. Not easy!
- These are tested in **first** and **second** stages respectively.
- Weak instruments can end up doing more harm than good.
5. Multicollinearity

- **Definition.** Two or more X’s in a regression exhibit a close linear relationship.

- **Consequences.** Moderate multicollinearity is not generally problematic and does not cause bias or inconsistency.

Severe multicollinearity can lead to inefficiency and erratic estimates. Consider regressing wages on age, and include age in both months and age in years.

- **Examples:** an issue in macroeconomics.

- **Solutions:** Could try dropping a variable to see if it makes a difference. Try to combine or rewrite X variables, eg oil price, inflation as X’s change to real oil price.
6. Heteroskedasticity

- **Definition** The variance of the error term changes in response to a change in the value of the independent variables.
- Eye balling data. Textbooks give formal tests.
- **Consequences.** Inefficient coefficient estimates. Biased standard errors. Unreliable hypothesis tests
- Mechanical solutions available. Weighted least squares (WLS) or Robust standard errors can be done on Stata.
7. Autocorrelation

- **Definition**. A relationship (positive or negative) between the values of the error in one period and the values of the error in another period. Textbooks give formal tests.

- **Consequences**. Inefficient coefficient estimates Biased standard errors Unreliable hypothesis tests
- **Mechanical solutions available.** HAC Robust standard errors can be done on Stata.
8. Time series data

- Hendry’s rainfall example in case study 2 has wide applicability as many macroeconomic time series grow over time and are **non-stationary**.
- Danger of **spurious** regressions is real. High $R^2$ and p-values have no economic meaning.
- Usual test statistics (t, F etc) are not valid if data are “non-stationary.”
- **Time series analysis** looks at the best way to model such data. Unit root tests look at the stationarity of the sample are the first step.
- Work with difference variables. OLS may not be the best way to model time series. VARS, cointegrations.
Case study Baumol (1986) and convergence

- Baumol (1986) tested if poorer countries grew quicker than richer countries: regression of growth rates on starting GDP. Convergence suggests a negative coefficient.
- Evidence of convergence.
- Only included countries that were rich 100 years ago who were collecting data.
- No data on today’s developing countries.
- Measurement error in the historical data.
- De long (1988) added more countries and found no evidence of convergence.
- Quality of findings were driven by the dataset.
Alviola et al. (2014) on obesity

- Impact of the number of fast-food restaurants on rates of obese children in Arkansas.
- Restaurants may geographically position themselves near consumers that are generally unconcerned about dietary health, and obesity. **Simultaneity**
- Use an instruments:
  - highway proximity: fast-food restaurants often choose locations to take advantage of business from highway travellers.
  - Persons aged 15 to 24: establishments typically locate in places where there is easy access to this age group.
Alviola et al. (2014) on obesity

- Also, omitted variables affect **both** school weight and the number of restaurants within the proximity of a school is very real.
- The instruments generated enough correlation to produce consistent estimates.
Case study 4. Angrist and Krueger (1990) on wages

- Schooling and wages are affected by the **ability** of the person. **Omitted variable.**
- Directly regressing wages rates on schooling will produce inconsistent estimates of model coefficients. $\text{cov}(X_i, \epsilon_i) \neq 0$
- Returns to schooling biased.
- Use **quarter of birth** as an instrument because part of the variation in school years is because of month of birth (with minimum leaving age laws).
Case study 4. Angrist and Krueger (1990) on wages

- e.g. if children are required to enter school in the September of the year in which they turn six and assume that December the 31st is the cut-off date, then children born in the first quarter will be 6 and 6.75 when they enter school, while those born in the fourth quarter will be 5 and .75.

- Wage effects of that part of the variation in school years is not due to the impact of omitted ability.

- But, the correlation between quarter of birth and education is fairly weak.

- The instrumental variables estimate of the rate of return to education is remarkably close to the ordinary least squares estimate, suggesting that there is little ability bias in conventional estimates of the return to education.
Levitt (1997) on police and crime

- When more police are hired, crime should decline.
- But...more police may be hired during crime waves. **Simultaneity**
- Increases in size of police force disproportionately concentrated in election years. An instrument.
- Need to be careful. However...can election cycles affect crime rates through other spending channels?
References


References


