

Econometrics is

Econometrics is

The estimation of relationships suggested by economic theory

Econometrics is

The estimation of relationships suggested by economic theory

The application of mathematical statistics to the analysis of
economic data

Keynes General Theory:

“Men are disposed as a rule and on average to increase their consumption as their income increases but not as much as the increase in their income”

Keynes General Theory:

“Men are disposed as a rule and on average to increase their consumption as their income increases but not as much as the increase in their income”

→ marginal propensity to consume < 1

Keynes General Theory:

“Men are disposed as a rule and on average to increase their consumption as their income increases but not as much as the increase in their income”

→ marginal propensity to consume < 1

Keynes General Theory:

“Men are disposed as a rule and on average to increase their consumption as their income increases but not as much as the increase in their income”

→ marginal propensity to consume < 1

and a **deterministic** mathematical model

- in this case a straight line given by

$$C = b_0 + b_1 Y \quad (1)$$

and

$$dC/dY = b_1 < 1$$

(b_0 and b_1 said to be *parameters* or *coefficients* of the equation)

Keynes General Theory:

“Men are disposed as a rule and on average to increase their consumption as their income increases but not as much as the increase in their income”

→ marginal propensity to consume < 1

and a **deterministic** mathematical model

- in this case a straight line given by

$$C = b_0 + b_1 Y \quad (1)$$

and

$$dC/dY = b_1 < 1$$

(b_0 and b_1 said to be parameters or *coefficients* of the equation)

So theory often gives an idea about the value of the parameter of interest
– but does not provide a definitive answer

In reality relationships between economic variables are not exact. Given data on consumption and income for a sample of individuals/time periods we would not expect all the observations to lie on the straight line implied by the theory in (1), because:

In reality relationships between economic variables are not exact. Given data on consumption and income for a sample of individuals/time periods we would not expect all the observations to lie on the straight line implied by the theory in (1), because:

- factors other than income affect consumption;
- agents with the same income have different tastes. (the more disaggregated the data the more individual heterogeneity between units of observations (regions, firms, individuals) increases

In reality relationships between economic variables are not exact. Given data on consumption and income for a sample of individuals/time periods we would not expect all the observations to lie on the straight line implied by the theory in (1), because:

- factors other than income affect consumption;
- agents with the same income have different tastes. (the more disaggregated the data the more individual heterogeneity between units of observations (regions, firms, individuals) increases

To allow for this **stochastic** variation, modify the deterministic model to include a random error (disturbance) term, u , to capture all factors which affect consumption but are not taken into account explicitly by the model.

$$C = b_0 + b_1 Y$$

To allow for this **stochastic** variation, modify the deterministic model to include a random error (disturbance) term, u , to capture all factors which affect consumption but are not taken into account explicitly by the model.

$$C = b_0 + b_1 Y$$

$$C = b_0 + b_1 Y + u$$

To allow for this **stochastic** variation, modify the deterministic model to include a random error (disturbance) term, u , to capture all factors which affect consumption but are not taken into account explicitly by the model.

$$C = b_0 + b_1 Y$$

$$C = b_0 + b_1 Y + u$$

This means that the model has statistical properties and now becomes a probabilistic rather than an exact (deterministic) description of the world and therefore requires a degree of evidence to accept or overturn it.

To allow for this **stochastic** variation, modify the deterministic model to include a random error (disturbance) term, u , to capture all factors which affect consumption but are not taken into account explicitly by the model.

$$C = b_0 + b_1 Y$$

$$C = b_0 + b_1 Y + u$$

This means that the model has statistical properties and now becomes a probabilistic rather than an exact (deterministic) description of the world and therefore requires a degree of evidence to accept or overturn it.

How much evidence is a matter of debate, but the role of econometrics is to try to assemble that evidence, to obtain estimates of the parameters of an economic model in order to try and validate or reject it at an *acceptable degree of probability*.

Formal mathematical economic modelling - such as (1) - is sometimes the start for econometric analysis, but often the theoretical underpinnings are much less formal.

Consider, as an example, the study of the determinants of earnings. Governments (and individuals) are interested in knowing what the returns (private and social) are from investments in education. It is therefore reasonable to try and find out quantify the returns using data and econometric tools

While formal economic theory (in this case human capital theory: Becker 1963) might specify a precise (quadratic) relationship between pay and education.

$$W = b_0 + b_1 \text{years of education} + b_2 \text{years of education}^2 + u$$

Consider, as an example, the study of the determinants of earnings. Governments (and individuals) are interested in knowing what the returns (private and social) are from investments in education. It is therefore reasonable to try and find out quantify the returns using data and econometric tools

While formal economic theory (in this case human capital theory: Becker 1963) might specify a precise (quadratic) relationship between pay and education.

$$W = b_0 + b_1 \text{years of education} + b_2 \text{years of education}^2 + u$$

Equally common sense and economic intuition might say that we would expect earnings and productivity to increase with the level of education.

Consider, as an example, the study of the determinants of earnings. Governments (and individuals) are interested in knowing what the returns (private and social) are from investments in education. It is therefore reasonable to try and find out quantify the returns using data and econometric tools

While formal economic theory (in this case human capital theory: Becker 1963) might specify a precise (quadratic) relationship between pay and education.

$$W = b_0 + b_1 \text{years of education} + b_2 \text{years of education}^2 + u$$

Equally common sense and economic intuition might say that we would expect earnings and productivity to increase with the level of education.

Often models based on explicit theory imply precise (**structural**) relationships between variables whereas models that rely on economic intuition are less encumbered by theoretical restrictions

Above all any relationship to be estimated must be **causal** ie

Above all any relationship to be estimated must be **causal** ie

- a) the direction of influence runs from the variable(s) on the right hand side of the equation to the variable on the left hand side

Above all any relationship to be estimated must be **causal** ie

- a) the direction of influence runs from the variable(s) on the right hand side of the equation to the variable on the left hand side
- b) the estimated impact is due to that variable alone and not from any unintended association with other variables not in the equation (confounding).

This is a fundamental issue in econometrics and much of the course is focussed on the techniques that help deal with this.

Above all any relationship to be estimated must be **causal** ie

- a) the direction of influence runs from the variable(s) on the right hand side of the equation to the variable on the left hand side
- c) the estimated impact is due to that variable alone and not from any unintended association with other variables not in the equation (confounding).

This is a fundamental issue in econometrics and much of the course is focussed on the techniques that help deal with this.

(In many cases econometrics tries to establish causality by holding other factors fixed but there are cases when other important factors are not observed so a different approach is needed).

The first step in any applied work is to find a suitable data set with which to amass information needed to test a hypothesis.

Most economic data come from non-experimental sources – social science researchers can rarely choose the level of a treatment, observe its outcome and compare the results with a control group. The problems associated with collecting and analysing non-experimental data underlie much of what econometrics is about.

(Computer exercises to help with this)

The first step in any applied work is to find a suitable data set with which to amass information needed to test a hypothesis.

Most economic data come from non-experimental sources – social science researchers can rarely choose the level of a treatment, observe its outcome and compare the results with a control group. The problems associated with collecting and analysing non-experimental data underlie much of what econometrics is about.

(Computer exercises to help with this)

Using the example above this means finding a single data source that has information on both individual pay and education levels.

The first step in any applied work is to find a suitable data set with which to amass information needed to test a hypothesis.

Most economic data come from non-experimental sources – social science researchers can rarely choose the level of a treatment, observe its outcome and compare the results with a control group. The problems associated with collecting and analysing non-experimental data underlie much of what econometrics is about.

(Computer exercises to help with this)

Using the example above this means finding a single data source that has information on both individual pay and education levels.

The next step – before doing any estimation - is to ensure that the data look sensible

(Eg. what units are the variables measured in, do the mean, minimum, maximum values of the variables look to be representative of the population under study).

The first step in any applied work is to find a suitable data set with which to amass information needed to test a hypothesis.

Most economic data come from non-experimental sources – social science researchers can rarely choose the level of a treatment, observe its outcome and compare the results with a control group. The problems associated with collecting and analysing non-experimental data underlie much of what econometrics is about.

(Computer exercises to help with this)

Using the example above this means finding a single data source that has information on both individual pay and education levels.

The next step – before doing any estimation - is to ensure that the data look sensible

(Eg. what units are the variables measured in, do the mean, minimum, maximum values of the variables look to be representative of the population under study).

“garbage in, garbage out”

The data set below is taken from the UK Labour Force Survey – a survey of around 60,000 households undertaken by the government every quarter and freely available to researchers. The LFS contains information on the pay and education (among other things)

A regression of hourly pay on years of education gives

```
reg hourpay yrsed
```

Source	SS	df	MS			
Model	2090.54471	1	2090.54471	Number of obs =	13424	
Residual	1917430.34	13422	142.857275	F(1, 13422) =	14.63	
Total	1919520.89	13423	143.002376	Prob > F =	0.0001	
				R-squared =	0.0011	
				Adj R-squared =	0.0010	
				Root MSE =	11.952	

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yrsed	-.0291423	.0076181	-3.83	0.000	-.0440748	-.0142098
_cons	12.79206	.1497493	85.42	0.000	12.49853	13.08559

Your intuition should tell you that this estimated coefficient looks very odd (negative much too small, implies 1 extra year of earnings is worth -3 pence an hour)

Inspection of the underlying data reveals that

```
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	13424	40.41783	12.15848	16	64
edage	13424	20.24888	13.54194	-8	97
hourpay	13424	12.37681	11.95836	3	607.26
sex	13424	1.524732	.4994066	1	2
yrsed	13424	14.24888	13.54194	-14	91

In this case the maximum (and mean) of years of education (yrsed) looks strange

- this is because in this dataset the age left education variable has missing value codes of 96 and 97 and -8 if respondents don't answer the question.

(so the years of education variable, yrsed = ageleft education – 6 is affected)

Removing all observations with these codes gives

```
su if edage>0 & edage<90
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	13011	41.0538	11.73547	16	64
edage	13011	17.97925	2.863167	10	44
hourpay	13011	12.573	12.07187	3	607.26
sex	13011	1.522558	.4995101	1	2
yrsed	13011	11.97925	2.863167	4	38

which looks more sensible.

As does the regression

```
reg hourpay yrsed if edage>0 & edage<90
```

Source	SS	df	MS			
Model	136113.541	1	136113.541	Number of obs =	13011	
Residual	1759833.61	13009	135.278162	F(1, 13009) =	1006.18	
Total	1895947.15	13010	145.729989	Prob > F =	0.0000	
				R-squared =	0.0718	
				Adj R-squared =	0.0717	
				Root MSE =	11.631	

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yrsed	1.129706	.0356146	31.72	0.000	1.059896	1.199516
_cons	-.9600234	.4386525	-2.19	0.029	-1.819846	-.1002003

which implies each extra year of education is worth an additional £1.13 an hour on pay

We also need to account for other potential influences on pay so that we don't make spurious correlations. Education generally increases with age. Also older workers tend to get paid more than younger workers.

The raw correlation coefficients make this clear.

```
corr if edage>0 & edage<90  
(obs=13011)
```

	age	edage	hourpay	sex	yrsed	lhw
age	1.0000					
edage	-0.1947	1.0000				
hourpay	0.0750	0.2679	1.0000			
sex	-0.0059	0.0054	-0.1271	1.0000		
yrsed	-0.1947	1.0000	0.2679	0.0054	1.0000	
lhw	0.1346	0.3846	0.7364	-0.1950	0.3846	1.0000

If didn't also account for the affect of age on pay, might mistakenly attribute its affect to education. Ordinary least squares (OLS), is a very common method of both separating out all the myriad influences on pay and establishing a ceteris paribus – other things equal – relationship. This is effectively the means by which a causal relationship between the **dependent** variable and a right hand side (independent) variable of interest is established

The basic regression is

```
reg hourpay yrsed if edage>0 & edage<90
```

Source	SS	df	MS			
Model	136113.541	1	136113.541	Number of obs =	13011	
Residual	1759833.61	13009	135.278162	F(1, 13009) =	1006.18	
Total	1895947.15	13010	145.729989	Prob > F =	0.0000	
				R-squared =	0.0718	
				Adj R-squared =	0.0717	
				Root MSE =	11.631	

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yrsed	1.129706	.0356146	31.72	0.000	1.059896	1.199516
_cons	-.9600234	.4386525	-2.19	0.029	-1.819846	-.1002003

]

Now, if control variables are added to the regression such that

```
. reg hourpay yrsed age sex if edage>0 & edage<90
```

Source	SS	df	MS			
Model	199012.993	3	66337.6642	Number of obs =	13011	
Residual	1696934.16	13007	130.463148	F(3, 13007) =	508.48	
Total	1895947.15	13010	145.729989	Prob > F =	0.0000	
				R-squared =	0.1050	
				Adj R-squared =	0.1048	
				Root MSE =	11.422	

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yrsed	1.240628	.0356579	34.79	0.000	1.170733	1.310522
age	.1353509	.0086997	15.56	0.000	.1182983	.1524035
sex	-3.090853	.2004807	-15.42	0.000	-3.483824	-2.697881
_cons	-3.139451	.6874401	-4.57	0.000	-4.486934	-1.791968

Controlling for other factors changes the estimated effect of education.

Also the interpretation of the years of education effect is that it is now a partial differential

$$\delta\text{Pay}/\delta\text{yrsed} = b_{\text{yrsed}}$$

holding age and gender fixed in this case.

Different **control variables** can give different conclusions about the size and significance of the causal relationship under investigation as can different **functional form** of the estimated model.

Why this is so

Which variables to include as controls

How to interpret the statistical significance of the results (and the regression output from the statistical package used to produce these results),

How to assess the statistical accuracy of the estimated relationship and test this against alternatives,

What to do about unobservable control variables

and assessing the appropriateness of the causality assumption

form the main subject matter of this course

In many ways we will explore in more detail many of the issues that were glossed over in your undergraduate econometrics courses while at the same time helping to turn you into applied economists capable of reading applied economics papers and of doing your own applied econometric studies