

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”

? marginal propensity to consume  $< 1$

and a **deterministic** mathematical model  
- in this case a straight line given by

$$C = b_0 + b_1Y \quad (1)$$

and  $dC/dY = b_1 < 1$

( $b_0$  and  $b_1$  said to be parameters of the equation)

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

- factors other than income affect consumption;
- individuals with the same income have different tastes.

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.

$$C = b_0 + b_1Y + u \quad (2)$$

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.

In practice this means trying to give answers to economic questions that require the analysis of data. 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.

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. Formal economic theory (in this case human capital theory: Becker 1963) might specify a precise (quadratic) **causal**<sup>1</sup> relationship between pay and education.

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

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

---

<sup>1</sup> Causality in this context means the direction of causality runs from education to pay and not the other way round. In many cases econometrics tries to establish causality by holding other factors fixed.

The first step is to find a suitable data set with which to amass information needed to test this hypothesis. Having obtained data on individual pay and education and ensured that the data look sensible, (do the mean, minimum, maximum values of the variables look to be representative of the population under study).

```
. su if agelfted
```

Variable	Obs	Mean	Std. Dev.	Min	Max
agelfted	16195	20.28577	15.14968	6	97
hourpay	16107	8.630641	6.184317	.06	120.19
female	16216	.5062284	.4999766	0	1
age	16216	38.46756	11.55726	16	64
yrsed	16195	14.28577	15.14968	1	91
yrsed2	16195	433.5818	1514.724	1	8281

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

This is because it is constructed as Age Left Education – 6 and the age left education variable has a missing value codes of 96 and 97

Removing all observations with this code gives

```
. su if agelfted<90
```

Variable	Obs	Mean	Std. Dev.	Min	Max
agelfted	15588	17.33699	2.538051	6	33
hourpay	15487	8.80566	6.211946	.06	120.19
female	15588	.5039774	.5000002	0	1
age	15588	39.2593	11.03901	16	64
yrsed	15588	11.33699	2.538051	1	27
yrsed2	15588	134.9686	66.95984	1	729

which looks more sensible.

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

```
. corr if agelfted<90
(obs=15487)
```

	agelfted	hourpay	female	age	yrsed	yrsed2
agelfted	1.0000					
hourpay	0.3490	1.0000				
female	0.0009	-0.2106	1.0000			
age	-0.1833	0.1247	-0.0278	1.0000		
yrsed	1.0000	0.3490	0.0009	-0.1833	1.0000	
yrsed2	0.9920	0.3427	-0.0088	-0.1532	0.9920	1.0000

If didn't 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 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 variable of interest is established.

```
. reg hourpay yrsed yrsed2 if agelfted<90
```

Source	SS	df	MS			
Model	73235.6905	2	36617.8453	Number of obs =	15487	
Residual	524342.387	15484	33.863497	F( 2, 15484) =	1081.34	
Total	597578.078	15486	38.5882783	Prob > F =	0.0000	
				R-squared =	0.1226	
				Adj R-squared =	0.1224	
				Root MSE =	5.8192	

  

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yrsed	1.389444	.1461662	9.51	0.000	1.102941	1.675947
yrsed2	-.0203976	.0055457	-3.68	0.000	-.0312678	-.0095274
_cons	-4.187382	.9206517	-4.55	0.000	-5.991967	-2.382797

OLS is one way of obtaining the average effect of the **independent variable** on the **dependent variable** (in this case the hourly pay rate)

This basic regression suggests that education raises hourly pay at the rate by  $1.39 - (2 \times 0.02) \times \text{yrsed}$  (measured in £) for each extra year of education. This is because the differential  $d\text{Pay}/d\text{yrsed} = b + 2\text{byrsed}$  so the effect is not constant but varies with the number of years of education.

So if years of ed. were 10, then 1 extra year is worth  
 $1.39 - .04 \times 10 = \text{£}0.99$  (99 pence an hour)

If years of ed. were 15, then 1 extra year is worth  
 $1.39 - .04 * 15 = \text{£}0.79$  (79 pence an hour)

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

```
. reg hourpay yrsed yrsed2 age female if agelfted<90
```

Source	SS	df	MS			
Model	124775.998	4	31193.9994	Number of obs =	15487	
Residual	472802.08	15482	30.5388244	F( 4, 15482) =	1021.45	
Total	597578.078	15486	38.5882783	Prob > F =	0.0000	
				R-squared =	0.2088	
				Adj R-squared =	0.2086	
				Root MSE =	5.5262	

  

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yrsed	2.717438	.1437637	18.90	0.000	2.435645	2.999232
yrsed2	-.0675497	.0054266	-12.45	0.000	-.0781864	-.056913
age	.118549	.0042063	28.18	0.000	.1103042	.1267938
female	-2.635273	.0890817	-29.58	0.000	-2.809884	-2.460662
_cons	-16.20252	.9636054	-16.81	0.000	-18.0913	-14.31374

Controlling for other factors changes the effect of education dramatically. The estimated effect in the example above is almost double that in the regression that contained no controls.

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

$$d\text{Pay}/d\text{yrsed} = b + 2b\text{yrsed}$$

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.

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.