

Simple 2 Variable Regression Analysis

Regression analysis is primarily concerned with quantifying the relationship between variables

Does more than measures of association like the correlation coefficient since it implies that the level of one variable directly influences the level of the other

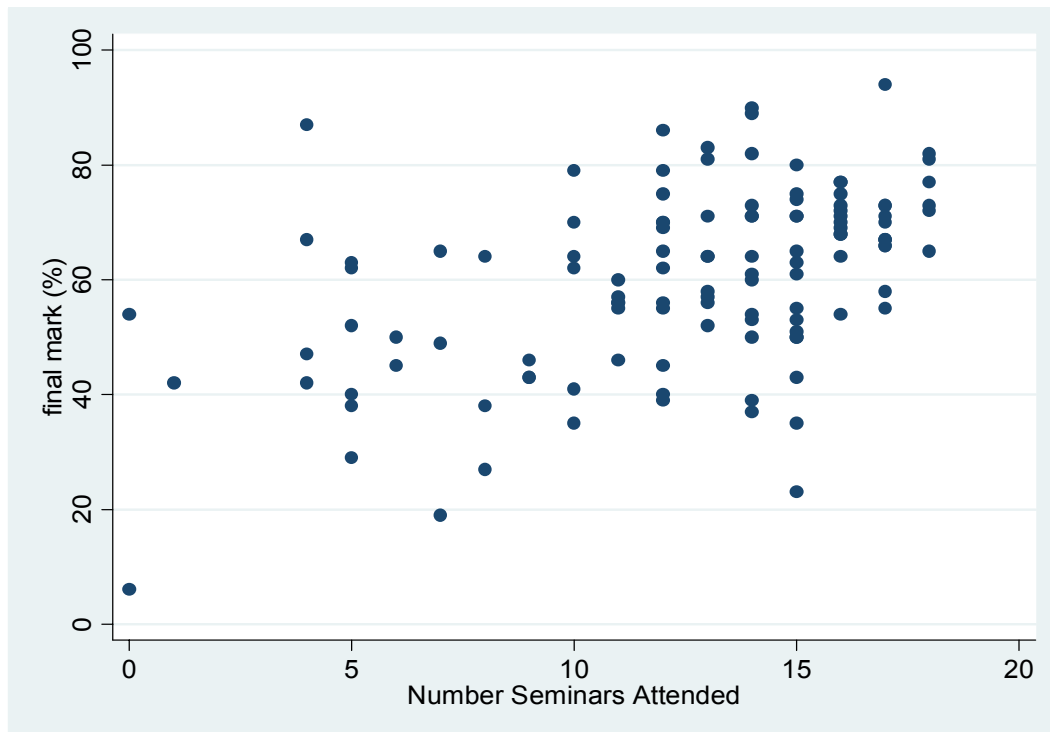
hence there is a **causal** relationship between the variable whose behaviour we would like to explain

- the **dependent** variable (usually denoted by the letter y)

and the **explanatory (or independent)** – usually denoted by the letter X - variable that we think can explain this behaviour

The **scatter diagram** exploring the relationship between the number of seminars and the final mark in the Ec2203 course looks like this

(each dot represents an individual mark and the number of seminars that individual attended during the year)



The Stata command to obtain this graph is

```
scatter mark num_sems
```

Note that what will be the dependent variable values appear on the Y axis and the explanatory variable values appear on the X axis

However this is still just an association and says nothing about how large that association is nor does it distinguish the direction of causality

Causality is assumed to go from the X variable (number of seminars) to the Y variable (final mark) and **not** the other way round

The idea is to summarise this relationship more formally and precisely

This is where regression analysis is useful.

If we fit a line through some data then this will give a predicted value for the dependent variable based on the value of the X variable and the values for the constant and the slope

The difference between the actual and predicted value gives the residual

$$u_i = y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 X_i$$

The larger the residual the worse the prediction

A residual exists for each observation in the data set

Intuitively the smaller the set of residuals the closer the line is to the actual values

To avoid the issue that positive residuals may offset negative residuals use the principle of **Ordinary Least Squares (OLS)**

- Try to fit a (straight) line through the data based on “minimising the sum of squared residuals”

Deriving OLS Estimates

Example: 2 variables Y and X
2 observations

Y	X
3	1
5	2

Try to fit a (straight) line of best fit using the OLS principle:

Minimise the sum of squared residuals

Given the equation of a straight line,

$$Y = b_0 + b_1X$$

we know

when $X = 1$, then the predicted value of Y

$$\hat{y} = \hat{b}_0 + \hat{b}_1(1) \quad (1)$$

and when $X = 2$, then the predicted value of Y

$$\hat{y} = \hat{b}_0 + \hat{b}_1(2) \quad (2)$$

It follows that the estimated residual for the 1st observation (where $y=3$ and $x=1$)

$$u_1 = y_1 - \hat{y}_1 = 3 - \hat{b}_0 - \hat{b}_1$$

and for the 2nd observation (where $y=5$ and $x=2$)

$$u_2 = y_2 - \hat{y}_2 = 5 - \hat{b}_0 - 2\hat{b}_1$$

OLS: minimise the sum of squared residuals

$$S = u_1^2 + u_2^2 = (3 - \hat{b}_0 - \hat{b}_1)^2 + (5 - \hat{b}_0 - 2\hat{b}_1)^2$$

$$S = (9 + \hat{b}_0^2 + \hat{b}_1^2 - 6\hat{b}_0 - 6\hat{b}_1 + 2\hat{b}_0\hat{b}_1)$$

$$+ (25 + \hat{b}_0^2 + 4\hat{b}_1^2 - 10\hat{b}_0 - 20\hat{b}_1 + 4\hat{b}_0\hat{b}_1)$$

$$S = (2\hat{b}_0^2 + 5\hat{b}_1^2 - 16\hat{b}_0 - 26\hat{b}_1 + 6\hat{b}_0\hat{b}_1 + 34) \quad (A)$$

To find values of \hat{b}_0 and \hat{b}_1 which minimise this sum, use simple calculus rules

$$1) \quad \frac{dS}{d\hat{b}_0} = 0 \quad \text{and} \quad 2) \quad \frac{dS}{d\hat{b}_1} = 0$$

$$\rightarrow 4\hat{b}_0 + 6\hat{b}_1 - 16 = 0 \quad 6\hat{b}_0 + 10\hat{b}_1 - 26 = 0$$

This gives 2 simultaneous equations

$$2\hat{b}_0 + 3\hat{b}_1 = 8$$

$$3\hat{b}_0 + 5\hat{b}_1 = 13$$

which can solve for unknown values of \hat{b}_0 and \hat{b}_1

Using rules for simultaneous equations, find that

$$\hat{b}_0 = 1 \text{ and } \hat{b}_1 = 2$$

So the estimated regression line becomes

$$\hat{Y} = 1 + 2X$$

ie the intercept (constant) with the y axis is at 1 and the slope of the straight line is 2

Now need to do this more generally so can apply the technique to any possible combination of (x, y) data pairs and any number of observations

If we wish to fit a (straight) line through N observations, then the OLS principle says choose \hat{b}_0 and \hat{b}_1 to minimise

$$S = u_1^2 + u_2^2 + \dots + u_N^2 = \sum_{i=1}^N u_i^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$\text{sub. in } \hat{y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

$$\begin{aligned} S &= (Y_1 - \hat{b}_0 - \hat{b}_1 X_1)^2 + \dots + (Y_n - \hat{b}_0 - \hat{b}_1 X_n)^2 \\ &= Y_1^2 + \hat{b}_0^2 + \hat{b}_1^2 X_1^2 - 2\hat{b}_0 Y_1 - 2\hat{b}_1 X_1 Y_1 + 2\hat{b}_0 \hat{b}_1 X_1 \\ &\quad + \dots \\ &\quad + Y_n^2 + \hat{b}_0^2 + \hat{b}_1^2 X_n^2 - 2\hat{b}_0 Y_n - 2\hat{b}_1 X_n Y_n + 2\hat{b}_0 \hat{b}_1 X_n \\ &= \sum Y_i^2 + N\hat{b}_0^2 + \hat{b}_1^2 \sum X_i^2 - 2\hat{b}_0 \sum Y_i - 2\hat{b}_1 \sum X_i Y_i + 2\hat{b}_0 \hat{b}_1 \sum X_i \end{aligned}$$

This is just a generalised version of (A) above

Again, find values of \hat{b}_0 and \hat{b}_1 which minimise this sum, using simple calculus rules

$$1) \quad \frac{dS}{d\hat{b}_0} = 0 \quad \text{and} \quad 2) \quad \frac{dS}{d\hat{b}_1} = 0$$

but now these 2 (1st order) minimisation conditions give

$$(1) \quad \frac{\partial S}{\partial \hat{b}_0} = 0 \Rightarrow 2N\hat{b}_0 - 2\sum Y_i + 2\hat{b}_1 \sum X_i = 0$$

and

$$(2) \quad \frac{\partial S}{\partial \hat{b}_1} = 0 \Rightarrow 2\hat{b}_1 \sum X_i^2 - 2\sum \sum X_i Y_i + 2\hat{b}_0 \sum X_i = 0$$

(1) and (2) are known as the **normal equations**

Using the fact that the sample means of Y and X

$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N} \Leftrightarrow N\bar{Y} = \sum_{i=1}^N y_i \quad \text{and}$$

$$\bar{X} = \frac{\sum_{i=1}^N x_i}{N} \Leftrightarrow N\bar{X} = \sum_{i=1}^N x_i$$

can re-write (1) as

$$2N\hat{b}_0 - 2N\bar{Y} + 2\hat{b}_1 N\bar{X} = 0$$

and so obtain the formula to calculate the OLS estimate of the intercept

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} \quad (3)$$

(*** learn this **)

Sub. this into (2) gives

$$\hat{b}_1 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{b}_1 \bar{X}) n \bar{X} = 0$$

$$\hat{b}_1 (\sum X_i^2 - n\bar{X}^2) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

Dividing both sides by 1/N

$$\hat{b}_1 \left(\frac{1}{N} \sum X_i^2 - \bar{X}^2 \right) = \frac{1}{N} \sum X_i Y_i - \bar{X}\bar{Y}$$

which gives the formula to calculate the OLS estimate of the slope

$$\hat{b}_1 \text{Var}(X) = \text{Cov}(X, Y)$$

$$\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (4)$$

(**** learn this ****)

(4) has a nice, clear intuitive meaning about the influence of the variable X on the size of the slope, since it shows that

- a) the greater the covariance between X and Y
- b) the smaller the variance of X

the larger the (absolute value of the) OLS estimate of \hat{b}_1

It is equally important to be able to interpret the effect of an estimated regression coefficient

Given OLS essentially passes a straight line through the data, then given

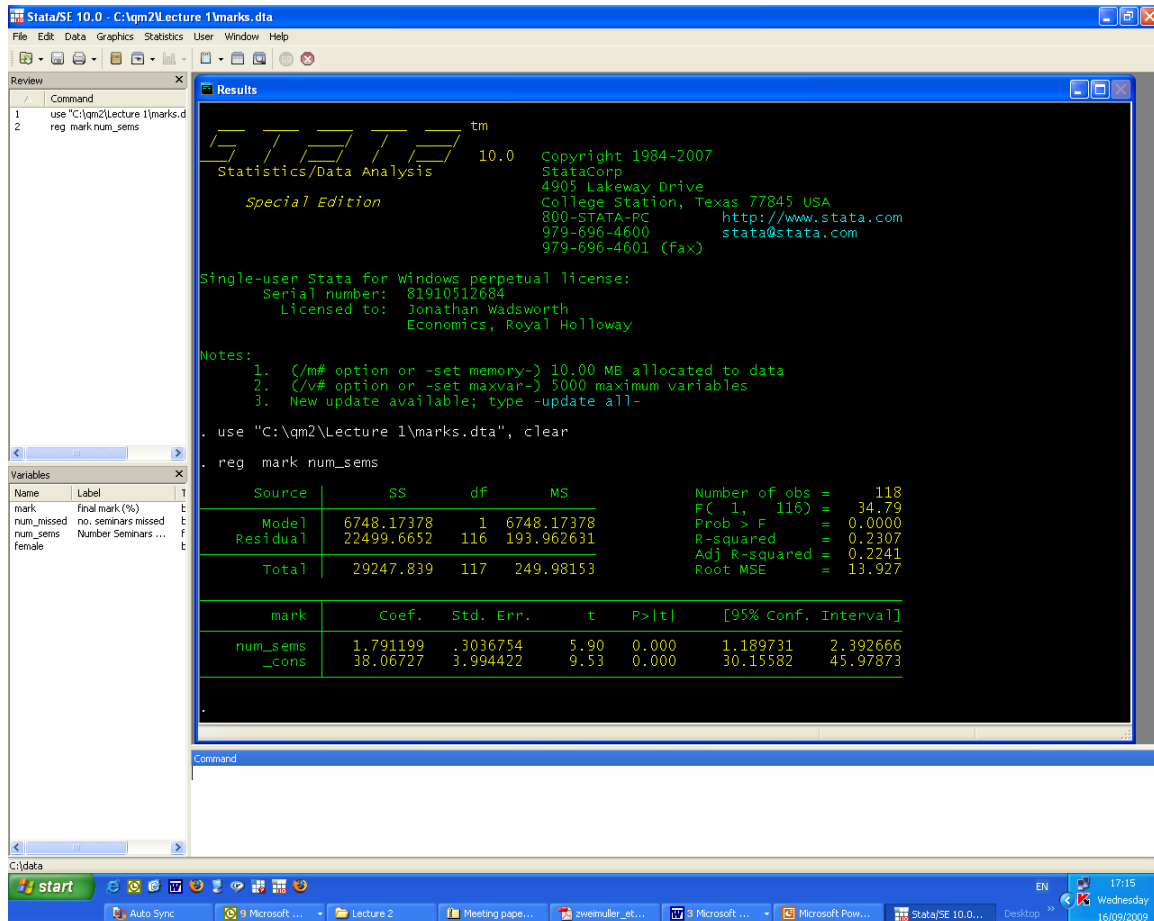
$$\hat{y} = \hat{b}_0 + \hat{b}_1 X$$

it follows that $\frac{dy}{dX} = \hat{b}_1$

So the OLS estimate of the slope will give an estimate of the *change* in the dependent variable y following a *unit change* in the level of the explanatory variable

$$dy = \hat{b}_1 dX$$

The Stata command to run an OLS regression and the resulting output looks like this



Make sure you know how to interpret the information given in the output

[To learn in Stata

Command

Function

use *datasetname.dta*

- read in the file called *datasetname*

describe

- describes the variables in the data set

tabulate *varname*

- tabulates the values and frequencies of the variable *varname*

sum *varname*

- *summarises* the means and standard deviation and other measures of dispersion of the variable *varname*

reg *depvar explanvar*

- does an OLS regression of *depvar* on *explanvar*