

# **Quantitative Research Methods**

**Week 3: -Measures of Association and  
Correlation  
-Introduction to regression**

# Measures of Association and Correlation

# Overview

- We have established how to **measure** whether there is a ***significant relationship*** between two variables
  - Recap: how?
- But there are other ways of measuring relationships eg ***strength and direction of the relationship***
- There are **statistical tests** to allow us to do this
- Sometimes it is also useful to present information in different ways
- Knowing the **strength and direction** of a relationship gives us the first opportunity to **predict data**, within a margin of error
- Bivariate relationships, at this stage, may be related without indicating **causality**
- At a later stage, we will consider causality

# Bivariate relationships

- Bivariate: between two variables
- We are interested in:
  - The magnitude (strength) of the relationship – **how closely related are the two variables?**
  - The direction of the relationship
  - is an increase in A correlated with an increase or a decrease in B?
  - The significance
  - how likely is it that the relationship occurred by chance?

# Examining data

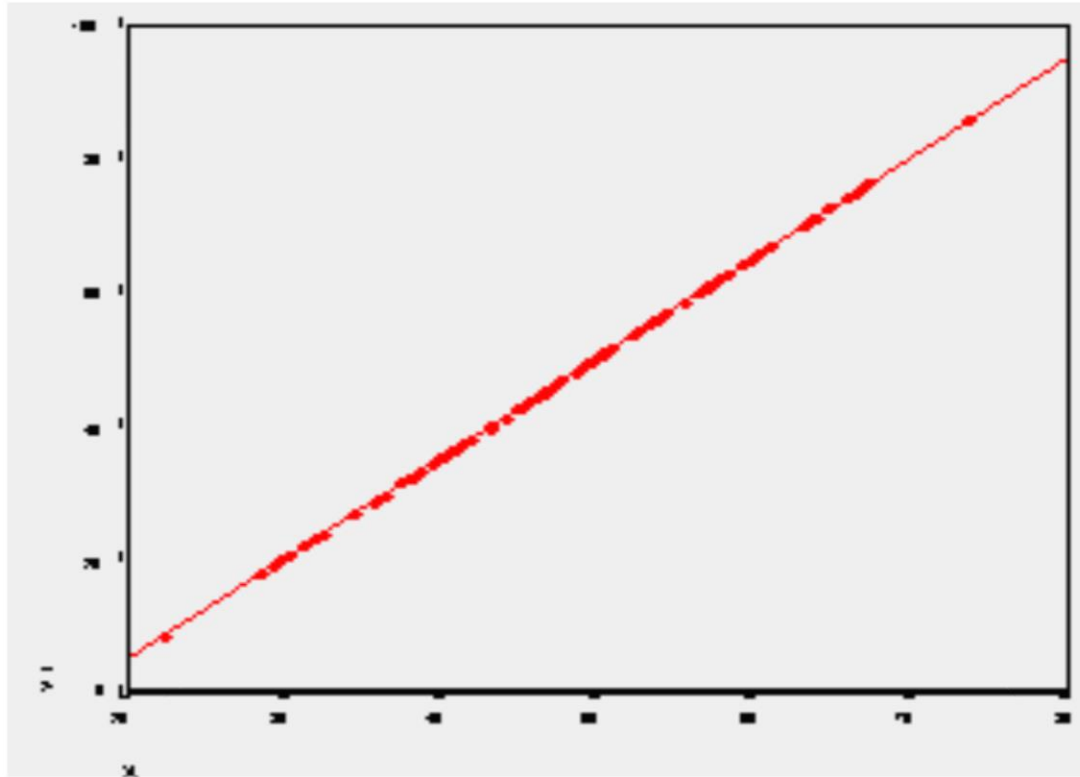
- **Previously** we've looked at **crosstabulations**; a crosstab is the best way to present relationships between nominal or ordinal data.
- However, **crosstabs are not appropriate for the study of scalar** data; imagine what a crosstab would look like with 100 categories!
- **In place of crosstabs**, when looking at interval data **we use scatter plots**. The x variable (independent) is located on the horizontal axis and the y variable (dependent) is located on the vertical axis

# Examining data

- When we have a large  $n$ , scatterplots can be extremely messy, because we have so many points (representing cases) on the graph
  - This can be reduced by using a random sample of the data, but the data as a whole will then need to be checked as well
- The correlation coefficient provides us with a summary of the relationship between the two variables (bivariate relationship). Correlation coefficients indicate the extent to which two variables covary
- Correlation coefficients can be tested for significance

# Scatterplot - example

$r=1.00$



This is an **example of a perfectly correlated relationship** – a change in X is accompanied by a uniform change in Y. This is called a regression line. Most relationships do not feature perfect correlation, so cases will be scattered either side of the regression line. The wider they are on either side, the less correlated the two variables are.

# Pearson Product-Moment Correlation Coefficient ( $r$ )

- The Pearson Product-Moment Correlation Coefficient ( $r$ ), aka correlation coefficient or Pearson's  $r$ , is a **measure** of the degree of linear **relationship** between two variables.
  - NB it does not distinguish between cause and effect.
- The Pearson's  $r$  is a number between -1 and 1:
  - 1 or -1 represents a perfect relationship, and 0 represents no relationship. Whether the coefficient is positive or negative indicates the *direction* of the relationship.

# Pearson Product-Moment Correlation Coefficient ( $r$ )

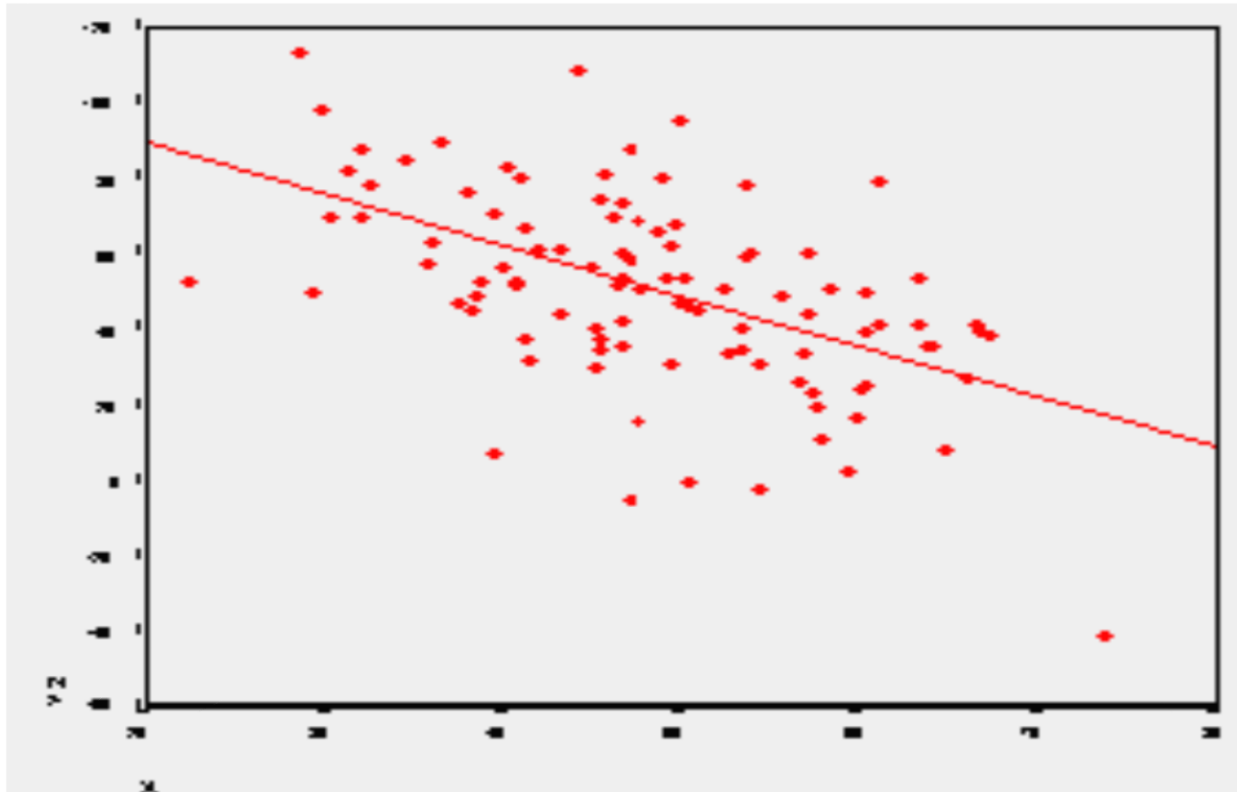
- A **positive sign** means that an upward change in one variable is accompanied by an upward change in the other variable.
  - Eg temperature and bottled water sales may be positively correlated
- A **negative sign** means that an upward change in one variable is accompanied by a downward change in the other variable.
  - Eg supply and demand may be negatively correlated: an increase in supply may result in a decrease in demand and vice versa.

# Pearson Product-Moment Correlation Coefficient ( $r$ )

- Eg temperature may be negatively correlated with thermal underwear sales
- The **sign of the coefficient has nothing to do with the strength of association.**
  - 0.75 is just as strong as +.75. -.90 is much stronger than +.45.

# Example

$r = -.54$



Looking at the graph above we can identify a **negative relationship**. The points are roughly scattered around the line- we see a moderate relationship.

# Rules of thumb for assessing the Pearson's correlation coefficient

0.80-1.00 Strong Association between Variables

0.60-0.79 Moderate-Strong Association

0.40-0.59 Weak-Moderate Association

0.30-0.39 Fairly Weak Association

0.20-0.29 Weak Association

0.00-0.19 Little, if any association

When there is no linear relationship, the correlation will be 0.00; when there is a perfect linear relationship, the correlation will be 1.00 or -1.00.

# Rules for using the correlation coefficient

## 1. Appropriate data

- Both of the variables, X and Y, must be measured by an interval scale.

- Do not use nominal data,

e.g. party of vote.

A dichotomous variable such as 0=male and 1=female may be used, because it is possible to conceptualise a scale.

Ordinal (ie rank) variables use a different measurement

- we will come back to this.

# Rules for using the correlation coefficient

## 2. Effect of Outliers

- An **outlier** which falls near the regression line would **increase** the size of the correlation coefficient.
- An **outlier** that falls some distance from the original regression line would **decrease** the size of the correlation coefficient.
- the **further away an outlier** is from the expected value, the **weaker the relationship** between the variables will be deemed to be (and vice versa)
- The **smaller the sample** size, the **greater the effect** of the outlier. If the sample size is large enough, the outlier will have little or no effect on the size of the correlation coefficient.

# Correlation and Causation

- **Beware! Correlation does not mean causation!**
  - **There may be an intervening variable outside your equation.**  
Eg the sale of **ice lollies** and **bottled water** are correlated, but both are caused by **hot weather** rather than each other
  - The **relationship** may be spurious, ie may simply exist by **chance**
    - Example: chocolate and Nobel prizes  
<http://www.bbc.co.uk/news/magazine-20356613>
  - It is not clear if X is influencing Y, or Y is influencing X, or X and Y are both influenced by Z, or if X and Y just happen to be similar
  - This is why meaningful hypotheses must always be generated from theory. Correlation simply means that as one variable changes so does the other.

# Spearman's Rank Order Correlation

- Remember that the Pearson's  $r$  is for interval/scale variables (Spearman's correlation coefficient, can be used when data is ranked and is a non-parametric analysis on ordinal data).
- If you are interested in studying ordinal data you should use the Spearman's Rank Correlation Coefficient
- This is similar to Pearson's  $r$  but it is based on the rank order of values and not their absolute value
- Spearman's rank is available in SPSS – we will see how to do it in the seminar

# Using correlation to predict

- “If two variables are associated we can predict **people’s characteristics** on one variable given a knowledge of their characteristics on another variable.
- Thus if we can predict more accurately given this knowledge than without, the **two variables must be correlated**. How much better we do so, acts as a measure of strength of association.” De Vaus pp. 167
- The key thing in prediction is that, unless the correlation coefficient is 1 or -1 (ie perfect correlation), there will be a degree of error

# Using correlation to predict

- The key thing in prediction is that, unless the correlation coefficient is 1 or -1 (ie perfect correlation), there will be a **degree of error**
- Both the Pearson's  $r$  and the Spearman's rank coefficient are based on the **proportional reduction of error**
- $E1 - E2 / E1$  where
  - $E1$  is the difference between the mean of  $Y$  and the observed value of  $Y$
  - $E2$  is the difference between the predicted value of  $Y$  given  $X$  and the observed value of  $Y$
- Understanding error does not make our predictions more accurate, but **it does make them more reliable**
  - We know how useful our predictions will be, the probability of being right and the likely margin of error
  - Understanding error is something we will come back to when looking at regression

# **Introduction to regression**

# Overview

- From correlation to regression
- Why is regression so exciting?
- The basic principles
- How to estimate the value of one variable if you know the value of the other
- Accounting for error
- Interpreting results
- Seminar – we will first do a walk through, then you try!

# From correlation to regression

- With correlations, we explored the relationship between two variables, X & Y
- X & Y can be plotted on a scatterplot
- If they are correlated – ie if a relationship exists between them - a regression line in the scatterplot shows the line of best fit

# From correlation to regression

- **Error** is an important part of determining the relationship between two variables
  - If the relationship is not **perfect**, **not all cases will sit on the line** of best fit
  - some will be above, some below. **The regression line minimises this error** by being as **close** as possible to the **true value**. When we do not know the true value, it is useful to know the margin of error to give us an idea of how far away we might be
- From this, we have the first element of prediction
- if we know X, then we can estimate Y, within a certain margin of error
- Regression is an extension of this principle

# Why is regression so exciting?

- It allows us to do two important things that we have not yet done in any meaningful way:
  - **Predict outcomes**  
if we know the relationship between  $X$  and  $Y$ , we can predict  $Y$  when we know  $X$
  - Get closer to determining causality
  - it shows what effect the independent variable has on the dependent variable

# Why is regression so exciting?

- When we get onto **multiple regression** next week, it gets even more exciting!
  - We can then consider **more than one independent** variable at a time, and can control for the effect of one independent variable when testing another
  - **Eg** if we want to know what impact age and sex have on voting, we can find out the effect that age has *independent of sex*, and the effect that sex has *independent of age*
- How cool is that?!

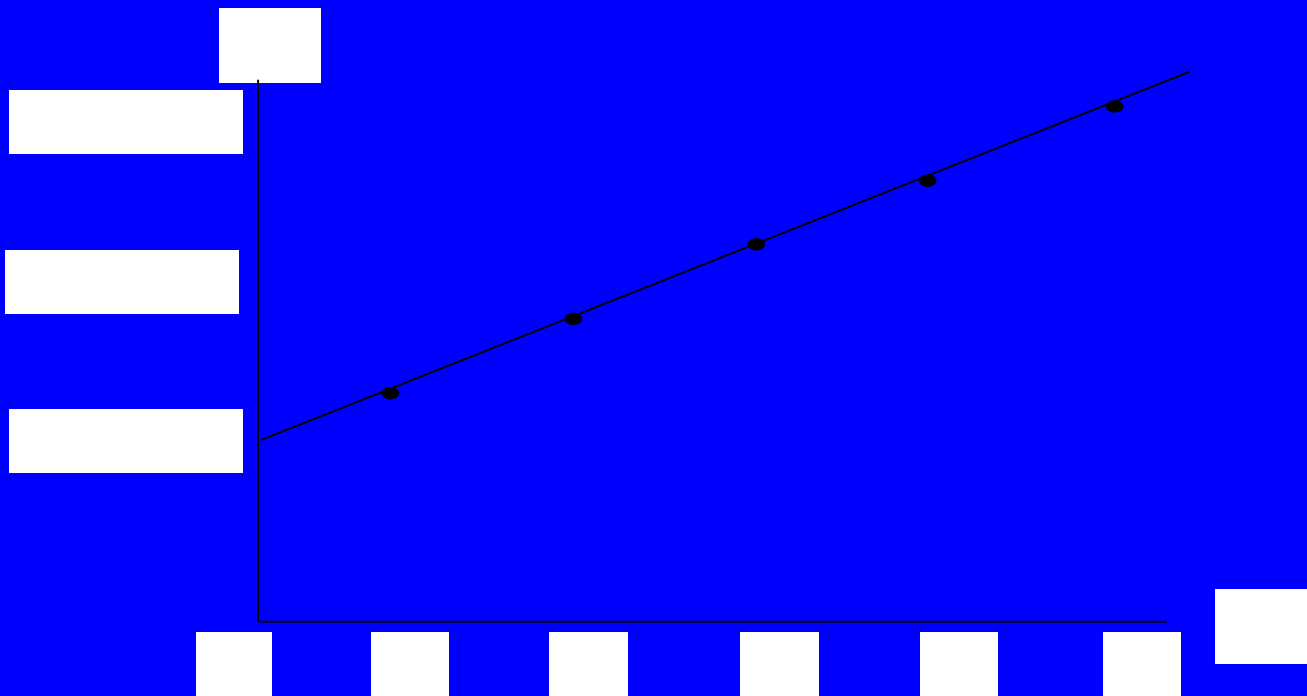
# The basic principles

- OLS regression allows us to predict the value of our dependent variable if we know the value of our independent variable, by drawing a line which gives us the best prediction with the least amount of error
- Regression is based on a simple formula:  
$$Y = a + bX$$

Where Y is the **dependent** variable and  
X is the **independent** variable  
a is the Y **intercept** (ie the value of Y when X = zero)  
– also sometimes referred to as the **constant**  
b is the **slope** of the line  
+ indicates a **positive** relationship  
- indicates a **negative** relationship (ie  $Y = a - bX$ )

This is how we would represent the following linear equation:

$$Y = 5 + 2X$$



$$Y = 5 + 2X$$

| X | Y  |
|---|----|
| 0 | 5  |
| 1 | 7  |
| 2 | 9  |
| 3 | 11 |
| 4 | 13 |
| 5 | 15 |

# Real world example

- A real world example of this linear relationship is this familiar formula:

$$Y = 32 + 9/5X$$

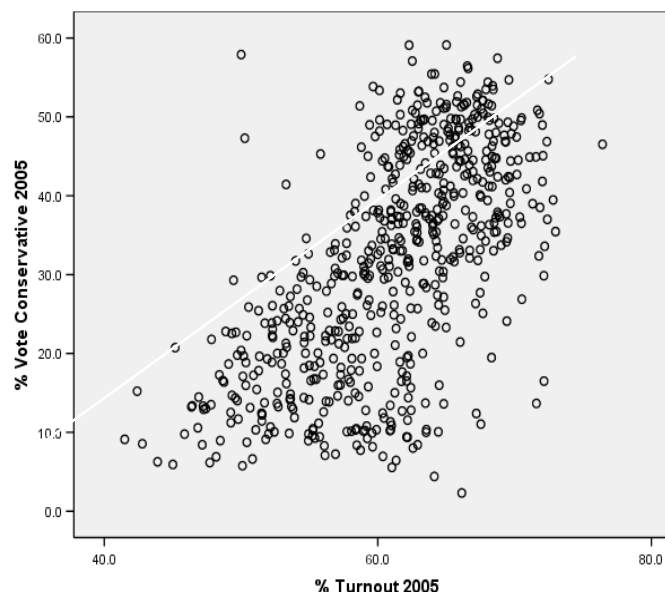
- Where the temperature in Fahrenheit (Y) is an exact linear function of the temperature in Celsius (X).
- ‘a’ is the constant – we always start with 32 (‘a’ = the value of our dependent variable when our independent variable is zero).
- ‘b’ (9/5) would be the slope of the line if we were to plot out the values.

# Another example

- Let us consider the following equation:  
$$Y = 5 + 0.6X$$
- The value of ' $a$ ' (5) indicates that when  $X$  is 0,  $Y$  is 5, so the line intercepts the  $Y$  axis at  $X=0$ ,  $Y = 5$
- The  $+$  indicates that the slope of the line is positive
- The value of  $b$  (0.6), the slope, is also known as the coefficient of the regression line.
- The value for  $b$  (the regression coefficient) quantifies any correlation between our dependent and independent variable. It indicates by how many units  $Y$  (the dependent variable) will change given a one-unit change in  $X$  (the independent variable).
- So when  $X$  increases by 1,  $Y$  increases by  $b$

# Error again!

- As social scientists, we rarely deal with exact linear functions. Most of the time our observations will have a range of values.
- Unless a relationship is perfectly correlated, **the regression line can only be an estimate – the true value may be lower or higher**
- **Ordinary least squares (OLS)** regression produces **coefficients** which have the **least amount of error**



# Error (cont)

- The **regression line** will provide a **range** of expected values for Y depending upon the value of X
- The **difference** between the **expected** value and the **actual** observed value is called the **error** term or residual
- Many times the **regression line will not actually touch *any* of the observation points** on our scatter plot, leaving a gap between our regression line and our observed value
- **Analysis** with SPSS provides the coefficient value which makes all the **residuals as small as possible**. This is known as **ordinary least squares regression (OLS)**

# Error (cont)

- In order to determine the line which provides the best fit, we **square the error** terms (residuals). We find the **variance**.
- We do not just add the values of each residual **because the negative and positive residuals** will cancel each other out. Instead we *square* the residuals to produce positive values for each observation (hence 'least squares').
- 'We *square* the residuals, rather than just *sum* them, because the sum of residuals for *any* line that passes through the point that is the mean for both the dependent and independent variable will equal zero.
- To eliminate the positive and negative signs, the residuals are squared so that we are only dealing with positive numbers' (G. Argyrous (2005), Statistics for Research, p. 166)

# Accounting for error

- The proper formula for regression has to include the element of error
- Think about the original equation:  
 $Y = a + bX$
- If this is **not a perfect relationship**, then the value of Y can only be an **estimate**
- We can represent Y with a hat –  $\hat{y}$  – to indicate that it is a predicted value rather than an actual value :  $\hat{y} = a + bX$
- Otherwise we have to incorporate error into the equation – the estimated value of Y will be the actual value of Y plus a degree of error

# Accounting for error

- The **less strong** the relationship, the **higher the margin of error**
- So we rewrite the equation as follows:
- **$Y = a + bX + e$** 
  - Where Y is the dependent variable and X is the independent variable
  - a = the intercept (the value of Y when X is zero)
  - b = the slope (the amount Y changes for an increase in X of 1)
  - **e = the error term (residual)**

# Using our results to make a prediction

- Knowing the **intercept** and the **slope**, we can now predict the value of Y for a given value of X and make the lowest error in each of our predictions
- Say we want to measure the impact of education on earnings
- Imagine that everyone earns a base salary of £5078, plus a premium of £732 for every year of formal education
- The following calculation produces an estimated salary for an employee with 10 years of formal education.

$$\begin{aligned}\hat{Y} &= 5078 + 732 X \\ &= 5078 + 732 (10) \\ &= 5078 + 7320\end{aligned}$$

$$\hat{Y} = \text{£}12,398$$

# Using our results to make a prediction

- **Warning!** The intercept is a number we should not always accept at face value
- Our calculations allow us to extrapolate from this data so we can *predict* what a person with 0 years of formal education would make, ie £5078 annually
- However, if we really were interested in **knowing the income of those with no education we would be much better off gathering** real data.

# Interpreting results

- We want to know the following things:
  - **Is the relationship positive or negative?** ( does an increase in X result in an increase or decrease in Y?)
    - Look at the predicted **slope**. If this is signed positively, the slope goes up, ie an increase in X results in an increase in Y. If it is signed negatively, an increase in X will result in a decrease in Y
  - **What is the predicted change in Y for an increase of 1 in X?** (ie the slope)
    - This is the coefficient. SPSS calculates this for us. If the relationship is  $Y = 2X$ , then when X increases by 1, Y will increase by 2, making the coefficient 2. If the coefficient is negative, so is the **slope**, ie when X goes up, Y goes down
  - **What is the value of the intercept?** what is the value of Y when X is zero?
    - This is the starting point of the equation, and in SPSS output it is called the **constant**
  - **What is the error term?**
    - SPSS reports this as the **standard error**. The true value should fall within the range created by the estimate plus or minus the error term

# Interpreting results (cont)

- **Is the relationship significant?** is it random or is it likely to be a true relationship? Remember, the less likely that it is to be random, the better, so a value of 0.05 or lower is good
  - This is easy to spot in output as we have seen it before
- **How strong is the relationship between X and Y?**
  - This is a little more complicated. In correlation, we used a Pearson's  $r$  to determine the strength of the relationship: 1 or -1 meant a perfect relationship, 0 meant no relationship
  - **In regression, we use a term called  $R^2$** , which works on similar principles

# $R^2$

- The  $R^2$  compares predictions made using the regression line to predictions made knowing only the mean value, and uses this to determine how much more useful it is to know the regression line than just the mean
- So the  **$R^2$  shows how much of the variance can be explained by the regression line**
- $R^2$  is measured between **0 and 1**: the closer to **1**, the **stronger** the relationship

# $R^2$

- When  $R^2 = 1$ , that means that **100% of the variation in Y can be explained by the variation in X**
- When  $R^2 = 0.52$ , this means that 52% of the variation in Y can be explained by the variation in X – ie the regression explains 52% of change in the dependent variable

# Interpreting results (3)

- The best way to interpret results is to look at some!
- So the seminar will begin collectively
- I will guide you through creating a regression, and then we can all interpret the results together
- After this, you should be able to interpret your own output, and that of other people's research