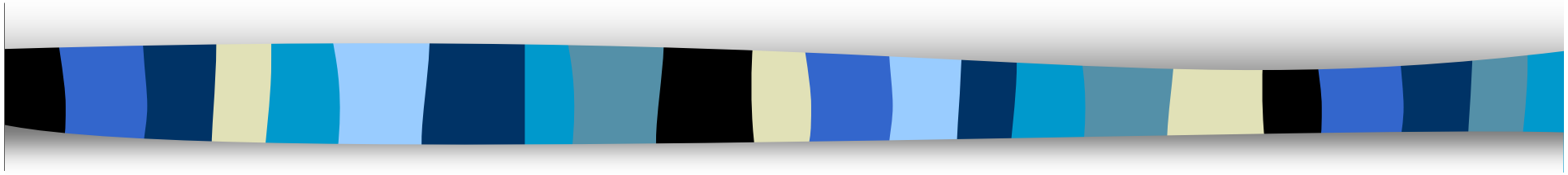
Graduate School 2008/2009

Social Science Statistics II

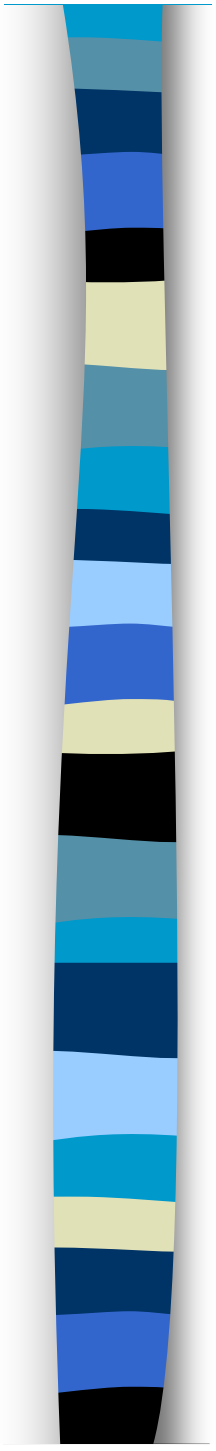
Gwilym Pryce

www.gpryce.com



SSS II

Lecture 1: Correlation and Regression



- Register
- Labs:

Aims and Objectives:

- Aim:
  - to introduce correlation coefficients and multiple regression
- Objectives:
  - by the end of this lecture students should be able to:
    - understand correlation coefficients and their limitations
    - understand intuitively the purpose of multiple regression
    - interpret coefficients and understand basic diagnostic output
    - be able to construct confidence intervals and hypothesis tests on regression coefficients

Plan

- 1. Covariance & Correlation Coefficients
- 2. Multiple Regression
- 3. Interpreting coefficients
- 4. Inference
- 5. Coefficient of Determination

1. Covariance and Correlation

- Simplest way to look at whether two variables are related is to look at whether they *co-vary*
  - *variance* of a singe variable represents the average amount that the data vary from the mean:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$= \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

- If we are interested in whether two variables are related, then we are interested in whether variations in one variable are met with corresponding variations in the other variable:
  - so when one variable deviates from its mean we would expect the other to deviate from its mean in a similar way
    - though not in the same direction if the variables are negatively related
- If there is no relationship then we would expect the changes in one variable to be independent of changes in the other.

Variance & Covariance

$$\text{var}(x) = s^2 = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

$$\text{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Example: Relationship between Feelings of Alienation and Monotonous Work

Worker:	1	2	3	4	5	Mean
x: Monotony score	4	5	1	2	5	3.4
y: Alienation score	19.4	27	7.7	13.4	29.6	19.42

$$\text{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\text{cov}(x, y) = \frac{(0.6)(-0.02) + (1.6)(7.38) + (-2.4)(-11.72) + (1.4)(-6.02) + (1.6)(10.18)}{4}$$

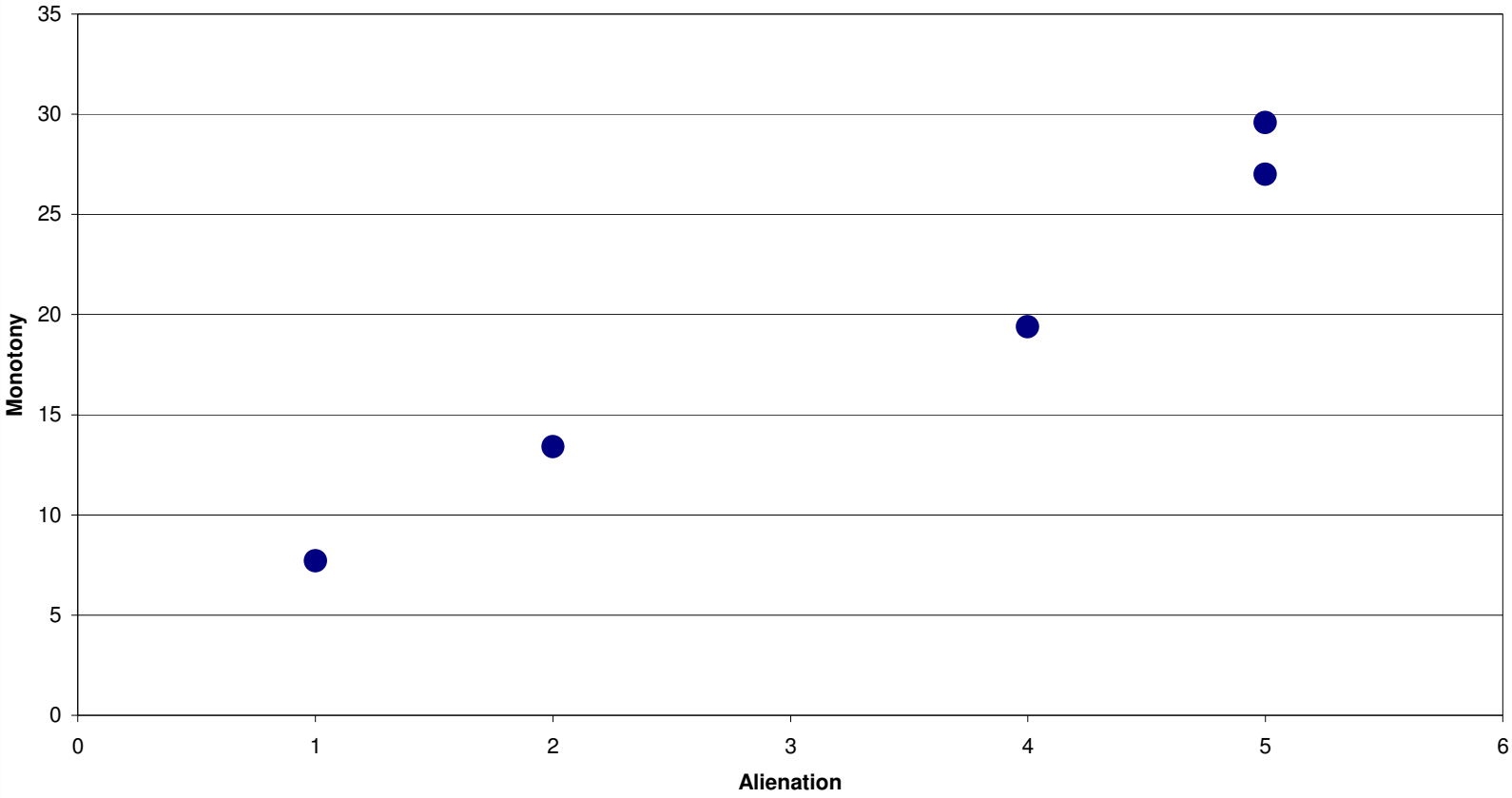
$$= \frac{(0.6)(-0.02) + (1.6)(7.38) + (-2.4)(-11.72) + (1.4)(-6.02) + (1.6)(10.18)}{4}$$

$$= \underline{\mathbf{12.992}}$$

NB Covariance between x and y can be written as $\text{cov}(x,y)$ or as σ_{xy}

- Positive covariance
  - indicates that as one variable deviates from the mean, the other variable deviates in the same direction
- Negative covariance
  - indicates that as one variable deviates from the mean in one direction, the other deviates from its mean in the opposite direction.

Work Monotony and Feelings of Alienation



The covariance scaling problem:

- The value of the covariance (like the variance) is sensitive to scale:
  - so if we divide monotony score by 100, we will get a different covariance value
  - Or, if *y* is miles per gallon, and *x* is average speed, then the cov(*x*,*y*) will be greater if one measures *x* in km per hour rather than miles per hour.

Correlation Coefficient

- One way round this scaling problem is to divide the covariance by the product of the standard deviations of *x* and *y*:

$$r(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- The simple correlation coefficient, *r*(*x*,*y*), has the same sign as the covariance but only varies between -1 and 1 and is unaffected by the scale used to measure the variables.

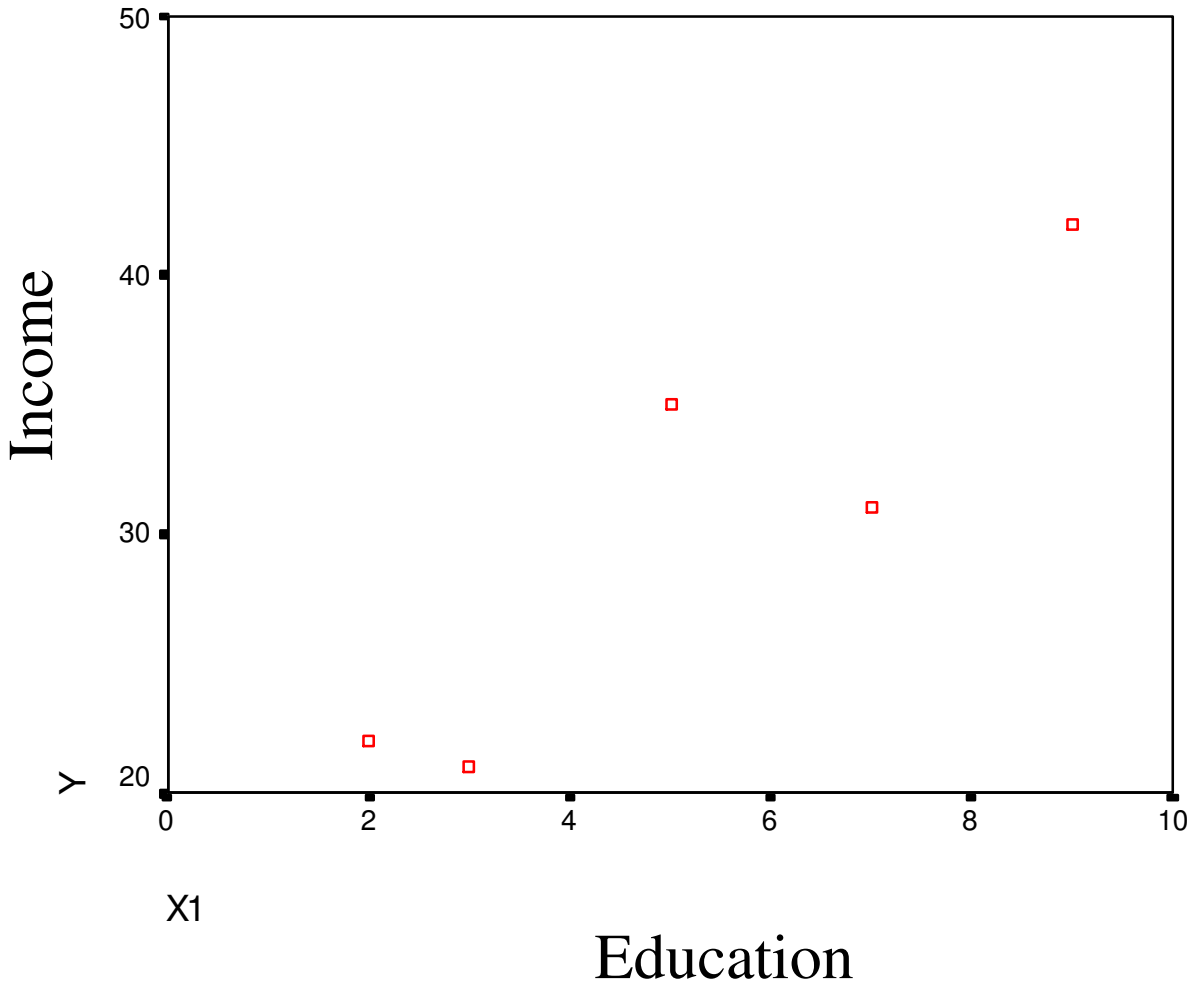
2. Multiple Regression

- The problem with simple correlation is that it does not allow you to control for the effect of other variables
  - e.g. there may be a strong simple correlation between income and education, but if one controls for IQ, then there may be a much smaller effect of education on income
    - (I.e. for a given IQ, there may not be a strong correlation between income and education)
- One way of overcoming this is to use multiple regression

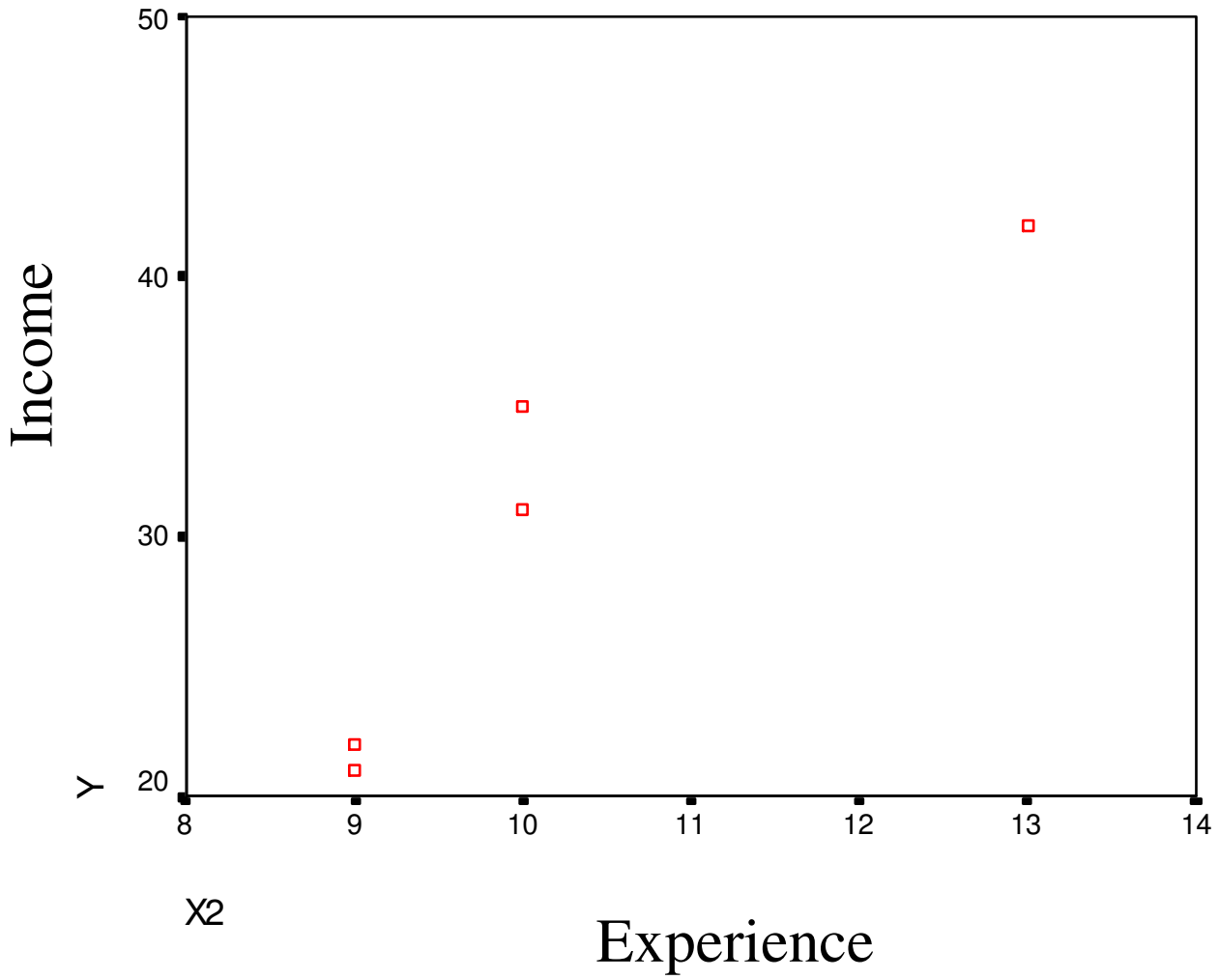
- Multiple regression is regression analysis when you have more than one explanatory variable.
  - E.g. sample of 5 persons randomly drawn from a large firm, data on annual salaries, years of education, and years of experience:
    - y = annual salary in £000s
    - x1 = years of education past secondary school
    - x2 = years of work experience

Y (Salary £000)	X1 (yrs of educ)	X2 (yrs of exp.)
35	5	10
22	2	9
31	7	10
21	3	9
42	9	13

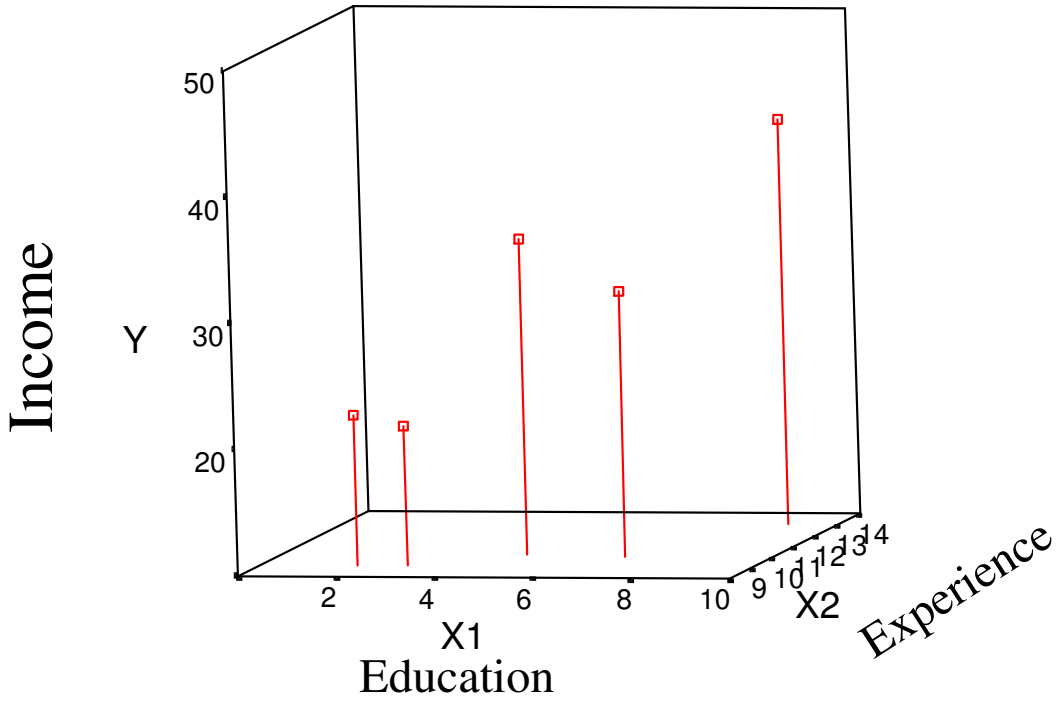
Two dimensional Plot: y on x1



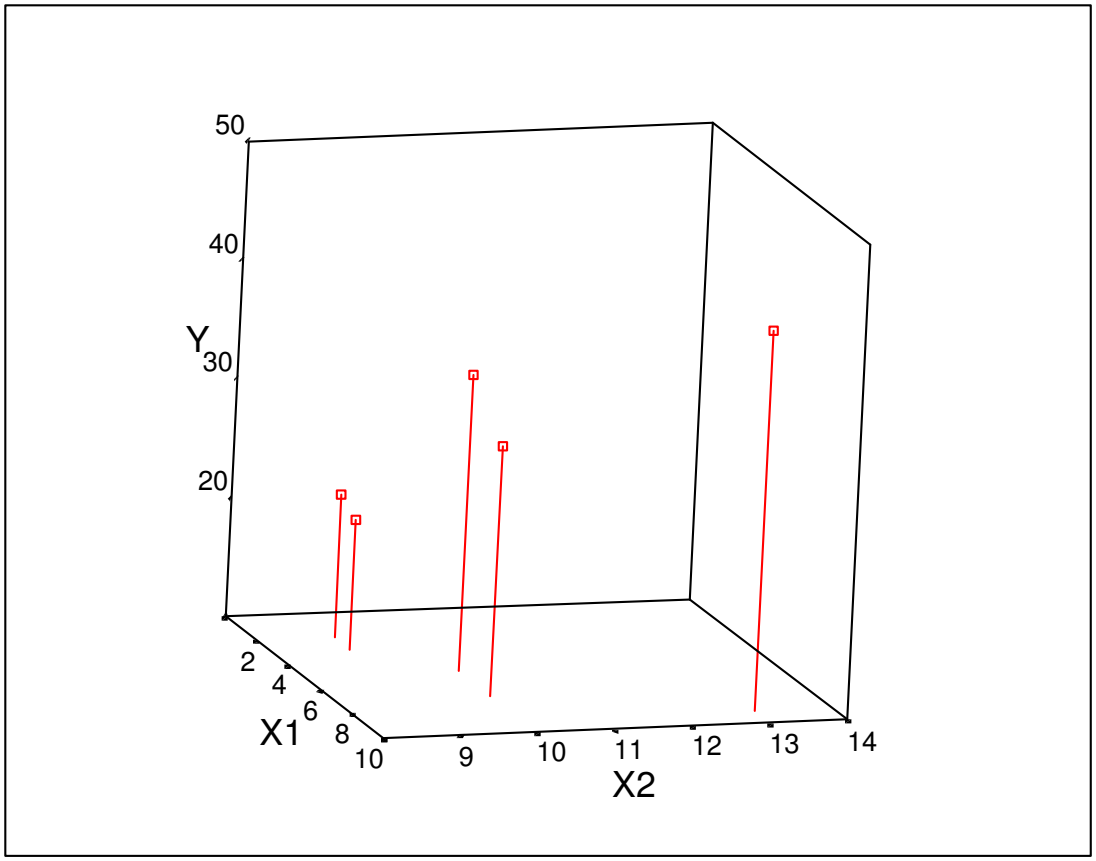
Two dimensional plot: y on x2



Three dimensional plot…



… same plot from a different angle



Regression Output:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.200	23.951		-.175	.877
Education	X1	1.450	1.789	.468	.811	.503
Experience	X2	2.633	3.117	.488	.845	.487

a. Dependent Variable: Y(measured in £000s)

Q/ What do the following items in this output mean:

- Dependent Variable: y
- B?
- Std. Error?
- Constant?
- t?
- Sig.?

3. Interpreting coefficients

- In a **simple regression** (one explanatory variable), the regression estimates the line of best fit: I.e it estimates the values of the intercept α and slope β to give, y^{hat}, predicted values of the dependent variable:

  $y^{hat} = a + bx$ where a and b are sample estimates of the population parameters α and β

- **<u>Multiple Regression:</u>** When you have two or more explanatory variables, the regression estimates the *plane* of best fit:

$$y^{hat} = a + b_1x_1 + b_2x_2$$

- (Beyond two explanatory variables, there is no simple graphical representation of the fitted surface)

- The slope coefficients represent the amount the dependent variable would increase for every additional unit of the explanatory variable.
  - In the income/education/experience example, the estimated equation of the plane was:

$$y^{hat} = -4.20 + 1.45x_1 + 2.63x_2$$

  - so for every extra year of education (x_1) income is predicted to rise by £1,450.
  - and for every extra year of experience (x_2) income is predicted to rise by £2,630.

- The estimates of the slopes and intercept are however subject to error...
  - (a) Model error:
    - Omitted variables or incorrect functional form.
  - (b) Sampling error:
    - Our sample may not be typical of the population as a whole
    - If we re-ran the regression on another random sample, the slope estimates would vary.
    - How much they vary from sample to sample is measured by the Standard Error of the slope.
  - (c) Measurement error:
    - Often unable to measure social science variables with precision

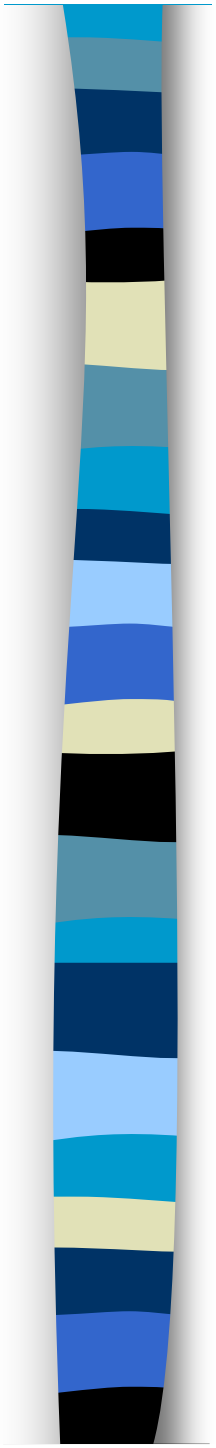
(a) Model Error

- There may unobserved factors, *u*, not included in the regression.
  - Even if we run the regression on the population, these omitted variables would mean that there would be an unexplained component in our model of y, leading to discrepancies between what the model predicts and our observations on y:

$\hat{y} = a + bx$ (1) *line, or "model", we have estimated*

$y = \hat{y} + u$ (2) *model prediction plus unexplained component.*

Sub (1) in (2): $\Rightarrow$ $y = \overbrace{a + bx} + u$



- The second source of error, (b) *Sampling Error*, is due to sampling variation, and we address this using inference (confidence intervals, hypothesis testing, sig...)
  - We assume we have a sample drawn at random from the population. If this is not the case, we need to consider methods that deal with *sample selection bias.*
- The third source of error, (c) *Measurement Error*, is considered in lecture 3.

4. Inference from regression

- Q/ What would happen to the estimate of the slope coefficient we obtained another random sample and re-ran the regression on that sample?

- The standard deviations of a, b_1 and b_2 are called *standard errors* since a, b_1 and b_2 are long run averages (*expected values*) of what you'd get if you run the regression on lots of different samples from the same population
  - in much the same way that the mean is the expected value of the population
  - a, b_1 and b_2 are the sample estimates of the slopes and intercept that you'd get from running a regression on the population

- the range of values you'd get for a parameter from repeated samples is called the *sampling distribution*
- The standard error reported in SPSS for a particular coefficient is an estimate of the standard deviation of the sampling distributions of the coefficient.

4.1 Confidence Intervals for regression coefficients:

- Population slope coefficient CI:

$$\beta = b \pm t_i SE_b$$

- The value of t_i will depend on what level of confidence we want and the ***df = n - k***
- Where:
  - *k* = number of coefficients being estimated including the constant
    = 1 + no. of variables in the regression.
- In this example, *n* = 5, so, *df* = 5 – 3 = 2
  - at 95% level of confidence & *df* = 2, t_i = 4.303
  - at 80% level of confidence & *df* = *2*, t_i = 1.886

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-4.200	23.951		-.175	.877
	X1	1.450	1.789	.468	.811	.503
	X2	2.633	3.117	.488	.845	.487

a. Dependent Variable: Y

95% Confidence interval for β_1 $= b_1 \pm 4.303 \text{ x } 1.789$

$= 1.45 \pm 7.698$

so we are 95% confident that β_1 lies between -6.248 and 9.148

This is a v. large interval due to v. small sample size and large standard errors.

4.2 Hypothesis tests on β_k

- The t-values provided in the SPSS output test the null hypothesis that $\beta_k = 0$.
  - *I.e.* that there is no relationship between *y* and x_i
- They are calculated by simply dividing the coefficient by its standard error.
  - *Sig.* gives us the associated 2-tail significance level for this test.
- In the above example, do you think we should accept or reject the null of no relationship for X1 and X2?

5. Partial Correlation Coefficients and the Coefficient of Determination

- In multiple regression we can calculate the partial correlation coefficient:
  - the correlation between y and x_2 controlling for the effect of x_1.
  - The square of the partial correlation coefficient is called the ***partial coefficient of determination***

Partial Coefficient of Determination

$$r^2_{yx_k} = \frac{t^2_{x_k}}{t^2_{x_k} + df}$$

A more commonly used measure is the coefficient of multiple determination…

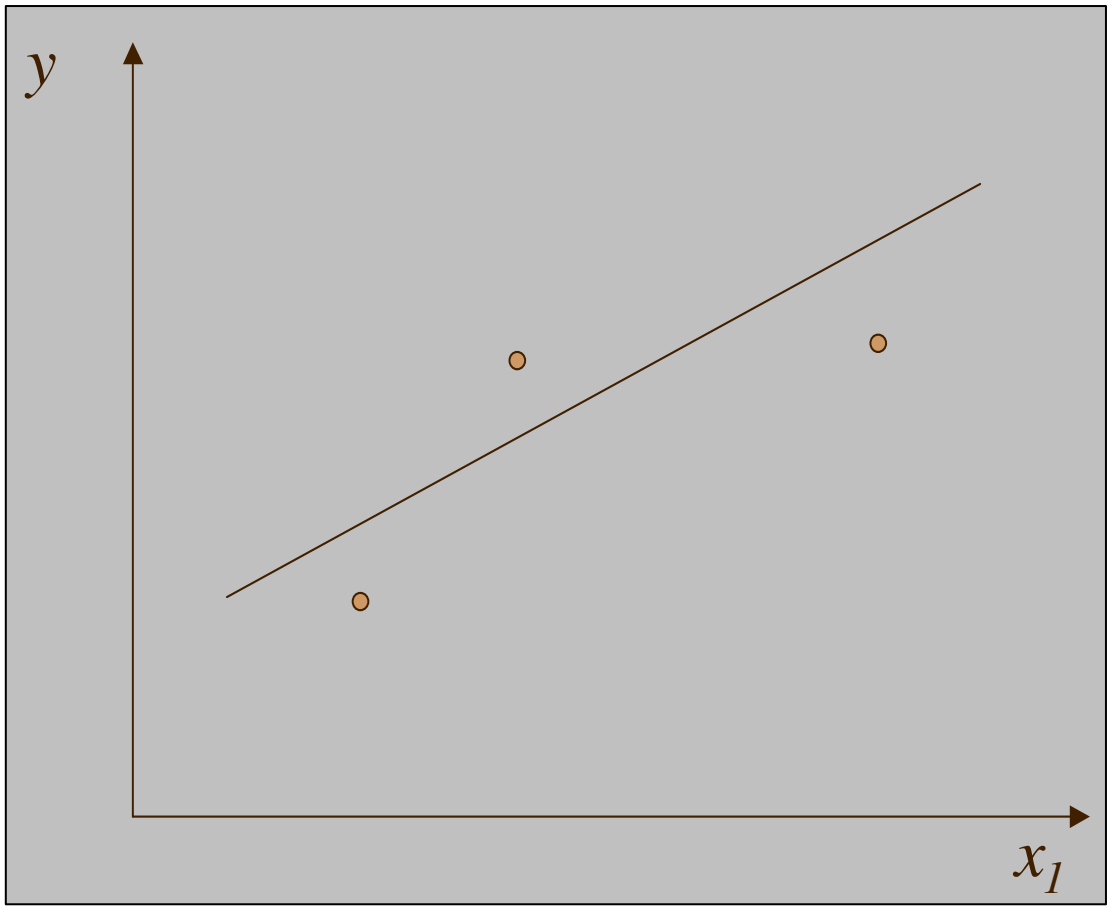
R^2 = Coefficient of Multiple Determination

- One useful measure that is worth looking at this stage is the Coefficient of Multiple Determination, or **R^2**.
  - This measures the proportion of the variation in *y* explained by *all* the explanatory variables together and is a good measure of the overall goodness of fit of the regression line or surface.
  - It varies between 0 and 1; the nearer it is to 1, the more of *y* that is being explained and so the better the goodness of fit.

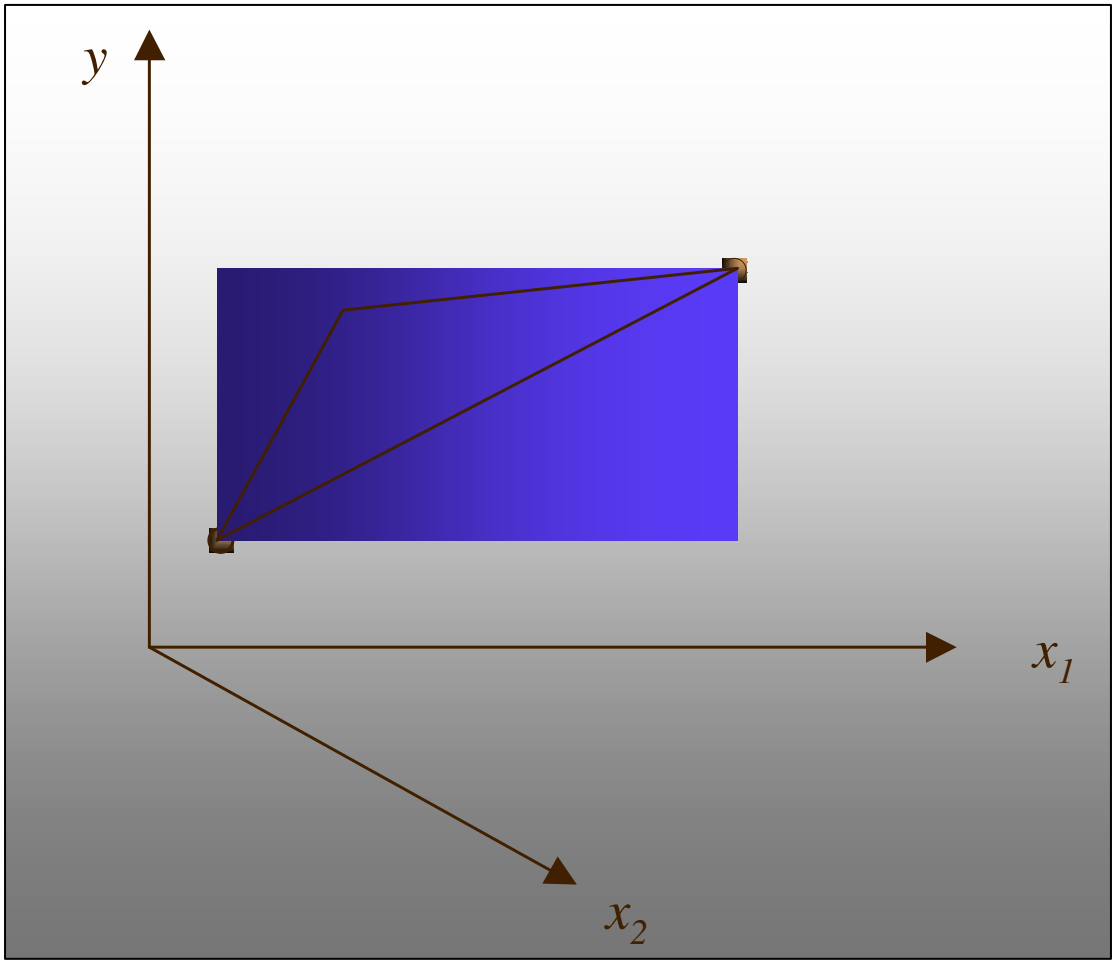
Adjusted R^2

- Each time an explanatory variable is added to the regression, the R^2 will rise even if there is no real gain in explanatory power.
  - This is because adding another "dimension" will always improve apparent goodness of fit even if the new dimension (I.e. variable) is not related to *y*
    - I.e. the R^2 will always increase as more variables are added, so the temptation is just to keep adding variables
    - this can be explained with a simple graphical e.g.

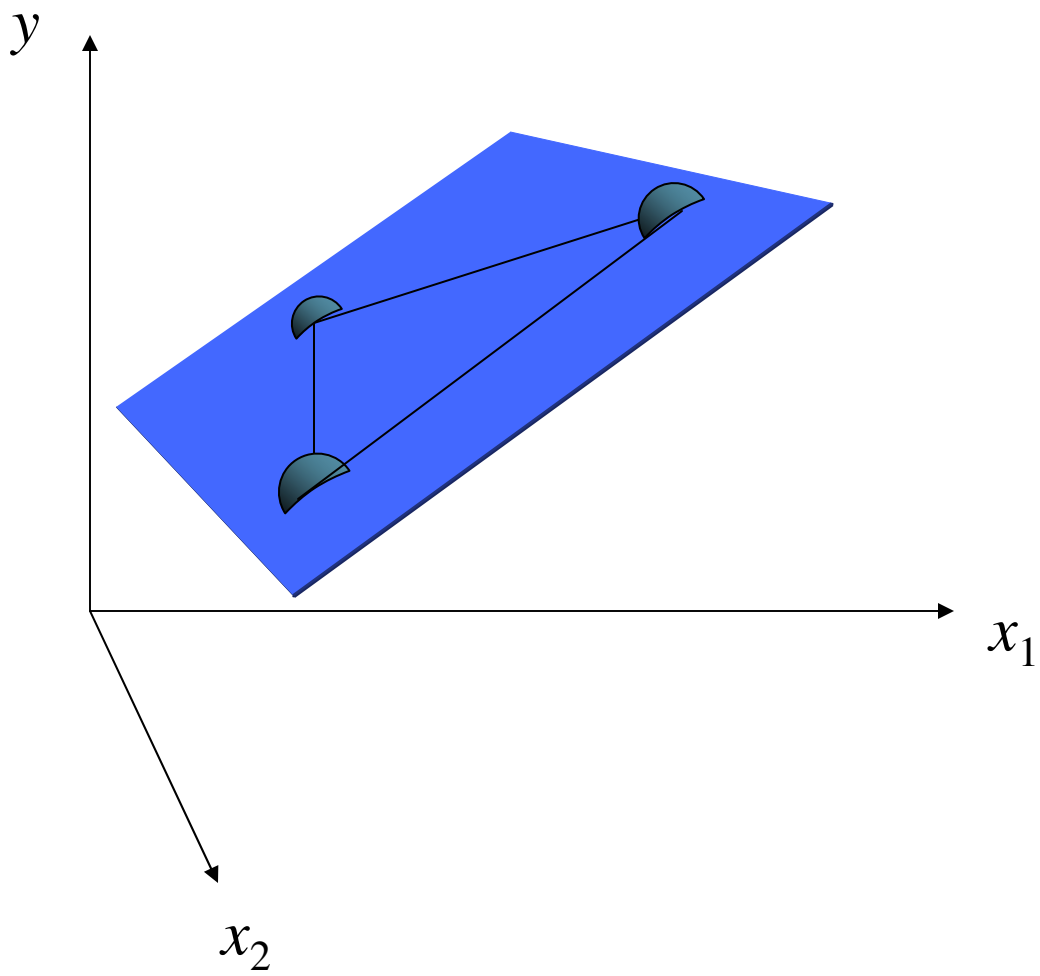
Consider 3 data points and 2 dimensions (i.e. 2 variables):

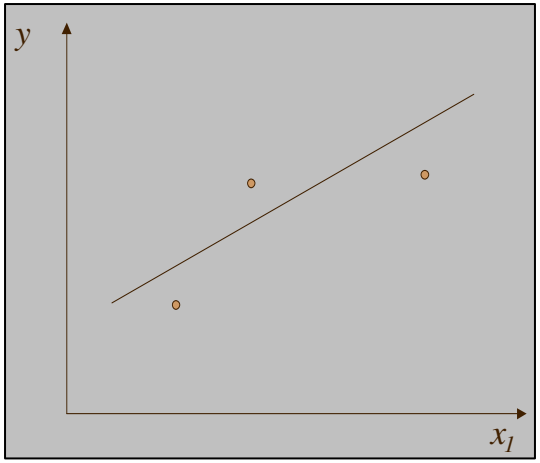


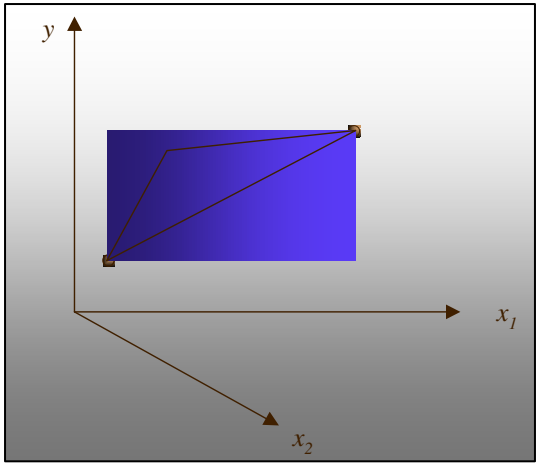
if you add another dimension (I.e variable), without adding any more observations, a plane can be fitted to connect the 3 points exactly:



NB: the above triangle is part of a larger plane...







- So, goodness of fit will appear to improve each time a variable is added, even if the new variable is totally unconnected to the dependent variable.
  - This is basically a *d.f.* problem: R^2 does not take into account the reduced *d.f.* when a variable is added without the inclusion of any additional observations

Degrees of freedom:

- Measures the number of independent pieces of information on which the precision of an estimate is based.
- It is calculated as the number of observations minus the number of additional parameters estimated for that calculation.
- In regression, every time you add an independent variable, the number of coefficients (i.e. parameters) you need to estimate increases by one, so the degrees of freedom falls by one.

- Since the Coefficient of Determination does not take into account changes to the degrees of freedom, we need to adjust the R^2 to control for the *df* effect of adding more variables...

- SPSS provides an adjusted R^2 measure with all regression output which takes into account the effect on *d.f.* of adding new variables:

$$\text{Adj. } R^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

- Thus, where more than one explanatory variable is included, you need to look at the **Adjusted R^2** rather than the R^2

Summary

- 1. Correlation Coefficients
- 2. Multiple Regression
  - OLS with more than one explanatory variable
- 3. Interpreting coefficients
  - b_k estimates how much $y\uparrow$ if $x_k\uparrow$ by one unit.
- 4. Inference
    - b_k only a sample estimate, thus distribution of b_k across lots of samples drawn from a given population
  - confidence intervals
  - hypothesis testing
- 5. Coefficient of Determination: R^2 and Adj R^2