Linear Regression

Starter

1. **(Review of last lesson)**
A group of students failed these marks in their Pure Maths and Statistics exams.

Student	A	B	C	D	E	F	G	H
Pure Maths	62	52	48	79	36	47	44	42
Statistics	74	66	52	71	56	73	40	57

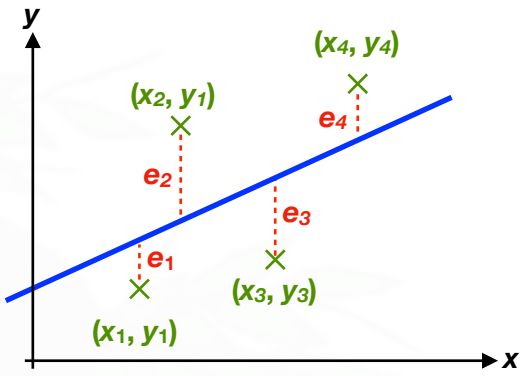
(a) Rank each set of marks.
(b) Calculate, to 4 d.p., Spearman's rank correlation coefficient.
(c) Test, at the $5\,\%$ level the hypothesis that there is a positive association between relative performance in the two exams.

Notes

So we've used Pearson's PMCC to show there is a linear relationship between two variables. The next job is to draw a line of best fit. Linear regression is one way of finding ***the*** line of best fit.

The idea behind least squares linear regression

The 4 points (x_1, y_1), (x_2, y_2), (x_3, y_3) and (x_4, y_4) are plotted on a scatter graph and a line of best fit, of the form $y = a + bx$, is drawn on the graph.



For each point (x_i, y_i) there is an error, e_i, associated with the point, which is the vertical distance to the line.

The least squares method finds the least value of the sum of the squares of the errors.

Since we are looking at ***vertical distances***, it is the least squares regression line of y ***on*** x. This is the regression line that OCR requires us to cover.

If we were looking at ***horizontal distances***, it is the least squares regression line of x ***on*** y — this is not required by OCR.

Formula for finding the least squares regression line of y ***on*** x

The least squares regression line, $y = a + bx$, is found using the formulae:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$a = \bar{y} - b\bar{x}$ where $\bar{x}$ and $\bar{y}$ are the means of the $x-$ and $y-$coordinates respectively.

These formulae are given in the formula booklet.

The formula for b is best considered using the formulae we found for PMCC:

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

These simplified formulae are in the formula booklet.

E.g. 1 The variables x and y are known to be linearly related. Fifty pairs of experimental observations of the two variables gave these results:

$$\sum x = 402.0,\ \sum y = 83.4,\ \sum xy = 680.2,\ \sum x^2 = 3238.2,\ \sum y^2 = 384.6.$$

(a) Obtain the regression line for y on x, giving constants to 4 sf.

(b) Estimate, to 4 sf, the value of y when $x = 7.8$.

Working

(a)
$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 680.2 - \frac{402.0 \times 83.4}{50} = 9.664$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 3238.2 - \frac{402.0^2}{50} = 6.12$$

$$b = \frac{S_{xy}}{S_{xx}} \approx 1.579 \text{ (4 s.f.)}$$

$$a = \bar{y} - b\bar{x} = \frac{83.4}{50} - 1.579 \times \frac{402.0}{50} = -11.03 \text{ (4 s.f.)}$$

$$y = 1.579x - 11.03$$

(b) When $x = 7.8$, $y = 1.579 \times 7.8 - 11.03 = 1.288$

E.g. 2 A farm food supplier monitors the number of hens kept, h, against the weekly consumption of food, f kg, for a sample of 10 smallholdings. The results are below:

$$\sum h = 360,\ \sum h^2 = 17362,\ \sum f = 286,\ \sum f^2 = 10928.94 \text{ and } \sum hf = 13773.6$$

(a) State which the independent variable is.

(b) Obtain the regression equation f on h in the form $f = a + bh$, giving constants to 3 s.f..

(c) Give a practical interpretation of the gradient b.

(d) If food costs £7.50 for a 25 kg bag, estimate the weekly cost of feeding 48 hens.

Using your calculator to find the regression line

Menu >> 6: Statistics >> 2: y=a+bx >> (Enter the data) >> AC >> OPTN >> 3: Regression Calc

Video (calculator): Special calculator function to find the equation of the regression line

E.g. 3 Find the regression line of y on x for the following data.

x	2	4	5	8	10
y	3	7	8	13	17

Disadvantages of least squares regression method

While the least squares regression method is a good method, outliers do have a disproportionate effect on the equation of the line.

The effect of linear coding

Linear coding of the linear regression can be carried out simply using substitution.

E.g. 4 For a set of data the linear regression line is given by $y = 9.4x - 7.8$. Find the regression line for s on t given that $y = 2s - 5$ and $x = 1 - 6t$.

Independent vs. dependent variables

An ***independent*** or ***control variable*** (or ***scientific constant***) is the variable changed or controlled in a scientific experiment to test the effects on the dependent variable. It is usually the $x-$axis of a graph.

A ***dependent*** (or response) variable is the variable being tested or measured in a scientific experiment. It is usually the $y-$axis of a graph.

Independent and dependent variables can be viewed in terms of cause and effect. When the independent variable is changed, this has an effect on the dependent variable. The independent variable is the one "controlled" by the experimenter and is usually the first variable to change chronologically.

The ***independent*** variable ***causes a change in*** the ***dependent*** variable.

For example, compare time spent studying vs. exam score. For a given exam, the former causes a change in the latter and so the independent variable is time spent studying and the exam score is the dependent variable. The test score cannot affect the time spent studying for a single exam.

For example, ice cream sales and temperature. When the temperature rises, ice cream sales rise so the sales are dependent on the temperature. So temperature is the independent variable and goes on the $x-$axis.

It is possible to have multiple dependent variables. For example, if a researcher is studying the effects of diet on health, the dependent variables could be blood pressure, weight, blood sugar, pulse etc. The independent variable is diet as this can be controlled.

However, if neither of the variables is controlled, there is no independent variable.

E.g. 5 Consider the following situations and identify the independent and dependent variables, if there are any:

(a) A researcher wants to see if there is a connection between test mark and revision time.

(b) A park ranger wonders whether a greater number of trees in a wood increases the number of squirrels present.

(c) A farmer wants to know whether putting different fertiliser on crops will make them grow more.

(d) The speed of a migrating bird was measured at one-minute intervals.

Working:

(a) "Revision time" is the independent variable as it can be controlled and comes before the "test mark". "Test mark" depends on revision time so is the dependent variable.

Video: Linear regression

Video: Interpreting lines of best fit

Linear regression EQ

Solutions to Starter and E.g.s

Exercise

p86 5C Qu 1i, 2i, 3-6, (7-8 red)

Summary

The least squares regression line of y on x, $y = a + bx$, is found using the formulae:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$a = \bar{y} - b\bar{x}$ where $\bar{x}$ and $\bar{y}$ are the means of the $x-$ and $y-$coordinates respectively.

These formulae are given in the formula booklet.

The formula for b is best considered using the formulae we found for PMCC:

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

These simplified formulae are in the formula booklet.