Contents 43

Regression and Correlation

Learning outcomes

You will learn how to explore relationships between variables and how to measure the strength of such relationships. You should note from the outset that simply establishing a relationship is not enough. You may establish, for example, a relationship between the number of hours a person works in a week and their hat size. Should you conclude that working hard causes your head to enlarge? Clearly not, any relationship existing here is not causal!

Regression

43.1

Introduction

Problems in engineering often involve the exploration of the relationship(s) between two or more variables. The technique of regression analysis is very useful and well-used in this situation. This Section will look at the basics of regression analysis and should enable you to apply regression techniques to the study of relationships between variables. Just because a relationship exists between two variables does not necessarily imply that the relationship is causal. You might find, for example that there is a relationship between the hours a person spends watching TV and the incidence of lung cancer. This does not necessarily imply that watching TV causes lung cancer.

Assuming that a causal relationship does exist, we can measure the strength of the relationship by means of a correlation coefficient discussed in the next Section. As you might expect, tests of significance exist which allow us to interpret the meaning of a calculated correlation coefficient.

Prerequisites

Before starting this Section you should ...

- have knowledge of Descriptive Statistics (HELM 36)
- be able to find the expectation and variance of sums of variables (HELM 39.3)
- understand the terms independent and dependent variables
- understand the terms biased and unbiased estimators

Learning Outcomes

On completion you should be able to ...

- define the terms regression analysis and regression line
- use the method of least squares for finding a line of best fit

1. Regression

As we have already noted, relationship(s) between variables are of interest to engineers who may wish to determine the degree of association existing between independent and dependent variables. Knowing this often helps engineers to make predictions and, on this basis, to forecast and plan. Essentially, regression analysis provides a sound knowledge base from which accurate estimates of the values of a dependent variable may be made once the values of related independent variables are known.

It is worth noting that in practice the choice of independent variable(s) may be made by the engineer on the basis of experience and/or prior knowledge since this may indicate to the engineer which independent variables are likely to have a substantial influence on the dependent variable. In summary, we may state that the principle objectives of regression analysis are:

(a) to enable accurate estimates of the values of a dependent variable to be made from known values of a set of independent variables;

(b) to enable estimates of errors resulting from the use of a regression line as a basis of prediction.

Note that if a regression line is represented as $y = f(x)$ where x is the independent variable, then the actual function used (linear, quadratic, higher degree polynomial etc.) may be obtained via the use of a theoretical analysis or perhaps a scatter diagram (see below) of some real data. Note that a regression line represented as $y = f(x)$ is called a **regression line of y on x**.

Scatter diagrams

A useful first step in establishing the degree of association between two variables is the plotting of a **scatter diagram**. Examples of pairs of measurements which an engineer might plot are:

(a) volume and pressure;

(b) acceleration and tyre wear;

(c) current and magnetic field;

(d) torsion strength of an alloy and purity.

If there exists a relationship between measured variables, it can take many forms. Even though an outline introduction to non-linear regression is given at the end of this Workbook, we shall focus on the **linear relationship** only.

In order to produce a good scatter diagram you should follow the steps given below:

1. Give the diagram a **clear title** and indicate exactly what information is being displayed;

2. Choose and **clearly mark** the axes;

3. Choose carefully and clearly mark the **scales** on the axes;

4. Indicate the **source** of the data.

Examples of scatter diagrams are shown below.

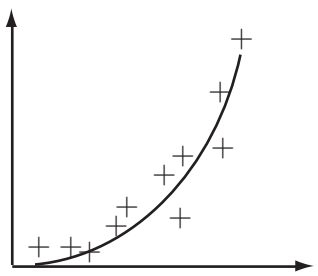


Figure 1

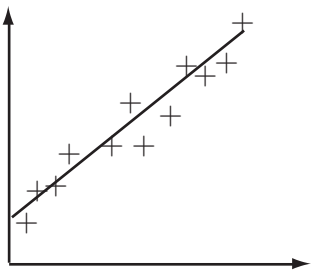


Figure 2

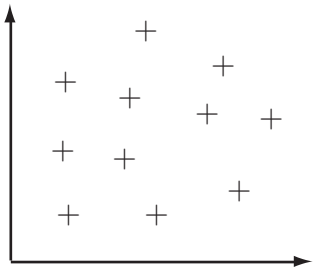


Figure 3

Figure 1 shows an association which follows a curve, possibly exponential, quadratic or cubic;

Figure 2 shows a reasonable degree of linear association where the points of the scatter diagram lie in an area surrounding a straight line;

Figure 3 represents a randomly placed set of points and no linear association is present between the variables.

Note that in Figure 2, the word 'reasonable' is not defined and that while points 'close' to the indicated straight line may be explained by random variation, those 'far away' may be due to assignable variation.

The rest of this Section will deal with linear association only although it is worth noting that techniques do exist for transforming many non-linear relationships into linear ones. We shall investigate linear association in two ways, firstly by using educated guess work to obtain a regression line 'by eye' and secondly by using the well-known technique called the method of least squares.

Regression lines by eye

Note that at a very simple level, we may look at the data and, using an 'educated guess', draw a line of regression 'by eye' through a set of points. However, finding a regression line by eye is unsatisfactory as a general statistical method since it involves guess-work in drawing the line with the associated errors in any results obtained. The guess-work can be removed by the method of least squares in which the equation of a regression line is calculated using data. Essentially, we calculate the equation of the regression line by minimising the sum of the squared vertical distances between the data points and the line.

The method of least squares - an elementary view

We assume that an experiment has been performed which has resulted in n pairs of values, say $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ and that these results have been checked for approximate linearity on the scatter diagram given below.

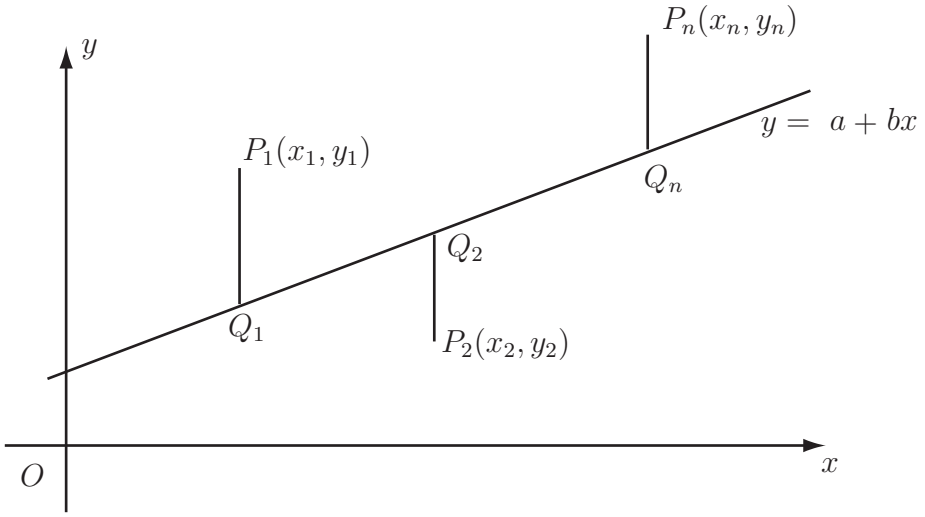


Figure 4

The vertical distances of each point from the line $y = a + bx$ are easily calculated as

$$y_1 - a - bx_1, \quad y_2 - a - bx_2, \quad y_3 - a - bx_3 \quad \cdots \quad y_n - a - bx_4$$

These distances are squared to guarantee that they are positive and calculus is used to minimise the sum of the squared distances. Effectively we are minimizing the sum of a two-variable expression and need to use partial differentiation. If you wish to follow this up and look in more detail at the technique, any good book (engineering or mathematics) containing sections on multi-variable calculus should suffice. We will not look at the details of the calculations here but simply note that the process results in two equations in the two unknowns m and c being formed. These equations are:

$$\sum xy - a\sum x - b\sum x^2 = 0 \qquad \text{(i)}$$

and

$$\sum y - na - b\sum x = 0 \qquad \text{(ii)}$$

The second of these equations (ii) immediately gives a useful result. Rearranging the equation we get

$$\frac{\sum y}{n} - a - b\frac{\sum x}{n} = 0 \quad \text{or, put more simply} \quad \bar{y} = a + b\bar{x}$$

where $(\bar{x}, \bar{y})$ is the mean of the array of data points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$.

This shows that the mean of the array always lies on the regression line. Since the mean is easily calculated, the result forms a useful check for a plotted regression line. Ensure that any regression line you draw passes through the mean of the array of data points.

Eliminating a from the equations gives a formula for the gradient b of the regression line, this is:

$$b = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \qquad \text{often written as} \qquad b = \frac{S_{xy}}{S_x^2}$$

The quantity S_x^2 is, of course, the variance of the x-values. The quantity S_{xy} is known as the **covariance** (of x and y) and will appear again later in this Workbook when we measure the degree of linear association between two variables.

Knowing the value of b enables us to obtain the value of a from the equation $\bar{y} = a + b\bar{x}$

Key Point 1

Least Squares Regression - y on x

The least squares regression line of y on x has the equation $y = a + bx$, where

$$b = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$ and a is given by the equation $a = \bar{y} - b\bar{x}$

It should be noted that the coefficients b and a obtained here will give us the regression line of y on x. This line is used to predict y values given x values. If we need to predict the values of x from given values of y we need the regression line of x on y. The two lines are not the same except in the (very) special case where all of the points lie exactly on a straight line. It is worth noting however, that the two lines cross at the point $(\bar{x}, \bar{y})$. It can be shown that the regression line of x on y is given by Key Point 2:

Key Point 2

Least Squares Regression - x on y

The regression line of x on y is

$$x = a' + b'y$$

where

$$b' = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} \quad \text{and} \quad a' = \bar{x} - b'\bar{y}$$

Example 1

A warehouse manager of a company dealing in large quantities of steel cable needs to be able to estimate how much cable is left on his partially used drums. A random sample of twelve partially used drums is taken and each drum is weighed and the corresponding length of cable measured. The results are given in the table below:

Weight of drum and cable (x) kg.	Measured length of cable (y) m.
30	70
40	90
40	100
50	120
50	130
50	150
60	160
70	190
70	200
80	200
80	220
80	230

Find the least squares regression line in the form $y = mx + c$ and use it to predict the lengths of cable left on drums whose weights are:

(i) 35 kg (ii) 85 kg (iii) 100 kg

In the latter case state any assumptions which you make in order to find the length of cable left on the drum.

Solution

Excel calculations give $\sum x = 700, \quad \sum x^2 = 44200, \quad \sum y = 1860 \quad \sum xy = 118600$ so that the formulae

$$b = \frac{\dfrac{\sum xy}{n} - \dfrac{\sum x}{n}\dfrac{\sum y}{n}}{\dfrac{\sum x^2}{n} - \left(\dfrac{\sum x}{n}\right)^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

give $a = -20$ and $b = 3$. Our regression line is $y = -20 + 3x$, so $y = 3x - 20$.

Hence, the required predicted values are:

$$y_{35} = 3 \times 35 - 20 = 85 \qquad y_{85} = 3 \times 85 - 20 = 235 \qquad y_{100} = 3 \times 100 - 20 = 280$$

all results being in metres.

To obtain the last result we have assumed that the linearity of the relationship continues beyond the range of values actually taken.

An article in the Journal of Sound and Vibration 1991 (151) explored a possible relationship between hypertension (defined as blood pressure rise in mm of mercury) and exposure to noise levels (measured in decibels). Some data given is as follows:

Noise Level (x)	Blood pressure rise (y)	Noise Level (x)	Blood pressure rise (y)
60	1	85	5
63	0	89	4
65	1	90	6
70	2	90	8
70	5	90	4
70	1	90	5
80	4	94	7
90	6	100	9
80	2	100	7
80	3	100	6

(a) Draw a scatter diagram of the data.

(b) Comment on whether a linear model is appropriate for the data.

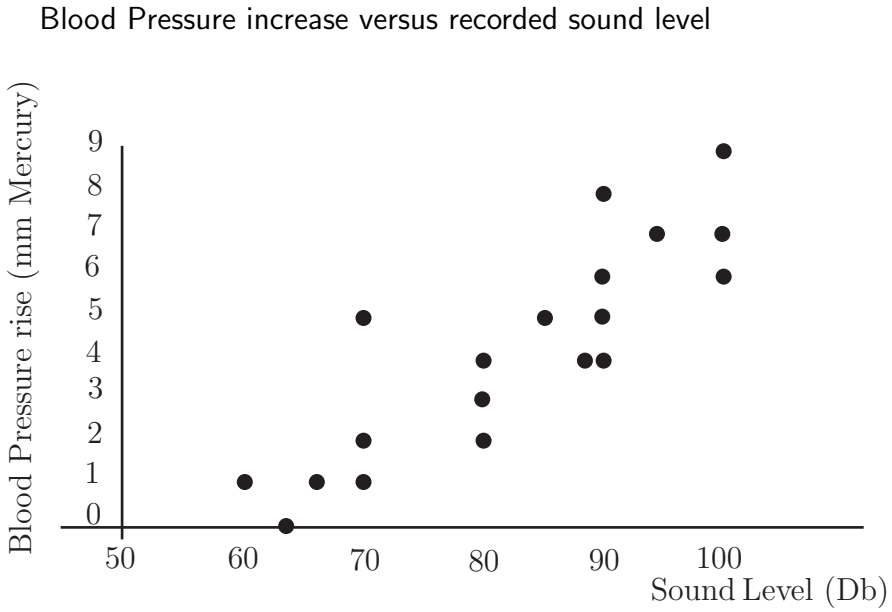
(c) Calculate a line of best fit of y on x for the data given.

(d) Use your regression line predict the expected rise in blood pressure for a exposure to a noise level of 97 decibels.

Your solution

Answer

(a) Entering the data into Microsoft Excel and plotting gives



(b) A linear model is appropriate.

(c) Excel calculations give $\sum x = 1656, \quad \sum x^2 = 140176, \quad \sum y = 86, \quad \sum xy = 7654$ so that $b = 0.1743$ and $a = -10.1315$. Our regression line is $y = 0.1743x - 10.1315$.

(d) The predicted value is: $y_{97} = 0.1743 \times 97 - 10.1315 = 6.78$ mm mercury.

The method of least squares - a modelling view

We take the dependent variable Y to be a random variable whose value, for a fixed value of x depends on the value of x and a random error component say e and we write

$$Y = \alpha + \beta x + e$$

Adopting the notation of conditional probability, we are looking for the expected value of Y for a given value of x. The expected value of Y for a given value of x is denoted by

$$\mathsf{E}(Y|x) = \mathsf{E}(\alpha + \beta x + e) = \mathsf{E}(\alpha + \beta x) + \mathsf{E}(e)$$

The variance of Y for a given value of x is given by the relationship

$$\mathsf{V}(Y|x) = \mathsf{V}(\alpha + \beta x + e) = \mathsf{V}(\alpha + \beta x) + \mathsf{V}(e), \quad \text{assuming independence.}$$

If $\mu_{Y|x}$ represents the true mean value of Y for a given value of x then

$$\mu_{Y|x} = \alpha + \beta x, \quad \text{assuming a linear relationship holds,}$$

is a straight line of mean values. If we now assume that the errors e are distributed with mean 0 and variance σ^2 we may write

$$\mathsf{E}(Y|x) = \mathsf{E}(\alpha + \beta x) + \mathsf{E}(e) = \alpha + \beta x \quad \text{since } \mathsf{E}(e) = 0.$$

and

$$\mathsf{V}(Y|x) = \mathsf{V}(\alpha + \beta x) + \mathsf{V}(e) = \sigma^2 \quad \text{since } \mathsf{V}(\alpha + \beta x) = 0.$$

This implies that for each value of x, Y is distributed with mean $\alpha + \beta x$ and variance σ^2. Hence when the variance is small the observed values of Y will be close to the regression line and when the variance is large, at least some of the observed values of Y may not be close to the line. Note that the assumption that the errors e are distributed with mean 0 may be made without loss of generality. If the errors had any other mean, we could subtract it and then add the mean to the value of c. The ideas are illustrated in the following diagram.

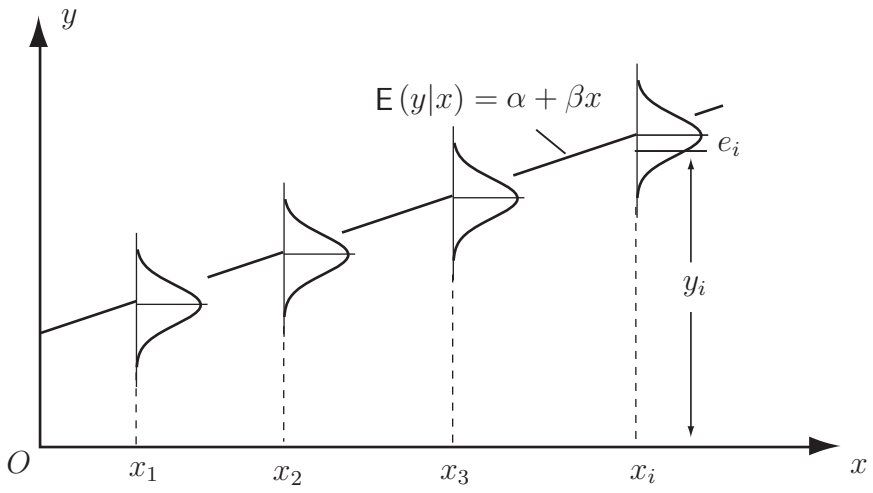


Figure 5

The regression line is shown passing through the means of the distributions for the individual values of x. The value of y corresponding to the x-value x_i can be represented by the equation

$$y_i = \alpha + \beta x_i + e_i$$

where e_i is the error of the observed value of y, that is the difference from its expected value, namely

$$\mathsf{E}(Y|x_i) = \mu_{y|x_i} = \alpha + \beta x_i$$

Now, if we estimate α and β with a and b, the *residual*, or estimated error, becomes

$$\hat{e}_i = y_i - a - bx_i$$

so that the sum of the squares of the residuals is given by

$$S = \sum \hat{e}_i^2 = \sum (y_i - a - bx_i)^2$$

and we may minimize the quantity S by using the method of least squares as before. The mathematical details are omitted as before and the equations obtained for b and a are as before, namely

$$b = \frac{\dfrac{\sum xy}{n} - \dfrac{\sum x}{n}\dfrac{\sum y}{n}}{\dfrac{\sum x^2}{n} - \left(\dfrac{\sum x}{n}\right)^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

Note that since the error e_i in the ith observation essentially describes the error in the fit of the model to the ith observation, the sum of the squares of the errors $\sum e_i^2$ will now be used to allow us to comment on the adequacy of fit of a linear model to a given data set.

Adequacy of fit

We now know that the variance $\mathrm{V}(Y|x) = \sigma^2$ is the key to describing the adequacy of fit of our simple linear model. In general, the smaller the variance, the better the fit although you should note that it is wise to distinguish between 'poor fit' and a large error variance. Poor fit may suggest, for example, that the relationship is not in fact linear and that a fundamental assumption made has been violated. A large value of σ^2 does not necessarily mean that a linear model is a poor fit.
It can be shown that the sum of the squares of the errors say SS_E can be used to give an unbiased estimator $\hat{\sigma}^2$ of σ^2 via the formula

$$\hat{\sigma}^2 = \frac{SS_E}{n-p}$$

where p is the number of independent variables used in the regression equation. In the case of simple linear regression $p = 2$ since we are using just x and c and the estimator becomes:

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

The quantity SS_E is usually used explicitly in formulae whose purpose is to determine the adequacy of a linear model to explain the variability found in data. Two ways in which the adequacy of a regression model may be judged are given by the so-called **Coefficient of Determination** and the **Adjusted Coefficient of Determination**.

The coefficient of determination

Denoted by R^2 , the Coefficient of Determination is defined by the formula

$$R^2 = 1 - \frac{SS_E}{SS_T}$$

where SS_E is the sum of the squares of the errors and SS_T is the sum of the squares of the totals given by $\sum(y_i - \hat{y}_i)^2 = \sum y_i^2 - n\bar{y}^2$. The value of R^2 is sometimes described as representing the amount of variability explained or accounted for by a regression model. For example, if after a particular calculation it was found that $R^2 = 0.884$, we could say that the model accounts for about 88% of the variability found in the data. However, deductions made on the basis of the value of R^2 should be treated cautiously, the reasons for this are embedded in the following properties of the statistic. It can be shown that:

(a) $0 \leq R^2 \leq 1$

(b) a large value of R^2 does not necessarily imply that a model is a good fit;

(c) adding a regressor variable (simple regression becomes multiple regression) *always* increases the value of R^2. This is one reason why a large value of R^2 does not necessarily imply a good model;

(d) models giving large values of R^2 can be poor predictors of new values if the fitted model does not apply at the appropriate x-value.

Finally, it is worth noting that to check the fit of a linear model properly, one should look at plots of residual values. In some cases, tests of goodness-of-fit are available although this topic is not covered in this Workbook.

The adjusted coefficient of determination

Denoted (often) by R^2_{adj}, the Adjusted Coefficient of Determination is defined as

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

where p is the number of variables in the regression equation. For the simple linear model, $p = 2$ since we have two unknown parameters in the regression equation, the intercept c and the coefficient m of x. It can be shown that:

(a) R^2_{adj} is a better indicator of the adequacy of predictive power than R^2 since it takes into account the number of regressor variables used in the model;

(b) R^2_{adj} does not necessarily increase when a new regressor variable is added.

Both coefficients claim to measure the adequacy of the predictive power of a regression model and their values indicate the proportion of variability explained by the model. For example a value of

R^2 or $R^2_{adj} = 0.9751$

may be interpreted as indicating that a model explains 97.51% of the variability it describes. For example, the drum and cable example considered previously gives the results outlined below with

$R^2 = 96.2$ and $R^2_{adj} = 0.958$

In general, R^2_{adj} is (perhaps) more useful than R^2 for comparing alternative models. In the context of a simple linear model, R^2 is easier to interpret. In the drum and cable example we would claim that the linear model explains some 96.2% of the variation it describes.

Drum & Cable (x)	x^2	Cable Length (y)	y^2	xy	Predicted Values	Error Squares
30	900	70	4900	2100	70	0.00
40	1600	90	8100	3600	100	100.00
40	1600	100	10000	4000	100	0.00
50	2500	120	14400	6000	130	100.00
50	2500	130	16900	6500	130	0.00
50	2500	150	22500	7500	130	400.00
60	3600	160	25600	9600	160	0.00
70	4900	190	36100	1330	190	0.00
70	4900	200	40000	14000	190	100.00
80	6400	200	40000	16000	220	400.00
80	6400	220	48400	17600	220	0.00
80	6400	230	52900	18400	220	100.00
Sum of x	Sum of x^2	Sum of y	Sum of y^2	Sum of xy		SSE =
$= 700$	$= 44200$	$= 1860$	$= 319800$	$= 118600$		1200.00
$b = 3$	$a = -20$	SST =		$R^2 =$		$R^2_{\text{adj}} =$
		31500		0.962		0.958

Use the drum and cable data given in Example 1 (page 7) and set up a spreadsheet to verify the values of the Coefficient of Determination and the Adjusted Coefficient of Determination calculated on page 12.

Your solution

Answer

As per the table on page 12 giving $R^2 = 0.962$ and $R^2_{adj} = 0.958$.

Significance testing for regression

Note that the results in this Section apply to the simple linear model only. Some additions are necessary before the results can be generalized.

The discussions so far pre-suppose that a linear model adequately describes the relationship between the variables. We can use a significance test involving the distribution to decide whether or not y is linearly dependent on x. We set up the following hypotheses:

$$H_0 : \ \beta = 0 \qquad \text{and} \qquad H_1 : \ \beta \neq 0$$

Key Point 3

Significance Test for Regression

The test statistic is

$$F_{test} = \frac{SS_R}{SS_E/(n-2)}$$

where $SS_R = SS_T - SS_E$ and rejection at the 5% level of significance occurs if

$$F_{test} > f_{0.05,1,n-2}$$

Note that we have one degree of freedom since we are testing only one parameter (m) and that n denotes the number of pairs of (x, y) values. A set of tables giving the 5% values of the F-distribution is given at the end of this Workbook (Table 1).

Example 2

Test to determine whether a simple linear model is appropriate for the data previously given in the drum and cable example above.

Solution

We know that

$$SS_T = SS_R + SS_E$$

where $SS_T = \sum y^2 - \dfrac{(\sum y)^2}{n}$ is the total sum of squares (of y) so that (from the spreadsheet above) we have:

$$SS_R = 31500 - 1200 = 30300$$

Hence

$$F_{test} = \frac{SS_R}{SS_E/(n-2)} = \frac{30300}{1200/(12-2)} = 252.5$$

From Table 1, the critical value is $f_{0.05,1,10} = 241.9$.

Hence, since $F_{test} > f_{0.05,1,10}$, we reject the null hypothesis and conclude that $\beta \neq 0$.

Regression curves

The Section should be regarded as introductory only. The reason for including non-linear regression is to demonstrate how the method of least squares can be extended to deal with cases where the relationship between variables is, for example, quadratic or exponential.

A regression curve is defined to be the curve passing through the expected value of Y for a set of given values of x. The idea is illustrated by the following diagram.

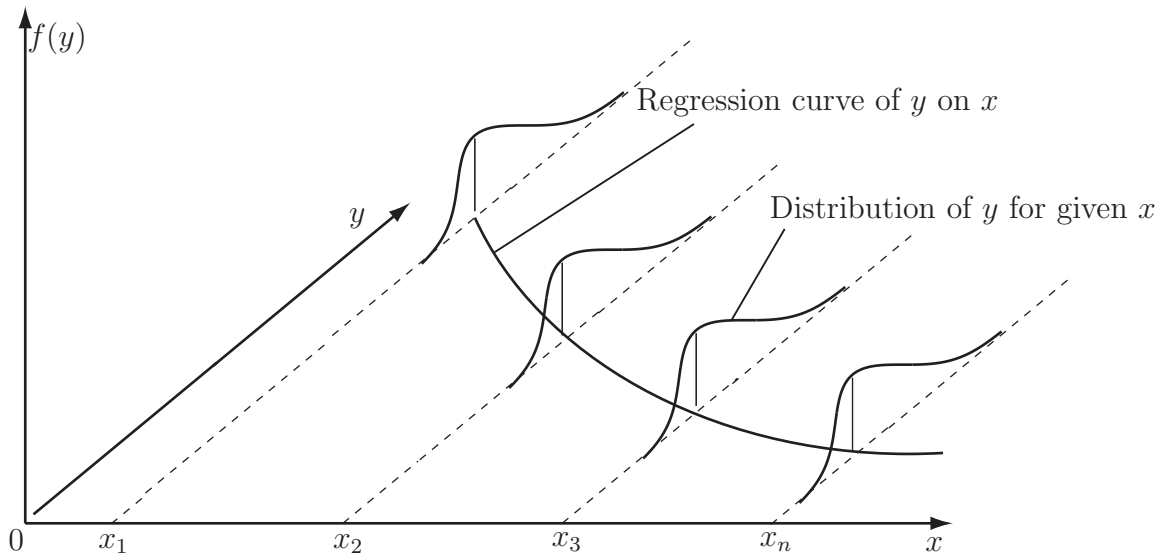


Figure 6

We will look at the quadratic and exponential cases in a little detail.

The quadratic case

We are looking for a functional relation of the form

$$y = \alpha + \beta x + \gamma x^2$$

and so, using the method of least squares, we require the values of a, b and c which minimize the expression

$$f(a, b, c) = \sum_{r=1}^{n} (y_r - a - bx_r - cx_r^2)^2$$

Note here that the regression described by the form

$$y = \alpha + \beta x + \gamma x^2$$

is actually a linear regression since the expression is linear in $\alpha,\ \beta$ and γ.
Omitting the subscripts and using partial differentiation gives

$$\frac{\partial f}{\partial a} = -2\sum (y - a - bx - cx^2)$$

$$\frac{\partial f}{\partial b} = -2\sum x(y - a - bx - cx^2)$$

$$\frac{\partial f}{\partial c} = -2\sum x^2(y - a - bx - cx^2)$$

At a minimum we require

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b} = \frac{\partial f}{\partial c} = 0$$

which results in the three linear equations

$$\sum y - na - b\sum x - c\sum x^2 = 0$$

$$\sum xy - a\sum x - b\sum x^2 - c\sum x^3 = 0$$

$$\sum x^2 y - a\sum x^2 - b\sum x^3 - c\sum x^4 = 0$$

which can be solved to give the values of a, b and c.

The exponential case

We use the same technique to look for a functional relation of the form

$$y = \alpha e^{\beta x}$$

As before, using the method of least squares, we require the values of a and b which minimize the expression

$$f(a, b) = \sum_{r=1}^{n} (y_r - a\mathrm{e}^{bx_r})^2$$

Again omitting the subscripts and using partial differentiation gives

$$\frac{\partial f}{\partial a} = -2\sum \mathrm{e}^{bx}(y - a\mathrm{e}^{bx})$$

$$\frac{\partial f}{\partial b} = -2\sum ax\mathrm{e}^{bx}(y - a\mathrm{e}^{bx})$$

At a minimum we require

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b} = 0$$

which results in the two non-linear equations

$$\sum y\mathrm{e}^{bx} - a\sum \mathrm{e}^{2bx} = 0$$

$$\sum xy\mathrm{e}^{bx} - a\sum x\mathrm{e}^{2bx}$$

which can be solved by iterative methods to give the values of a and b.

Note that it is possible to combine (for example) linear and exponential regression to obtain a regression equation of the form

$$y = (\alpha + \beta x)\mathrm{e}^{\gamma x}$$

The method of least squares may then be used to find estimates $a,\ b,\ c$ of $\alpha,\ \beta,\ \gamma$.

Correlation

43.2

Introduction

While medical researchers might be interested in knowing the answers to questions such as 'Is age related to blood pressure?' engineers might be interested in knowing the answers to questions such as 'Is the shear strength of a weld related to its diameter?' or 'Is the rate of wear of a petrol engine related to its operating temperature?' As you already know (from reading the introduction to Section 43.1 concerning the topic of regression), statisticians measure the strength of a relationship between two variables by using a quantity called the correlation coefficient. As you might expect, tests exist which allow us to interpret the meaning of a calculated correlation coefficient.

Prerequisites

Before starting this Section you should ...

- have knowledge of Descriptive Statistics as presented in HELM 36
- have knowledge of Hypothesis Testing based on the t-distribution as presented in HELM 41
- have knowledge of Regression as presented in Section 43.1

Learning Outcomes

On completion you should be able to ...

- explain what is meant by the term correlation coefficient
- perform a statistical test in order to interpret the possible meaning of a correlation coefficient

1. Correlation

So far we have assumed that we have a random variable Y related to an independent variable x which can be measured with some accuracy. In the equation below, the dependent variable Y is a random variable whose value, for a fixed value of x depends on a random error component say e and we have

$$Y = mx + c + e$$

In some situations, both X and Y are random variables and you should note that we can still use a regression line of y on x if we are required to predict values of y from observations made on x. In this case the variables x and y play different roles. In correlation, the two variables are interchangeable. Examples involving two random variables often quoted are the shear strength (y) and diameter of spot welds (x) (neither can be precisely controlled) and the bending moment (y) and shear (x) at the fixed point of a beam as illustrated below

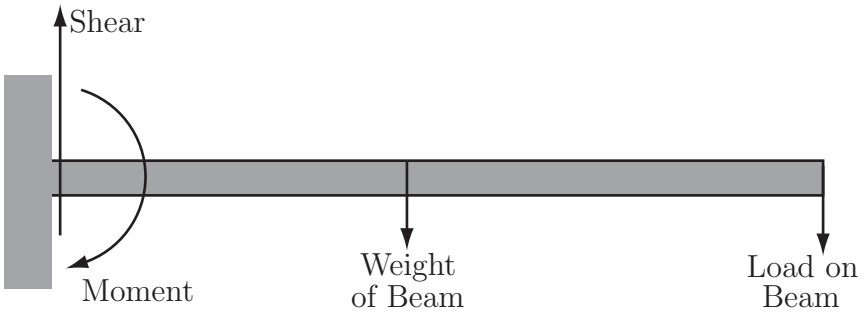


Figure 6

Again, neither variable (shear or moment) can be precisely controlled, each is a random variable. In cases such as these, we turn to the correlation coefficient (sometimes called Pearson's coefficient of correlation or simply Pearson's r) defined as

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

where σ_{xy} is the covariance between X and Y and σ_x and σ_y are the standard deviations of X and Y. We need to express this formula in terms of quantities which facilitate the easy calculation of the correlation coefficient.

Key Point 4

Pearson's Coefficient of Correlation, *r*

In terms of corresponding sample values (x, y),

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

Further, it can also be shown that $-1 \leq r \leq 1$ and that:

(a) $r = -1$ represents perfect negative correlation with all (x, y) lying on a straight line with negative gradient;

(b) $r = 1$ represents perfect positive correlation with all (x, y) lying on a straight line with positive gradient;

(c) $r = 0$ represents the situation where either there is no *linear* relationship between the variables or that any relationship existing is non-linear.

The calculation of Pearson's r

The worked example below shows the setting out of a table which will facilitate the easy calculation of Pearson's r.

Example 3

Find the value of Pearson's r for the following set of data obtained by reading seven torque values (x) from an electric motor using current (y).

Student	1	2	3	4	5	6	7
x-Value	16	14	12	10	8	6	4
y-Value	12	8	16	14	4	10	6

Solution

The calculation is done as follows:

x	y	x^2	y^2	xy
16	12	256	144	192
14	8	196	64	112
12	16	144	256	192
10	14	100	196	140
8	4	64	16	32
6	10	36	100	60
4	6	16	36	24
$\sum x = 70$	$\sum y = 70$	$\sum x^2 = 812$	$\sum y^2 = 812$	$\sum xy = 752$

Substituting in the formula we developed for r gives the result:

$$r = \frac{752 \times 7 - 70 \times 70}{\sqrt{(7 \times 812 - 70^2)(7 \times 812 - 70^2)}} = 0.46$$

In practice, one would set up a spreadsheet or use a specialist statistical software package to do the calculations.

Comment

Any value of r calculated says something about the degree of correlation present between the two independent random variables present in the calculation. In order to give real meaning to the value of the correlation coefficient we should test the significance of the value of r, in this case 0.46.

The significance of Pearson's r

In order to test the significance of a calculated value of r we assume that both x and y are normally distributed and set up the hypotheses:

$$H_0: \ \rho = 0 \qquad H_1: \ \rho \neq 0$$

where ρ is the 'true' value of the population correlation. If the assumption of normality is false the test must not be used. We know that the value of $-1 \leq r \leq 1$ and we wish to know whether our correlation coefficient is significantly different to zero.

Key Point 5

Significance of Pearson's r

It can be shown that the test statistic

$$r_{test} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

calculated from a sample of n pairs of values, follows a t-distribution with $n-2$ degrees of freedom.

Note that many authors simply miss out the modulus sign and ignore the sign of r should it be negative. The test statistic is then written

$$r_{test} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

and critical values depending on the level of significance required are read off from t-tables in the usual way. A copy of t-distribution tables is included at the end of this Workbook (Table 2).

Example 4

Test the significance of the value of r obtained from Example 3 concerning electric motor torque values. Use the 5% level of significance.

Solution

The sample size is 7 so we have 5 degrees of freedom. The value of r_{test} is given by

$$r_{test} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.46 \times \sqrt{7-2}}{\sqrt{1-0.46^2}} = 1.158$$

From Table 2, the critical value for a two-sided test at the 5% level of significance is 2.571. In this case, since $1.158 < 2.571$ we cannot reject the null hypothesis at the 5% level of significance and conclude that for the motor under investigation, there is no evidence of a relationship between torque produced and current used.

Hooke's law relates the extension of a spring under load to its extended length. The following results were obtained experimentally.

Load (N)	2	5	8	11	15
Extension (mm)	2	23	62	119	223

Calculate Pearson's r and test its significance at the 5% level. What conclusion can you draw?

Your solution

Answer

Setting up a spreadsheet to do the calculations gives:

Load (x)	Exten. (y)	xy	x^2	y^2
2	2	4	4	4
5	23	115	25	529
8	62	496	64	3844
11	119	1309	121	14161
15	223	3345	225	49729
Sum$(x) =$	Sum$(y) =$	Sum$(xy) =$	Sum$(x^2) =$	Sum$(y^2) =$
41	429	5269	439	68267

$r = 0.97379629 \qquad r_{\text{test}} = 7.41645174$

Hence, since the critical value for a two-sided t-test at the 5% level read off from tables is 3.182 we see that since $7.416 > 3.182$ we can reject the null hypothesis at the 5% level and conclude that the correlation coefficient is significantly different from zero.

Comments on interpretation

Some care should always be taken when interpreting results obtained from correlation coefficient calculations.

(a) A high correlation does not necessarily imply that a causal relationship exists between the variables considered. For example, it may be that a high degree of correlation exists between the number of road accidents in a particular city and the number of late trains arriving at a station in another city both over the same time period. In general one would not expect to find a causal relation between the variables involved. Similar comments apply to, for example, water hardness and average income for towns in the UK.

(b) When considering the behaviour of two variables, one should realize that it is possible that both variables may change because of the influence of a third variable. An example often quoted in this context is the Gas law

$$\frac{PV}{T} = \text{constant}$$

where say, pressure and volume may change because of a change in temperature.

(c) A low value of the correlation coefficient does not necessarily imply that no relationship exists between the variables being considered. Remember that the correlation coefficient is indicative of a *linear relationship only* and that a low or zero value of r may indicate that a non-linear relationship exists. For example a set of points lying on the curve $y = x^2$ *might* (see the Tasks below) result in a zero value of r.

Write down five (x, y) points (symmetrical about zero) lying on the parabola $y = x^2$. Show that the correlation coefficient between x and y is zero.

Your solution

x	y	xy	x^2	y^2
Sum$(x) =$	Sum$(y) =$	Sum$(xy) =$	Sum$(x^2) =$	Sum$(y^2) =$

Answer

Let the five points be (for example) $(-2,4),(-1,1),(0,0),(1,1),(2,4)$

x	y	xy	x^2	y^2
-2	4	-8	4	16
-1	1	-1	1	1
0	0	0	0	0
1	1	1	1	1
2	4	8	4	16
Sum$(x)=$	Sum$(y)=$	Sum$(xy)=$	Sum$(x^2)=$	Sum$(y^2)=$
0	10	0	10	34

The value of r is given by

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{\left(n\sum x^2-(\sum x)^2\right)\left(n\sum y^2-(\sum y)^2\right)}}=\frac{5\times 0-0\times 10}{\sqrt{(5\times 10-0^2)(5\times 34-10^2)}}=0$$

Write down five (x,y) points (all involving positive values of x and y) lying on the parabola $y=x^2$. Show that the correlation coefficient between x and y is non-zero.

Your solution

x	y	xy	x^2	y^2
Sum$(x)=$	Sum$(y)=$	Sum$(xy)=$	Sum$(x^2)=$	Sum$(y^2)=$

Answer

Let the five points be (for example) $(0,0),(1,1),(2,4),(3,9),(4,16)$,

x	y	xy	x^2	y^2
0	0	0	0	0
1	1	1	1	1
2	4	8	4	16
3	9	27	9	81
4	16	64	16	256
Sum$(x)=$	Sum$(y)=$	Sum$(xy)=$	Sum$(x^2)=$	Sum$(y^2)=$
10	30	100	30	354

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{\left(n\sum x^2-\left(\sum x\right)^2\right)\left(n\sum y^2-\left(\sum y\right)^2\right)}}=\frac{5\times 100-10\times 30}{\sqrt{(5\times 30-10^2)(5\times 354-30^2)}}=0.959$$

Spearman's coefficient of correlation

There are times when data cannot be expressed in terms of numbers directly. For example, an audio engineer might be asked to give an opinion on the quality of sound produced by three sets of speakers. The results will represent a judgement made by the engineer. The engineer could adopt a set of criteria including, for example, the clarity of the treble, the power of the base and the ability of the speakers to distinguish between instruments. Suppose the results are as follows:

Test Item	Rating	Rank Order
Speaker Pair B	9/10	1
Speaker Pair A	8/10	2
Speaker Pair C	5/10	3

Note that the results are not numeric in an arithmetic sense so you cannot do meaningful arithmetic using the results. In order to see this, just ask what a calculation based on the ranks such as

$$\frac{1+2^2}{3}$$

would actually mean. The answer is, of course, nothing!

During your career as an engineer you may be asked to rank data in a similar way to that outlined above. You may be asked to assess the work of colleagues for promotion purposes or give an opinion on the visual appeal of alternative designs of manufactured objects such as mobile telephones, food containers or television sets.

Assigning numbers to data in order of size (often called ranking methods) can also be useful if one does not wish to make assumptions about the nature of the distributions underlying the data. (For example whenever at least one of the distributions describing the behaviour of the variables may not be normal.) In order to check the level of correlation between results obtained by ranking data we calculate Spearman's coefficient of correlation.

Key Point 6

Spearman's Coefficient of Correlation, R

$$R = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

where $D = R_X - R_y$ is the difference of the rank R_X of an item according to variable X and rank R_Y of the item according to variable Y.

The formula indicates that the differences of each pair of ranked values are to be found, squared and summed. It is worth noting that even though it is not obvious, Spearman's coefficient is just Pearson's coefficient applied to ranks.

The calculation of Spearman's R

The following worked example illustrates the procedure.

Example 5

A production engineer is asked to grade, on the basis of 12 criteria A to L, a junior colleague who has applied for promotion. In order to try to ensure that he treats the colleague fairly, the engineer repeats his gradings after a few days. On the basis of the results below, can you conclude that the results are consistent? The gradings are percentages.

Criterion	First Grading(X)	R_X	Second Grading(Y)	R_Y
A	55	8	75	7
B	53	9	80	6
C	78	3	89	4
D	50	10	63	11
E	48	11	67	10
F	61	7	69	9
G	66	6	73	8
H	76	4	93	2
I	85	2	87	5
J	90	1	95	1
K	69	5	92	3
L	45	12	59	12

Solution

The calculation may be set out as follows:

Criterion	R_X	R_Y	$D = R_X - R_Y$	D^2
A	8	7	1	1
B	9	6	3	9
C	3	4	-1	1
D	10	11	-1	1
E	11	10	1	1
F	7	9	-2	4
G	6	8	-2	4
H	4	2	2	4
I	2	5	-3	9
J	1	1	0	0
K	5	3	2	4
L	12	12	0	0
				$\sum D^2 = 38$

Substituting in the formula for R gives the value

$$R = 1 - \frac{6 \times 38}{12 \times 143} = 0.87$$

Note that we have not made any attempt to interpret the meaning of this figure of 0.87. Methods for doing this are discussed below.

The significance of spearman's R

Like Pearson's r the value of R may be shown to lie in the range $-1 \leq R \leq 1$ and in order to test the significance of a calculated value of R we set up the hypotheses

$$H_0 : \ \rho = 0 \qquad H_1 : \ \rho \neq 0$$

Key Point 7

Significance of Spearman's R

We wish to know whether our correlation coefficient is significantly different to zero. It can be shown that for large samples, the test statistic

$$R_{test} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

calculated from a sample of n pairs of values, follows a t-distribution with $n-2$ degrees of freedom.

Critical values depending on the level of significance required are read from t-tables. When dealing with Spearman's coefficient of correlation, the size of the sample is important. Different authors recommend different minimum sample sizes, a common recommendation being a *minimum* of $n = 10$. Even though they are not used here, you should note that tables are available which allow us to read critical values corresponding to small sample sizes.

Example 6

A production engineer is asked to grade, on the basis of 12 criteria (say) A to L a junior colleague who has applied for promotion. He repeats his gradings after a few days. The results (calculated in Example 5) gave a value of $R = 0.87$. Test at the 5% level to determine whether the results are consistent.

Solution

The calculation is:

$$R_{test} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{0.87 \times \sqrt{12-2}}{\sqrt{1-0.87^2}} = 5.580$$

The 5% critical value for a two sided test read from tables is 2.228 and since $5.580 > 2.228$ we conclude that we must reject the null hypothesis that the correlation coefficient is zero.

As a result of two tests given to 10 students studying laboratory safety, the students were placed in the following class order.

Student	Test 1	Test 2
A	2	3
B	4	5
C	3	7
D	5	9
E	1	10
F	6	2
G	8	6
H	7	8
I	9	4
J	10	1

Use Spearman's R to discuss the consistency of their performances. Can you make any meaningful comment regarding the two tests as a means of assessing laboratory safety?

Your solution

Answer
Setting up the hypotheses

$$H_0: \ R = 0 \qquad H_1: \ R \neq 0$$

and doing the appropriate calculations using a spreadsheet gives:

Test 1	Test 2	D	D^2
2	3	-1	1
4	5	-1	1
3	7	-4	16
5	9	-4	16
1	10	-9	81
6	2	4	16
8	6	2	4
7	8	-1	1
9	4	5	25
10	1	9	81
			sum $= 242$
	$R = -0.4666667$	$R_{\text{test}} = 1.49240501$	

From t-tables it may be seen that the critical value (8 degrees of freedom) at the 5% level of significance is 2.306. Since $1.492 < 2.306$ we cannot reject the null hypothesis that there is no correlation between the results. This implies that the performances of the students on the tests may not be related and we should question at least one of the tests as a means of assessing laboratory safety. One could, of course, question the usefulness of both tests!

As part of an educational research project, twelve engineering students were given an intelligence test (IQ score) at the start of their first year course. At the end of the first year their results in engineering science (ES score) were noted down on the expectation that they would correlate with the results of the intelligence test. The results were as follows:

Student	1	2	3	4	5	6	7	8	9	10	11	12
IQ Score	135	120	125	135	125	140	135	140	135	140	120	135
ES Score	85	74	76	90	85	87	94	98	81	91	76	74

Calculate Pearson's r for these data. Can you conclude that there is a linear relationship between IQ scores and ES scores? You may assume that the IQ scores and the ES scores are each normally distributed.

Your solution

Answer

Setting up the hypotheses

$$H_0: \ R = 0 \qquad H_1: \ R \neq 0$$

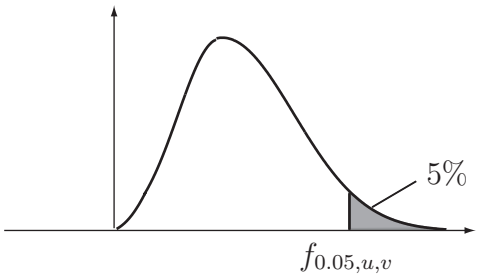
and doing the appropriate calculations using a spreadsheet gives:

$IQ(x)$	$ES(y)$	xy	x^2	y^2
135	85	11475	18225	7225
120	74	8880	14400	5476
125	76	9500	15625	5776
135	90	12150	18225	8100
125	85	10625	15625	7225
140	87	12180	19600	7569
135	94	12690	18225	8836
140	98	13720	19600	9604
135	81	10935	18225	6561
140	91	12740	19600	8281
120	76	9120	14400	5776
135	74	9990	18225	5476
sum$x = 1585$	sum$y = 1011$	sum$xy = 134005$	sum$x^2 = 209975$	sum$y^2 = 85905$

$$r = 0.696 \qquad r_{\text{test}} = 3.065$$

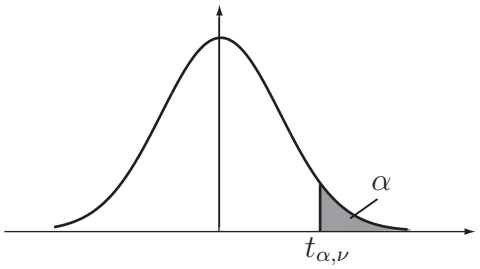
From t-tables it may be seen that the critical value (10 degrees of freedom) at the 5% level of significance is 1.812. Since $3.065 > 1.812$ we reject the null hypothesis that there is no linear association between the results. This implies that the performances of the students on the ES tests is linearly related to their IQ scores.

Table 1: Upper 5% points of the F distribution



	Degrees of Freedom for the Numerator (u)														
v	1	2	3	4	5	6	7	8	9	10	20	30	40	60	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	248.0	250.1	251.1	252.2	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.45	19.46	19.47	19.48	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.62	8.59	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.75	5.72	5.69	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.53	4.46	4.43	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.87	3.81	3.77	3.74	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.38	3.34	3.30	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.15	3.08	3.04	3.01	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.86	2.83	2.79	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.70	2.66	2.62	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.65	2.57	2.53	2.49	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.54	2.47	2.43	2.38	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.46	2.38	2.34	2.30	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.39	2.31	2.27	2.22	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.33	2.25	2.20	2.16	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.28	2.19	2.15	2.11	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.23	2.15	2.10	2.06	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.19	2.11	2.06	2.02	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.16	2.07	2.03	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	2.04	1.99	1.95	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.10	2.01	1.96	1.92	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.07	1.98	1.94	1.89	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.05	1.96	1.91	1.86	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.03	1.94	1.89	1.84	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.01	1.92	1.87	1.82	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	1.99	1.90	1.85	1.80	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	1.97	1.88	1.84	1.79	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	1.96	1.87	1.82	1.77	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	1.94	1.85	1.81	1.75	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	1.93	1.84	1.79	1.74	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.84	1.74	1.69	1.64	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.75	1.65	1.59	1.53	1.39
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.57	1.46	1.39	3.32	1.00

Table 2: Critical points of student's t distribution



α	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
v										
1	.325	1.000	3.078	6.314	12.706	31.825	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.902	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.514	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.487
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291