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!

Time allocation

You are expected to spend approximately ten hours of independent study on the material presented in this workbook. However, depending upon your ability to concentrate and on your previous experience with certain mathematical topics this time may vary considerably.

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 . . .

① study Descriptive Statistics using Workbook 36

② make sure you can find the expectation and variance of sums of variables using Workbook 39.3

③ understand the terms independent and dependent variables.

④ understand the terms biased and unbiased estimators.

Learning Outcomes

After completing this Section you should be able to . . .

✓ understand what is meant by the terms regression analysis and regression line.

✓ understand 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 dependant 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 for 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.

In this work, even though an outline introduction to non-linear regression is given at the end of the 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.

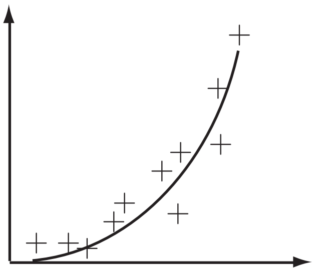


Figure 1

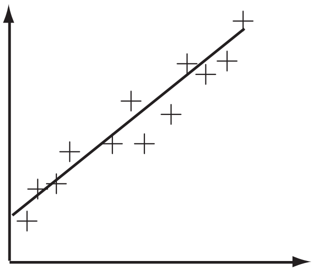


Figure 2

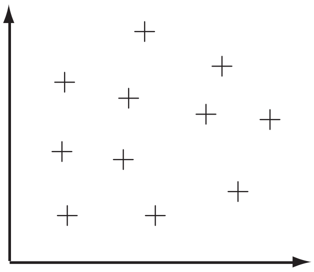


Figure 3

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 unit 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.

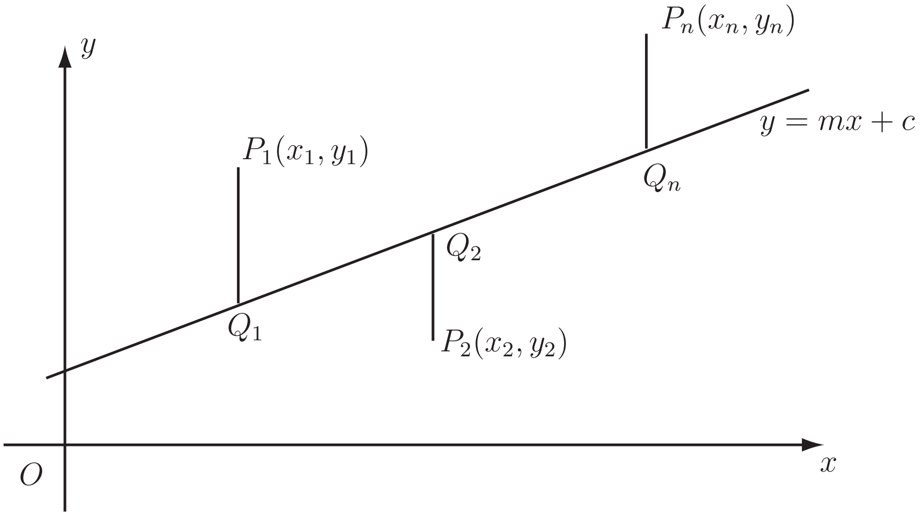
Regression Lines by Eye

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

(i) 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.



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

$$y_1 - mx_1 - c, \quad y_2 - mx_2 - c, \quad y_3 - mx_3 - c \quad \cdots \quad y_n - mx_n - c$$

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:

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

and

$$\sum y - m\sum x - nc = 0 \qquad \text{(ii)}$$

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

$$\frac{\sum y}{n} - m\frac{\sum x}{n} - c = 0$$

or, put more simply

$$\bar{y} = m\bar{x} + c$$

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 c from the equations gives a formula for the gradient m of the regression line, this is:

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

This is often written as

$$m = \frac{S_{xy}}{S_x^2}$$

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 m enables us to obtain the value of c from the equation

$$\bar{y} = m\bar{x} + c$$

Summary

The least squares regression line of y on x has the equation $\bar{y} = m\bar{x} + c$. Remember that:

$$m = \frac{\dfrac{\sum xy}{n} - \dfrac{\sum x}{n}\dfrac{\sum y}{n}}{\dfrac{\sum x^2}{n} - \left(\dfrac{\sum x}{n}\right)^2} \text{ and that c is given by the equation } c = \bar{y} - m\bar{x}$$

It should be noted that the coefficients m and c 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

$$x = m'y + c'$$

where

$$m' = \frac{\dfrac{\sum xy}{n} - \dfrac{\sum x}{n}\dfrac{\sum y}{n}}{\dfrac{\sum y^2}{n} - \left(\dfrac{\sum y}{n}\right)^2} \text{ and } c' = \bar{x} - m'\bar{y}$$

Example 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 of 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 $\sum x = 700$, $\sum x^2 = 44200$, $\sum y = 1860$ $\sum xy = 118600$ so that the formulae

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

give $m = 3$ and $c = -20$. Our regression line is $y = 3x - 20$.

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.

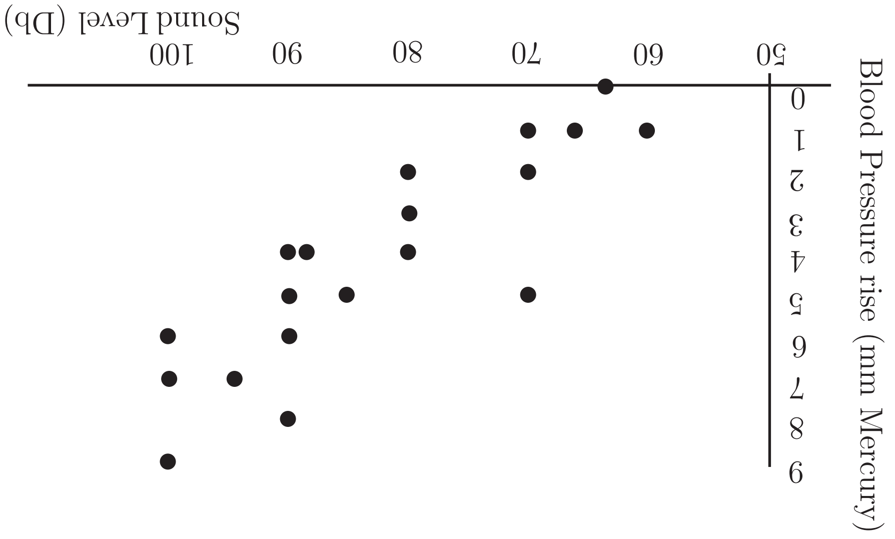
(c) Calculate a line of best fit of y on x for the data given.

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

Your solution

(a) Entering the data into Excel and plotting gives

Blood Pressure increase versus recorded sound level



(b) A linear model is appropriate.

(c) Excel calculations give $\sum x = 1656$, $\sum x^2 = 140176$, $\sum y = 86$, $\sum xy = 7654$ so that $m = 0.1743$ and $c = -10.1315$. 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

(ii) A Modelling View

We take the dependent variable Y to be 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

$$Y = mx + c + e$$

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

$$E(Y|x) = E(mx + c + e) = E(mx + c) + E(e)$$

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

$$V(Y|x) = V(mx + c + e) = V(mx + c) + V(e), \quad \text{assuming independence.}$$

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

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

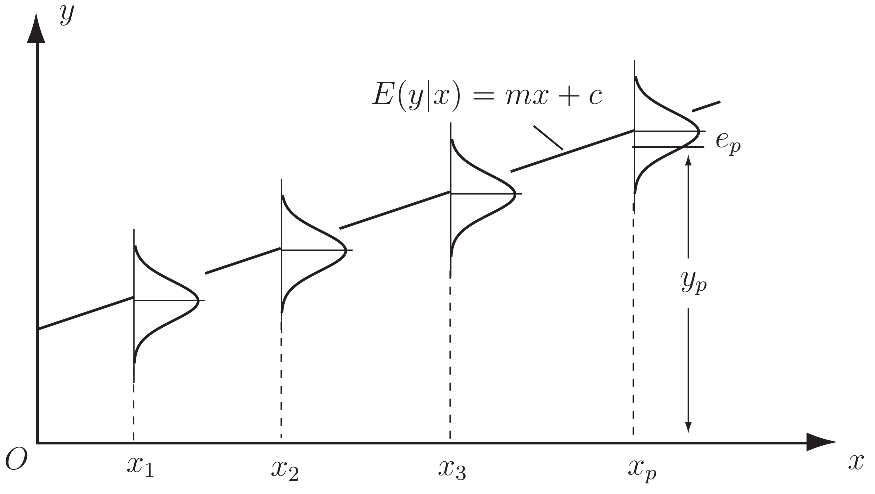
is a straight line of mean values. If we now assume that the errors e are distributed with mean 0 and variance σ^2 we may write

$$E(Y|x) = E(mx + c) + E(e) = mx + c \quad \text{since } E(e) = 0.$$

and

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

This implies that for each value of x, Y is distributed with mean $mx + c$ 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 and variance σ^2 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.



The regression line is shown passing through the means of the distributions for the individual values of x. Any randomly selected value of y may be represented by the equation

$$y_p = mx_p + c + e_p$$

where e_p is the error of the observed value of y its expected value, namely

$$E(Y|x_p) = \mu_{y|x_p} = mx_p + c$$

Note that

$$e_p = y_p - mx_p - c$$

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

$$S = \sum e_p^2 = \sum (y_p - mx_p - c)^2$$

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 m and c are as before, namely

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

Note that since the error e_p in the pth observation essentially describes the error in the fit of the model to the pth observation, the sum of the squares of the errors $\sum e_p^2$ 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 $V(Y|x) = \sigma^2$ 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 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 $\sum(y_r - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$. The value of R^2 is sometimes referred loosely 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 regressor variables x and 1. 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

$$R^2 \qquad \text{or} \qquad R^2_{adj} = 0.9751$$

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

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

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^2	Cable Length	y^2	xy	Predicted	Error
(x)		(y)			Values	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
$m = 3$	$c = -20$	SST =		$R^2 =$		$R^2_{\text{adj}} =$
		31500		0.962		0.958

Use the drum and cable data given originally and set up a spreadsheet to verify the values of the Coefficient of Variation and the Adjusted Coefficient of Variation.

Your solution

As per the table above 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:

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

It may be shown that 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.

Example 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 $SS_T = \sum y^2 - \dfrac{\left(\sum y\right)^2}{n}$ 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 tables, 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 $m \neq 0$.