**CHAPTER SIX**

**STATISTICAL REGRESSION AND CORRELATION**

**Objectives:**

After completing this chapter you should be able to:

* Define regression and correlation.
* Understand and Compute regression equation.
* Make a scatter plot of the given data and plot a regression line on the scatter plot.

**6.1 What is a regression model?**

Regression analysis is almost certainly the most important tool at the econometrician’s disposal. But what is regression analysis? In very general terms, regression is concerned with describing and evaluating the relationship between a given variable and one or more other variables. More specifically, regression is an attempt to explain movements in a variable by reference to movements in one or more other variables.

To make this more concrete, denote the variable whose movements the regression seeks to explain by **Y** and the variables which are used to explain those variations by **X**1, **X**2, . . . , **X**k .

Hence, in this relatively simple setup, it would be said that variations in k variables (the xs) cause changes in some other variable, y. This chapter will be limited to the case where the model seeks to explain changes in only one variable y.

**6.2 Regression versus Correlation**

All readers will be aware of the notion and definition of correlation. The correlation between two variables measures the degree of linear association between them. If it is stated that y and x are correlated, it means that y and x are being treated in a completely symmetrical way. Thus, it is not implied that changes in x cause changes in y, or indeed that changes in y cause changes in x. Rather, it is simply stated that there is evidence for a linear relationship between the two variables, and that movements in the two are on average related to an extent given by the correlation coefficient.

In regression, the dependent variable (y) and the independent variable ( s) (Xs) are treated very differently. The y variable is assumed to be random or ‘stochastic’ in some way, i.e. to have a probability distribution. The X variables are, however, assumed to have fixed (‘non-stochastic’) values in repeated samples. Regression as a tool is more flexible and more powerful than correlation.

**6.3 Simple regression**

For simplicity, suppose for now that it is believed that y depends on only one x variable. Again, this is of course a severely restricted case, three examples of the kind of relationship that may be of interest include:

* How asset returns vary with their level of market risk
* Measuring the long-term relationship between stock prices and dividends
* Constructing an optimal hedge ratio.

Suppose that a researcher has some idea that there should be a relationship between two variables y and x, and that financial theory suggests that an increase in x will lead to an increase in y . A sensible first stage to testing whether there is indeed an association between the variables would be to form a scatter plot of them. In this case, it appears that there is an approximate positive linear relationship between x and y which means that increases in x are usually accompanied by increases in y , and that the relationship between them can be described approximately by a straight line. It would be possible to draw by hand onto the graph a line that appears to fit the data. The intercept and slope of the line fitted by eye could then be measured from the graph. However, in practice such a method is likely to be laborious and inaccurate.

It would therefore be of interest to determine to what extent this relationship can be described by an equation that can be estimated using a defined procedure. It is possible to use the general equation for a straight line to get the line that best ‘fits’ the data. The researcher would then be seeking to find the values of the parameters or coefficients, α and β, which would place the line as close as possible to all of the data points taken together.

**y = α + βx**

However, this equation (y = α + βx) is an exact one. Assuming that this equation is appropriate, if the values of α and β had been calculated, then given a value of x, it would be possible to determine with certainty what the value of y would be. Imagine a model which says with complete certainty what the value of one variable will be given any value of the other!

Charles Spearman’s coefficient of correlation (or rank correlation) is the technique of determining the degree of correlation between two variables in case of ordinal data where ranks are given to the different values of the variables. The main objective of this coefficient is to determine the extent to which the two sets of ranking are similar or dissimilar. This coefficient is determined as under:

Spearman's coefficient of correlation (or r) = 1-6di2**/**n(n2-1)

 Where: di=difference between ranks of ith pair of the two variables

 n=number of pairs of observations

Karl Pearson’s coefficient of correlation (or simple correlation) is the most widely used method of measuring the degree of relationship between two variables. This coefficient assumes the following:

(i) that there is linear relationship between the two variables;

(ii) that the two variables are casually related which means that one of the variables is independent and the other one is dependent; and

(iii) a large number of independent causes are operating in both variables so as to produce a normal distribution.

Karl Pearson’s coefficient of correlation can be worked out thus.

Karl Pearson’s coefficient of correlation (or r) \* = (Xi-Xbar)(Yi-Ybar)**/**n.ðx.ðy

Alternatively, the formula can be written as:

 (This applies when we take zero as the assumed mean for both variables, X and Y.)

Where Xi = ith value of X variable

 Xbar = mean of X

 Yi = ith value of Y variable

 Ybar = Mean of Y

 n = number of pairs of observations of X and Y

 ðx = Standard deviation of X

 ðy = Standard deviation of Y

In case we use assumed means (Ax and Ay for variables X and Y respectively) in place of true

means, then Karl Person’s formula is reduced to:

 Example 1. Two points on the straight line that represents Mr. Jones’s salary are(x 1 ; y 1) = (5 ; 140 000) and (x2 ; y2) = (9 ; 216 000).

 The slope of the line is given by

b= y2-y1 = 216,000-140,000 = 76,000 = **19,000**

 x2-x1 9-5 4

Two points on the straight line that represents Mr. Brown’s salary are (x1 ; y1) = (6 ; 160 000) and (x2 ; y2) = (9 ; 244 000). The slope of the line is

244,000-160,000 = **28,000**

 9-6

Two points on the straight line that represents Mr. Smith’s salary are (x1 ; y1) = (7 ; 150 000)

and (x2 ; y2) = (9 ; 210 000).The slope of the line is

 b = 30 000.

Mr. Smith’s salary increased at the highest rate.

 A. The equation is y = a + bx with b = 28000, thus y = a + 28000x.

 Substitute point (6 ; 160 000) into the equation:

 160 000 = a + 28 000 ×6

 a + 168 000 = 160 000

 a = −8 000.

 Thus y = −8000 + 28000x.

 B. The year 2008 corresponds to x = 10.

Substitute x = 10 into y = −8000 + 28000x

Which gives

y = −8 000 + 28 000 × 10

= 272 000.

In 2008 he will make R272 000.

Example 2. It looks as if there is a positive linear correlation between average interest rate and yearly investment. This means that if the average interest rate increases, then yearly investment will also increase.

 Average interest rate

A. You must do these calculations directly on your calculator. We are only showing mathematical how your calculator came to the answer.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year(i) | Average interest ( xi)  | Yearly investment (yi )  | xi2 | xiyi | yi2 |
| 1 | 13.8  | 1,060  | 190.44  | 14,628  | 1,123,600 |
| 2 | 14.5  | 940 | 210.25  | 13,630  | 883,600 |
| 3 | 13.7  | 920 | 187.69  | 12,604  | 846,400 |
| 4 | 14.7  | 1,110  | 216.09  | 16,317  | 1,232,100 |
| 5 | 14.8  | 1,550  | 219.04  | 22,940  | 2,402,500 |
| 6 | 15.5  | 1,850  | 240.25  | 28,675  | 3,422,500  |
| 7 | 16.2  | 2,070  | 262.44  | 33,534  | 4,284,900 |
| 8 | 15.9  | 2,030  | 252.81  | 32,277  | 4,120,900 |
| 9 | 14.9  | 1,780  | 222.01  | 26,522  | 3,168,400 |
| 10 | 15.1  | 1,420  | 228.01  | 21,442  | 2,016,400 |
| n=10  | 149.1  | 14,730  | 2,229.03  | 222,569  | 23,501,300 |

$$r=n\sum\_{i=1}^{n}xiyi-\sum\_{i=1}^{n}xi\sum\_{i=1}^{n}yi$$

 $\sqrt{n\sum\_{i=1}^{n}xi2-(\sum\_{i=1}^{n}xi)2}\sqrt{n\sum\_{i=1}^{n}yi2-(\sum\_{i=1}^{n}yi)2}$

 =10(22,569)-(149.1)(14,730)

 $\sqrt{10\left(2,229.03\right)-\left(149.1\right)2 \sqrt{10\left(23,501,300\right)\left(147,730\right)2}}$

 =24,447/32,759.8161

 =**0.8989**

b. The coefficient of determination is r2=0.89892=0.8080. This means that almost 81% of the variation in yearly investments can be declared by the average interest rate.

 C. The equation of the straight line is y=a+bx, where

 b= $\sum\_{i=1}^{10}xiyi-\sum\_{i=1}^{10}xi\sum\_{i=1}^{10}yi$

 $\sqrt{n\sum\_{i=1}^{10}xi2-(\sum\_{i=1}^{10}xi)2}$

 =10(22,569)-(149.1)(14,730)

 $\sqrt{10\left(2,229.03\right)-\left(149.1\right)2 }$

 =24,447/59.49=**494.99**

**And** a=$ n\sum\_{i=1}^{10}yi-n\sum\_{i=1}^{10}xi$

 n n

=14,730- (494.99)(149.1) = **-5907.30**

 10 10

Thus y=-5907.30+494.99x

Example 3. The following table shows the number of loans approved for different amounts during the second half of 2008.

|  |  |
| --- | --- |
|  Amount of loan in Birr 100,000 (x)  | Number of loans (y)  |
| 2 | 45 |
| 3 | 250 |
| 4 | 250 |
| 5 | 175 |
| 6 | 125 |

Example 4. A study was undertaken at eight garages to determine how the resale value of a car is affected by its age. The following data was obtained:

|  |  |  |
| --- | --- | --- |
| Garage | Age of car (in years) |  Resale value (in Birr) |
| 1 | 1 | 41,250 |
| 2 | 6 | 10,250 |
| 3 | 4 | 24,310 |
| 4 | 2 | 38,720 |
| 5 | 5 | 8,740 |
| 6 | 4 | 26,110 |
| 7 | 1 | 38,650 |
| 8 | 2 | 36,200 |

The garage manager suspects a linear relationship between the two variables. Fit a curve of the form y = a + bx to the data.

The equation for the regression line is

y = 48 644,17− 6 596, 93,The correlation coefficient is r = −0, 9601

**Summary**

In studying the relationship between two variables x and y, each of n observed (x, y) data pairs is plotted as a point on a graph to obtain a scatter plot. If the scatter plot indicates a linear tendency, we fit a straight line equation of the form y= a + bx

To the n data points as follows: first compute the slope(regression coefficient) b with the formula: b= n(xy)-(x)(y)

 n(x2)-(x)2

Next, compute the constant term(the y intercept) **a** with the formula

a=y-b(x)/n

Then write the equation by substituting the computed values of a and b in to the expression

 y= a +bx

 The expression ybar=a+bx is called an estimating equation, or a *regression equation*.

Correlation analysis is carried out when interest centers on whether or not a relationship exists, rather than on obtaining an estimating(regression) equation. Then the sample correlation coefficient r= n(xy)-(x)(y)

 $\sqrt{n\sum\_{}^{}x2-(\sum\_{}^{}x)2}\sqrt{n\sum\_{}^{}y2-(\sum\_{}^{}y)2}$

is completed.

The coefficient of determination is: r2= (n$(\sum\_{}^{}xy)-(\sum\_{}^{}x)(\sum\_{}^{}y))2$

 (n($\sum\_{}^{}x2)-(\sum\_{}^{}x)2)(n\left(\sum\_{}^{}y2\right)-(\sum\_{}^{}y)2)$

The coefficient of determination is interpreted by saying that r2 is the proportion of the variation in the dependent variable that is explained by the variation in the independent variable.

$\sqrt{n\sum\_{i=1}^{n}xi2-(\sum\_{i=1}^{n}xi)2}\sqrt{n\sum\_{i=1}^{n}yi2-(\sum\_{i=1}^{n}yi)2}$***Exercises***

1. Explain the concept of regression and point out its usefulness in dealing with business problems.

2. Distinguish between correlation and regression.

3. Point out the role of regression analysis in business decision making. What are the important properties of regression coefficients?

4. The following data give the hardness(x) and the tensile strength(y) of 7 samples of metal in certain units. Find the linear regression equation of y on x.

X 146 152 158 164 170 176 182

Y 75 78 77 89 82 85 86

5. Joe brown collected data on car weights in thousands of pounds x and miles per gallon y. The (x,y) data pairs are (2.0,33),(2.4,30),(2.8,28),(3.2,23), and(3.6,19).(a)Make a scatter plot of a data.(b) Compute a and b, then write the equation of the regression line.(c) plot the regression line on the scatter plot in a.