REGRESSION ANALYSIS

1 INTRODUCTION

Regression analysis attempts to establish the nature of the relationship between variables that are related *causally*, that is, to study the *functional* relationship between the variables and thereby provide a mechanism for *prediction* or forecasting. The method of regression essentially attempts to establish a mathematical equation relating the *response* variable, usually denoted by *Y*, and one or more *predictors*, denoted by $X_1, X_2, \ldots, X_n$. The response and predictor variables are also respectively known as the *dependent* and *independent* variables.

2 SIMPLE REGRESSION

In the case of *bivariate* data, there is only *one* independent variable. If *X* is the independent variable and *Y* is the dependent one, the objective is thus to determine an equation of the form $\hat{Y} = f(X)$ where $f(X)$ can be *linear* or *non-linear* in nature. (The reason for the circumflex accent (or 'hat') on the *Y* symbol will be explained in Section 8.2 below.)

Consider a very simple example in an attempt to explain how the procedure works – let us say that it is suspected that 'temperature' has a direct effect on the 'length' of a metal rod. It is known that an increase in temperature will increase the length of the rod.

We may perform an experiment in the laboratory as follows: heat the rod continuously and record its length at randomly chosen temperatures. Once it has attained a reasonably high temperature, let it cool down and, during the cooling process, record its length again at randomly chosen temperatures. (Note that same temperatures need not give the same length!) Another way of proceeding is to preset the temperatures at which the lengths are to be recorded.

In one such experiment, the results were recorded in the **Table 2.1** below.

Temperature (^oC)	13	50	63	58	20	78	39	55	29	62
Length (cm)	5.10	5.68	5.85	5.74	5.25	5.98	5.59	5.73	5.46	5.81

Table 2.1

It is obvious from the experiment that the length of the metal rod depends on its temperature. Thus, 'temperature' is the independent variable and 'length' is the dependent variable. These will be respectively denoted by *X* and *Y*. The first step in the determination of the equation $\hat{Y} = f(X)$ is to plot the above results in

the form of points (x, y) on graph. The resulting figure is known as a *scatter diagram* and its importance is that it gives us an idea of whether the points describe a linear or non-linear relationship between X and Y.

Fig. 2.2 and **2.3** below display possibilities of *linear* and *exponential* relationships respectively. However, for the purpose of this course, we will only consider linear relationships.

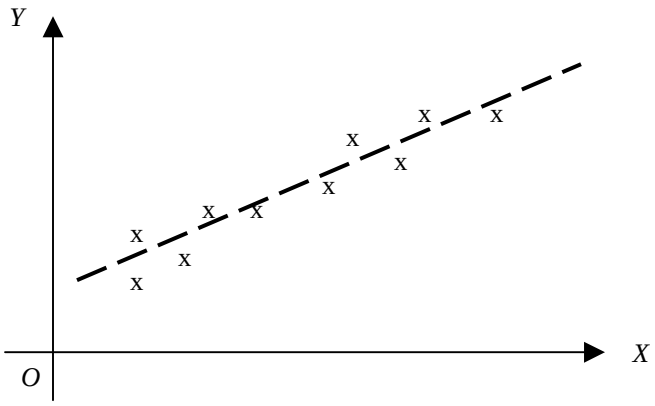

Fig. 2.2 Linear relationship

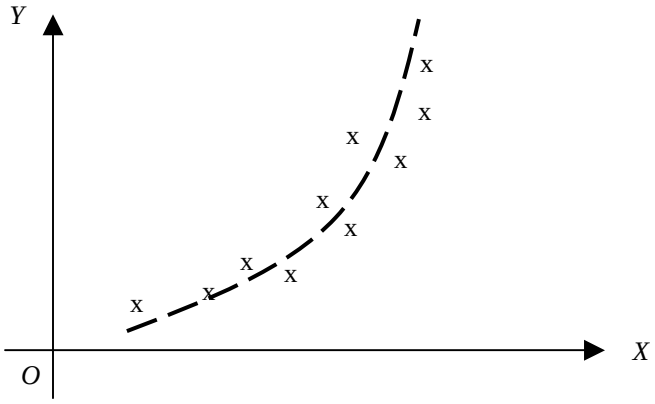

Fig. 2.3 Exponential relationship

3 SIMPLE LINEAR REGRESSION

In this case, the mathematical relationship between X and Y is given as $\hat{Y} = a + bX$. This is similar to the usual equation of a line in Mathematics, that is, $y = mx + c$, except for the notation. Here, a is the y-intercept of the regression line of Y on X and b is its gradient. a and b are known as the *regression coefficients* and are constants to be determined so that prediction is possible.

The 'hat' over the response variable Y is used to emphasise on the fact that the y-value is an *expected* or *theoretical* value, hence, a *forecast*. Note also that, if we had several independent variables, the relationship would be a *multiple regression* model.

4 THE METHOD OF LEAST SQUARES (OLS)

The *method of (ordinary) least squares* extrapolates from observed values of a relationship to a *functional* relationship, especially to a curve that best fits a given set of data. The equation connecting the variables is sought by the minimisation of the sum of the squares of the differences between the observed and the theoretical values (*residuals* or errors).

The above definition is more clearly explained by the following scatter diagram:

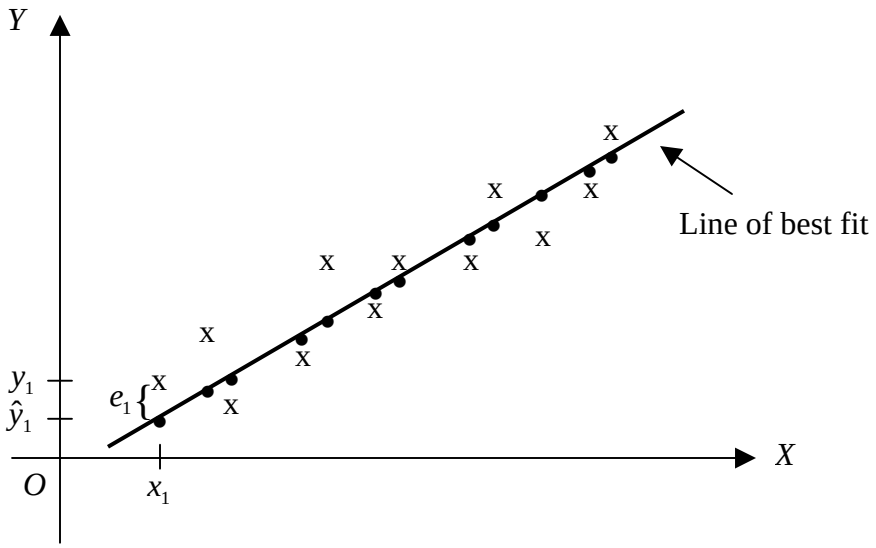

Fig. 4.1 The method of least squares

The points indicated by crosses in **Fig. 4.1** above are values obtained by observation (in an experiment) whereas those indicated by black dots are their *expected* counterparts. For example, one of the observed pairs was the point (x_1, y_1); it can be seen that, for that same x-value, there exists an expected y-value, denoted by $\hat{y}_1$. We could interpret $\hat{y}_1$ as the value, which should have been theoretically obtained, but for experimental or systematic errors. The difference between the i^{th} observed value and its corresponding expected value is known as a *residual* or, simply, an error, mathematically equal to $|\hat{y}_i - y_i|$ and denoted by e_i (e_1 is indicated in **Fig. 4.1**).

The principle of the least-squares method is based on minimising the *sum of the squares of the residuals*, $\sum e_i^2$. The intricate mathematical manipulations being beyond the scope of this course, it will simply be mentioned that, for the regression line of *Y* on *X*, the method of least-squares yield two *normal equations* from which the regression coefficients can be determined.
These equations are given by

$$\sum y = na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

Solving them simultaneously will give

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

Using the data from our rod-heating experiment, with the *x*-values rearranged in ascending order for better understanding, we have **Table 4.2**:

Temperature (^oC)	13	20	29	39	50	55	58	62	63	78
Length (cm)	5.10	5.25	5.46	5.59	5.68	5.73	5.74	5.81	5.85	5.98

Table 4.2

Using a pocket calculator, we may summarise the above information as

$$\sum x = 467 \quad \sum x^2 = 25717 \quad \sum xy = 2674.93 \quad \sum y = 56.19 \quad \sum y^2 = 316.4141$$

[*The reader is advised to check the above sums on a calculator in regression mode.*]

Thus, the coefficients of regression line of *length* on *temperature* are

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{(10)(2674.93) - (467)(56.19)}{(10)(25717) - (467)^2} = 0.013.$$

$$a = \frac{\sum y - b\sum x}{n} = \frac{56.19 - 467b}{10} = 5.011$$

so that the regression line of length on temperature is $\hat{Y} = 5.011 + 0.013X$. *The values of a and b can also be found from a calculator.*

Interpretation of regression coefficients

The constant a, the value of y when $x = 0$, represents the length of the rod at 0^o *C*, that is 5.011 cm, whereas b (the gradient) is the rate of change of length with temperature, that is, for every rise of 1^o *C*, the length of the rod increases by 0.013 cm.

5 PREDICTION

The ultimate aim of regression being prediction, once the equation of the regression line of Y on X has been determined, we can use it to find values of Y for given values of X.

5.1 Interpolation

If prediction is made for those x-values lying *within* the range of values of X given in the table, the process is known as *interpolation*. *Note that the value of Y can also be found by graphical drawing and from a calculator.*

For example, at a temperature of 50^o *C*, the expected length of the rod would be $5.011 + 0.013(50) = 5.66$ cm (correct to two decimal places). Note that this is not necessarily equal to the observed length at the same temperature, that is, 5.68 cm (from Table **2.1**).

5.2 Extrapolation

Extrapolation, in a sense, real *forecasting*, may be done to a certain extent but may prove to be quite risky sometimes. If the given value of X lies *outside* the range of values in the table, then we should be very careful when predicting its corresponding Y-value. This is because the behaviour of Y is known only for the given values in the table. It is quite possible that the relationship between X and Y does not follow the same pattern for values other than those given in the table.

However, concessions may be made for values of X lying very near the *least* and *greatest* values in the table on the assumption that it is very probable that, near those regions, Y will follow an identical pattern as described by the regression equation. We should bear in mind that the *further* the given value of X is from the table values, the *less reliable* will be the forecast.

For example, at a temperature of 80^o *C*, the expected length of the rod would be $5.011 + 0.013(80) = 6.05$ cm (correct to two decimal places). This is a *very reliable* forecast since it is quite close to the maximum x-value of 78^o *C*. The idea of reliability is easily understood if we were asked to forecast the length of the rod at a temperature of $1\ 000\ 000^o$ *C*. Substitution in the regression equation

will yield a hypothetical length of 130.18 m! We very well know that, at this temperature, the rod will have already melted ☺.

Example 1

A large field of maize was divided into six plots of equal area and each plot fertilised with a different concentration of fertiliser. The yield of maize from each plot is shown below.

Concentration (oz m^{-2})	0	1	2	3	4	5
Yield (tonnes)	15	22	31	40	48	54

(a) Obtain the equation of the regression line for yield on concentration, giving the values of the coefficients to 2 decimal places.

(b) Interpret the regression coefficients in the context of the problem.

(c) Use the regression line to obtain the yield when the concentration is 3 oz m^{-2}. State precisely what is being estimated by this value.

(d) State any reservations you would have about making an estimate from the regression equation of the expected yield per plot if 7 oz m^{-2} of fertiliser is applied.

Solution

Since the yield depends on the concentration of fertiliser, it is the dependent variable and will be denoted by *Y*. If we denote concentration by *X*, then the information from the table can be summarised as

Concentration (oz m^{-2})	0	1	2	3	4	5
Yield (tonnes)	15	22	31	40	48	54

$n = 6 \quad \sum x = 15 \quad \sum x^2 = 55 \quad \sum xy = 666 \quad \sum y = 210 \quad \sum y^2 = 8490$

(a) From the formulae of *a* and *b*, we have

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{(6)(666) - (15)(210)}{(6)(55) - (15)^2} = 8.06\,.$$

$$a = \frac{\sum y - b\sum x}{n} = \frac{210 - 15b}{6} = 14.86$$

The equation of the regression line of Y on X is $Y = 14.86 + 8.06X$.

(b) The minimum yield per plot (without any fertiliser) is 14.86 oz m^{-2}.
For every extra oz m^{-2} of fertilizer added, there is an additional yield of 8.06 tonnes per plot.

(b) When $X = 3$, $Y = 14.86 + (8.06)(3) = 39.04$ tonnes. We have just calculated the yield for a plot which has a concentration of 3 oz m^{-2} of fertiliser.

(d) When $X = 7$, $Y = 14.86 + (8.06)(7) = 71.28$ tonnes. This estimate is not very reliable since there is no guarantee that Y will have the same behaviour for values of X outside the given range in the table. One would think that there should be a ceiling for the yield per plot. Otherwise, the yield would grow indefinitely, which is a very unrealistic situation.

Example 2

A company monitored the number of days (x) of business trips taken by executives of the company and the corresponding claims *(£y*) they submitted to cover the total expenditure of these trips. A random sample of 10 trips gave the following results.

x (days)	10	3	8	17	5	9	14	16	21	13
y (£)	116	39	85	159	61	94	143	178	225	134

(a) Find an equation of the regression line of Y on X in the form $Y = a + bX$.

(b) Interpret the slope b and intercept a of your line.

(c) Find the expected expenditure of a trip lasting 11 days.

(d) State, giving a reason, whether or not you would use the line to find the expected expenditure of a trip lasting 2 months.

Solution

The information in the table above is summarised as follows:

$$n = 6 \quad \sum x = 116 \quad \sum x^2 = 1630 \quad \sum xy = 1728 \quad \sum y = 1234 \quad \sum y^2 = 180754$$

(a) From the formulae of a and b, we have
The equation of the regression line of Y on X is $Y = 8.64 + 9.89X$.

(b) The minimum allowance for any trip is £ 8.64.
For every extra day, there is an additional expenditure of £ 9.89.

(c) When $X = 11$, $Y = 8.64 + (9.89)(11) =$ £ 117.43. A trip lasting 11 days is expected to cost £ 117.43.

(d) 2 months are approximately equal to 60 days. Since 60 is well outside the given range of values of X, it is not advisable to use the line to make any forecast. There is no guarantee that the same policy of allocating £ 9.89 per day is applicable. For example, there could be special package deals for longer trips.