UNIT 4 BIVARIATE AND MULTIPLE REGRESSION

Structure

4.0 INTRODUCTION

Psychologists, as other scientists, are also interested in prediction. Since our domain of enquiry relates with human behaviour, our predictions are associated with human behaviour. We are interested in knowing how human beings will behave provided we have some information about them. It is not that we all the time depend on theories such as psychoanalysis, behaviourism or cognitive in order to predict human behaviour. There are also statistical methods which can help predict certain phenomenon of human behaviour. We would study in this unit the statistical methods that can be used for the purpose of prediction. These statistical methods are called Regression. We will first learn the concept of regression, then learn how to plot the relationship between variables, and learn to work out The Regression Equation. We will also deal with how far we can be accurate in predicting with the help of regression equation by the help of tests of significance. Finally we will be dealing with how to interpret regression and deal with also Multiple regression, that is, which variables influence a particular phenomenon.

4.1 OBJECTIVES

After completing this unit, you will be able to:

- Describe and explain concept of regression correlation;
- Explain, describe and differentiate between bivariate regression and multiple regresion;

- Describe and explain concept of multiple correlation;
- Develop a regression equation;
- Compute the a and b of bivariate regression by using OLS;
- Test the significance of regression;
- Interpret regression results;
- Apply the regression techniques to the real data;
- Explain Multiple regression; and
- Use Multiple regression in real data.

4.2 BIVARIATE AND MULTIPLE REGRESSION

We always see that the meteorology department predicts the rain, the economists predict the outcome of a particular policy, financial experts predict the share market, the election experts predict the outcome of voting and so on. Similarly using statistical method, psychologists too can predict certain human behaviours. For example, one can predict the examination marks after writing the examination by checking what questions we have attempted and how we have faired in it and what marks we can expect for each question and so on.

Most of us are interested in this exercise of prediction. On many occasions, we also predict the behaviour of our friends, colleagues and family. Predicting and trying to speculate about what might happen in future is integral part of human curiosity. While there are many theories of psychology and personality that would help us predict behaviours, one can also predict a certain phenomenon in terms of statistics. This method is called regression in statistics.

The simplest form of the regression is simple linear regression (at times also called as bivariate regression). Carl Frederick Gauss discovered a method of least squares (1809) and later on developed Gauss-Markov theorem (1821). Sir Fransis Galton contributed to the method of regression and also gave the name.

Let us see what this is all about. Let us say we have data on two variables (Y and X), and we create an equation, called regression equation, which later on helps us in predicting the score of one variable (Y) by simply using the scores on another variable (X). Let us learn about the utility of the regression analysis, how to do it, how to test the significance and issues surrounding it.

Regression analysis tries to predict Y variable from X variable. In the general form, it tries to predict Y from a X_1, X_2, …, X_k, where k is number of predictor variables. Initially we will learn about two variable prediction, one of which is a predictor and the other one will be predicted. Then we will look at the general form of Regression.

Just think of the variables that can be used in prediction in psychology. Look at the following statements. (See the box below)

- Stress leads to health deterioration.
- Openness increases creativity.
- Extraversion increases social acceptance.

- Social support influences coping with mental health problems.
- Stigma about mental illness decides the help seeking behaviour.
- Parental intelligence leads to child's intelligence
- Attitude to job and attrition depends on affective commitment to the organisation.

What do you see in common in all these statements?

All the statements above have two variables.

One of the variables can potentially predict the other variable.

4.2.1 Predicting One Variable from Another

Let us now consider the problem of prediction. How to predict Y from X.

The Y variable is called the dependent variable. It is also called as criterion variable. It is the variable that has to be predicted.

The X variable is a called as independent variable. It is also called as predictor variable (Please note that in experimental psychology we define independent variable as the variable that is manipulated by the experimenter, whereas in regression the term is used less strictly. In Regression, the independent variable is not manipulated by the researcher or experimenter.)

If X is predicting Y, then typically it is said that 'X is regressed on Y'.

Let's identify the X and Y in our statements given in the box.

In the first statement Stress (X) lead to the health (Y) deterioration

In the second statement, Openness (X) increases the creativity (Y).

In the third statement, Extroversion (X) increases the social acceptance (Y).

In fourth statement, Social support (X) influences the coping with mental health problems (Y).

In the fifth statement, Stigma about mental illness (X) decided the help seeking behaviour (Y).

In the sixth statement, Parental intelligence (X) leads to child's intelligence (Y).

In the last statement, Attitude (Y) to job and attrition depends on affective commitment(X) to the organisation

Before we learn how to do the regression, we shall quickly browse through the basic concepts in regression analysis.

4.2.2 Plotting the Relationship

We have already learned to plot the scatter plots. We shall try to plot a scatter and try to understand regression graphically.

The perfect relationship

Look at the following example. You have data of five swimmers on two variables, hours of practice per day (X) and time taken (Y).

Swimmer	hours of practice per day	Time taken (in seconds)
A	1	50
B	2	45
C	3	40
D	4	35
E	5	30

Now plot the relationship between them as a scatter. You know how to do that. We have now tried to draw a line that passes through all the data points in the scatter. And we have successfully done it.

Looking at figure 4.1 you realise that as the number of hours spent in practice increase the time taken is reducing. There is a perfect linear relationship between them. This means that you can draw a line on the scatter that passes through all the data points on the scatter.

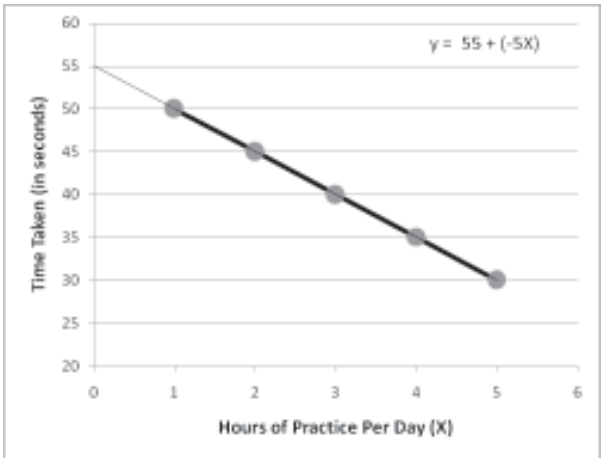

Fig. 4.1: Figure showing the data between number of hours spent in practice and time take.

For this data, the slope of the line can be calculated by using a simple technique.

$$Slope = \frac{Y_2 - Y_1}{X_2 - X_1} \qquad \text{(eq. 4.1)}$$

Where Y_2 and Y_1 are any two points on Y axis and X_2 and X_1 are corresponding two points on X axis.

For example, take $Y_2 = 45$ and $Y_1 = 40$ and corresponding X_2 and X_1 are 2 and 3. The slope is

$$Slope = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{45 - 40}{2 - 3} = \frac{5}{-1} = -5 \qquad \text{(eq. 4.2)}$$

The slope of the line is – 5.

The point at which the line passes through the Y axis (the Y intercept of the line) is 55.

Now, if we ask about the unknown score, 6 hours of practice per day, then the predicted X score is 25 seconds (which is very close the world record).

How have we obtained it? we have solved it for a equation of straight line. That equation is

$$Y = a + bX \qquad \text{(eq. 4.3)}$$

Where a = point where the line passes the Y axis and

b = is a slope of the line.

We have $a = 55$ and $b = -5$. So for X = 6 the Y will be

(eq. 4.4)

The Imperfect Relationship.

But the problem is the real data will not be so systematic and all data points in scatter will not fall on a straight line.

Look at the following example of the stigma and visits to mental health professionals. The Table 4.2 shown below display the data of stigma and number of appointments missed to mental health professional.

Table 4.2: Data of stigma and number of appointments missed to mental health professional

Patient	Stigma scores	Number of appointments missed
1	60	5
2	50	2
3	70	9
4	73	6
5	64	9
6	68	4
7	56	3
8	54	8
9	49	3
10	66	11

This data was obtained from ten patients who are suffering due to mental illness. The data was collected on King, Show and others (2007) Stigma scale and the data were obtained on number of visits missed by the patients. The data is plotted in the scatter plot below.

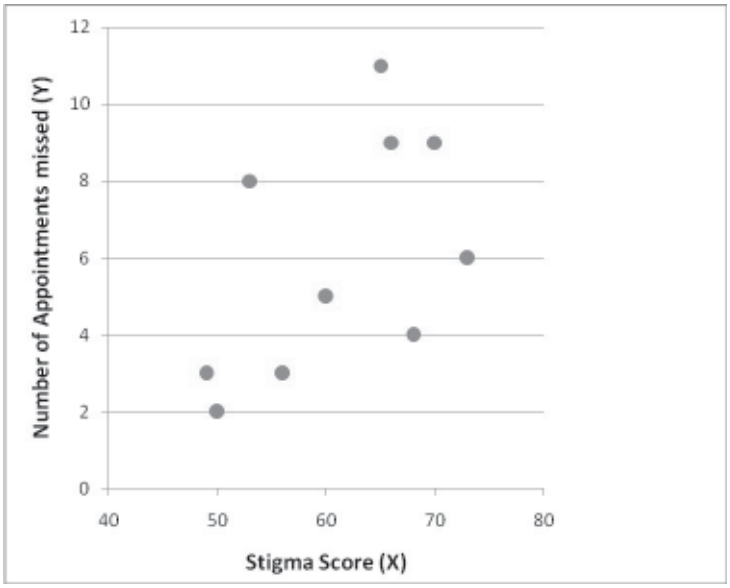

Fig. 4.2: Scatter showing the relationship between stigma and number of appointments missed

Now you will realise that it is not possible to draw a straight line that passes through all the data points. Then how to know the relationship between X and Y and then predict the scores of Y from scores of X. How to draw the straight line for this data? This is a problem one would face with real data. The linear regression analysis solves this problem.

4.2.3 Mean, Variance and Covariance: Building Blocks of Regression

In order to understand the building blocks of regression we must describe some of the terms such as (i) the mean, (ii) variance, and (iii) covariance which are presented in the following section.

i) Mean

Mean of variable X (symbolised as $\bar{X}$) is sum of scores ($\sum_{i=1}^{n} X_i$) divided by number of observations (n). The mean is calculated in following way.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(eq. 4.5)}$$

You have learned this in the first block. We will need to use this as a basic element to compute correlation.

ii) Variance

The variance of a variable X (symbolised as S_X^2) is the sum of squares of the deviations of each X score from the mean of X ($\sum (X - \bar{X})^2$) divided by number of observations (n).

$$^{2}_{X} = \frac{\sum (X - \bar{X})^2}{n} \qquad \text{(eq. 4.6)}$$

You have already learned that standard deviation of variable X , symbolised as S_X , is square root of variance of X, symbolised as .

iii) Covariance

The covariance between X and Y (Cov_{XY} or S_{XY}) can be stated as

$$Cov_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} \qquad \text{(eq. 4.7)}$$

Covariance is a number that indicates the association between two variables. To compute covariance, deviation of each score on X from its mean ($\bar{X}$) and deviation of each score on Y from its mean ($\bar{Y}$) is initially calculated.

Then products of these deviations are obtained.

Then, these products are summated.

This sum gives us the numerator for covariance.

Divide this sum by number of observations (*n*).

The resulting number is covariance.

4.2.4 The Regression Equation

The regression equation can be written as

$$Y = \alpha + \beta X + \varepsilon \quad \text{(eq. 4.8)}$$

Where,

Y = dependent variable or criterion variable

á = the population parameter for the y-intercept of the regression line, or regression coefficient (r=óy/ óx)

â = population slope of the regression line or regression coefficient (r*óx/ óy)

å = the error in the equation or residual

The value of á and â are not known, since they are values at the level of population. The population level value is called the parameter. It is virtually impossible to calculate parameter. So we have to estimate it. The two parameters estimated are á and â. The estimator of the á is 'a' and the estimator for â is 'b'. So at the sample level equation can be written as

$$Y = a + bX + e \quad \text{(eq. 4.9)}$$

Where,

Y = the scores on Y variable

X = scores on X variable

a = the Y-intercept of the regression line for the sample or regression constant in sample

b = the slope of the regression line or regression coefficient in sample

e = error in prediction of the scores on Y variable, or residual

$$\hat{Y} = a + bX \quad \text{(eq. 4.10)}$$

Where, Y = predicted value of Y in sample. This value is not an actual value but the value of Y that is predicted using the equation $\hat{Y} = a + bX$. So we can write error as by substituting the in the earlier equation.

$$S_x^2 \quad \text{(eq. 4.11)}$$

$$Y - \hat{Y} = e \quad \text{(eq. 4.12)}$$

This is a useful expression. We shall use it while computing the statistical significance of the regression and will also be useful for understanding the least squares.

4.2.5 Ordinary Least Squares (OLS)

Just recall the data between the stigma scores and number of appointments missed by the person. Now, if we have to draw the straight line that will explain the relationship between the stigma scores and number of appointments missed, then there will be many such lines possible. Out of them, which line we should consider as the best fit line?

It is not possible to draw a straight line that will pass through all the points. And many lines are possible with the earlier equation $Y = a + bX + e$

This problem is solved by the method of least squares or ordinary least squares (OLS).

One easy way to judging how good the line is, is to know how close various values of $\hat{Y}$ are to corresponding values of Y, which means to check how close predicted value ($\hat{Y}$) is to the actual value of the Y.

These predicted values are computed by using the various values of X in the data.

But how to decide what is the best fit?

One logical solution to this problem is to look at the error term, e. the e is defined as

$$Y - \hat{Y} = e$$

Which means,

$$Y - (a + bX) = e \qquad \text{(eq. 4.13)}$$

The $Y - \hat{Y}$ is the error in prediction of the Y.

This error is called as an obtained residual of the regression.

The best line is the one that minimises this residual.

Some of the predicted values of Y will be higher than the actual value of Y and some would be lower, and hence the sum of residual will be zero.

In order to take care of this problem, the summation is not done over the $Y - \hat{Y}$ Instead the $\sum\left(Y - \hat{Y}\right)^2$ is summated. An attempt to *minimise the sum of the squared errors* — minimise the $\sum\left(Y - \hat{Y}\right)^2$ is made, this is called as least squares.

Calculation of a and b

The values for *a* and *b* that minimises the sum of the squared errors — minimise the $\sum\left(Y - \hat{Y}\right)^2$ need to be calculated. The *b* can be calculated as follows.

$$b = \frac{Cov_{XY}}{S_X^2} \qquad \text{(eq. 4.14)}$$

Where,

Cov_{XY} = covariance between X and Y. This is given by the formula $\sum$ (X-X)(Y-Y) / N S_X^2 = variance of X

The *b* is covariance of X and Y divided by the variance of X. it can be rewritten as

$$b = \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}}{S_X^2} \qquad \text{(eq. 4.15)}$$

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{nS_X^2} \qquad \text{(eq. 4.16)}$$

The *a* can be calculated as follows by using our earlier equation.

$$\bar{Y} = a + b\bar{X} \qquad \text{(eq. 4.17)}$$

$$a = \bar{Y} - b\bar{X} \qquad \text{(eq. 4.18)}$$

Once we know how to calculate *a* and *b*, then we can solve the problem of regression. Let's now solve the example we have started with. The example was about the predicting the number of appointments missed by the patient (Y) by using the Stigma scale scores (X). The data is as follows:

Table 4.3: Table showing the computation of *a* and *b*.

Patient	**Stigma scores**	**Number of appointments missed**	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	60	5	-1	-1	1	1	1
2	50	2	-11	-4	121	16	44
3	70	9	9	3	81	9	27
4	73	6	12	0	144	0	0
5	64	9	3	3	9	9	9
6	68	4	7	-2	49	4	-14
7	56	3	-5	-3	25	9	15
8	54	8	-7	2	49	4	-14
9	49	3	-12	-3	144	9	36
10	66	11	5	5	25	25	25
X	$\sum X$ =610	$\sum Y$ =60			$\sum (X - \bar{X})^2$ = 648	$\sum (Y - \bar{Y})^2$ = 86	$\sum (X - \bar{X})(Y - \bar{Y})$ = 129
n = 10	$\bar{X} = 61$	$\bar{Y} = 6$					

$$S_X = \sqrt{\sum (X - \bar{X})^2 / n} = 8.50$$

$$S_Y = \sqrt{\sum (Y - \bar{Y})^2 / n} = 2.93$$

$$Cov_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} = \frac{129}{10} = 12.9$$

$$b = \frac{Cov_{XY}}{S_X^2} = \frac{12.9}{8.50^2} = \frac{12.9}{64.8} = 0.1991$$

$$a = \bar{Y} - b\bar{X} = 6 - (0.1991 \times 61) = -6.144$$

Step 1. You need scores of subjects on two variables. We have scores on ten subjects on two variables, the Stigma scores (X) and number of appointments missed (Y).

Then list the pairs of scores on two variables in two columns.

The order will not make any difference.

Remember, same individuals' two scores should be kept together.

Label the predictor variable as X and criterion as Y.

Step 2. Compute the mean of variable X and variable Y. It was found to be 61 and 6 respectively.

Step 3. Compute the deviation of each X score from its mean ($\overline{X}$) and each Y score from its own mean ($\overline{Y}$). This is shown in the column labeled as $X - \overline{X}$ and $Y - \overline{Y}$. As you have learned earlier, the sum of these columns has to be zero.

Step 4. Compute the square of $X - \overline{X}$ and $Y - \overline{Y}$. This is shown in next two columns labeled as $(X - \overline{X})^2$ and $(Y - \overline{Y})^2$. Then compute the sum of these squared deviations of X and Y.

The sum of squared deviations for X is 648 and for Y it is 86.

Divide them by *n* to obtain the standard deviations for X and Y. The was found to be 8.49. Similarly, the S_y was found to be 3.09.

Step 5. Compute the cross-product of the deviations of X and Y. These cross-products are shown in the last column labeled as $(X - \overline{X})$ and $(Y - \overline{Y})$. Then obtain the sum of these cross-products. It was found to be 129. Now, we have all the elements required for computing *b*.

Step 6. Compute the covariance between X and Y, which turned out to be 12.9.

Step 7. Compute the *b* value by dividing the covariance XY (Cov_{XY}) by the variance of . We compute S_x^2 by taking *n* as a denominator the S_x^2 value is 64.8. The b is found to be 0.1991. Now we can easily compute the *a* which is

$$a = \overline{Y} - b\overline{X} = 6 - (0.1991 \times 61) = -6.144 .$$

Once the *a* and *b* are computed, we can write the regression equation to get the predicted values of Y as follows:

$$\hat{Y} = a + bX \qquad \text{(eq. 4.19)}$$

$$\hat{Y} = -6.144 + (0.1991 \times X)$$

Now we can compute the predicted values for each of the X value. For example the predicted value for the first X value (60) is as follows:

$$5.80 = -6.144 + (0.1991 \times 60)$$

In this way you can compute the predicted Y value for each of the X score. Now you realise that this value is not Y value but the predicted Y value obtained from X.

Now look at the table below. It gives the X, Y and Predicted Y values.

Table 4.4: Table showing the computation of the significance for the *b*, the slope of the line

Ss	Stigma scores (X)	Number of appointments missed (Y)	Predicted value of Y	Residual $Y-\hat{Y}=e$	Residual $(Y-\hat{Y})^2=e$	Variance explained $\hat{Y}-\overline{Y}$	Variance explained Squared $(\hat{Y}-\overline{Y})^2$
1	60	5	5.80	-0.80	0.64	-0.20	0.04
2	50	2	3.81	-1.81	3.28	-2.19	4.80
3	70	9	7.79	1.21	1.46	1.79	3.21
4	73	6	8.39	-2.39	5.71	2.39	5.71
5	64	9	6.60	2.40	5.77	0.60	0.36
6	68	4	7.39	-3.39	11.52	1.39	1.94
7	56	3	5.00	-2.00	4.02	-1.00	0.99
8	54	8	4.61	3.39	11.52	-1.39	1.94
9	49	3	3.61	-0.61	0.37	-2.39	5.71
10	66	11	7.00	4.00	16.04	1.00	0.99
Sum	610	60	60	0	60.32	0	25.68

With the availability of residual, we can obtain the sum of squared residual. The sum of squared residual is 60.32. This is the minimum value that can be obtained if a straight line is drawn for the relationship between X and Y.

There is no other line than can give value as small as this.

So this line is considered as a best fit line.

The mean of Y is 6. So we can now obtain an interesting expression. This expression is $\hat{Y}-\overline{Y}$.

This will provide us the amount of variance in Y explained by the predicted value of Y which is $\hat{Y}$.

The sum of this difference is bound to be zero. So we square the difference.

The sum of square of the difference between predicted value of Y and mean of Y is given below:

$$\sum(\hat{Y}-\overline{Y})^2 ,$$

This is the amount of variance explained in the Y by the predicted value of the Y. This can be expressed as follows:

Total Variance in Y = Variance Explained by Regresion + Residual variance

(eq. 4.20)

This can be written as

$$Y-\overline{Y}=(\hat{Y}-\overline{Y})+(Y-\hat{Y}) \quad \text{(eq. 4.21)}$$

$$\sum Y-\overline{Y}=\sum(\hat{Y}-\overline{Y})+\sum(Y-\hat{Y}) \quad \text{(eq. 4.22)}$$

Since the summation of these differences are zero, we square the difference. The equation can be rewritten as

$$\sum (Y-\bar{Y})^2 = \sum (\hat{Y}-\bar{Y})^2 + \sum (Y-\hat{Y})^2 \qquad \text{(eq. 4.23)}$$

Where,

$\sum (Y-\bar{Y})^2$ = Total variance in Y. Total sum of squares (SS_T).

$\sum (\hat{Y}-\bar{Y})^2$ = Variance in Y explained by X. Sum of squares explained ($SS_{Regression}$).

$\sum (Y-\hat{Y})^2$ = variance in Y not explained by X. Residual sum of squares ($SS_{Residual}$).

$$SS_{Total} = SS_{Regresion} + SS_{Residual} \qquad \text{(eq. 4.24)}$$

Look at the figure below. You will understand the division of SS_{Total} into $SS_{Regression}$ and $SS_{Residual}$.

It shows that the distance between the $\bar{Y}$ and Y is total deviation of that Y value from $\bar{Y}$. This is shown as $(Y-\bar{Y})$.

From this total deviation or variation, the explained variation is distance between $\bar{Y}$ and the predicted Y value. This is shown as $\hat{Y}-\bar{Y}$.

This is explained by the regression line. The distance that regression equation fails to explain is between Y and predicted value of Y. This distance is residual or remaining variance that regression equation cannot explain. This is shown as $Y-\hat{Y}$.

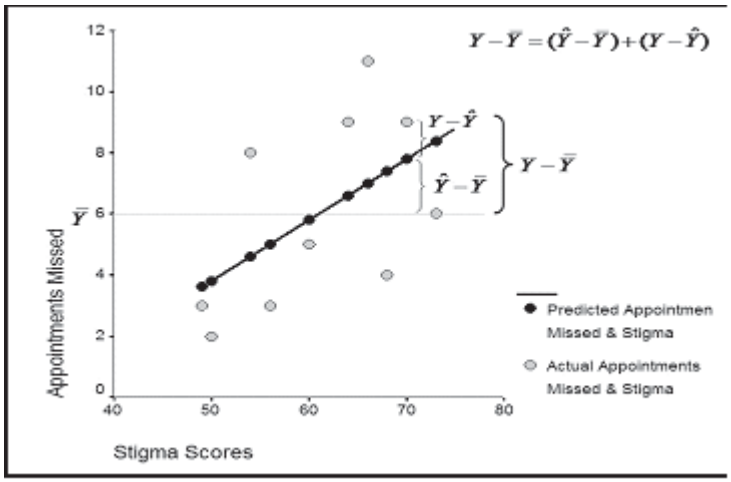

Fig. 4.3: The figure showing the scatter of X and Y, the regression line, and also explains the variance explained, residual and total.

4.2.6 Significance Testing of b

The *F*-distribution is employed to test the significance of the *b*.

The slope or regression coefficient obtained on sample is an estimator of the population slope or population regression coefficient called as â.

We already have completed the basics for computing the *F*-distribution.

$$F = \frac{S^2_{Between}}{S^2_{Within}} \qquad \text{(eq. 4.25)}$$

In case of regression, the same formula is used. The sum of squares total, sum of squares regression, and sum of squares residual have already been computed. We will use them now. Look at the table below.

Table 4.5: Table showing the computation of significance of *b*.

Source	Sum of Squares	*df*	S^2	F
Regression	$\sum(\hat{Y}-\bar{Y})^2$	k	$\frac{\sum(\hat{Y}-\bar{Y})^2}{k}$	$\frac{S^2_{Regresion}}{S^2_{Residual}}$
Residual	$\sum(Y-\hat{Y})^2$	$n-k-1$	$\frac{\sum(Y-\hat{Y})^2}{n-k-1}$	
Total	$\sum(Y-\bar{Y})^2$	$n-1$		

Where, n = sample size, and k = number of independent variables.

The null and the alternative hypothesis tested are as follows:

The F is computed for our example.

Table 4.6: Table showing the computation of the *F*-statistics for the data.

Source	Sum of Squares	*df*	S^2	F
Regression	25.68	1	25.68	$\frac{25.68}{7.54}=3.41$
Residual	60.32	8	7.54	
Total	86	9		

The *F*-value needs to be tested for its significance. The *F*-value at numerator *df* = 1, and denominator *df* = 8 at 0.05 level is 5.31. The obtained value of the *F* is smaller than the tabled value of the F. This means that we need to accept the null hypothesis which states that the $\hat{a} = 0$.

This might look surprising for some of you. But one thing we need to understand is the fact that the sample size (*n*) for this example is very small. Given that small *n*, the ability to reject the false null hypothesis is not so good and that's the reason we are accepting this null hypothesis.

4.2.7 Accuracy of Prediction

The present example has not turned out to be significant. But we will continue to discuss the issues in regression. How accurate we are in predicting Y from X is one of the important issues. We will look at various measures that tell us about the accuracy of prediction. We will continue to use this example considering it as significant even when it is not.

Standard Error of Measurement:

The standard error of estimate provides us an estimate of the error in the estimation. It can be calculated as follows:

$$s_{Y.X}=\sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}}=\sqrt{\frac{SS_{Residual}}{df}} \qquad \text{(eq. 4.26)}$$

The standard error in our example can be computed using the formula as follows:

$$s_{Y.X} = \sqrt{\frac{SS_{Residual}}{df}} = \sqrt{\frac{60.32}{8}} = 7.54$$

Percentage of Variance Explained: r^2

The r^2 can be used as a measure of amount of variance X explained in Y. This shows the proportion of variance explained from total variance. Look at the following equation.

$$r^2 = \frac{SS_{Regression}}{SS_{Total}} \qquad \text{(eq. 4.27)}$$

$$r^2 = \frac{\sum(\hat{Y} - \bar{Y})}{\sum(Y - \bar{Y})} \qquad \text{(eq. 4.28)}$$

$$r^2 = \frac{\sum(\hat{Y} - \bar{Y})}{\sum(Y - \bar{Y})} = \frac{25.68}{86} = 0.299$$

Which means that 29.9 percent variance in Y is explained by X. This 'explained variance' around 30 percent is a good amount of variance considering the unreliability of psychological variables.

Indeed, the square root of the r^2, will give us the correlation between the X and Y.

Proportional Improvement in Prediction

The Proportional Improvement in Prediction (PIP) is one of the measure of accuracy. It is calculated as follows:

$$\text{PIP} = 1 - \sqrt{(1 - r^2)} \qquad \text{(eq. 4.29)}$$

In case of our example,

$$\text{PIP} = 1 - \sqrt{(1 - r^2)} = 1 - \sqrt{(1 - .299} = 0.162$$

The PIP value for our example is 0.162. So the proportional improvement in prediction is .162.

4.2.8 Assumption Underlying Regression

Some of the important assumptions for doing the regression analysis are as follows:

i) Independence among the pairs of score.

   This assumption implies that the scores of any two observations (subjects in case of most of psychological data) are not influenced by each other. Each pair of observation is independent. This is assured when different subjects provides different pairs of observation.

ii) The variance of the error terms is constant for each value of X.

iii) The relationship between X and Y is linear.

iv) The error terms follow the normal distribution with a mean zero and variance one.

v) Independence of Error Terms. The error terms are independent. They are uncorrelated.

vi) The population of X and the population of Y follow normal distribution and the population pair of scores of X and Y has a normal bivariate distribution.

This assumption(vi) states that the population distribution of both the variables (X and Y) is normal. This also means that the pair of scores follows bivariate normal distribution. This assumption can be tested by using statistical tests for normality.

4.2.9 Interpretation of Regression

The linear regression analysis provides us with lot of information about the data. This information need to be carefully interpreted. The intercept (á), the slope (â), the r^2, the *F*-value, need to be interpreted. Let us take these one by one .

The Intercept (á)

The intercept of regression line is the point at which regression line passes through the y-axis. This point is called as *a* in the sample and á in the population.

One straightforward interpretation of the *a* is it is a regression constant. It is that value, which we need to add into *b*X in order to get the predicted value of Y.

The other way of understanding the intercept is, intercept of regression line is that value of Y when the X value is zero. This interpretation looks intuitive. The correctness of this interpretation depends on whether we have sufficient X values near zero. In our example, the X was Stigma scores. The lowest value of the stigma scores was 49. Obviously we do not have any scores of X that are near zero. So the interpretation is unwarranted. This is due to two reasons.

i) We have not taken the complete range of the X values since we are studying the group of patients.

ii) The real zero X value is almost near impossible and the value of intercept 6.144, which is a Y-value when X is zero, is defiantly not possible.

iii) Nobody would miss -6 appointments. The best is not a single appointment is missed. So this interpretation is not applicable.

Slope (â):

The slope parameter is most important part of regression. The slope is called as regression coefficient. This also has straightforward interpretation.

The slope is that change in the Y when X changes by one unit.

So *rate of change* interpretation is common interpretation of the slope.

In our example, the slope value is 0.1991. This means that if the score on Stigma scale increases by one unit, the number of appointments missed will change by a value of .20.

This would also mean that as the score increases by 5 scale points on the Stigma scale, the person is likely to miss one appointment.

The r^2.:

The r^2 is the value that gives us the percentage of the variance X explains in Y.

The value is .299 in our example.

This means that roughly 30 percent variance in Y can be explained by X.

This can also be understood as proportional reduction in error.

Since we obtain r^2 by dividing the $SS_{Total} - SS_{Error}$ by the $SS_{Total.}$

This tells us about how much of the error is reduced.

The F ratio

The *F*-statistics is computed to test a null hypothesis â = 0.

If the *F*-value is statistically significant then the null hypothesis â = 0 is rejected.

Otherwise one has to accept the null hypothesis. If the null hypothesis is accepted, then there is no need to do the rest of the statistics.

This clearly means that X cannot linearly predict the Y.

However, in our example, the sample size is too small.

So the power of the statistical test is also small.

4.3 STANDARDISED REGRESSION ANALYSIS AND STANDARDISED COEFFICIENTS

We have learned about doing the regression analysis with X and Y. In previous chapters we have also learned to calculate the Z-scores. The Z is a standard score of a variable. To remind you, the Z can be calculated for each of the variable with the formula given below:

$$Z = \frac{X - \bar{X}}{S}$$

The mean of the Z is zero and the standard deviation is one.

Now, instead of predicting Y from X, we calculate the Z scores for both X and Y.

They will be denoted as Z_X and Z_Y.

Now we carry out the regression on standard variables than on unstandardised variables.

The regression equation will be

$$Z_Y = a + bZ_X + e \qquad \text{(eq. 4.30)}$$

Now, the intercept term is completely redundant in this equation because when we take the standard variable (that is Z) then the Y-intercept of the regression line is by default becomes *zero*. so the equation reduces to

$$Z_Y = bZ_X + e \qquad \text{(eq. 4.31)}$$

The beta value obtained in this regression equation is quite interesting.

Let us recall the correlation coefficient.

The correlation coefficient $r = Cov_{XY} / S_X S_Y$.

Now, with both the variable being standardised, the S_X, S_Y and S_X^2, will all be equal to one.

The slope for regression is calculated as $b = Cov_{XY} / S_X^2$.

Now you will realise that the slope (b) is equal to the r, correlation coefficient.

4.4 MULTIPLE REGRESSION

When we have multiple predictors than a single predictor variable, the regression carried out is called as multiple regression.

So we have a dependent variable and a set of independent variables. Suppose we have X_1, X_2, X_3, … up to X_k as k independent variables, and Y as a dependent variable, then the regression equation for sample can be written as:

$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_k K_k \qquad \text{(eq. 4.32)}$$

The same equation for the population can be written as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k K_k \qquad \text{(eq. 4.33)}$$

Look at the following data. The data is about three variables, number of appointments missed, stigma scores, and the distance between the hospital and home.

Generally, one would expect that if the stigma is high, then the appointments would be missed. Similarly if the hospital is far away, then the appointments may be missed.

Table 4.7: Table of the data for appointments missed, stigma scores and distance of the hospital from home for 10 patients

Appointments Missed (Y)	Stigma Scores (X_1)	Distance of the hospital (X_2)
2	40	2
3	43	5
4	45	4
5	46	7
6	60	9
7	63	5
8	69	2
9	54	8
11	70	6
11	62	9

The equation for which we carry out the regression analysis is as follows:

$$\text{Appointmnets Missed } (Y) = a + b_1 \text{ Stigma Score} + b_2 \text{ Distance from Home} + e$$

We will solve the numerical for this problem. I shall directly provide you with the answer.

The Multiple R^2 for this problem is 0.81. which means that 81 percent information in appointments missed is explained by these two variables.

The adjusted value for the same is .76.

The value of intercept is –7.88.

The slope for stigma is 0.22 and

The slope for distance is 0.40.

The results of significance testing are as follows:

Table 4.8: Table showing the significance testing and the ANOVA summary

Source	Sum of Squares	*Df*	Mean Square	F	Sig.
Regression	73.279	2	36.640	14.981	.003
Residual	17.121	7	2.446		
Total	90.400	9			

The obtained F-value tells us that the overall model we have tested for is turning out to be significant. We can actually test the significance of each of the b separately. When that is done, the b of stigma turned out to be significant ($t = 4.61$, $p < .01$) but the distance did not ($t = 1.93$, $p > .05$).

Here too the size of the sample appears to be the problem leading to non significant results.

The multiple regression equation can be solved hierarchically or directly.

When the equation is solved directly, all the predictors are entered into the equation simultaneously.

When the equation is solved hierarchically, then the predictors are entered one after another depending on the theory or simply depending on their statistical ability to predict the Y.

The multiple regression is very useful technique in psychological research.

4.5 LET US SUM UP

Now we know how to solve the problem of prediction in psychological research. We can develop suitable regression equation and test it against the data. We can test the predictability, amount of information in dependent explained by independent, etc. this technique is very informative.

When we do regression, it does not mean that the causality appears in the equation. It is not a function of statistics. It has to come from theory.

We now know how to set up a multiple regression equation. Though we do not know how to do the calculations, we can understand the results of multiple regression.

4.6 UNIT END QUESTIONS

Given below are some problems with Answers

1) A researcher was interested in predicting marks obtained in the first year of the college form the marks obtained in the high school. He collected data of 15 individuals which is given below. Find out the Independent Variable and Dependent Variable.

   Write regression equation, calculate *a* and *b*, plot the scatter and straight line, write null and alternative hypothesis, determine significance, and comment on the accuracy of the prediction.

School marks	College marks
67	65
45	50
65	60
60	71
55	54
53	49
59	58
64	69
67	75
69	73
70	64
58	66
63	62
71	65
74	78

2) A researcher was interested in predicting general satisfaction of people form perceived social support. She collected data of 10 individuals which is given below. Find out the IV and DV, Write regression equation, calculate *a* and *b*, plot the scatter and straight line, write null and alternative hypothesis, determine significance, and comment on the accuracy of the prediction.

Satisfaction with Life	Perceived Social Support
7	7
6	6
5	6
8	3
9	6
7	4
6	4
3	2
11	9
8	5

3) A researcher was interested in predicting stage performance form social anxiety. She collected data of 10 individuals which is given below. Find out the IV and DV, Write regression equation, calculate *a* and *b*, plot the scatter and straight line, write null and alternative hypothesis, determine significance, and comment on the accuracy of the prediction.

Stage Performance	Social Anxiety
9	11
7	9
6	11
10	7
10	11
9	9
9	8
5	7
14	13
10	9

4) A researcher was interested in predicting attitude to working condition form affective commitment to job. She collected data of 12 individuals which is given below. Find out the IV and DV, Write regression equation, calculate *a* and *b*, plot the scatter and straight line, write null and alternative hypothesis, determine significance, and comment on the accuracy of the prediction.

Attitude to Work	Affective Commitment
5	10
7	13
4	8
5	9
7	14
9	16
3	10
2	6
8	16
7	13
6	9
9	8

Answers:

1) $r = .78$, $r^2 = .608$, $a = 9.27$, $b = .87$, $SS_{Regression} = 641.75$, $SS_{Residual} = 413.19$, $F = 20.19$.

2) $r = .64$, $r^2 = .41$, $a = 3.41$, $b = .69$, $SS_{Regression} = 17.98$, $SS_{Residual} = 26.02$, $F = 5.53$.

3) $r = .51$, $r^2 = .26$, $a = 2.7$, $b = .65$, $SS_{Regression} = 14.67$, $SS_{Residual} = 42.22$, $F = 2.78$.

4) $r = .67$, $r^2 = .45$, $a = .958$, $b = .458$, $SS_{Regression} = 25.21$, $SS_{Residual} = 30.79$, $F = 8.19$.

4.7 SUGGESTED READINGS

Garrett, H.E. (19). *Statistics In Psychology And Education*. Goyal Publishing House, New Delhi.

Guilford, J.P.(1956). *Fundamental Statistics in Psychology and Education.* McGraw Hill Book company Inc. New York.