Regression Analysis

11

Learning Objectives

After reading this chapter, you would be able to understand:

- The basic concepts of regression analysis.
- Estimation of regression coefficients and testing the significance of these coefficients.
- Estimation of coefficient of determination and testing its significance.
- How to perform multiple regression analysis.
- Heteroscedasticity, multicollinearity and autocorrelation.

Introduction to Regression Analysis

In business decision-making, it becomes crucial to understand and examine the impact of one or more variables on a particular variable. For example, we may be interested to study the impact of advertising expenditure and R&D expenditure on sales. Hence, it is important for the firms to understand the nature of relationship between the variables. In such a situation, regression analysis is used. Regression is the most widely used technique among all statistical tools to study the cause and effect relationship among variables. The genesis of the concept of regression is attributed to Sir Francis Galton (1885), who, while analyzing the heights of parents and sons, found that if the parents were very tall, their children tended to be tall, but shorter than the parents as their height was moderating toward the mean. This phenomenon is termed as regressing or stepping back to the mean. Since then, this approach has widely been used not only in economics and business but also in physical, natural, and other social sciences.

Primarily, regression analysis examines the functional relationship among the variables. The key objective of the regression analysis is to estimate the response values of a dependent variable on the basis of the values of the independent variable(s). That is why, the dependent variable is also called the response variable, whereas the independent variables are called explanatory variables or predictors. In simple regression analysis, only one independent variable is considered to estimate the relationship between the dependent and the independent variable. For example, if we take sales as the dependent variable and advertising expenditure as the independent variable, this becomes a case for simple regression analysis. Multiple regression analysis involves more than one independent variable. In this case, sales can be taken as dependent variable and advertising expenditure and R&D expenditure can be considered as the explanatory variables. Hence, we can say that regression technique aims at examining the functional relationship among the variables and then estimating the value of the dependent variable on the basis of the values of the independent variable(s). This chapter presents the discussion on linear regression.

The Simple Regression Model

In this model, the values of the dependent variable (Y) are linearly related to the values of the explanatory variable (X). The equation that represents the population simple regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

where Y_i is the value of the dependent variable. β_0 and β_1 are the parameters of the population regression equation. These parameters indicate the intercept and the slope of the regression equation, respectively. X_i represents the value of the independent variable and U_i indicates error term in the regression line.

The parameters β_0 and β_1 in the population regression model are estimated by the sample regression coefficients b_0 and b_1 as given in the following sample regression equation:

$$Y_i = b_0 + b_1 X_i + e_i$$

On the other hand, the estimated linear regression equation based on sample data is given by:

$$\hat{Y}_i = b_0 + b_1 X_i$$

It is important to note here that in the sample regression model, Y_i are the observed values of the dependent variable. $\hat{Y}$, on the other hand, indicates the estimated values of the dependent variable. In the case of simple regression model, it is good to plot the two variables to examine whether they have linear relationship. Figure 11.1 provides the graphical presentation of Y_i and X_i.

$$\hat{Y}_i = b_0 + b_1 X_i$$

In Figure 11.1, the straight line indicates the estimated values of sales (Y_i) for different values of advertising expenditure (X_i), whereas the individual scattered values are the observed values of the sales (Y_i) at different values of advertising expenditure (X_i). The straight line shown in the graph is the line of best fit which passes through all observed data points in such a way that the sum of all errors or residuals (difference between the observed and estimated value of Y) is equal to zero and the sum of squared errors is minimized. The line of best fit is estimated with the help of ordinary least squares (OLS) approach.

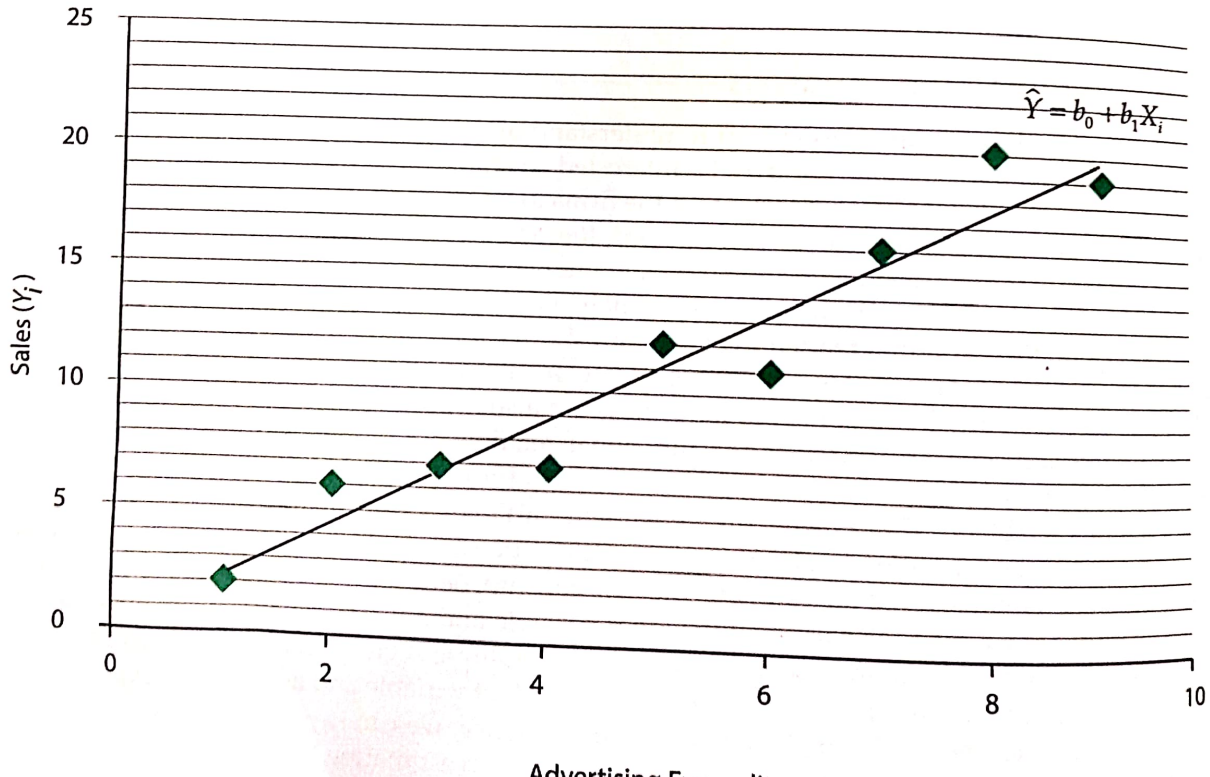


Figure 11.1: Scatter Plot of Data on Two Variables.

Ordinary Least Squares Approach

OLS approach is used to estimate the linear regression equation:

$$Y = b_0 + b_1 X_i + e_i \quad \text{Sample linear regression equation}$$

$$\hat{Y} = b_0 + b_1 X_i \quad \text{Estimated equation}$$

Using this approach, the regression coefficients are estimated in such a manner that ensures the line of best fit. In other words, the following two conditions are met:

(i) $\sum e_i = 0$, where $e_i = Y - \hat{Y}$. Hence, $\sum e_i = \sum (Y_i - b_0 - b_1 X) = 0$

(ii) $\sum e_i^2 = \sum (Y - \hat{Y})^2$ is minimum, where $\sum e_i^2 = \sum (Y_i - b_0 - b_i X_i)^2$

In the second condition, the sum of the squared deviations between the estimated and actual values of the dependent variable is minimized. This is known as least square criterion.

The principle of least squares ensures the selection of that line for which the sum of squares of differences between the observed values and the estimated values is minimum. Using maxima–minima approach, b_0 and b_1 are derived in such a manner that $\sum e_i^2$ is minimum. The values of b_0 and b_1 are determined by solving the following equations:

(i) $\sum e_{i=0}$

(ii) $\dfrac{d\left(\sum e_i^2\right)}{db_1} = 0$ (putting first-order derivative of $\sum e_i^2$ equal to zero)

Without getting into details of differential calculus, the following two equations, known as the normal equations, can be stated as:

$$\sum Y_i = \sum b_0 + \sum b_1 X_i \quad \text{or} \quad \sum Y_i = n b_0 + b_1 \sum X_i$$
$$\sum Y_i X_i = \sum b_0 X_i + \sum b_1 X_i^2 \quad \text{or} \quad \sum Y_i X_i = b_0 \sum X_i + b_1 \sum X_i^2$$

Further, from these equations, the formulae for b_0 and b_1 are derived that satisfy the least squares criterion, as follows:

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$b_0 = \bar{Y} - b\bar{X}_i$$

Using the above formulae, the regression coefficients are estimated and the following regression equation is formulated.

$$\hat{Y} = b_0 + b_1 X_i$$

This equation can be used to estimate the value of the dependent variable for every given value of explanatory variable. It is important to note that regression can be used for predicting the values of the dependent variable. For example, if we want to see the likely impact of the advertising expenditure on sales in future, it can be studied using regression analysis.

Assumptions of Regression Analysis

While conducting the regression analysis, we make certain assumptions about the error term in the regression model. Violation of these assumptions makes the regression coefficients unreliable, and hence the conclusions from regression analysis are not useful.

1. The error term μ_i is normally distributed with mean zero and standard deviation σ, at each value of X_i:

$$\mu_i \sim N(0, \sigma^2),\ i = 1, 2, \ldots, n$$

2. The variance of μ_i, denoted by σ^2, is constant for all values of X_i. This phenomenon is known as homoscedasticity. The violation of this is known as heteroscedasticity.

3. The different values of μ_i are independent of each other. This assumption becomes much more important while working with time series data, as in time series data the error terms are more likely to be associated with each other. The violation of this assumption leads to the problem of autocorrelation in regression.
4. The different independent variables (in the case of multiple regression) are independent of each other. The violation of this assumption leads to the problem of multicollinearity in regression.
5. Every predicted value of Y comes from the normal distribution of Y for each value of variable X (as shown in Figure 11.2).

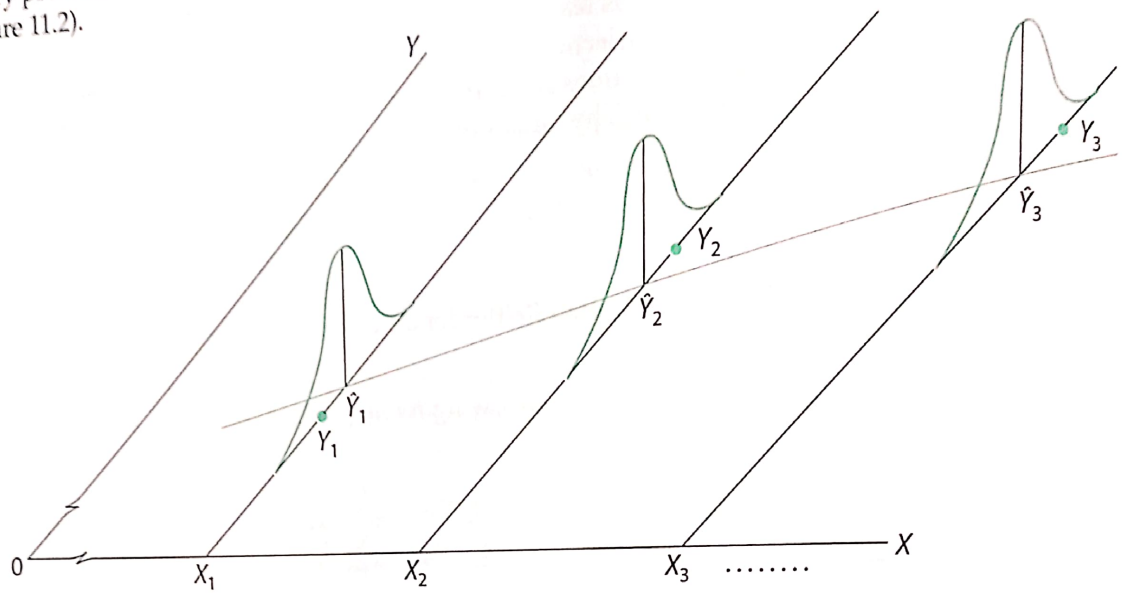


Figure 11.2: Simple Linear Regression of Y on X.

Case Study 1

Fashion Planet

Fashion Planet is a retail chain in the business of casual and formal wear. Using the pull strategy, the company has been investing heavily in advertising to attract customers. The management of this company is interested to know how the advertising expenditure has been affecting their sales in the past. For this, a research agency was hired, who collected data from the finance department of the company. The research agency collected information on the sales ($1,000s) and advertising expenditure ($1,000s) for the last 30 months. Data are given in Table 11.1.

Table 11.1: Monthly Sales and Advertising Expenditure of Fashion Planet ($1,000s).

Month	Sales (Y)	Advertising Expenditure (X)	Month	Sales (Y)	Advertising Expenditure (X)
1	49	14	16	52	15
2	40	9	17	54	15
3	50	15	18	58	16
4	46	13	19	56	15
5	44	12	20	60	18
6	52	15	21	49	14
7	54	15	22	40	9
8	58	16	23	50	15
9	56	15	24	46	13

(Continued)

Table 11.1: (Continued).

Month	Sales (Y)	Advertising Expenditure (X)	Month	Sales (Y)	Advertising Expenditure (X)
10	60	18	25	44	10
11	49	14	26	52	15
12	40	9	27	54	15
13	50	15	28	58	16
14	46	13	29	56	13
15	44	12	30	63	16

The research agency was also entrusted with the task of developing a model for forecasting the future sales for any particular level of advertising budget. In order to estimate the nature of relationship between the two variables, following calculations were done (Table 11.2):

Table 11.2: Calculations for OLS Estimation.

Month	Sales (Y)	Advertising Expenditure (X)	XY	X^2
1	49	14	686	196
2	40	9	360	81
3	50	15	750	225
4	46	13	598	169
5	44	12	528	144
6	52	15	780	225
7	54	15	810	225
8	58	16	928	256
9	56	15	840	225
10	60	18	1,080	324
11	49	14	686	196
12	40	9	360	81
13	50	15	750	225
14	46	13	598	169
15	44	12	528	144
16	52	15	780	225
17	54	15	810	225
18	58	16	928	256
19	56	15	840	225
20	60	18	1,080	324
21	49	14	686	136
22	40	9	360	81

(Continued)

Table 11.2: (Continued).

Month	Sales (Y)	Advertising Expenditure (X)	XY	X^2
23	50	15	750	225
24	46	13	598	169
25	44	10	440	100
26	52	15	780	225
27	54	15	810	225
28	58	16	928	256
29	56	13	728	169
30	63	16	1,008	256
Total	**1,530**	**420**	**21,808**	**6,042**

From the above calculations, we get:

$$n = 30,$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{1,530}{30} = 51$$

$$\bar{X} = \frac{\sum X}{n} = \frac{420}{30} = 14$$

$$\sum X^2 = 6,042$$

$$\sum XY = 21,808$$

Now the regression coefficients would be calculated as follows:

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$b_1 = \frac{30(21,808) - 420(1,530)}{30(6,042) - (420)^2} = 2.395$$

$$b_0 = \bar{Y} - b_1 \bar{X}_1$$

$$b_0 = 51 - 2.395(14) = 17.47$$

The estimated model is:

$$Y = 17.47 + 2.4X$$

From the estimated relationship, it can be said that \$1,000 expenditure in advertising is expected to bring increase in sales by \$2,400. This model can also be used for forecasting. For example, if the company decides to invest \$5,000 in the advertising in the next month, the total sales expected can be calculated as:

$$\text{Sales} = 17.47 + 2.4(5) = 29.4$$

The sales are likely to be to the tune of \$29,400.

Testing the Significance of Regression Coefficients

After estimating the model, the next step is to test the significance of the regression coefficient. Testing the significance of b_1 becomes much more important as it determines whether X is a significant explanatory variable of Y. Whether the two variables have cause and effect relationship in the population of observations is an important question to be answered. This exercise is done with the help of hypothesis testing, where a conclusion is drawn about the population parameter on the basis of sample evidence. The significance of regression coefficient is tested using t-test in hypothesis testing. In order to perform this test, apart from the value of the regression coefficient (b_1), we also need the standard error (SE) of the estimate (b_1). The concept of standard error is similar to the concept of standard deviation. In standard deviation, we consider the difference between the different values of a variable and their mean. However, standard error is a measure of difference between all possible values of the statistics (b_1) arising from different samples and the population parameter (β_1). This measure is based on the sampling distribution of the statistics. Therefore, we can say that if we take all

possible samples (say k) of Y and X from the population and compute b_1 for each sample, then we will have k number of b_1 values. This distribution of b_1 will have mean as β_1 (population mean). The standard deviation of a statistic is termed as standard error. The standard error of b_1 is calculated as follows:

$$SE(b_1) = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}} \Big/ \sum_{i=1}^{n}(X_i - \bar{X})^2$$

Let us calculate the standard error of b_1 in the case of our problem on Fashion Planet (Table 11.3).

Table 11.3: Calculations for Testing Significance of b_1.

Sales (Y)	Advertising Expenditure (X)	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$X-\hat{X}$	$(X-\hat{X})^2$
49	14	51.07	–2.07	4.28	0	0
40	9	39.07	0.98	0.86	–5	25
50	15	53.47	–3.47	12.04	1	1
46	13	48.67	–2.67	7.13	–1	1
44	12	46.27	–2.27	5.15	–2	4
52	15	53.47	–1.47	2.16	1	1
54	15	53.47	0.53	0.28	1	1
58	16	55.87	2.13	4.54	2	4
56	15	53.47	2.53	6.4	1	1
60	18	60.67	–0.67	0.45	4	16
49	14	51.07	–2.07	4.28	0	0
40	9	39.07	0.93	0.86	–5	25
50	15	53.47	–3.47	12.04	1	1
46	13	48.67	–2.67	7013	–41	1
44	12	46.27	–2.27	5.15	–2	4
52	15	53.47	–1.47	2.16	1	1
54	15	55.87	0.53	0.28	1	1
58	16	53.47	2.13	4.54	2	4
56	15	60.67	2.53	6.4	1	1
60	18	51.07	–0.67	0.45	4	16
49	14	39.07	–2.07	4.28	0	0
40	9	53.47	0.93	0.86	–5	25
50	15	48.67	–3.47	12.04	1	1
46	13	41.47	–2.67	7.13	–1	1
44	10	53.47	2.53	6.4	–4	16
52	15	53.47	–1.47	2.16	1	1
54	15	55.87	0.53	0.28	1	1
58	16		2.13	4.54	2	4

(Continued)

Table 11.3: (Continued).

Sales (Y)	Advertising Expenditure (X)	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$X-\hat{X}$	$(X-\hat{X})^2$
56	13	48.67	7.33	53.73	–1	1
63	16	55.87	7.13	50.84	2	4
1,530	420		228.87	228.87		

$$SE(b_1)=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y})^2}{n-2}\Big/\sum_{i=1}^{n}(X_i-\bar{X})^2}=\sqrt{\frac{\frac{228.87}{30-2}}{162}}=\sqrt{\frac{228.87}{162\times 28}}=0.225$$

Now, in order to test the significance of b_1, the steps that would be followed are explained in Figure 11.3.

Now, in the case of Fashion Planet study, the significance of b_1 is tested as follows:

Null hypothesis: $\beta_1 = 0$

The process of hypothesis testing is for this is divided into five steps.

Null Hypothesis	$\beta_1 = 0$	Null hypothesis states that there is no relationship between X and Y and hence regression coefficient is not significant.
Alternate Hypothesis	$\beta_1 \neq 0$	Alternate hypothesis states that there is relationship between X and Y and hence regression coefficient is significant.
Level of Significance	$\alpha = 0.05$	Level of confidence is assumed as 95 percent.

Test-Statistic

Under the validity of null hypothesis, the test-statistic would be:

$$t=\frac{b_1-E(b_1)}{SE(b_1)}$$

$$t=\frac{b_1-\beta_1}{SE(b_1)}$$

Where t-statistic follows t-distribution with $n-2$ degrees of freedom.

Conclusion

If $|t| > t_{\alpha,d.f}$ where d.f $= n-2$, i.e. if the calculated value of t is more than the tabulated value of t at 5% level of significance with $n-2$ degrees of freedom, then null hypothesis is rejected and vice versa. If null hypothesis is rejected, this infers that the regression coefficient is significant and hence X is a significant explanatory variable of Y. If null hypothesis is accepted, this leads to the conclusion that the regression coefficient is not significant; hence there is no significant linear relationship between the variables. The decision of acceptance or rejection of null hypothesis can also be taken with calculated level of significance, which is also known as p-value. If calculated p-value, which is the calculated level of significance is less than the theoretical or assumed level of significance i.e. α, then the null hypothesis is rejected and vice versa. The comparison of p-value and α is provided while discussing the SPSS results on regression later in this chapter.

Figure 11.3: Significance Testing Procedure for Regression Coefficient.

Alternate hypothesis: $\beta_1 \neq 0$

Level of significance: $\alpha = 0.05$

Test-statistic: Under the validity of null hypothesis, the test-statistic would be:

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{2.4}{0.225} = 10.67$$

where $b_1 = 2.4, \beta_1 = 0$ (under the validity of H_0) and SE $(b_1) = 0.225$.

Conclusion

Now the tabulated value of t at 5% level of significance with 28 degrees of freedom is 2.048.

Hence, we found that the calculated value of t is more than the tabulated value of t; hence, null hypothesis is rejected. We conclude that advertising expenditure is a significant explanatory variable. Hence, we can say the following estimated model is a good fit:

$$Y = 17.47 + 2.4X$$

Coefficient of Determination (R^2)

Coefficient of determination measures the strength of the fitted OLS equation. It explains what proportion of the variations in the dependent variable is explained by the independent variable. If we found that β_1 is not zero, this infers that a linear relationship exists between X and Y. R^2 is a measure of strength of this relationship and is known as coefficient of determination.

The total variation of Y can be decomposed into two parts—explained variation and unexplained variation.

Total variation = explained variation + unexplained variation

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

$$SST = SSR + SSE$$

Explained variation is also called variation due to the regression and the unexplained variation is also known as variation due to error term. The coefficient of determination is calculated as follows:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Calculation of Coefficient of Determination in Fashion Planet Study

In our study, the calculation of R^2 requires first the calculation of SSR and SST. So, let us calculate these terms first (Table 11.4).

Table 11.4: Calculations of R^2.

Sales (Y)	Advertising Expenditure (X)	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(\hat{Y}_i - \bar{Y})$	$(\hat{Y}_i - \bar{Y})^2$
49	14	–2	4	0	0
40	9	–11	121	–11.98	143.4
50	15	–1	1	2.4	5.74
46	13	–5	25	–2.4	5.74
44	12	–7	49	–4.79	22.94
52	15	1	1	2.4	5.74
54	15	3	9	2.4	5.74
58	16	7	49	4.79	22.94
56	15	5	25	2.4	5.74
60	18	9	81	9.58	91.78
49	14	–2	4	0	0
40	9	–11	121	–11.98	143.4
50	15	–1	1	2.4	5.74
46	13	–5	25	–2.4	5.74
44	12	–7	49	–4.79	22.94
52	15	1	1	2.4	5.74
54	15	3	9	2.4	5.74
58	16	7	49	4.79	22.94
56	15	5	25	2.4	5.74
60	18	9	81	9.58	91.78
49	14	–2	4	0	0
40	9	–11	121	–11.98	143.4
50	15	–1	1	2.4	5.74
46	13	–5	25	–2.4	5.74
44	10	–7	49	–9.58	91.78
52	15	1	1	2.4	5.74
54	15	3	9	2.4	5.74
58	16	7	49	4.79	22.94
56	13	5	25	–2.4	5.74
63	16	12	144	4.79	22.94
1,530	**420**		**1,158**		**929.24**

$$SST = \text{total variation} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
$$= 1,158$$

$$SSR = \text{explained variation} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$
$$= 929.24$$

Hence,

$$R^2 = \frac{929.24}{1,158} = 0.802$$

Therefore, we can say that around 80% variations of sales (Y) are explained or captured by the advertising expenses (X). In other words, we can say that 80% variations in Y can be attributed to X.

Testing the Significance of Coefficient of Determination (R^2)

After estimating the regression equation, the significance of the regression coefficient was tested with the help of hypothesis testing (Fig. 11.4). Similarly, once the coefficient of determination is estimated, its significance is to be tested to ascertain whether the estimated line is a good fit. It is important to note that when we test the significance of an individual regression coefficient of an independent variable, the objective is to determine whether this independent variable is a significant explanatory variable. And when we test the significance of R^2, the purpose is to test the significance of the overall model. Whether the estimated equation is a good fit or not is judged by checking the significance of the coefficient of determination, as this measure determines the strength of the estimated model.

The significance of R^2 is tested with the help of analysis of variance (ANOVA), a tool that involves the use of F-test. Before we do ANOVA calculations, let us set up the hypothesis testing framework for it.

The following five-step process would be followed:

Null Hypothesis	$R^2 = 0$	Coefficient of determination is not significant. Hence, the estimated equation is not a good fit.
Alternate Hypothesis	$R^2 \neq 0$	Coefficient of Determination is significant.
Level of Significance	$\alpha = 0.05$	Level of confidence is assumed as 95 percent.

Test Statistic

Under the validity of null hypothesis, the test statistic would be:

$$F = \frac{SSR/(k-1)}{SSE/(n-k)}$$

where F statistic follows F-distribution with v_1 and v_2 degrees of freedom, where $v_1 = (k-1)$ and $v_2 = (n-k)$.

Here k represents the total number of variables in the model and n represents the number of observations or the sample size. In simple regression analysis, $k = 2$ (one dependent and one independent variable).

Conclusion

If $F > F_\alpha$ at v_1 and v_2 degrees of freedom, i.e. if the calculated value of F is more than the tabulated value of F at 5% level of significance with v_1 and v_2 degrees of freedom, then null hypothesis can be rejected and vice versa. If null hypothesis is rejected, this infers that the coefficient of determination is significant and the estimated regression equation is also significant, and hence the estimated line is the best fit line. If null hypothesis is not rejected, this leads to the conclusion that the regression equation is not significant. More discussion on this is presented in multiple regression analysis, wherein we have more than one explanatory variable in the regression model.

Figure 11.4: Hypothesis Testing Process for Coefficient of Determination.

In order to calculate the value of *F*-statistic, we have to construct the ANOVA table as given in Fig. 11.5:

ANOVA Table

Source of Variation	Sum of Squares	*DF*	Mean Square	*F*
Regression	*SSR*	$k-1$	$MRS=\frac{SSR}{k-1}$	$F=\frac{MSR}{MSE}$
Error	*SSE*	$n-k$	$MSE=\frac{SSE}{n-k}$	
Total	*SST*	$n-1$		

if $F < F_{k-1,\ n-k}$, Null hypothesis accepted otherwise rejected

Figure 11.5: Calculations for Analysis of Variance (ANOVA).

In the above table, SSR represents variation due to regression, SSE represents variation due to error term, and SST is the total variation. The degrees of freedom for regression and error term are $(k-1)$ and $(n-k)$, respectively. The degrees of freedom for total model are $(n-1)$. Mean square error is calculated by dividing the sum of square with its respective degree of freedom. Finally, *F*-value is calculated by dividing the MSR by MSE. This *F*-value is compared with the tabulated or critical *F*-value, and conclusion is drawn about the significance of the coefficient of determination.

Now, let us do these calculations for our study on Fashion Planet.

We have already calculated earlier SSR and SST as given below:

$$\text{SST} = 1{,}158, \quad \text{SSR} = 929.24$$

Now,

$$\text{SST} = \text{SSR} + \text{SSE}$$

Hence,

$$\text{SSE} = \text{SST} - \text{SSR}$$

$$\text{SSE} = 1{,}158 - 929.24 = 228.76$$

Now, the mean square measures are calculated as follows:

$$\text{MSR} = \frac{929.24}{(2-1)} = 929.24$$

$$\text{MSE} = \frac{228.76}{(30-2)} = 8.171$$

Now,

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{929.24}{8.17} = 113.73$$

Now, from the statistical table on *F*-test, we can find that *F*-value (at 5% level of significance with 1 and 28 degrees of freedom) is 4.19. Since the calculated value of *F*-statistics is more than the tabulated value, we can conclude that the null hypothesis is rejected and the estimated regression line is significant.

Adjusted R^2

Adjusted R^2 measure is a better version of R^2 (coefficient of determination). This is calculated by using following formula.

$$\bar{R}^2 = 1-(1-R^2)\frac{(n-1)}{(n-k)}$$

where $\bar{R}^2$, is the adjusted coefficient of determination. The total number of variables in the model is indicated by k, and n is the size of the sample. This concept is more useful in multiple regression, where we have more than one independent variables. The adjusted R^2 is an improved measure over coefficient of determination.

Use of Excel for Simple Linear in Regression

Let us consider the data from Fashion Planet case study for using Excel for regression analysis. Following steps are required to perform regression analysis with Excel.

Step 1: Go to **Data** menu in Excel and click on **Data Analysis.**
Step 2: In **Data Analysis** window, click on **Regression** (Figure 11.6).
Step 3: Enter data for **Input Y Range** and **Input X Range.** Tick on **Labels** (Figure 11.7)
Step 4: Click **Ok.**

The Excel Output is given in Table 11.5

In the Excel output, since the calculated level of significance for ANOVA and regression coefficient is below 5%, we reject the null hypotheses for R^2 as well as the regression coefficient. The estimated regression line represents significant relationship between sales as a dependent variable and advertising expenditure as a driver of sales. Our results corroborate with manual calculations.

Month	Sales (Y)	Advertising expenditure (X)
1	49	14
2	40	9
3	50	15
4	46	13
5	44	12
6	52	15
7	54	15
8	58	16
9	56	15
10	60	18

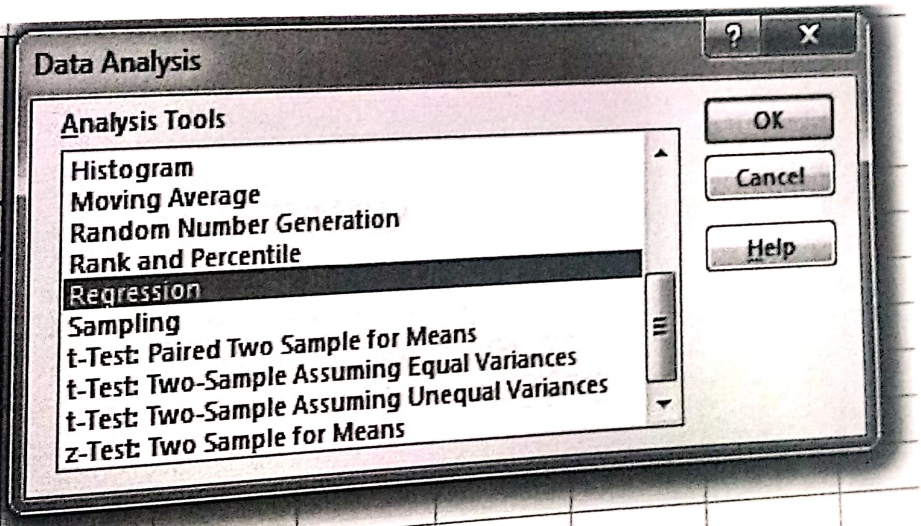


Figure 11.6: Use of Excel for Simple Linear Regression.

Month	Sales (Y)	Advertising expenditure (X)
1	49	14
2	40	9
3	50	15
4	46	13
5	44	12
6	52	15
7	54	15
8	58	16
9	56	15
10	60	18
11	49	14
12	40	9
13	50	15
14	46	13
15	44	12

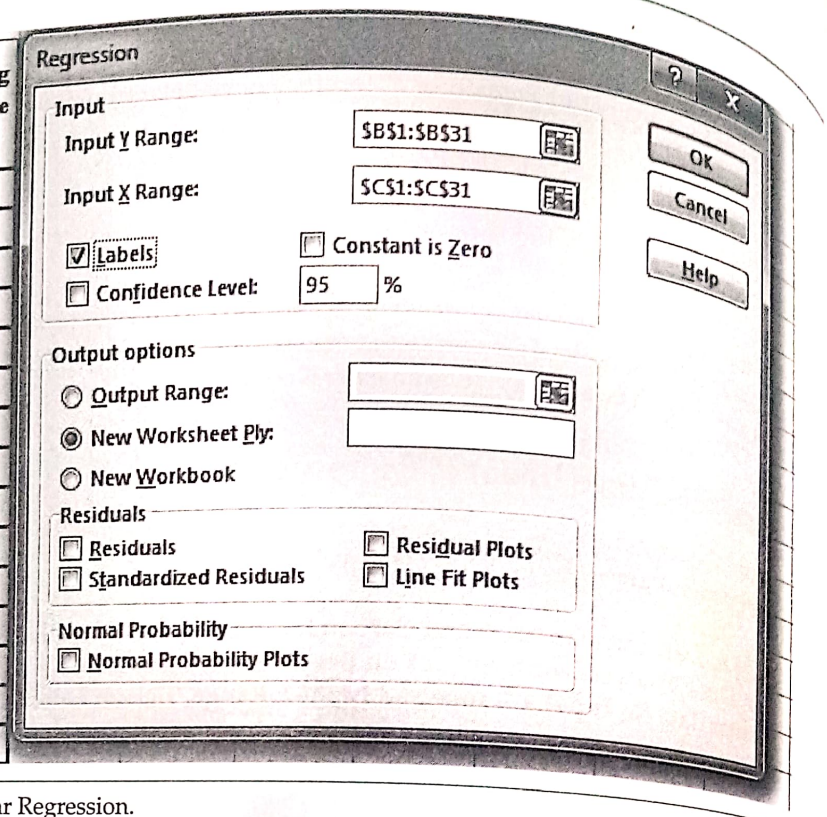


Figure 11.7: Use of Excel for Simple Linear Regression.

Table 11.5: Excel Output for Simple Linear Regression.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.895818318							
R Square	0.802490458							
Adjusted R Square	0.795436546							
Standard Error	2.858046594							
Observations	30							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	929.2839506	929.2839506	113.7653028	2.2751E-11			
Residual	28	228.7160494	8.168430335					
Total	29	1158						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.4691358	3.18670257	5.481884619	7.44218E-06	10.9414715	23.99680011	10.9414715	23.99680011
Ad. Exp (X)	2.395061728	0.224549348	10.66608189	2.2751E-11	1.935093241	2.855030216	1.935093241	2.855030216

Use of SPSS in Regression

In SPSS Data Editor window, two worksheets—**Data View** and **Variable View**—appear. First, the variables have to be created in the **Variable View** and then data can be entered in the **Data View,** as shown in Figs. 11.8 and 11.9.

In our study on Fashion Planet, two variables—Sales and AdEx (for Advertising Expenditure)—are created in the Variable View, as shown in Figure 11.10. Further, the data are entered in the Data View as presented in Figure 11.11.

It is important to note that in Figure 11.10, the variables can be given a short name in the Name column in the Variable View, whereas in the Label column the variables can be given a detailed description.

Table 11.6: Regression Output for

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.896[a]	.802	.795	2.85805

[a]Predictors: (Constant), Advertising Expenditure

ANOVA[a]

Model		Sum of Squares	DF	Mean Square	F	Sig.
1	Regression	929.284	1	929.284	113.765	.000[b]
	Residual	228.716	28	8.168		
	Total	1158.000	29			

[a]Dependent Variable: Sales
[b]Predictors: (Constant), Advertising Expenditure

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	17.469	3.187		5.482	.000
	Advertising Expenditure	2.395	.225	.896	10.666	.000

[a]Dependent Variable: Sales

Multiple Regression Analysis

Multiple regression analysis is a study of relationship between the variables when the number of independent variables is more than one. In this analysis, the values of the dependent variable (Y) are related to the values of a set of the explanatory variables (X_1, X_2, X_3, …, X_k). The equation that represents the simple regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + U$$

where β_0 is the intercept parameter of the population regression equation and β_i, $i = 1, 2, \ldots, k$, are the regression coefficient parameters of regression equation. X_i represents the value of the independent variable and U indicates error in the regression line.

The corresponding sample regression equation is given by:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + e,$$

where b_0 is the intercept of the sample regression equation and b_i, $i = 1, 2, \ldots, k$, are the regression coefficients of the sample regression equations. These coefficients $b_1, b_2, \ldots, b_k$ are the estimates of $\beta_1, \beta_2, \ldots, \beta_k$, and e is the error term for sample regression model. Now, the estimated sample regression equation would be:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

The correspondence between population and sample regression model is shown in Figure 11.14.

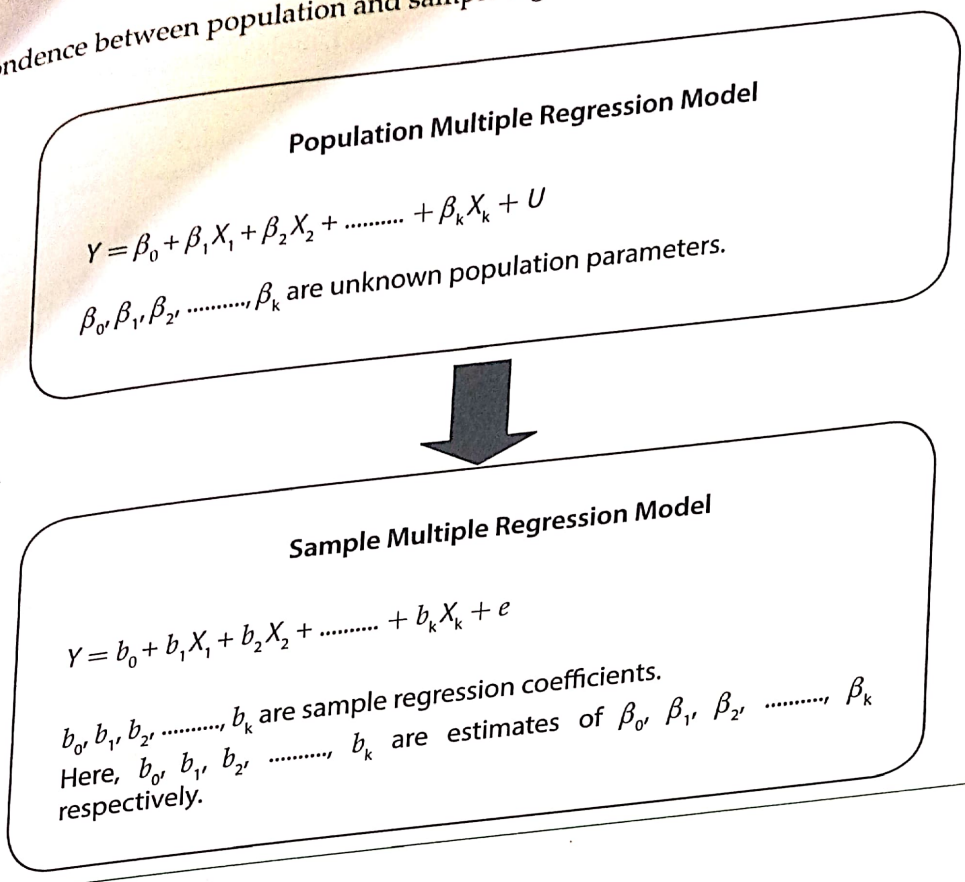


Figure 11.14: Multiple Regression Analysis.

Estimation of Multiple Regression Model

As in the case of simple regression model, OLS technique is used to estimate multiple regression models also. Using this approach, the various parameters ($\beta_0, \beta_1, \beta_2, \ldots, \beta_k$) of the population regression equation are estimated by sample regression coefficients, namely, $b_0, b_1, b_2, \ldots, b_k$.

According to OLS criterion, Σe_i^2 is minimized. Hence, in multiple regression analysis also, minimization of $\Sigma(Y_i - \hat{Y}_i)^2$ will lead to best fit estimators for multiple independent variables model.

The manual estimation process of a multiple regression model involves use of matrix algebra, which is beyond the scope of this book. Therefore, for multiple regression, we will use **Excel and SPSS** to get the regression output.

Using Excel for Multiple Regression

Let us study the case of Fashion Planet in a multivariate framework where we also have data on quality expenditure. Let us take quality expenditure (million Rs.) as another independent variable in our previous model of sales as dependent variable and advertising expenditure as the only independent variable. Hence, now, we have two independent variables in our multiple regression model.

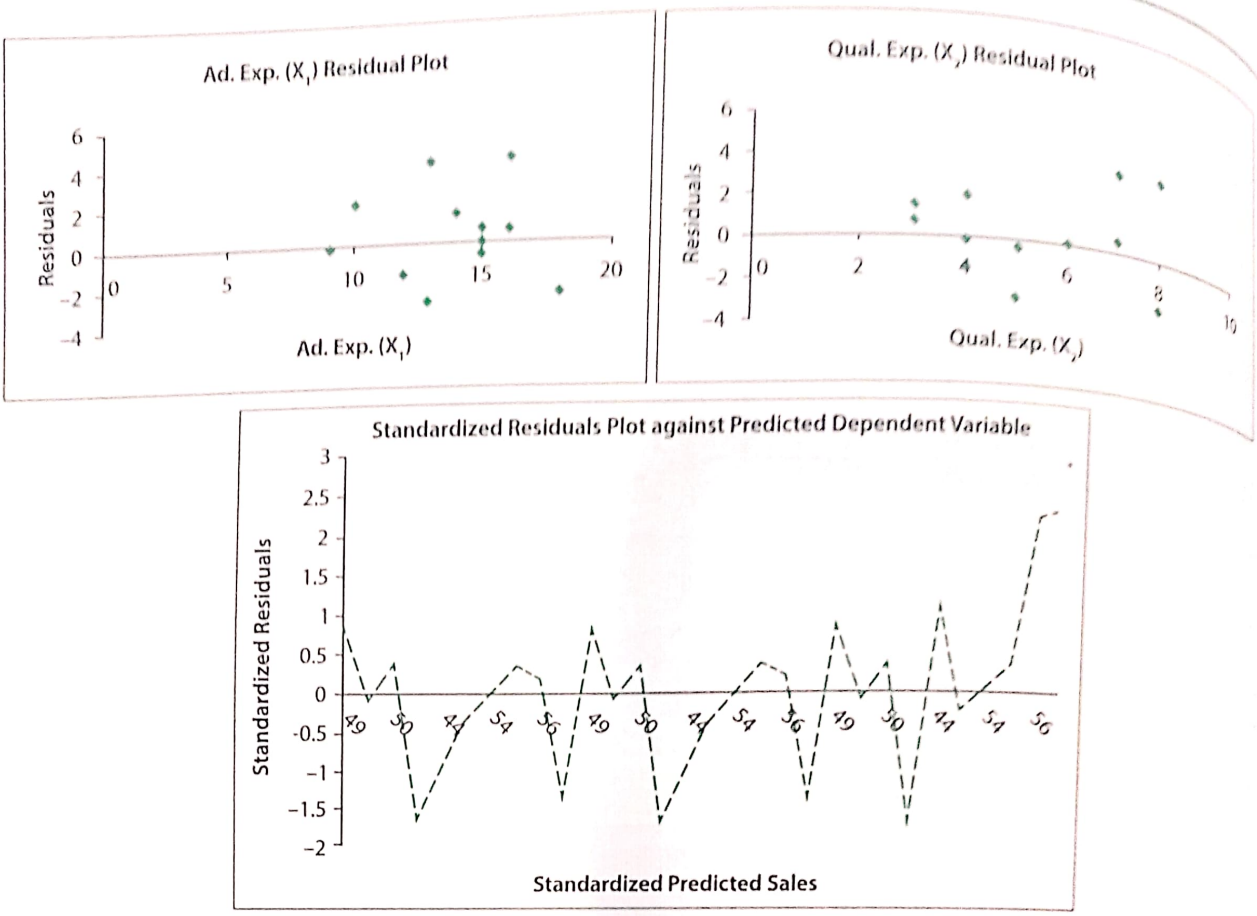


Figure 11.17: Residual Plots in Multiple Regression.

The Excel output for regression in Table 11.7 shows that the p-value of both the regression coefficients is below the level of significance (5%). It may be noted that p-value is the calculated level of significance. If it is less than theoretical level of significance, known as α, then null hypothesis of insignificant relationship between the explanatory variable and the dependent variable is rejected. In our case, rejection of hypotheses leads to the inference that both the advertising expenditure and quality expenditure are significant drivers of sales. Further, in our results, we also found that the plot of residuals against each explanatory variables is not showing a specific pattern. This is a desirable scenario indicating absence of heteroscedasticity. When we plotted the standardized residuals against standardized predicted sales (Figure 11.17), the absence of any trend or pattern in the standardized scores indicated presence of homoscedasticity.

Now, we will move towards SPSS. As an advanced software, it provides much deeper and wider insights in regression analysis.

Using SPSS for Multiple Regression

Considering the case study of Fashion Planet, data on sales are shown as dependent variable and those on advertising expenditure and quality expenditure are shown as two explanatory variables in the IBM SPSS Data Editor (Figure 11.18). The SPSS commands are shown in Figures 11.19 and 11.20. The SPSS output for multiple regression analysis is given in Table 11.8.

Regression output is shown in Figure 11.22. From the table of coefficients, we can see that both the regression coefficients are significant as the sig. value (calculated level of significance) for both the variables is 0%, which is less than the assumed level of significance (α as 5%). The betas are the standardized coefficients obtained by running regression using normalized variables. A variable is normalized by subtracting its mean from each of its values and dividing it by its standard deviation, so that the mean and variance of the normalized variable are 0 and 1, respectively. A standardized regression coefficient indicates how much dependent variable will change when there is one standard deviation change in the independent variable. The relative importance of independent variables in explaining the variation of the dependent variable is indicated by betas.

normal *P–P* plot for standardized residuals is closer to the straight line. Normal *P–P* plot checks the normality of a variable. The plotted points should be closer to the straight line. Considerable departure would suggest that normality assumption is not met. In our case, the assumption of normality of the error term is not violated.

The estimated regression equation can be written as:

$$\text{Sales} = 17.985 + 1.779\ (\text{Ad. Exp.}) + 1.56\ (\text{Qual. Exp.})$$

The results infer that 1 million rupees investment in advertising expenditure is likely to increase sales by 1.779 million rupees and 1 million rupees investment in quality expenditure is expected to increase sales by 1.56 million rupees.

Discussion on other assumptions such as multicollinearity and absence of autocorrelation is presented later in this chapter after discussion on stepwise regression.

Stepwise Regression Analysis

In stepwise regression, single variable at a time is added to or dropped from the model. It can be termed as forward stepwise regression or backward stepwise regression. In forward stepwise regression analysis, one explanatory variable is added to the model at a time in the process of selecting variables for the final model. In backward stepwise regression analysis, all the variables are included in the beginning, and then we remove insignificant variables by dropping only one insignificant variable at a time. Let us apply forward stepwise regression for a company when we have data on sales, advertising expenditure, R&D expenditure, sales person, number of showroom, number of customers, and time period as normal and recession to the problem of Galaxy Garments. The SPSS command for forward stepwise regression is demonstrated in Figure 11.22 and output is given in Table 11.9.

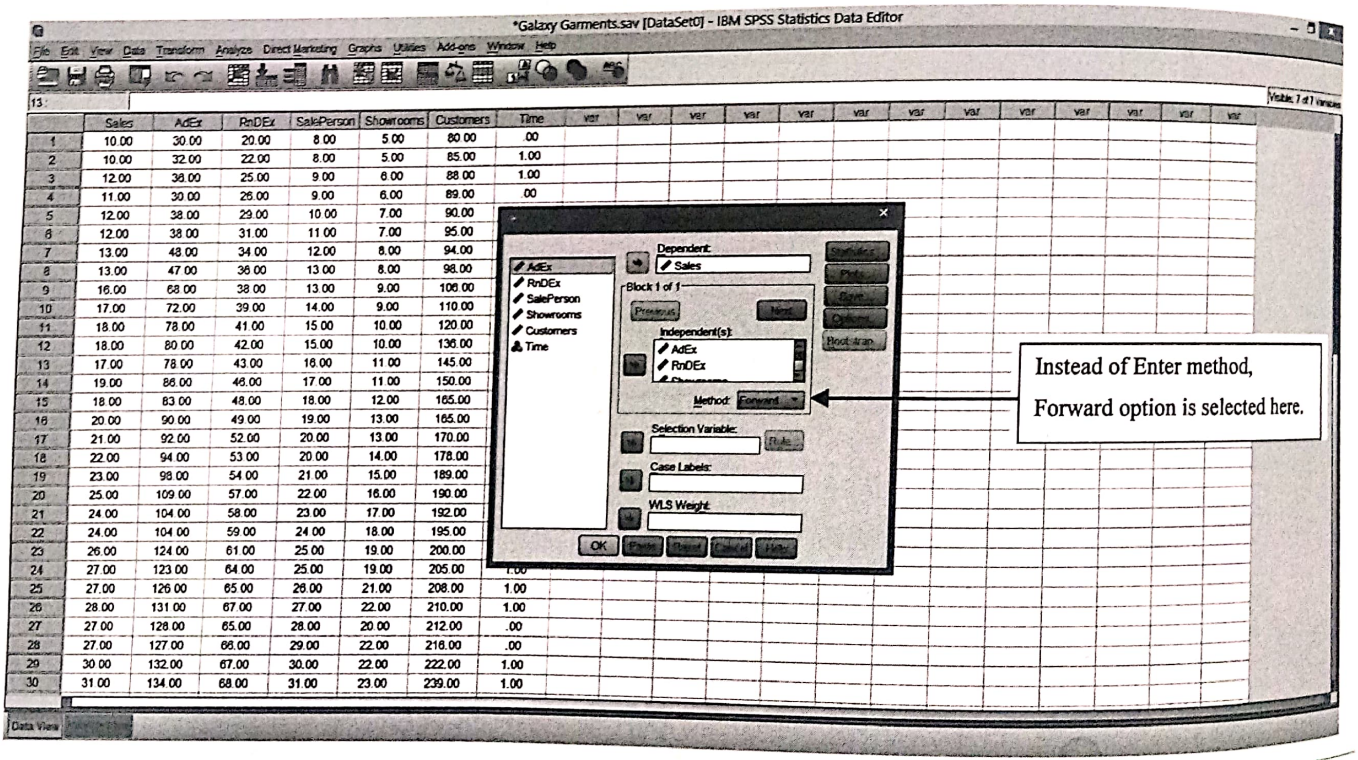


Figure 11.22: Using SPSS for Forward Stepwise Regression.

As it can be seen from the results, model 4 contains all the significant explanatory variables with $R^2 = 99.1\%$. The significance of R^2 for each model is tested and the results are presented in Table 11.9. In each stage, a variable is selected for inclusion in the model on the basis of the partial correlation coefficient. The variable with highest partial correlation coefficient occupies position in the model. R&D expenditure is selected first of all. If we look at the table for excluded variables, the partial correlation for advertising expenditure (0.540) is highest. Therefore, it is included in the second model. The list of variables excluded in each model is presented in the table for excluded variables. From the results, our estimated equation can be written as:

TABLE I.

SIGNIFICANT VALUES $t_\nu(\alpha)$ of t-Distribution
(TWO-TAIL AREAS)
$P[\mid t \mid > t_\nu(\alpha)] = \alpha$

d.f.	*Probability (Level of Significance)*					
(ν)	0·50	0·10	0·05	0·02	0·01	0·001
1	1·00	6·31	12·71	31·82	63·66	636·62
2	0·82	2·92	4·30	6·97	6·93	31·60
3	0·77	2·35	3·18	4·54	5·84	12·94
4	0·74	2·13	2·78	3·75	4·60	8·61
5	0·73	2·02	2·57	3·37	4·03	6·86
6	0·72	1·94	2·45	3·14	3·71	5·96
7	0·71	1·90	2·37	3·00	3·50	5·41
8	0·71	1·86	2·31	2·90	3·36	5·04
9	0·70	1·83	2·26	2·82	3·25	4·78
10	0·70	1·81	2·23	2·76	3·17	4·59
11	0·70	1·80	2·20	2·72	3·11	4·44
12	0·70	1·78	2·18	2·68	3·06	4·32
13	0·69	1·77	2·16	2·65	3·01	4·22
14	0·69	1·76	2·15	2·62	2·98	4·14
15	0·69	1·75	2·13	2·60	2·95	4·07
16	0·69	1·75	2·12	2·58	2·92	4·02
17	0·69	1·74	2·11	2·57	2·90	3·97
18	0·69	1·73	2·10	2·55	2·88	3·92
19	0.69	1.73	2.09	2.54	2.86	3.88
20	0·69	1·73	2·09	2·53	2·85	3·85
21	0·69	1·72	2·08	2·52	2·83	3·83
22	0·69	1·72	2·07	2·51	2·82	3·79
23	0·69	1·71	2·07	2·50	2·81	3·77
24	0·69	1·71	2·06	5·49	2·80	3·75
25	0·68	1·71	2·06	2·49	2·79	3·73
26	0·68	1·71	2·06	2·48	2·78	3·71
27	0·68	1·70	2·05	2·47	2·77	3·69
28	0·68	1·70	2·05	2·47	2·76	3·67
29	0·68	1·70	2·05	2·46	2·76	3·66
30	0·68	1·70	2·04	2·46	2·75	3·65
∞	0·67	1·65	1·96	2·33	2·58	3·29

Next, the researcher determined the values of Σx, Σy, Σx^2, and Σxy.

Hospital	Number of Beds x	FTEs y	x^2	xy
1	23	69	529	1587
2	29	95	841	2755
3	29	102	841	2958
4	35	118	1225	4130
5	42	126	1764	5292
6	46	125	2116	5750
7	50	138	2500	6900
8	54	178	2916	9612
9	64	156	4096	9984
10	66	184	4356	12,144
11	76	176	5776	13,376
12	78	225	6084	17,550
	$\Sigma x = 592$	$\Sigma y = 1692$	$\Sigma x^2 = 33,044$	$\Sigma xy = 92,038$

Using these values, the analyst solved for the sample slope (b_1) and the sample y-intercept (b_0).

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 92,038 - \frac{(592)(1692)}{12} = 8566$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33,044 - \frac{(592)^2}{12} = 3838.667$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{8566}{3838.667} = 2.232$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{n} = \frac{1692}{12} - (2.232)\frac{592}{12} = 30.888$$

The least squares equation of the regression line is

$$\hat{y} = 30.888 + 2.232x$$

The slope of the line, $b_1 = 2.232$, means that for every unit increase of x (every additional bed), y (number of FTEs) is predicted to increase by 2.232. Even though the y-intercept helps the analyst sketch the graph of the line by being one of the points on the line (0, 30.888), it has limited usefulness in terms of this solution because $x = 0$ denotes a hospital with no beds. On the other hand, it could be interpreted that a hospital has to have at least 31 FTEs to open its doors even with no patients—a sort of "fixed cost" of personnel.

AN INDIAN ADAPTATION

Tenth Edition

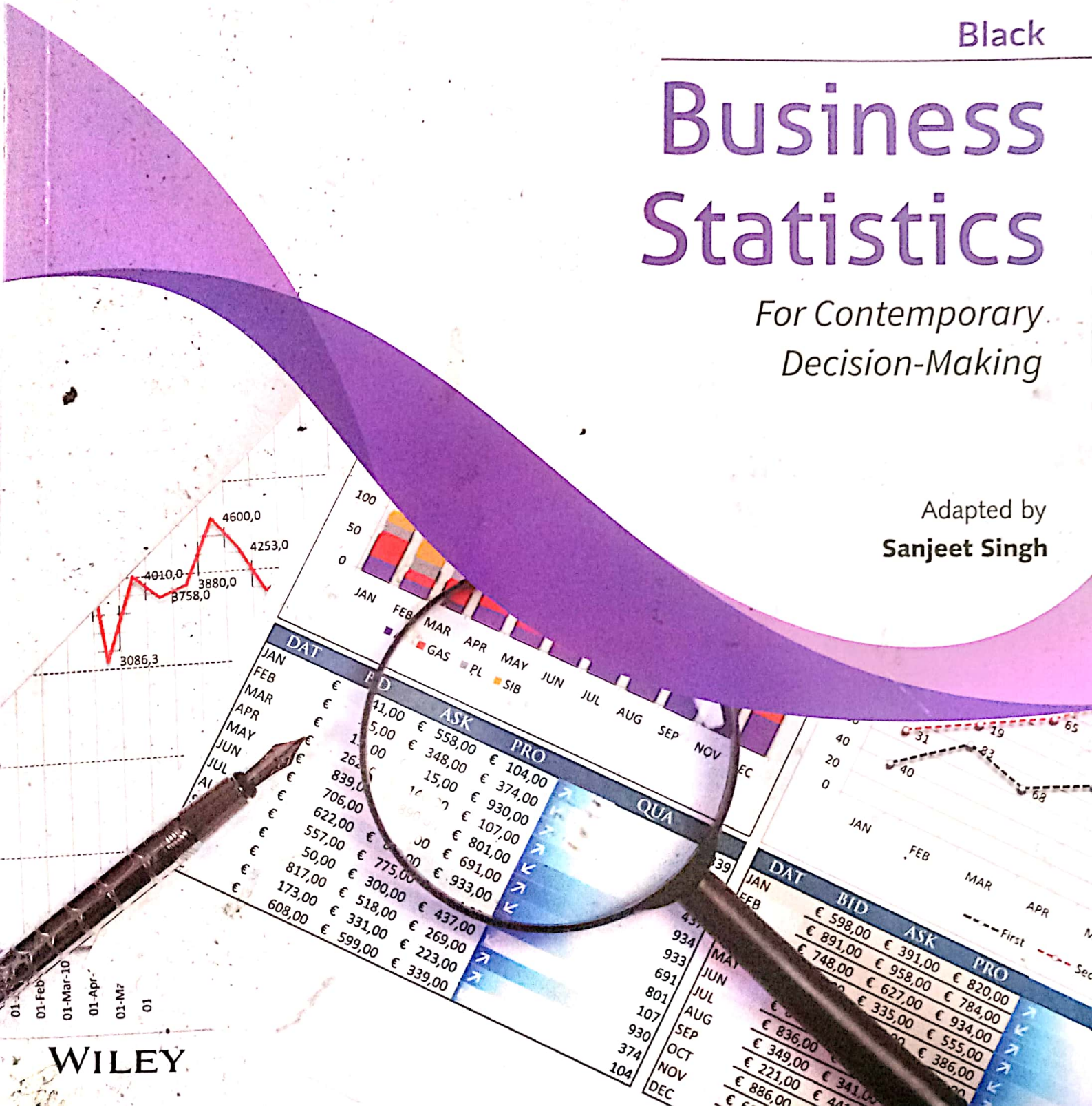
Black

Business Statistics

For Contemporary Decision-Making

Adapted by

Sanjeet Singh



WILEY