Chapter 8
Multiple Regression

8.1 Introduction

Multiple regression analysis is one of the dependence technique in which the researcher can analyze the relationship between a single-dependent (criterion) variable and several independent variables. In multiple regression analysis, we use independent variables whose values are known or fixed (non-stochastic) to predict the single-dependent variable whose values are random (stochastic). In multiple regression analysis, our dependent and independent variables are metric in nature; however, in some situations, it is possible to use non-metric data as independent variable (as dummy variable).

Gujarati and Sangeetha (2008) defined regression as:

'It is concerned with the study of the dependence of one variable, the *dependent variable*, on one or more other variables, the *explanatory variables*, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) value of the later'.

8.2 Important Assumptions of Multiple Regression

1. Linearity—the relationship between the predictors and the outcome variable should be linear
2. Normality—the errors should be normally distributed—technically normality is necessary only for the *t*-tests to be valid, estimation of the coefficients (errors are identically and independently distributed
3. Homogeneity of variance (homoscedasticity)—the error variance should be constant
4. Independence (no autocorrelation)—the errors associated with one observation are not correlated with errors of any other observation
5. There is no multicollinearity or perfect correlation between independent variables.

S. Sreejesh et al., *Business Research Methods*,
DOI: 10.1007/978-3-319-00539-3_8,

Additionally, there are issues that can arise during the analysis that, while strictly speaking, are not assumptions of regression, are none the less, of great concern to regression analysis. These are

1. Outliers; it is an observation whose dependent variable value is unusual given its values on the predictor variable (independent variable).
2. Leverage; an observation with an extreme value on a predictor variable is called a point with high leverage.
3. Influence; an observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outliers.

8.3 Multiple Regression Model with Three Independent Variables

One of the well-known supermarket chains (ABC group) in the country has adopted an aggressive marketing decision particularly to increase the sales of its own private brands in the last 19 months. Recently, the company decided to investigate its product sales in the last 19 months. In the last 19 months, the company has invested lot of money in three strategic areas: Advertisement, marketing (excluding advertisement and distribution) and its distribution network. The company decided to do a multiple regression analysis to predict the impact of advertisement, marketing, and distribution expenses on its sales (Table 8.1a).

8.4 Multiple Regression Equation

A multiple regression equation with three independent variables is given below:

$$Y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t' \tag{1}$$

$$\begin{aligned} Sales_t = {} & \beta_1(constant) + \beta_2(Advertisement\ Ex.)_t + \beta_3(Marketing\ Ex.)_t \\ & + \beta_4(Distribution\ Ex.)_t \\ & + u_t' \end{aligned} \tag{2}$$

Here, Y_t is the value of the dependent variable (here it is sales) on time period t, β_1 is the intercept or average value of dependent variable when all the independent variables are absent. $\beta_2 \beta_3$, *and* β_4, are the slope of sales (partial regression coefficients) with respect independent variables like advertisement expenses, marketing expenses, and distribution expenses holding other variables constant. For example, the coefficient value β_2 implies that one unit change (increase or

Table 8.1a Sales, advertising, marketing, and distribution expenses

Months	Sales (In lakhs)	Advertising expenses (In lakhs)	Marketing expenses (In lakhs)	Distribution expenses (In lakhs)
1	9324.6	9	129.8	139.9
2	11870.8	20.2	206.1	124.7
3	15118.6	9.8	105.1	169.9
4	19406.4	17.2	53	483.9
5	21715.4	11.5	65	495
6	28270.2	38.9	68.5	618.4
7	41960.1	41.9	81	850.3
8	64647.5	139.9	203	1273.2
9	77826.3	344.9	439.6	1624.6
10	83059.5	451.6	767.7	1538.3
11	78855.6	656.3	1680	1474.9
12	94407	882	1638	1732
13	90615	1051	1376	1594
14	92313	1170	2063	1588
15	92038	1676	2361	2041
16	111281	2518	354	1188
17	134859	2044	195	1133
18	151252	2257	234	1069
19	174580	3389	234	1376

decrease) of in advertisement will lead to β_2 unit time changes (increase or decrease) in sales holding other variables constant. u_t' is the random error in Y, for time period t.

8.5 Regression Analysis Using SPSS

How to Check Linearity Assumption.mp4

Step 1 Open the data file named **Supermarket.sav** (Fig. 8.1).

Step 2 Go to Analyze => Regression =>Linear to get the Linear Regression window as given in Fig. 8.2.

Step 3 Click the dependent variable **Sales** from the left panel of the Linear Regression window into dependent variable (right panel) and other three variables into Independent window (Fig. 8.3).

Step 4 Click the **Statistics** option and select **Estimates**, **Model fit**, and **Descriptives,** then click on **Continue** to get the main window of Linear Regression (Fig. 8.4).

Step 5 Go to the main window of linear regression and click **OK** (Fig. 8.5).

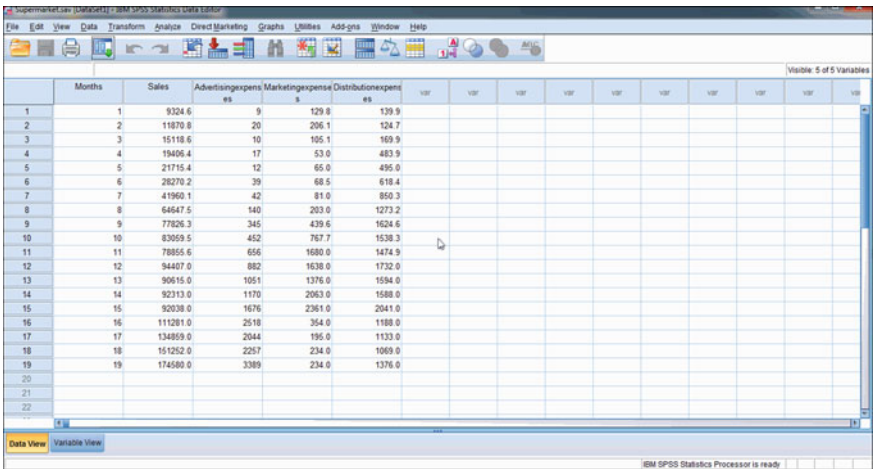

Fig. 8.1 SPSS data view window

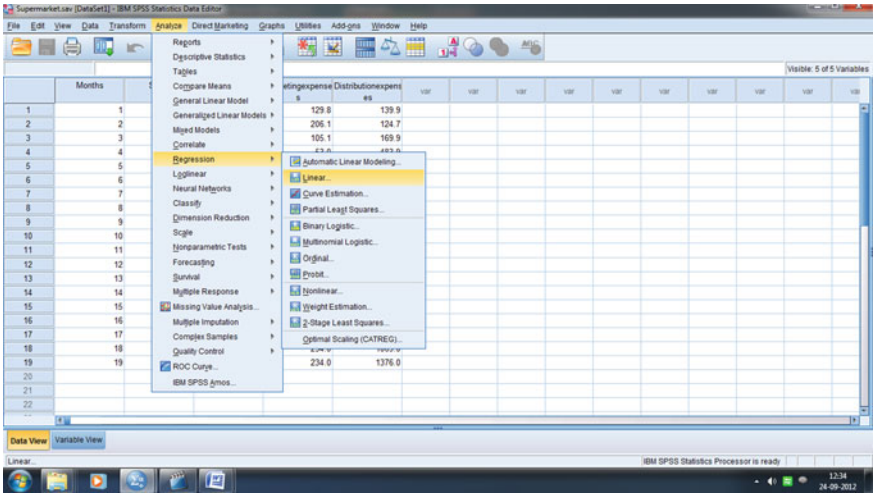

Fig. 8.2 SPSS linear regression window

8.6 Output Interpretation for Regression Analysis

Table 8.1b in SPSS regression output shows the model summary, which provides the value of R (Multiple Correlation), R^2 (Coefficient of Determination) and Adjusted R^2 (R^2 adjusted with Degrees of Freedom). In this model, R has a value of 0.970. This value represents the multiple correlation between dependent and independent variables. The value of R^2 shows all the three independent variables

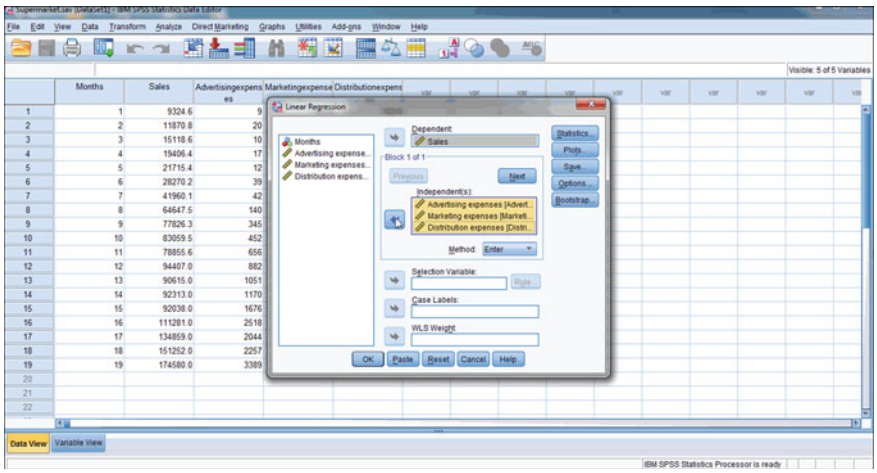

Fig. 8.3 SPSS linear regression window

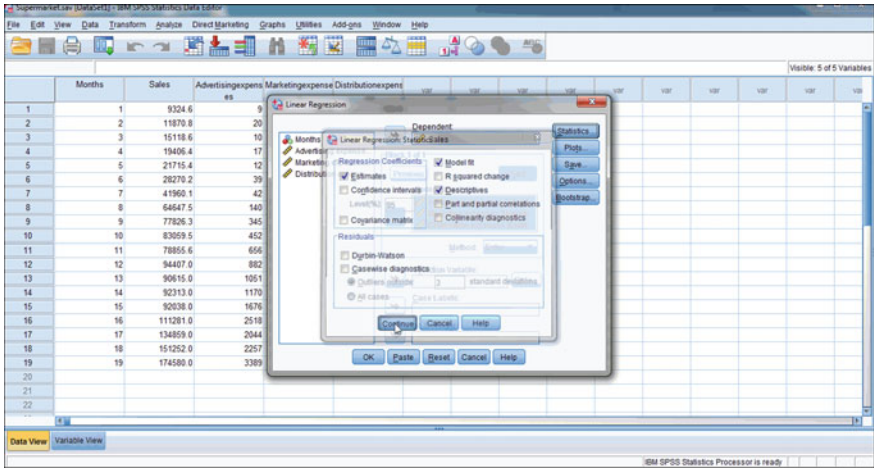

Fig. 8.4 Linear regression statistics window

can account for 94 % of the variation in sales. In other words, if the researcher would like to explain the contribution of all these three expenses on sales, looking at the R^2 it is possible. This means that around 6 % of the variation is sales cannot be explained by all these expenses. Therefore, it can be concluded that there must be other variables that have influence on sales.

Table 8.2 reports an analysis of variance (ANOVA). This table shows all the sums of squares associated with regression. The regression sum of square explains the sum of squares explained by the model or all the independent variables. Residual sum of squares explains the sum of squares for the residual or

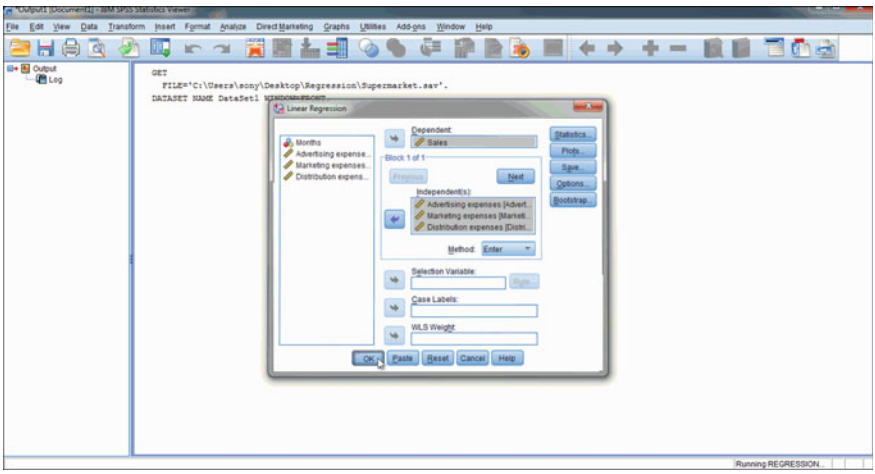

Fig. 8.5 SPSS linear regression window

unexplained part. Total sum of squares explains the sum of squares of the dependent variable. The third column shows the associated degrees of freedom for each sum of squares. The mean sum of squares for the regression and residuals are calculated by dividing respective sum of squares by its degrees of freedom. The most important part in this table is F value, which is calculated by taking the ratio of mean square of regression and mean square of residual. For this model, the F value is 78.742, which is significant ($p < .01$). This result tells us that there is less than a 0.1 % chance that an F-ratio this large would happen if the null hypothesis were true. Therefore, looking at the ANOVA table, we can infer that our regression model results in significantly better prediction of sales.

Looking at the ANOVA explained in Table 8.2, we cannot make inference about the predictive ability of individual independent variables. Table 8.3 provides details about the model parameters. Looking at the beta vales and its significance, one can interpret the significance of each predictor on the dependent variable. The value 6908.926 is the constant term which is β_1 in Eqs. 1 and 2. This can be interpreted as when no money is spent on all these three areas (advertising, marketing, and distribution) or $X_2\, X_3\, X_4 = 0$, the model predicts that average sales would be 6908.92 (remember our unit of measurement is in lakhs). The coefficient value for advertising expenses is 33.56(β_2) is the partial regression coefficient for advertising expenses. This value represents the change in the outcome associated with the unit change in the predictor or independent variable, while other variables

Table 8.1b Model summary

Model	R	R^2	Adjusted R^2	Std. error of the estimate
1	0.970[a]	0.940	0.928	13093.8291

[a] Predictors: (constant), distribution expenses, advertising expenses, marketing expenses

Table 8.2 ANOVA[a]

Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	40500692519.872	3	13500230839.957	78.742	0.000[b]
	Residual	2571725425.134	15	171448361.676		
	Total	43072417945.006	18			

[a] Dependent variable: sales
[b] Predictors: (constant), distribution expenses, advertising expenses, marketing expenses

Table 8.3 Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients	*t*	Sig.
		B	Std. error	Beta		
1	(Constant)	6908.926	6840.615		1.010	0.329
	Advertising expenses	33.569	3.545	0.709	9.468	0.000
	Marketing expenses	−15.625	6.203	−0.244	−2.519	0.024
	Distribution expenses	43.485	9.002	0.524	4.831	0.000

[a] Dependent variable: sales

hold constant. Therefore, it can be interpreted that if our independent variable is increased by one unit (here advertising expenses), then our model predicts that 33.56 unit times change in depended variable (here sales) occurs while holding other variables like marketing expenses and distribution expenses constant. As our unit of measurement for the advertising expenses were in lakhs, it can be interpreted that an increase in advertising expenses of Rs. 1 lakhs will increase the sales 33,56000 lakhs (100000 * 33.569) holding other expenses constant. In the same fashion, one can also interpret the other coefficients. The negative sign of the coefficients indicates an inverse relationship between dependent and independent variables.

Standard Error Column explains the standard error associated with each estimate or coefficients. The standardized coefficients column shows the standardized coefficient values for each estimate in which the unit of measurement is common. These coefficients can be used for explaining the relative importance of each independent variable when the unit of measurement is different for different independent variables. Looking at the coefficients, one can infer that advertising expense is the most important predictor followed by distribution expenses.

The last two columns show *t*-value and associated probability. The *t*-value can be calculated as unstandardized coefficients divided by its respective standard error. The *t*-test tells us whether the β-value is different from 0 or not. The last column of the Table 8.3 shows the exact probability that the observed value of *t* would occur if the value of β in the population were 0. If the probability value is less than 0.05, then the researcher agree that result reflect a genuine effect or β is different from 0. From the table, it is evident that for all the three independent variables, the probability value is less than that the assumed 0.05 level, and so we

can say that in all the three cases, the coefficient values are significantly different from zero or it significantly contributes to the model.

8.7 Examination of Major Assumptions of Multiple Regression Analysis

8.7.1 Examination of Residual

Examining the residual provide useful insights in examining the appropriateness of the underlying assumptions and regression model fitted A **residual** is the difference between the observed value of Y_i and the value predicted by the regression equation $\hat{Y}_i$. Residuals are used in the calculation of several statistics associated with regression. ***Without verifying that your data have met the regression assumptions, the results may be misleading.***

8.7.2 Test of Linearity

When we do linear regression, we assume that the relationship between the response variable and the predictors is linear. This is the assumption of linearity. If this assumption is violated, the linear regression will try to fit a straight line to data that does not follow a straight line. Checking the linear assumption in the case of simple regression is straightforward, since we only have one predictor. All we have to do is a scatter plot between the each response variable (independent variable) and the predictor (dependent variable) to see if nonlinearity is present, such as a curved band or a big wave-shaped curve. The examination of linearity can be examined through the following video.

How to Check Normality Assumption.mp4

8.7.3 Test of Normality

The assumption of a normally distributed error term can be examined by constructing a histogram of the residuals. A visual check reveals whether the distribution is normal. It is also useful to examine the normal probability of plot of standardized residuals compared with expected standardized residuals from the normal distribution. If the observed residuals are normally distributed, they will fall on the 45-degree line. Additional evidence can be obtained by determining the

percentages of residuals falling within ± 2 *SE* or ± 2.5 *SE*. More formal assessment can be made by running the tests: Shapiro–Wilk, Kolmogorov–Smirnov, Cramer–von Mises and Anderson–Darling.[1]

How to Autocorrelation Assumption.mp4

8.7.4 Test of Homogeneity of Variance (Homoscedasticity)

The assumption of constant variance of the error term can be examined by plotting the residuals against the predicted values of the dependent variable, $\hat{Y}_i$. If the pattern is not random, the variance of the error term is not constant. See the video ***How to check Normality Assumption***.

8.7.5 Test of Autocorrelation

How to Check No Multicollinearity Assumption.mp4

A plot of residuals against time, or the sequence of observations, will throw some light on the assumption that the error terms are uncorrelated or no autocorrelation. A random pattern should be seen if this assumption is true. A more formal procedure for examining the correlations between the error terms is the Durbin–Watson test (Applicable only for time series data).

8.7.6 Test of Multicollinearity

The presence of multicollinearity or perfect linear relationship between independent variables can be identified using different methods. These methods are:

1. VIF (Variance-Inflating factor): As a rule of thumb, If the VIF value exceeds 10, which will happen only if correlation between independent variables exceeds 0.90, that variable is said to be highly collinear (Gujarati and Sangeetha 2008).

[1] *Null hypothesis* the observations are normally distributed, *alternative hypothesis* not normally distributed.

2. TOL (Tolerance): The closer the TOL to zero, the greater the degree of collinearity of the variables (Gujarati and Sangeetha 2008).
3. Conditional Index (CI): If CI exceeds 30, there is severe multicollinearity (Gujarati and Sangeetha 2008).
4. Partial Correlations: High partial correlation between independent variables also shows the presence of multicollinearity.

How to Check No Multicollinearity Assumption.mp4

8.7.7 Questions

Examine the following fictitious data

Model	R	R^2	Adjusted R square	Std. error of the estimate
1	0.863	0.849	0.850	13.8767

1. Which of the following statements can we *not* say?

   (a) The standard error is an estimate of the variance of y, for each value of x.
   (b) **In order to obtain a measure of explained variance, you need to square the correlation coefficient**.
   (c) The correlation between x and y is 86 %.
   d) The correlation is good here as the data points cluster around the line of fit quite well. So prediction will be good.
   (e) The correlation between x and y is 85 %.

2. The slope of the line is called:

   (a) **Which gives us a measure of how much y changes as x changes**.
   (b) Is the point where the regression line cuts the vertical axis.
   (c) A correlation coefficient indicates the variability of the points around the regression line in the scatter diagram.
   (d) None of the above.
   (e) The average value of the dependent variable.

3. Using some fictitious data, we wish to predict the musical ability for a person who scores 8 on a test for mathematical ability. We know the relationship is positive. We know that the slope is 1.63 and the intercept is 8.41. What is their predicted score on musical ability?

(a) 80.32
(b) −4.63
(c) **21.45**
(d) 68.91
(e) 54.55

4. We have a negative relationship between number of drinks consumed and number of marks in a driving test. One individual scores 3 on number of drinks consumed, another individual scores 5 on number of drinks consumed. What will be their respective scores on the driving test if the intercept is 18 and the slope 3?

   (a) It is not possible to predict from negative relationships.
   (b) Driving test scores (*Y*-axis) will be 51 and 87 [individual who scored 5 on drink consumption].
   (c) Driving test scores (*Y*-axis) will be 27 [individual who scored 3 on drink consumption] and 33 [individual who scored 5 on drink consumption].
   (d) **Driving test scores (*Y*-axis) will be 9 [individual who scored 3 on drink consumption] and 3 [individual who scored 5 on drink consumption**].
   (e) None of these.

5. You are still interested in whether problem-solving ability can predict the ability to cope well in difficult situations; whether motivation can predict coping and whether these two factors together predict coping even better. You produce some more results.

   Dependent variable coping skills in difficult situations

	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	Std. error	Beta		
Constant	−0.466	0.241		1.036	0.302
Problem	0.200	0.048	0.140	2.082	0.030
Motivation	0.950	0.087	0.740	10.97	0.000

   Which of the following statements is incorrect?

(a) As motivation increases by one standard deviation, coping skills increases by almost three quarters of a standard deviation (0.74). Thus, motivation appears to contribute more to coping skills than problem solving.
(b) As motivation increases by one unit coping skills increases by 0.95.

(c) **The *t*-value for problem solving is 2.082 and the associated probability is 0.03. This tells us the likelihood of such a result arising by sampling error, assuming the null hypothesis is true, is 97 in 100.**
(d) Problem solving has a regression coefficient of 0.20. Therefore, as problem solving increases by one unit coping skills increases by 0.20.
(e) None of these.