CHAPTER 6 LINEAR REGRESSION AND CORRELATION

Contents

Objectives: In business and economic applications, frequently interest is in relationships between two or more random variables, and the association between variables is often approximated by postulating a linear functional form for their relationship.

After working through this chapter, you should be able to:

(i) understand the basic concepts of regression and correlation analyses;

(ii) determine both the nature and the strength of the linear relationship between two variables;

(iii) extend the simple regression techniques to examine decision-making situations where multiple regression can be used to make predictions;

(iv) interpret outputs of multiple regression from computer packages.

6.1 Introduction

This chapter presents some statistical techniques to analyze the association between two variables and develop the relationship for prediction.

6.2 Curve Fitting

Very often in practice a relation is found to exist between two (or more) variables.

It is frequently desirable to express this relationship in mathematical form by determining an equation connecting the variables.

To aid in determining an equation connecting variables, a first step is the collection of data showing corresponding values of the variables under consideration.

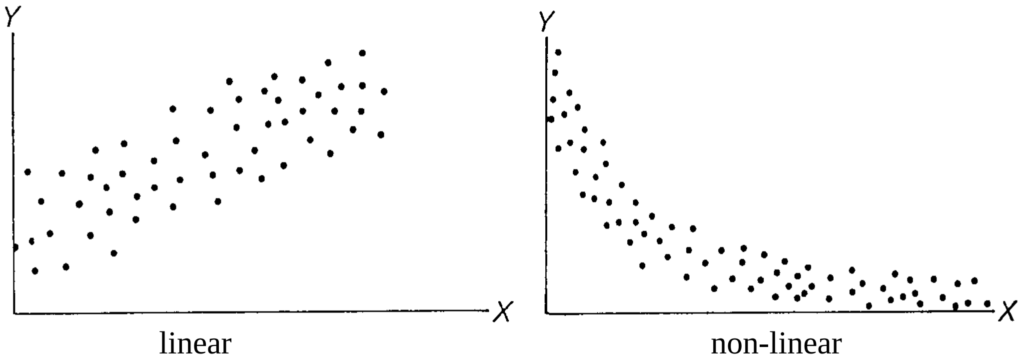

Figure 1. Scatter diagram

6.3 Fitting a Simple Linear Regression Line

To determine from a set of data, a line of best fit to infer the relationship between two variables.

6.3.1 The Method of Least Squares

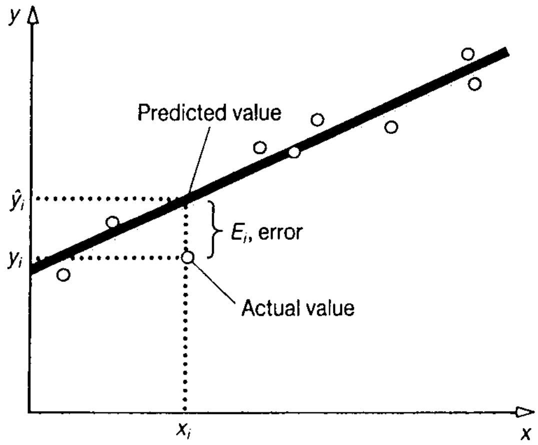

Figure 2. Sample observations and the sample regression line

Determining the line of "best fit":

$$\hat{y} = a + bx$$

by minimizing $\sum E_i^2$.

To minimize $\sum E_i^2$, we apply calculus and find the following "normal equations":

$$\sum y = na + b\sum x \qquad (1)$$
$$\sum yx = a\sum x + b\sum x^2 \qquad (2)$$

Solve (1) and (2) simultaneously, we have:

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

Notes:

1. The formula for calculating the slope b is commonly written as

$$b = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2}$$

which the numerator and denominator then reduce to formulas

$$\begin{aligned}\sum(x-\overline{x})(y-\overline{y}) &= \sum(xy-\overline{x}y-x\overline{y}+\overline{x}\,\overline{y}) \\ &= \sum xy-\overline{x}\sum y-\overline{y}\sum x+\sum\overline{x}\,\overline{y} \\ &= \sum xy-n\overline{x}\,\overline{y}-n\overline{x}\,\overline{y}+n\overline{x}\,\overline{y} \\ &= \sum xy-n\overline{x}\,\overline{y}\end{aligned}$$

and

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum\left(x^2-2x\overline{x}+\overline{x}^2\right) \\ &= \sum x^2-2\overline{x}\sum x+\sum\overline{x}^2 \\ &= \sum x^2-2n\overline{x}^2+n\overline{x}^2 \\ &= \sum x^2-n\overline{x}^2\end{aligned}$$

respectively; and $a=\overline{y}-b\overline{x}$ is the y-intercept of the regression line.

2. When the equation $\hat{y}=a+bx$ is calculated from a sample of observations rather than from a population, it is referred as a sample regression line.

Example 1

Suppose an appliance store conducts a 5-month experiment to determine the effect of advertising on sales revenue and obtains the following results

Month	Advertising Expenditure (in \$1,000)	Sales Revenue (in \$10,000)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Find the sample regression line and predict the sales revenue if the appliance store spends 4.5 thousand dollars for advertising in a month.

From the data, we find that

$$n=5,\ \sum x=15,\ \sum y=10,\ \sum xy=37,\ \sum x^2=55.$$

Hence $\bar{x}=\dfrac{\sum x}{n}=\dfrac{15}{5}=3$ and $\bar{y}=\dfrac{\sum y}{n}=\dfrac{10}{5}=2$.

Then the slope of the sample regression line is

$$b=\frac{\sum xy-n\bar{x}\,\bar{y}}{\sum x^2-n\bar{x}^2}$$

$$=$$

$$=$$

and the y-intercept is

$$a=\bar{y}-b\bar{x}$$

$$=$$

$$=$$

The sample regression line is thus

$$\hat{y}=$$

So if the appliance store spends 4.5 thousand dollars for advertising in a month, it can expect to obtain $\hat{y}=$ $=$ ten-thousand dollars as sales revenue during that month.

Example 2

Obtain the least squares prediction line for the data below:

	y_i	x_i	x_i^2	$x_i y_i$	y_i^2
	101	1.2	1.44	121.2	10201
	92	0.8	0.64	73.6	8464
	110	1.0	1.00	110.0	12100
	120	1.3	1.69	156.0	14400
	90	0.7	0.49	63.0	8100
	82	0.8	0.64	65.6	6724
	93	1.0	1.00	93.0	8649
	75	0.6	0.36	45.0	5625
	91	0.9	0.81	81.9	8281
	105	1.1	1.21	115.5	11025
Sum	959	9.4	9.28	924.8	93569

$$b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = \frac{10(924.8) - (9.4)(959)}{10(9.28) - (9.4)^2} = \frac{233.4}{4.44} = 52.568$$

$$a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \frac{959}{10} - 52.568\left(\frac{9.4}{10}\right) = 46.486$$

Therefore, $\hat{y} = 46.486 + 52.568x$

Example 3

Find a regression curve in the form $y = a + b\ln x$ for the following data:

x_i	1	2	3	4	5	6	7	8
y_i	9	13	14	17	18	19	19	20

$\ln x_i$	0	0.693	1.099	1.386	1.609	1.792	1.946	2.079
y_i	9	13	14	17	18	19	19	20

$\sum \ln x_i = 10.604$ $\sum \left(\ln x_i\right)^2 = 17.518$

$\sum y_i = 129$ $\sum \left(\ln x_i\right) y_i = 189.521$

$$b = \frac{n\sum(\ln x)y - \left(\sum \ln x\right)\left(\sum y\right)}{n\sum(\ln x)^2 - \left(\sum \ln x\right)^2} = \frac{8(189.521) - (10.604)(129)}{8(17.518) - (10.604)^2} = 5.35$$

$$a = \frac{\sum y}{n} - b\frac{\sum \ln x}{n} = \frac{129}{8} - 5.35\left(\frac{10.604}{8}\right) = 9.03$$

Therefore, $\hat{y} = 9.03 + 5.35 \ln x$

6.4 Linear Correlation Analysis

Correlation analysis is the statistical tool that we can use to determine the degree to which variables are related.

6.4.1 Coefficient of Determination, r^2

Problem: how well a least squares regression line fits a given set of paired data?

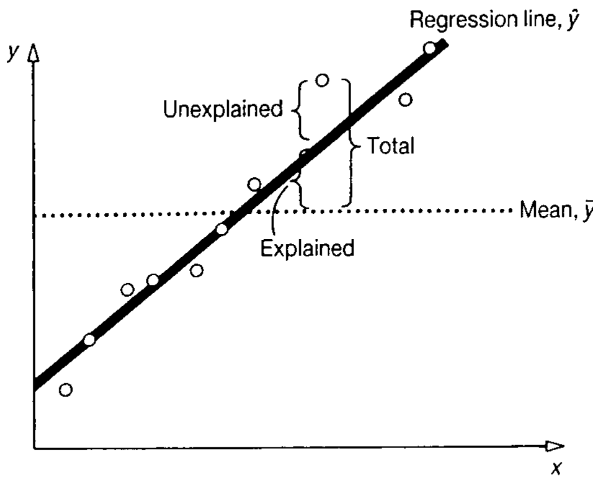

Figure 3. Relationships between total, explained and unexplained variations

Variation of the y values around their own mean = $\sum(y - \bar{y})^2$

Variation of the y values around the regression line = $\sum(y - \hat{y})^2$

Regression sum of squares = $\sum(\hat{y} - \bar{y})^2$

We have:

$$\sum\left(y-\bar{y}\right)^2 = \sum\left(\hat{y}-\bar{y}\right)^2 + \sum\left(y-\hat{y}\right)^2$$

$$\Rightarrow \qquad 1 = \frac{\sum\left(\hat{y}-\bar{y}\right)^2}{\sum\left(y-\bar{y}\right)^2} + \frac{\sum\left(y-\hat{y}\right)^2}{\sum\left(y-\bar{y}\right)^2}$$

$$\Rightarrow \qquad \frac{\sum\left(\hat{y}-\bar{y}\right)^2}{\sum\left(y-\bar{y}\right)^2} = 1 - \frac{\sum\left(y-\hat{y}\right)^2}{\sum\left(y-\bar{y}\right)^2}.$$

Denoting $\dfrac{\sum\left(\hat{y}-\bar{y}\right)^2}{\sum\left(y-\bar{y}\right)^2}$ by r^2, then

$$r^2 = 1 - \frac{\sum\left(y-\hat{y}\right)^2}{\sum\left(y-\bar{y}\right)^2}.$$

r^2, the coefficient of determination, is the proportion of variation in y explained by a sample regression line.

For example, r^2 =0.9797; that is, 97.97% of the variation in y is due to their linear relationship with x.

6.4.2 Correlation Coefficient

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

and $-1 \le r \le 1$.

Notes:

The formulas for calculating r^2 (sample coefficient of determination) and r (sample coefficient of correlation) can be simplified in a more common version as follows:

$$r^2 = \frac{\left(\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)\right)^2}{\sum\left(x-\bar{x}\right)^2\sum\left(y-\bar{y}\right)^2} = \frac{\left(\sum xy - n\bar{x}\,\bar{y}\right)^2}{\left(\sum x^2 - n\bar{x}^2\right)\left(\sum y^2 - n\bar{y}^2\right)}$$

$$r = \sqrt{r^2} = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2\sum\left(y-\bar{y}\right)^2}} = \frac{\sum xy - n\bar{x}\,\bar{y}}{\sqrt{\left(\sum x^2 - n\bar{x}^2\right)\left(\sum y^2 - n\bar{y}^2\right)}}$$

Since the numerator used in calculating r and b are the same and both denominators are always positive, r and b will always be of the same sign. Moreover, if r=0 then b=0; and vice versa.

Example 4

Calculate the sample coefficient of determination and the sample coefficient of correlation for example 1. Interpret the results.

From the data we get

$$n=5,\ \sum x=15,\ \sum y=10,\ \sum xy=37,\ \sum x^2=55,\ \sum y^2=26.$$

Then, the coefficient of determination is given by

$$r^2 = \frac{\left(\sum xy - n\bar{x}\,\bar{y}\right)^2}{\left(\sum x^2 - n\bar{x}^2\right)\left(\sum y^2 - n\bar{y}^2\right)}$$

$$=$$

$$=$$

$$=$$

and

$$r = \qquad\qquad =$$

$r^2 =$ implies that of the sample variability in sales revenue is explained by its linear dependence on the advertising expenditure. $r =$ indicates a very strong positive linear relationship between sales revenue and advertising expenditure.

Example 5

Interest rates (x) provide an excellent leading indicator for predicting housing starts (y). As interest rates decline, housing starts increase, and vice versa. Suppose the data given in the accompanying table represent the prevailing interest rates on first mortgages and the recorded building permits in a certain region over a 12-year span.

	Year					
	1985	1986	1987	1988	1989	1990
Interest rates (%)	6.5	6.0	6.5	7.5	8.5	9.5
Building permits	2165	2984	2780	1940	1750	1535
	Year					
	1991	1992	1993	1994	1995	1996
Interest rates (%)	10.0	9.0	7.5	9.0	11.5	15.0
Building permits	962	1310	2050	1695	856	510

(a) Find the least squares line to allow for the estimation of building permits from interest rates.

(b) Calculate the correlation coefficient r for these data.

(c) By what percentage is the sum of squares of deviations of building permits reduced by using interest rates as a predictor rather than using the average annual building permits $\bar{y}$ as a predictor of y for these data?

6.5 Spearman's Rank Correlation

Occasionally we may need to determine the correlation between two variables where suitable measures of one or both variables do not exist.

However, variables can be ranked and the association between the two variables can be measured by r_s:

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$, where d is the difference of rank between x and y.

$-1 \le r_s \le 1$

if r_s closes to 1: strong positive association
if r_s closes to -1: strong negative association
if r_s closes to 0: no association

Notes:

1. The two variables must be ranked in the same order, giving rank 1 either to the largest (or smallest) value, rank 2 to the second largest (or smallest) value and so forth.

2. If there are ties, we assign to each of the tied observations the mean of the ranks which they jointly occupy; thus, if the third and fourth ordered values are identical we assign each the rank of $\frac{3+4}{2} = 3.5$, and if the fifth, sixth and seventh ordered values are identical we assign each the rank of $\frac{5+6+7}{3} = 6$.
3. The ordinary sample correlation coefficient r can also be used to calculate the rank correlation coefficient where x and y represent ranks of the observations instead of their actual numerical values.

Example 6

Calculate the rank correlation coefficient r_s for example 1.

Month (1)	Value x (2)	rank (x) (3)	Value y (4)	rank (y) (5)	d (6)=(3)-(5)	d^2 (7)
1	1	1	1	1.5	-0.5	0.25
2	2	2	1	1.5	0.5	0.25
3	3	3	2	3.5	-0.5	0.25
4	4	4	2	3.5	0.5	0.25
5	5	5	4	5	0	0

By formula

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$=$

$=$

r_s = indicates a correlation between the rankings of advertising expenditure and sales revenue. Note that if we apply the ordinary formula of correlation coefficient r to calculate the correlation coefficient of the rankings of the variables in example 6, the result would be slightly different. Since

$$n\text{=5, } \sum \text{rank}(x) \text{=15, } \sum \text{rank}(y) \text{=15, } \sum (\text{rank}(x))(\text{rank}(y)) \text{=54,}$$
$$\sum (\text{rank}(x))^2 \text{=55, } \sum (\text{rank}(y))^2 \text{=54,}$$

then r =

which is very close to the result of r_s.

Example 7

Calculate the Spearman's rank correlation, r_s, between x and y for the following data:

y_i	$\text{rank}(y_i)$	x_i	$\text{rank}(x_i)$	$(\text{rank}(y_i) - \text{rank}(x_i))^2$
52		10		
54		14		
47		6		
42		8		
49		6		
38		4		
50		8		
49		8		

Example 8

The data in the table represent the monthly sales and the promotional expenses for a store that specializes in sportswear for young women.

Month	Sales (in $1,000)	Promotional expenses (in $1,000)
1	62.4	3.9
2	68.5	4.8
3	70.2	5.5
4	79.6	6.0
5	80.1	6.8
6	88.7	7.7
7	98.6	7.9
8	104.3	9.0
9	106.5	9.2
10	107.3	9.7
11	115.8	10.9
12	120.1	11.0

(a) Calculate the coefficient of correlation between monthly sales and promotional expenses.

(b) Calculate the Spearman's rank correlation between monthly sales and promotional expenses.

(c) Compare your results from part a and part b. What do these results suggest about the linearity and association between the two variables?

6.6 Multiple Regression and Correlation Analysis

We may use more than one independent variable to estimate the dependent variable, and in this way, attempt to increase the accuracy of the estimate. This process is called multiple regression and correlation analysis. It is based on the same assumptions and procedures we have encountered using simple regression. The principal advantage of multiple regression is that it allows us to use more of the information available to us to estimate the dependent variable. Sometimes the correlation between two variables may be insufficient to determine a reliable estimating equation. Yet, if we add the data from more independent variables, we may be able to determine an estimating equation that describes the relationship with greater accuracy.

Considering the problem of estimating or predicting the value of a dependent variable *y* on the basis of a set of measurements taken on *p* independent variables $x_1, \ldots, x_p$, we shall assume a theoretical equation of the form:

$$\mu_{y|x_1,\ldots,x_p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p ,$$

where $\beta_0, \ldots, \beta_p$ are coefficient parameters to be estimated from the data. Denoting these estimates by $b_0, \ldots, b_p$, respectively, we can write the sample regression equation in the form:

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_p x_p ,$$

The coefficients in the model are estimated by the least-squares method. For a random sample of size *n* (i.e. *n* data points), the least-squares estimates are obtained such that the residual sum of squares (*SSE*) is minimized, where

$$SSE = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip} \right)^2 .$$

With only two independent variables (i.e. *p* = 2) the sample regression equation reduces to the form:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

The least-squares estimates b_0, b_1 and b_2 are obtained by solving the following normal equations simultaneously:

$$nb_0 + b_1\sum x_1 + b_2\sum x_2 = \sum y$$
$$b_0\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 = \sum x_1y$$
$$b_0\sum x_2 + b_1\sum x_1x_2 + b_2\sum x_2^2 = \sum x_2y$$

Example 9

A placement agency would like to predict the salary of senior staff (y) by his years of experience (x_1) and the number of employees he supervises (x_2). A random sample of 12 cases is selected and the observations are shown in the following table.

Salary ('000)	Year of experience	Number of employees supervised
62	10	175
65	12	150
72	18	135
70	15	175
81	20	150
77	18	200
72	19	180
77	22	225
75	20	175
90	21	275
82	19	225
95	23	300

$$\sum x_1 = 217 \quad \sum x_2 = 2365 \quad \sum y = 918$$
$$\sum x_1^2 = 4093 \quad \sum x_2^2 = 494375 \quad \sum x_1x_2 = 44025$$
$$\sum x_1y = 16947 \quad \sum x_2y = 185230 \quad \sum y^2 = 71230$$

The normal equations are:

$$12b_0 + 217b_1 + 2365b_2 = 918$$
$$217b_0 + 4093b_1 + 44025b_2 = 16947$$
$$2365b_0 + 44025b_1 + 494375b_2 = 185230$$

When we solve these three equations simultaneously, we get the least-squares estimates of the regression coefficients.

Alternatively, SPSS is used to analyze this set of sample data and gives the following results:

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Number of employee supervised, Year of experience[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Salary ('000)

Proportion of variability in y explained by the model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.931[a]	.866	.836	3.865

a. Predictors: (Constant), Number of employee supervised, Year of experience

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	868.563	2	434.281	29.073	.000[a]
	Residual	134.437	9	14.937		
	Total	1003.000	11			

a. Predictors: (Constant), Number of employee supervised, Year of experience

b. Dependent Variable: Salary ('000)

Test statistic for the significance of the regression model.

The probability of F greater than 29.073 equals to 0.000, therefore the regression model is significant at 5% level of significance.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	33.703	5.765		5.846	.000
	Year of experience	1.371	.364	.563	3.770	.004
	Number of employee supervised	9.136E-02	.028	.485	3.250	.010

a. Dependent Variable: Salary ('000)

Estimates of regression coefficients

Standard errors of estimates

Test statistics of regression coefficients

All are less than 0.05, hence, all the regression coefficients differ from zero significantly at 5% level of significance.

From the above results, the fitted regression equation is

$$\hat{y} = 33.703 + 1.371x_1 + 0.09x_2$$

Similar to the simple linear regression model, the sum of squares identity

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

will also hold.

Denote

total sum of squares, $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$,

regression sum of squares, $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ and

residual sum of squares, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$,

the identity becomes $SST = SSR + SSE$.

The coefficient of determination, R^2, is evaluated by

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

which states the percentage of variation of y that can be explained by the multiple linear regression model.

Given a fixed sample size n, R^2 will generally increase as more independent variables are included in the multiple regression equation. However, the additional independent variables may not contribute significantly to the explanation of the dependent variable.

6.6.1 Inferences on the parameters

The significance of individual regression coefficients can be tested. All the b_i's are assumed normally distributed with mean β_i.

The null hypothesis and alternative hypothesis are

$H_0 : \beta_i = 0$ (i.e. x_i is not a significant explanatory variable)

$H_1 : \beta_i \neq 0$ (i.e. x_i is a significant explanatory variable)

We can test these hypotheses using the t-test. The test statistic

$$t = \frac{b_i}{s.e.(b_i)}$$

follows the t-distribution with n-p-1 degrees of freedom. Note that $s.e.(b_i)$ is the standard error of b_i.

Using the SPSS results of Example 9 again, the standard errors of b_1 and b_2 are 0.364 and 0.028 respectively and the corresponding test statistics are 3.770 and 3.250. The significances of b_1 and b_2 are 0.004 and 0.01 respectively, and hence we reject H_0 and conclude that both independent variables (i.e. x_1 and x_2) are significant explanatory variables of y at 5% level of significance.

6.6.2 Analysis of Variance (ANOVA) approach

The analysis of variance approach is used to test for the significance of the multiple linear regression model. The null hypothesis and alternative hypothesis are

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (i.e. y does not depend on the x_i's)
$H_1:$ at least one $\beta_i \neq 0$ (i.e. y depends on at least one of the x_i's)

After evaluating the sum of squares, the ANOVA table is constructed as follows:

Source	SS	df	MS	F
Regression	SSR	p	$MSR = SSR/p$	MSR/MSE
Residual	SSE	n-p-1	$MSE = SSE/(n\text{-}p\text{-}1)$	
Total	SST	n-1		

The test statistic $F = \dfrac{MSR}{MSE}$ follows the F distribution with p and n-p-1 degrees of freedom under the null hypothesis. If $F > F_{\alpha,p,n-p-1}$, there is evidence to reject the null hypothesis.

Example 9 has the F-statistic = 29.073 with significance 0.000, therefore the multiple regression equation is highly significant.

6.6.3 Multicollinearity in Multiple Regression

In multiple-regression analysis, the regression coefficients often become less reliable as the degree of correlation between the independent variables increases. If there is a high level of correlation between some of the independent variables, we have a problem that statisticians call *multicollinearity*.

Multicollinearity might occur if we wished to estimate a firm's sales revenue and we used both the number of salespeople employed and their total salaries. Because the values associated with these two independent variables are highly correlated, we need to use only one set of them to make our estimate. In fact, adding a second variable that is correlated with the first distorts the values of the regression coefficients.

EXERCISE: LINEAR REGRESSION AND CORRELATION

1. The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Find the equation of the regression line.

(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report but was ill at the time of the final examination.

2. (a) From the following information draw a scatter diagram and by the method of least squares draw the regression line of best fit.

Volume of sales (thousand units), x	5	6	7	8	9	10
Total expenses (thousand \$), y	74	77	82	86	92	95

(b) What will be the total expenses when the volume of sales is 7,500 units?

(c) If the selling price per unit is \$11, at what volume of sales will the total income from sales equal the total expenses?

3. The following data show the unit cost of producing certain electronic components and the number of units produced:

Lot size, x	50	100	250	500	1000
Unit cost, y	\$108	\$53	\$24	\$9	\$5

It is believed that the regression equation is of the form

$$y = ax^b .$$

By simple linear regression technique or otherwise estimate the unit cost for a lot of 400 components.

4. Two variables x and y are related by the law:

$$y = \alpha x + \beta x^2 .$$

State how α and β can be estimated by the simple linear regression technique.

5. Compute and interpret the correlation coefficient for the following grades of 6 students selected at random.

Mathematics grade	70	92	80	74	65	83
English grade	74	84	63	87	78	90

6. The following table shows a traffic-flow index and the related site costs in respect of eight service stations of ABC Garages Ltd.

Site No.	Traffic-flow index	Site cost (in 1000)
1	100	100
2	110	115
3	119	120
4	123	140
5	123	135
6	127	175
7	130	210
8	132	200

(a) Calculate the coefficient of correlation for this data.
(b) Calculate the coefficient of rank correlation.

7. As a result of standardized interviews, an assessment was made of the IQ and the attitude to the employing company of a group of six workers. The IQ's were expressed as whole numbers within the range 50-150 and the attitudes are assigned to five grades labeled 1, 2, 3, 4 and 5 in order of decreasing approval. The results obtained are summarized below:

Employee	A	B	C	D	E	F
IQ	127	85	94	138	104	70
Attitude score	2	4	3	1	2	5

Is there evidence of an association between the two attributes?

8. For the following multiple regression equation:

$$\hat{y} = 50 - 2x_1 + 7x_2 \; with \; R^2 = 0.40$$

(a) Interpret the meaning of the slopes.
(b) Interpret the meaning of the Y intercept.
(c) Interpret the meaning of the coefficient of multiple determination R^2.

9. The following ANOVA summary table was obtained from a multiple regression model with two independent variables.

Source	**Degrees of freedom**	**Sum of squares**	**Mean squares**	***F***
Regression	2	30		
Error	10	120		
Total	12	150		

(a) Determine the mean square that is due to regression and the mean square that is due to error.

(b) Determine the computed ***F*** statistic.

(c) Determine whether there is a significant relationship between Y and the two explanatory variables at the 0.05 level of significance.

10. Given the following information from a multiple regression analysis

$$n = 25, b_1 = 5, b_2 = 10, S_{b_1} = 2, S_{b_2} = 8 \text{, where } S_{b_i} = \text{standard error of } b_i$$

(a) Which variable has the largest slope in units of a ***t*** statistic?

(b) At the 0.05 level of significance, determine whether each explanatory variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables that should be included in this model.

11. Amy trying to purchase a used Toyota car has researched the prices. She believes the year of the car and the number of miles the car has been driven both influence the purchase price. Data are given below for 10 cars with the price (***Y***) in thousands of dollars, year (***X_1***), and miles driven (***X_2***) in thousands.

(Y) Price ($ thousands)	(X_1) Year	(X_2) Miles (thousands)
2.99	1987	55.6
6.02	1992	18.4
8.87	1993	21.3
3.92	1988	46.9
9.55	1994	11.8
9.05	1991	36.4
9.37	1992	28.2
4.2	1988	44.2
4.8	1989	34.9
5.74	1991	26.4

(a) Using SPSS, fit the least-squares equation that best relates these three variables.

(b) Amy would like to purchase a 1991 Toyota with about 40,000 miles on it. How much do you predict she will pay?

12. Steven Reich, a statistics professor in a leading business school, has a keen interest in factors affecting students' performance on exams. The midterm exam for the past semester had a wide distribution of grades, but Steven feels certain that several factors explain the distribution: He allowed his students to study from as many different books as they liked, their IQs vary, they are of different ages, and they study varying amounts of time for exams. To develop a predicting formula for exam grades, Steven asked each student to answer, at the end of the exam, questions regarding study time and number of books used. Steven's teaching records already contained the IQs and ages for the students, so he compiled the data for the class and ran a multiple regression with a computer package. The output from Steven's computer run was as follows:

Predictor	Coef	Stdev	t-ratio	p
Constant	-49.948	41.55	-1.20	0.268
HOURS	1.06931	0.98163	1.09	0.312
IQ	1.36460	0.37627	3.63	0.008
BOOKS	2.03982	1.50799	1.35	0.218
AGE	-1.79890	0.67332	-2.67	0.319
s = 11.657	R-sq = 76.7%			

(a) What is the least squares regression equation for these data?
(b) What percentage of the variation in grades is explained by this equation?
(c) What grade would you expect for a 21-year-old student with an IQ of 113, who studied 5 hours and used three different books?

13. Refer to Q12. The following additional output was provided by the computer when Steven ran the multiple regression:

Analysis of Variance

Source	DF	SS	MS	F	p
Regression	4	3134.42	783.60		
Error	7	951.25	135.89		
Total	11	4085.67			

(a) What is the observed value of ***F***?
(b) At a significance level of 0.05, what is the appropriate critical value of ***F*** to use in determining whether the regression as a whole is significant?
(c) Based on your answers to part (a) and part (b), is the regression significant as a whole?

14. A New Canada-based commuter airline has taken a survey of its 15 terminals and has obtained the following data for the month of February, where

SALES = total revenue based on number of tickets sold (in thousands of dollars)
PROMOT = amount spent on promoting the airline in the area (in thousands of dollars)
COMP = number of competing airlines at that terminal
FREE = the percentage of passengers who flew free (for various reasons)

Sales($)	**Promot($)**	**Comp**	**Free**
79.3	2.5	10	3
200.1	5.5	8	6
163.2	6.0	12	9
200.1	7.9	7	16
146.0	5.2	8	15
177.7	7.6	12	9
30.9	2.0	12	8
291.9	9.0	5	10
160.0	4.0	8	4
339.4	9.6	5	16
159.6	5.5	11	7
86.3	3.0	12	6
237.5	6.0	6	10
107.2	5.5	10	4
155.0	3.5	10	4

Predictor	Coef	Stdev	t-ratio	p
Constant	172.34	51.38	3.35	0.006
PROMOT	25.950	4.877	5.32	0.000
COMP	-13.238	3.686	-3.59	0.004
FREE	-3.041	2.342	-1.30	0.221

(a) Use the above computer output to determine the least-squares regression equation for the airline to predict sales.

(b) Do the percentage of passengers who fly free cause sales to decrease significantly? State and test appropriate hypothesis. Use $\alpha = 0.05$.

15. Alex Yeung, manager of Star Shine's Diamond and Jewellery Store, is interested in developing a model to estimate consumer demand for his rather expensive merchandise. Because most customers buy diamonds and jewelry on credit, Alex is sure that two factors that must influence consumer demand are the current annual inflation rate and the current lending rate at the leading banks in UK. Explain some of the problems that Alex might encounter if he were to set up a regression model based on his two predictor variables.