CHAPTER 20

Stage 1
Problem definition

Stage 2
Research approach developed

Stage 3
Research design developed

Stage 4
Fieldwork or data collection

Stage 5
Data preparation and analysis

Stage 6
Report preparation and presentation

Correlation and regression

Objectives

After reading this chapter, you should be able to:

1 discuss the concepts of product moment correlation, partial correlation and part correlation, and show how they provide a foundation for regression analysis;

2 explain the nature and methods of bivariate regression analysis and describe the general model, estimation of parameters, standardised regression coefficient, significance testing, prediction accuracy, residual analysis and model cross-validation;

3 explain the nature and methods of multiple regression analysis and the meaning of partial regression coefficients;

4 describe specialised techniques used in multiple regression analysis, particularly stepwise regression, regression with dummy variables, and analysis of variance and covariance with regression;

5 discuss non-metric correlation and measures such as Spearman's rho and Kendall's tau.

Correlation is the simplest way to understand the association between two metric variables. When extended to multiple regression, the relationship between one variable and several others becomes more clear.

Overview

Chapter 19 examined the relationship among the *t* test, analysis of variance and covariance, and regression. This chapter describes regression analysis, which is widely used for explaining variation in market share, sales, brand preference and other marketing results. This is done in terms of marketing management variables such as advertising, price, distribution and product quality. Before discussing regression, however, we describe the concepts of product moment correlation and partial correlation coefficient, which lay the conceptual foundation for regression analysis.

In introducing regression analysis, we discuss the simple bivariate case first. We describe estimation, standardisation of the regression coefficients, and testing and examination of the strength and significance of association between variables, prediction accuracy, and the assumptions underlying the regression model. Next, we discuss the multiple regression model, emphasising the interpretation of parameters, strength of association, significance tests and examination of residuals.

We then cover topics of special interest in regression analysis, such as stepwise regression, multicollinearity, relative importance of predictor variables, and cross-validation. We describe regression with dummy variables and the use of this procedure to conduct analysis of variance and covariance. We begin with two examples that illustrate applications of regression analysis.

example

GlobalCash Project

Multiple regression

In the GlobalCash Project, multiple regression analysis was used to develop a model that explained 'bank preference' in terms of respondents' evaluations of the banks in their own countries, through four choice criteria. The dependent variable was the preference for individual banks. The independent variables were the evaluations of each bank on balance reporting, domestic payments and collections, international payments and collections, and managing currencies. The results indicated that all the factors of the choice criteria, except managing currencies, were significant in explaining bank preference. The coefficients of all the variables were positive, indicating that higher evaluations on each of the significant factors led to higher preference for that bank. The model had a good fit and good ability to predict bank preference. ■

example

Regression rings the right bell for Avon[1]

Avon Products were having significant problems with their sales staff. The company's business, dependent on sales representatives, was facing a shortage of sales representatives without much hope of getting new ones. Regression models were developed to reveal the possible variables that were fuelling this situation. The models revealed that the most significant variable was the level of the appointment fee that reps paid for materials. With data to back up its actions, the company lowered the fee. This resulted in an improvement in the recruitment and retention of sales reps. ■

These examples illustrate some of the uses of regression analysis in determining which independent variables explain a significant variation in the dependent variable of interest, the structure and form of the relationship, the strength of the relationship, and predicted values of the dependent variable. Fundamental to regression analysis is an understanding of the product moment correlation.

Product moment correlation

In marketing research, we are often interested in summarising the strength of association between two metric variables, as in the following situations:

- How strongly are sales related to advertising expenditures?
- Is there an association between market share and size of the sales force?
- Are consumers' perceptions of quality related to their perceptions of prices?

Product moment correlation (r)
A statistic summarising the strength of association between two metric variables.

In situations like these, the **product moment correlation (r)**, is the most widely used statistic, summarising the strength of association between two metric (interval or ratio scaled) variables, say X and Y. It is an index used to determine whether a linear or straight line relationship exists between X and Y. It indicates the degree to which the variation in one variable, X, is related to the variation in another variable, Y. Because it was originally proposed by Karl Pearson, it is also known as the *Pearson correlation coefficient* and also referred to as *simple correlation*, *bivariate correlation* or merely the *correlation coefficient*. From a sample of n observations, X and Y, the product moment correlation, r, can be calculated as

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Division of the numerator and denominator by $n - 1$ gives

$$r = \frac{\sum_{i=1}^{n}\dfrac{(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}}{\sqrt{\sum_{i=1}^{n}\dfrac{(X_i - \overline{X})^2}{n-1} \sum_{i=1}^{n}\dfrac{(Y_i - \overline{Y})^2}{n-1}}}$$

$$= \frac{COV_{xy}}{S_x S_y}$$

Covariance
A systematic relationship between two variables in which a change in one implies a corresponding change in the other (COV_{xy})

In these equations, $\overline{X}$ and $\overline{Y}$ denote the sample means, and S_x and S_y the standard deviations. COV_{xy}, the **covariance** between X and Y, measures the extent to which X and Y are related. The covariance may be either positive or negative. Division by S_xS_y achieves standardisation so that r varies between –1.0 and +1.0. Note that the correlation coefficient is an absolute number and is not expressed in any unit of measurement. The correlation coefficient between two variables will be the same regardless of their underlying units of measurement.

As an example, suppose that a researcher wants to explain attitudes towards a respondent's city of residence in terms of duration of residence in the city. The attitude is measured on an 11-point scale (1 = do not like the city, 11 = very much like the city), and the duration of residence is measured in terms of the number of years the respondent has lived in the city. In a pre-test of 12 respondents, the data shown in Table 20.1 are obtained.

Table 20.1 Explaining attitude towards the city of residence

Respondent number	*Attitude toward the city*	*Duration of residence*
1	6	10
2	9	12
3	8	12
4	3	4
5	10	12
6	4	6
7	5	8
8	2	2
9	11	18
10	9	9
11	10	17
12	2	2

The correlation coefficient may be calculated as follows:

$$\begin{aligned}\bar{X} &= (10 + 12 + 12 + 4 + 12 + 6 + 8 + 2 + 18 + 9 + 17 + 2)/12\\ &= 9.333\end{aligned}$$

$$\begin{aligned}\bar{Y} &= (6 + 9 + 8 + 3 + 10 + 4 + 5 + 2 + 11 + 9 + 10 + 2)/12\\ &= 6.583\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= (10 - 9.33)(6 - 6.58) + (12 - 9.33)(9 - 6.58)\\ &\quad + (12 - 9.33)(8 - 6.58) + (4 - 9.33)(3 - 6.58)\\ &\quad + (12 - 9.33)(10 - 6.58) + (6 - 9.33)(4 - 6.58)\\ &\quad + (8 - 9.33)(5 - 6.58) + (2 - 9.33)(2 - 6.58)\\ &\quad + (18 - 9.33)(11 - 6.58) + (9 - 9.33)(9 - 6.58)\\ &\quad + (17 - 9.33)(10 - 6.58) + (2 - 9.33)(2 - 6.58)\\ &= -0.3886 + 6.4614 + 3.7914 + 19.0814 + 9.1314 + 8.5914\\ &\quad + 2.1014 + 33.5714 + 38.3214 - 0.7986 + 26.2314 + 33.5714\\ &= 179.6668\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= (10 - 9.33)^2 + (12 - 9.33)^2 + (12 - 9.33)^2 + (4 - 9.33)^2\\ &\quad + (12 - 9.33)^2 + (6 - 9.33)^2 + (8 - 9.33)^2 + (2 - 9.33)^2 + (18 - 9.33)^2\\ &\quad + (9 - 9.33)^2 + (17 - 9.33)^2 + (2 - 9.33)^2\\ &= 0.4489 + 7.1289 + 7.1289 + 28.4089 + 7.1289 + 11.0889 + 1.7689\\ &\quad + 53.7289 + 75.1689 + 0.1089 + 58.8289 + 53.7289\\ &= 304.6668\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= (6 - 6.58)^2 + (9 - 6.58)^2 + (8 - 6.58)^2 + (3 - 6.58)^2 + (3 - 6.58)^2\\ &\quad + (10 - 6.58)^2 + (4 - 6.58)^2 + (5 - 6.58)^2 + (2 - 6.58)^2 + (11 - 6.58)^2\\ &\quad + (9 - 6.58)^2 + (10 - 6.58)^2 + (2 - 6.58)^2\\ &= 0.3364 + 5.8564 + 2.0164 + 12.8164 + 11.6964 + 6.6564 + 2.4964\\ &\quad + 20.9764 + 19.5364 + 5.8564 + 11.6964 + 20.9764\\ &= 120.9168\end{aligned}$$

Thus,

$$r = \frac{179.6668}{\sqrt{(304.6668)(120.9168)}}$$

$$= 0.9361$$

In this example, $r = 0.9361$, a value close to 1.0. This means that respondents' duration of residence in the city is strongly associated with their attitude towards the city. Furthermore, the positive sign of r implies a positive relationship; the longer the duration of residence, the more favourable the attitude and vice versa.

Since r indicates the degree to which variation in one variable is related to variation in another, it can also be expressed in terms of the decomposition of the total variation (see Chapter 19). In other words,

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

$$= \frac{SS_x}{SS_y}$$

$$= \frac{\text{total variation} - \text{error variation}}{\text{total variation}}$$

$$= \frac{SS_y - SS_{error}}{SS_y}$$

Hence, r^2 measures the proportion of variation in one variable that is explained by the other. Both r and r^2 are symmetric measures of association. In other words, the correlation of X with Y is the same as the correlation of Y with X. It does not matter which variable is considered to be the dependent variable and which the independent. The product moment coefficient measures the strength of the linear relationship and is not designed to measure non-linear relationships. Thus $r = 0$ merely indicates that there is no linear relationship between X and Y. It does not mean that X and Y are unrelated. There could well be a non-linear relationship between them, which would not be captured by r (see Figure 20.1).

When computed for a population rather than a sample, the product moment correlation is denoted by the Greek letter rho, ρ. The coefficient r is an estimator of ρ. Note that the calculation of r assumes that X and Y are metric variables whose distributions have the same shape. If these assumptions are not met, r is deflated and

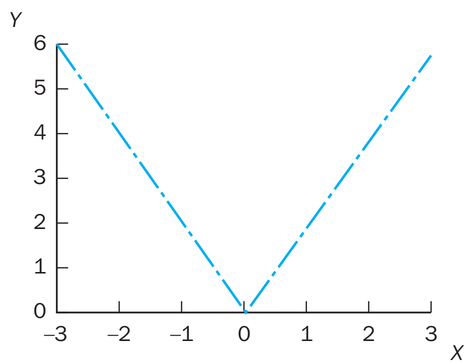


Figure 20.1
A non-linear relationship for which $r = 0$

underestimates ρ. In marketing research, data obtained by using rating scales with a small number of categories may not be strictly interval. This tends to deflate r, resulting in an underestimation of ρ.[2]

The statistical significance of the relationship between two variables measured by using r can be conveniently tested. The hypotheses are

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

The test statistic is

$$t = r\left[\frac{n-1}{1-r^2}\right]^{\frac{1}{2}}$$

which has a t distribution with $n - 2$ degrees of freedom.[3] For the correlation coefficient calculated based on the data given in Table 20.1,

$$t = 0.9361\left[\frac{12-2}{1-(0.9361^2)}\right]^{\frac{1}{2}}$$
$$= 8.414$$

and the degrees of freedom = 12 – 2 = 10. From the t distribution table (Table 4 in the Statistical Appendix), the critical value of t for a two-tailed test and $\alpha = 0.05$ is 2.228. Hence, the null hypothesis of no relationship between X and Y is rejected. This, along with the positive sign of r, indicates that attitude towards the city is positively related to the duration of residence in the city. Moreover, the high value of r indicates that this relationship is strong.

In conducting multivariate data analysis, it is often useful to examine the simple correlation between each pair of variables. These results are presented in the form of a correlation matrix, which indicates the coefficient of correlation between each pair of variables. Usually, only the lower triangular portion of the matrix is considered. The diagonal elements all equal 1.00, since a variable correlates perfectly with itself. The upper triangular portion of the matrix is a mirror image of the lower triangular portion, since r is a symmetric measure of association. The form of a correlation matrix for five variables, V_1 to V_5 is as follows:

	V_1	V_2	V_3	V_4	V_5
V_1					
V_2	0.5				
V_3	0.3	0.4			
V_4	0.1	0.3	0.6		
V_5	0.2	0.5	0.3	0.7	

Although a matrix of simple correlations provides insights into pairwise associations, sometimes researchers want to examine the association between two variables after controlling for one or more other variables. In the latter case, partial correlation should be estimated.

Partial correlation

Partial correlation coefficient
A measure of the association between two variables after controlling or adjusting for the effects of one or more additional variables.

Whereas the product moment or simple correlation is a measure of association describing the linear association between two variables, a **partial correlation coefficient** measures the association between two variables after controlling for or adjusting for the effects of one or more additional variables. This statistic is used to answer the following questions:

- How strongly are sales related to advertising expenditures when the effect of price is controlled?
- Is there an association between market share and size of the sales force after adjusting for the effect of sales promotion?
- Are consumers' perceptions of quality related to their perceptions of prices when the effect of brand image is controlled?

As in these situations, suppose that a researcher wanted to calculate the association between *X* and *Y* after controlling for a third variable, *Z*. Conceptually, one would first remove the effect of *Z* from *X*. To do this, one would predict the values of *X* based on a knowledge of *Z* by using the product moment correlation between *X* and *Z*, r_{xz}. The predicted value of *X* is then subtracted from the actual value of *X* to construct an adjusted value of *X*. In a similar manner, the values of *Y* are adjusted to remove the effects of *Z*. The product moment correlation between the adjusted values of *X* and the adjusted values of *Y* is the partial correlation coefficient between *X* and *Y*, after controlling for the effect of *Z*, and is denoted by $r_{xy \cdot z}$. Statistically, since the simple correlation between two variables completely describes the linear relationship between them, the partial correlation coefficient can be calculated by a knowledge of the simple correlations alone, without using individual observations.

$$r_{xy \cdot z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2}\ \sqrt{1 - r_{yz}^2}}$$

To continue our example, suppose that the researcher wanted to calculate the association between attitude towards the city, *Y*, and duration of residence, X_1, after controlling for a third variable, importance attached to weather, X_2. These data are presented in Table 20.2.

Table 20.2 Explaining attitude towards the city of residence, including 'importance attached to weather'

Respondent number	*Attitude towards the city*	*Duration of residence*	*Importance attached to weather*
1	6	10	3
2	9	12	11
3	8	12	4
4	3	4	1
5	10	12	11
6	4	6	1
7	5	8	7
8	2	2	4
9	11	18	8
10	9	9	10
11	10	17	8
12	2	2	5

The simple correlations between the variables are

$$r_{yx_1} = 0.9361 \qquad r_{yx_2} = 0.7334 \qquad r_{x_1x_2} = 0.5495$$

The required partial correlation may be calculated as follows:

$$r_{xy_1 \cdot x_2} = \frac{0.9361 - (0.5495)(0.7334)}{\sqrt{1 - (0.5495)^2}\ \sqrt{1 - (0.7334)^2}}$$

$$= 0.9386$$

As can be seen, controlling for the effect of importance attached to weather has little effect on the association between attitude towards the city and duration of residence.

Partial correlations have an *order* associated with them that indicates how many variables are being adjusted or controlled for. The simple correlation coefficient, r, has a zero order, because it does not control for any additional variables while measuring the association between two variables. The coefficient $r_{xy \cdot z}$ is a first-order partial correlation coefficient, because it controls for the effect of one additional variable, Z. A second-order partial correlation coefficient controls for the effects of two variables, a third-order for the effects of three variables, and so on. The higher-order partial correlations are calculated similarly. The $(n + 1)$th order partial coefficient may be calculated by replacing the simple correlation coefficients on the right side of the preceding equation with the nth order partial coefficients.

Partial correlations can be helpful for detecting spurious relationships (see Chapter 18). The relationship between X and Y is spurious if it is solely because X is associated with Z, which is indeed the true predictor of Y. In this case, the correlation between X and Y disappears when the effect of Z is controlled. Consider a case in which consumption of a breakfast cereal brand (C) is positively associated with income (I), with $r_{ci} = 0.28$. Because this brand was popularly priced, income was not expected to be a significant factor. Therefore, the researcher suspected that this relationship was spurious. The sample results also indicated that income is positively associated with household size (H), $r_{hi} = 0.48$, and that household size is associated with breakfast cereal consumption, $r_{ch} = 0.56$. These figures seem to indicate that the real predictor of breakfast cereal consumption is not income but household size. To test this assertion, the first-order partial correlation between cereal consumption and income is calculated, controlling for the effect of household size. The reader can verify that this partial correlation, $r_{ci \cdot h}$, is 0.02, and the initial correlation between cereal consumption and income vanishes when the household size is controlled. Therefore, the correlation between income and cereal consumption is spurious. The special case when a partial correlation is larger than its respective zero-order correlation involves a suppressor effect (see Chapter 18).[4]

Part correlation coefficient
A measure of the correlation between Y and X when the linear effects of the other independent variables have been removed from X (but not from Y).

Another correlation coefficient of interest is the **part correlation coefficient**. This coefficient represents the correlation between Y and X when the linear effects of the other independent variables have been removed from X but not from Y. The part correlation coefficient, $r_{y(x \cdot z)}$, is calculated as follows:

$$r_{y(x \cdot z)} = \frac{r_{xy} - r_{yz} r_{xz}}{\sqrt{\qquad 1}}$$

The part correlation between attitude towards the city and the duration of residence, when the linear effects of the importance attached to weather have been removed from the duration of residence, can be calculated as

$$r_{y(x_1 \cdot x_2)} = \frac{0.9361 - (0.5495)(0.7334)}{\sqrt{1 - (0.5495)^2}}$$

$$= 0.63806$$

The partial correlation coefficient is generally viewed as more important than the part correlation coefficient. The product moment correlation, partial correlation and part correlation coefficient all assume that the data are interval or ratio scaled. If the data do not meet these requirements, the researcher should consider the use of non-metric correlation.

example

Selling ads to home shoppers[5]

Advertisements play a very important role in forming attitudes and preferences for brands. In general, it has been found that for low-involvement products, attitude towards the advertisement mediates brand cognition (beliefs about the brand) and attitude towards the brand. What would happen to the effect of this mediating variable when products are purchased through a home shopping network? Home Shopping Budapest in Hungary conducted research to assess the impact of advertisements towards purchase. A survey was conducted in which several measures were taken, such as attitude towards the product, attitude towards the brand, attitude towards the ad characteristics and brand cognitions. It was hypothesised that in a home shopping network, advertisements largely determined attitude towards the brand. To find the degree of association of attitude towards the ad with both attitude towards the brand and brand cognition, a partial correlation coefficient could be computed. The partial correlation would be calculated between attitude towards the brand and brand cognitions after controlling for the effects of attitude towards the ad on the two variables. If attitude towards the ad is significantly high, then the partial correlation coefficient should be significantly less than the product moment correlation between brand cognition and attitude towards the brand. Research was conducted which supported this hypothesis. Then Saatchi & Saatchi designed the ads aired on Home Shopping Budapest to generate positive attitude towards the advertising. This turned out to be a major competitive weapon for the network.

Non-metric correlation

Non-metric correlation
A correlation measure for two non-metric variables that relies on rankings to compute the correlation.

At times the researcher may have to compute the correlation coefficient between two variables that are non-metric. It may be recalled that non-metric variables do not have interval or ratio scale properties and do not assume a normal distribution. If the non-metric variables are ordinal and numeric, Spearman's rho, ρ_s, and Kendall's tau, τ, are two measures of **non-metric correlation** which can be used to examine the correlation between them. Both these measures use rankings rather than the absolute values of the variables, and the basic concepts underlying them are quite similar. Both vary from −1.0 to +1.0.

In the absence of ties, Spearman's ρ_s yields a closer approximation to the Pearson product moment correlation coefficient, r, than does Kendall's τ. In these cases, the absolute magnitude of τ tends to be smaller than Pearson's r. On the other hand, when the data contain a large number of tied ranks, Kendall's τ seems more appropriate. As a rule of thumb, Kendall's τ is to be preferred when a large number of cases fall into a relatively small number of categories (thereby leading to a large number of ties). Conversely, the use of Spearman's ρ_s is preferable when we have a relatively larger number of categories (thereby having fewer ties).[6]

The product moment as well as the partial and part correlation coefficients provide a conceptual foundation for bivariate as well as multiple regression analysis.

Regression analysis

Regression analysis
A statistical procedure for analysing associative relationships between a metric-dependent variable and one or more independent variables.

Regression analysis is a powerful and flexible procedure for analysing associative relationships between a metric-dependent variable and one or more independent variables. It can be used in the following ways:

1. To determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.
2. To determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship.
3. To determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables.
4. To predict the values of the dependent variable.
5. To control for other independent variables when evaluating the contributions of a specific variable or set of variables.

Although the independent variables may explain the variation in the dependent variable, this does not necessarily imply causation. The use of the terms dependent or criterion variables and independent or predictor variables in regression analysis arises from the mathematical relationship between the variables. These terms do not imply that the criterion variable is dependent on the independent variables in a causal sense. Regression analysis is concerned with the nature and degree of association between variables and does not imply or assume any causality. Bivariate regression is discussed first, followed by multiple regression.

Bivariate regression

Bivariate regression
A procedure for deriving a mathematical relationship, in the form of an equation, between a single metric-dependent variable and a single metric-independent variable.

Bivariate regression is a procedure for deriving a mathematical relationship, in the form of an equation, between a single metric-dependent or criterion variable and a single metric-independent or predictor variable. The analysis is similar in many ways to determining the simple correlation between two variables. Since an equation has to be derived, however, one variable must be identified as the dependent variable and the other as the independent variable. The examples given earlier in the context of simple correlation can be translated into the regression context.

- Can variation in sales be explained in terms of variation in advertising expenditures? What is the structure and form of this relationship, and can it be modelled mathematically by an equation describing a straight line?
- Can the variation in market share be accounted for by the size of the sales force?
- Are consumers' perceptions of quality determined by their perceptions of price?

Before discussing the procedure for conducting bivariate regression, we define some important statistics associated with bivariate regression analysis.

Bivariate regression model. The basic regression equation is $Y_i = \beta_0 + \beta_1 X_i + e_i$, where Y = dependent or criterion variable, X = independent or predictor variable, β_0 = intercept of the line, β_1 = slope of the line, and e_i is the error term associated with the ith observation.

Coefficient of determination. The strength of association is measured by the coefficient of determination, r^2. It varies between 0 and 1 and signifies the proportion of the total variation in Y that is accounted for by the variation in X.

Estimated or predicted value. The estimated or predicted value of Y_i is $\hat{Y}_i = a + bx$, where $\hat{Y}_i$ is the predicted value of Y_i, and a and b are estimators of β_0 and β_1, respectively.

Regression coefficient. The estimated parameter b is usually referred to as the non-standardised regression coefficient.

Scattergram. A scatter diagram, or scattergram, is a plot of the values of two variables for all the cases or observations.

Standard error of estimate. This statistic, the SEE, is the standard deviation of the actual Y values from the predicted $\hat{Y}$ values.

Standard error. The standard deviation of b, SE_b, is called the standard error.

Standardised regression coefficient. Also termed the beta coefficient or beta weight, this is the slope obtained by the regression of Y on X when the data are standardised.

Sum of squared errors. The distances of all the points from the regression line are squared and added together to arrive at the sum of squared errors, which is a measure of total error, Σe_j^2.

***t* statistic.** A t statistic with $n - 2$ degrees of freedom can be used to test the null hypothesis that no linear relationship exists between X and Y, or $H_0 : \beta_1 = 0$, where

$$t = \frac{b}{\text{SE}_b}$$

Conducting bivariate regression analysis

The steps involved in conducting bivariate regression analysis are described in Figure 20.2.

Plot the scatter diagram

Suppose that the researcher wants to explain attitudes towards the city of residence in terms of the duration of residence (see Table 20.2). In deriving such relationships, it is often useful to first examine a scatter diagram. A scatter diagram, or scattergram, is a

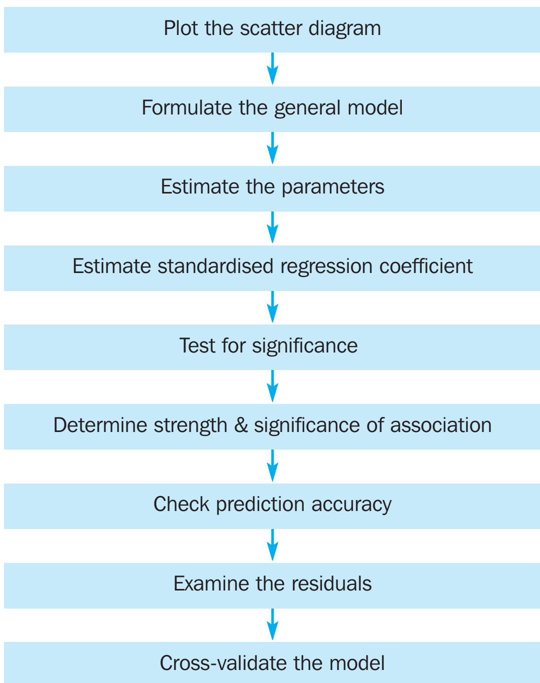


Figure 20.2
Conducting bivariate regression analysis

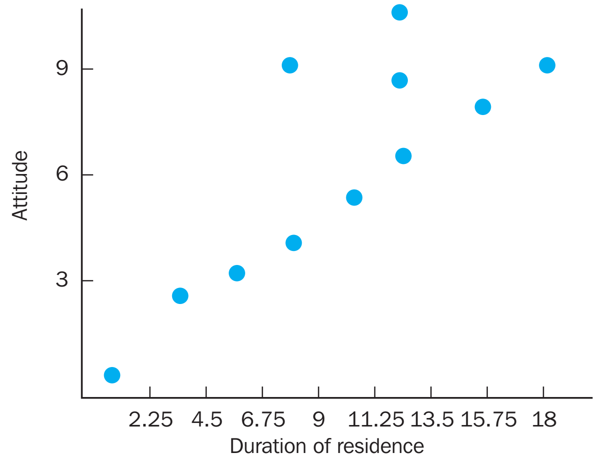


Figure 20.3
Plot of attitude with duration

plot of the values of two variables for all the cases or observations. It is customary to plot the dependent variable on the vertical axis and the independent variable on the horizontal axis. A scatter diagram is useful for determining the form of the relationship between the variables. A plot can alert the researcher to patterns in the data or to possible problems. Any unusual combinations of the two variables can be easily identified. A plot of Y (attitude towards the city) against X (duration of residence) is given in Figure 20.3. The points seem to be arranged in a band running from the bottom left to the top right. One can see the pattern: as one variable increases, so does the other. It appears from this scattergram that the relationship between X and Y is linear and could be well described by a straight line. How should the straight line be fitted to best describe the data?

The most commonly used technique for fitting a straight line to a scattergram is the **least squares procedure**. This technique determines the best-fitting line by minimising the vertical distances of all the points from the line. The best-fitting line is called the regression line. Any point that does not fall on the regression line is not fully accounted for. The vertical distance from the point to the line is the error, e_j (see Figure 20.4). The distances of all the points from the line are squared and added together to arrive at the sum of squared errors, which is a measure of total error, Σe_j^2.

Least squares procedure
A technique for fitting a straight line to a scattergram by minimising the vertical distances of all the points from the line.

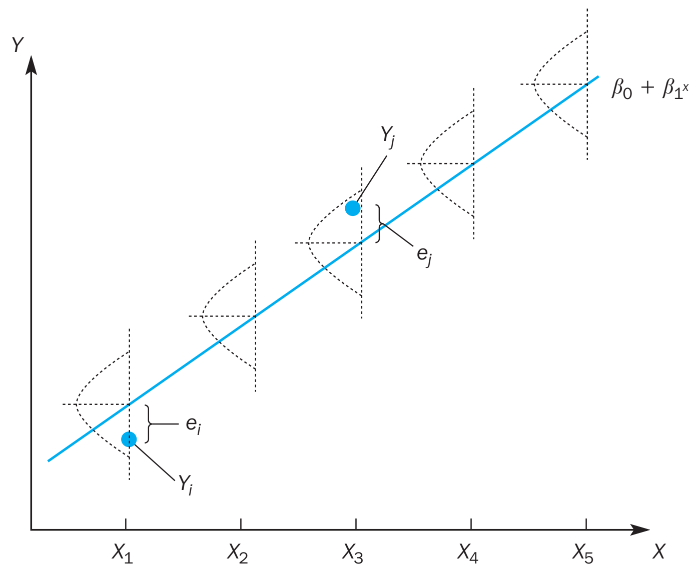


Figure 20.4
Bivariate regression

In fitting the line, the least squares procedure minimises the sum of squared errors. If Y is plotted on the vertical axis and X on the horizontal axis, as in Figure 20.4, the best fitting line is called the regression of Y on X, since the vertical distances are minimised. The scatter diagram indicates whether the relationship between Y and X can be modelled as a straight line and, consequently, whether the bivariate regression model is appropriate.

Formulate the general model

In the bivariate regression model, the general form of a straight line is

$$Y = \beta_0 + \beta_1 X$$

where Y = dependent or criterion variable
X = independent or predictor variable
β_0 = intercept of the line
β_1 = slope of the line.

This model implies a deterministic relationship in that Y is completely determined by X. The value of Y can be perfectly predicted if β_0 and β_1 are known. In marketing research, however, very few relationships are deterministic. Thus, the regression procedure adds an error term to account for the probabilistic or stochastic nature of the relationship. The basic regression equation becomes

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

where e_i is the error term associated with the ith observation.[7] Estimation of the regression parameters, β_0 and β_1, is relatively simple.

Estimate the parameters

In most cases, β_0 and β_1 are unknown and are estimated from the sample observations using the equation

$$\hat{Y}_i = a + bx_i$$

Estimated or predicted value
The value $Y_i = a + b_x$, where a and b are, respectively, estimators of β_0 and β_1, the corresponding population parameters.

where $\hat{Y}_i$ is the **estimated or predicted value** of Y_i, and a and b are estimators of β_0 and β_1 respectively. The constant b is usually referred to as the non-standardised regression coefficient. It is the slope of the regression line, and it indicates the expected change in Y when X is changed by one unit. The formulae for calculating a and b are simple.[8] The slope, b, may be computed in terms of the covariance between X and Y (COV_{xy}) and the variance of X as

$$b = \frac{COV_{xy}}{S_x^2}$$

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

$$= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

The intercept, a, may then be calculated using

$$a = \bar{Y} - b\bar{X}$$

For the data in Table 20.2, the estimation of parameters may be illustrated as follows:

$$\begin{aligned}\sum_{i=1}^{12} X_iY_i &= (10)(6) + (12)(9) + (12)(8) + (4)(3) + (12)(10) + (6)(4) + (8)(5) + (2)(2) \\ &\quad + (18)(11) + (9)(9) + (17)(10) + (2)(2) \\ &= 917\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{12} X_i^2 &= 10^2 + 12^2 + 12^2 + 4^2 + 12^2 + 6^2 + 8^2 + 2^2 + 18^2 + 9^2 + 17^2 + 2^2 \\ &= 1350\end{aligned}$$

It may be recalled from earlier calculations of the simple correlation that

$$\bar{X} = 9.333$$
$$\bar{Y} = 6.583$$

Given $n = 12$, b can be calculated as

$$\begin{aligned}b &= \frac{917 - (12)(9.333)(6.583)}{1350 - (12)(9.333)^2} \\ &= 0.5897 \\ a &= \bar{Y} - b\bar{X} \\ &= 6.583 - (0.5897)(9.333) \\ &= 1.0793\end{aligned}$$

Note that these coefficients have been estimated on the raw (untransformed) data. Should standardisation of the data be considered desirable, the calculation of the standardised coefficients is also straightforward.

Estimate the standardised regression coefficient

Standardisation is the process by which the raw data are transformed into new variables that have a mean of 0 and a variance of 1 (Chapter 17). When the data are standardised, the intercept assumes a value of 0. The term *beta coefficient* or *beta weight* is used to denote the standardised regression coefficient. In this case, the slope obtained by the regression of Y on X, B_{yx}, is the same as the slope obtained by the regression of X on Y, B_{xy}. Moreover, each of these regression coefficients is equal to the simple correlation between X and Y:

$$B_{yx} = B_{xy} = r_{xy}$$

There is a simple relationship between the standardised and non-standardised regression coefficients:

$$B_{yx} = b_{yx}\left(\frac{S_x}{S_y}\right)$$

For the regression results given in Table 20.3, the value of the beta coefficient is estimated as 0.9361.

Once the parameters have been estimated, they can be tested for significance.

Table 20.3 Bivariate regression

Multiple R	0.93608
R^2	0.87624
Adjusted R^2	0.86387
Standard error	1.22329

Analysis of variance			
	df	*Sum of squares*	*Mean square*
Regression	1	105.95222	105.95222
Residual	10	14.96444	1.49644
F = 70.80266	Significance of F = 0.0000		

Variables in the equation					
Variable	*b*	SE_b	β	*T*	*Sig. of T*
Duration	0.58972	0.07008	0.93608	8.414	0.0000
(Constant)	1.07932	0.74335		1.452	0.1772

Test for significance

The statistical significance of the linear relationship between X and Y may be tested by examining the hypotheses

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

The null hypothesis implies that there is no linear relationship between X and Y. The alternative hypothesis is that there is a relationship, positive or negative, between X and Y. Typically, a two-tailed test is done. A t statistic with $n - 2$ degrees of freedom can be used, where

$$t = \frac{b}{SE_b}$$

and SE_b denotes the standard deviation of b, called the *standard error*.[9] The t distribution was discussed in Chapter 18.

Using a software package, the regression of attitude on duration of residence, using the data shown in Table 20.2, yielded the results shown in Table 20.3. The intercept, a, equals 1.0793, and the slope, b, equals 0.5897. Therefore, the estimated equation is

$$\text{attitude } (\hat{Y}) = 1.0793 + 0.5897 \text{ (duration of residence)}$$

The standard error or standard deviation of b is estimated as 0.07008, and the value of the t statistic, t = 0.5897/0.0701 = 8.414, with $n - 2$ = 10 degrees of freedom. From Table 4 in the Statistical Appendix, we see that the critical value of t with 10 degrees of freedom and α = 0.05 is 2.228 for a two-tailed test. Since the calculated value of t is larger than the critical value, the null hypothesis is rejected. Hence, there is a significant linear relationship between attitude towards the city and duration of residence in the city. The positive sign of the slope coefficient indicates that this relationship is positive. In other words, those who have lived in the city for a longer time have more positive attitudes towards the city.

Determine strength and significance of association

A related inference involves determining the strength and significance of the association between Y and X. The strength of association is measured by the coefficient of determination, r^2. In bivariate regression, r^2 is the square of the simple correlation coefficient obtained by correlating the two variables. The coefficient r^2 varies between 0 and 1. It signifies the proportion of the total variation in Y that is accounted for by the variation in X. The decomposition of the total variation in Y is similar to that for analysis of variance (Chapter 19). As shown in Figure 20.5, the total variation SS_y may be decomposed into the variation accounted for by the regression line, SS_{reg}, and the error or residual variation, SS_{error} or SS_{res}, as follows:

$$SS_y = SS_{reg} + SS_{res}$$

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

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

The strength of the association may then be calculated as follows:

$$r^2 = \frac{SS_{reg}}{SS_y}$$

$$= \frac{SS_y - SS_{res}}{SS_y}$$

To illustrate the calculations of r^2, let us consider again the effect of attitude towards the city on the duration of residence. It may be recalled from earlier calculations of the simple correlation coefficient that

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

$$= 120.9168$$

The predicted values $(\hat{Y})$ can be calculated using the regression equation

$$\text{attitude } (\hat{Y}) = 1.0793 + 0.5897 \text{ (duration of residence)}$$

For the first observation in Table 20.2, this value is

$$(\hat{Y}) = 1.0793 + (0.5897 \times 10) = 6.9763$$

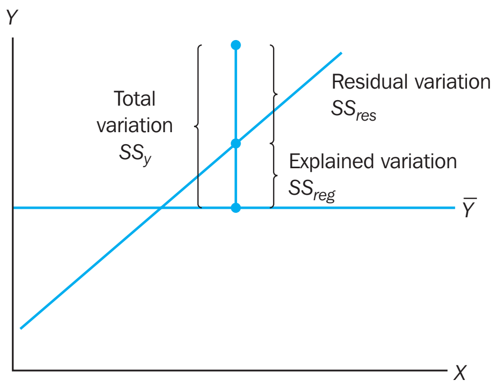


Figure 20.5
Decomposition of the total variation in bivariate regression

For each successive observation, the predicted values are, in order, 8.1557, 8.1557, 3.4381, 8.1557, 4.6175, 5.7969, 2.2587, 11.6939, 6.3866, 11.1042, 2.2587. Therefore,

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

$$\begin{aligned}
&= (6.9763 - 6.5833)^2 + (8.1557 - 6.5833)^2 + (8.1557 - 6.5833)^2 \\
&\quad + (3.4381 - 6.5833)^2 + (8.1557 - 6.5833)^2 + (4.6175 - 6.5833)^2 \\
&\quad + (5.7969 - 6.5833)^2 + (2.2587 - 6.5833)^2 + (11.6939 - 6.5833)^2 \\
&\quad + (6.3866 - 6.5833)^2 + (11.1042 - 6.5833)^2 + (2.2587 - 6.5833)^2 \\
&= 0.1544 + 2.4724 + 2.4724 + 9.8922 + 2.4724 + 3.8643 + 0.6184 \\
&\quad + 18.7021 + 26.1182 + 0.0387 + 20.4385 + 18.7021 \\
&= 105.9466
\end{aligned}$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\begin{aligned}
&= (6 - 6.9763)^2 + (9 - 8.1557)^2 + (8 - 8.1557)^2 + (3 - 3.4381)^2 \\
&\quad + (10 - 8.1557)^2 + (4 - 4.6175)^2 + (5 - 5.7969)^2 + (2 - 2.2587)^2 \\
&\quad + (11 - 11.6939)^2 + (9 - 6.3866)^2 + (10 - 11.1042)^2 + (2 - 2.2587)^2 \\
&= 14.9644
\end{aligned}$$

It can be seen that $SS_y = SS_{reg} + SS_{res}$. Furthermore,

$$\begin{aligned}
r^2 &= \frac{SS_{reg}}{SS_y} \\
&= \frac{105.9466}{120.9168} \\
&= 0.8762
\end{aligned}$$

Another equivalent test for examining the significance of the linear relationship between X and Y (significance of b) is the test for the significance of the coefficient of determination. The hypotheses in this case are

$$H_0: R^2_{pop} = 0$$
$$H_1: R^2_{pop} > 0$$

The appropriate test statistic is the F statistic

$$F = \frac{SS_{reg}}{SS_{res}/(n-2)}$$

which has an F distribution with 1 and $n - 2$ degrees of freedom. The F test is a generalised form of the t test (see Chapter 18). If a random variable is t distributed with n degrees of freedom, then t^2 is F distributed with 1 and n degrees of freedom. Hence, the F test for testing the significance of the coefficient of determination is equivalent to testing the following hypotheses:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

or

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

From Table 20.3, it can be seen that

$$r^2 = \frac{105.9522}{105.9522 + 14.9644}$$
$$= 0.8762$$

which is the same as the value calculated earlier. The value of the F statistic is

$$F = \frac{105.9522}{14.9644/10}$$
$$= 70.8027$$

with 1 and 10 degrees of freedom. The calculated F statistic exceeds the critical value of 4.96 determined from Table 5 in the Statistical Appendix. Therefore, the relationship is significant at $\alpha = 0.05$, corroborating the results of the t test. If the relationship between X and Y is significant, it is meaningful to predict the values of Y based on the values of X and to estimate prediction accuracy.

Check prediction accuracy

To estimate the accuracy of predicted values, $\hat{Y}$, it is useful to calculate the standard error of estimate, *SEE*. This statistic is the standard deviation of the actual Y values from the predicted $\hat{Y}$ values.

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

$$SEE = \sqrt{\frac{SS_{res}}{n-2}}$$

or, more generally, if there are k independent variables

$$SEE = \sqrt{\frac{SS_{res}}{n-k-1}}$$

SEE may be interpreted as a kind of average residual or average error in predicting Y from the regression equation.[10]

Two cases of prediction may arise. The researcher may want to predict the mean value of Y for all the cases with a given value of X, say X_0, or predict the value of Y for a single case. In both situations, the predicted value is the same and is given by $\hat{Y}$, where

$$\hat{Y} = a + bX_0$$

But the standard error is different in the two situations, although in both situations it is a function of *SEE*. For large samples, the standard error for predicting the mean value of Y is $SEE/\sqrt{n}$ and for predicting individual Y values it is *SEE*. Hence, the construction of confidence intervals (see Chapter 15) for the predicted value varies, depending upon whether the mean value or the value for a single observation is being predicted. For the data given in Table 20.3, *SEE* is estimated as follows:

$$SEE = \sqrt{\frac{14.9644}{12-2}}$$
$$= 1.22329$$

The final two steps in conducting bivariate regression, namely examination of residuals and model cross-validation, are considered later, and we now turn to the assumptions underlying the regression model.

Assumptions

The regression model makes a number of assumptions in estimating the parameters and in significance testing, as shown in Figure 20.4:

1. The error term is normally distributed. For each fixed value of *X*, the distribution of *Y* is normal.[11]
2. The means of all these normal distributions of *Y*, given *X*, lie on a straight line with slope *b*.
3. The mean of the error term is 0.
4. The variance of the error term is constant. This variance does not depend on the values assumed by *X*.
5. The error terms are uncorrelated. In other words, the observations have been drawn independently.

Insights into the extent to which these assumptions have been met can be gained by an examination of residuals, which is covered in the next section on multiple regression.[12]

Multiple regression

Multiple regression
A statistical technique that simultaneously develops a mathematical relationship between two or more independent variables and an interval-scaled dependent variable.

Multiple regression involves a single dependent variable and two or more independent variables. The questions raised in the context of bivariate regression can also be answered via multiple regression by considering additional independent variables:

- Can variation in sales be explained in terms of variation in advertising expenditures, prices and level of distribution?
- Can variation in market shares be accounted for by the size of the sales force, advertising expenditures and sales promotion budgets?
- Are consumers' perceptions of quality determined by their perceptions of prices, brand image and brand attributes?

Additional questions can also be answered by multiple regression:

- How much of the variation in sales can be explained by advertising expenditures, prices and level of distribution?
- What is the contribution of advertising expenditures in explaining the variation in sales when the levels of prices and distribution are controlled?
- What levels of sales may be expected given the levels of advertising expenditures, prices and level of distribution?

example

Global brands, local ads[13]

Europeans welcome brands from other countries, but when it comes to advertising, they seem to prefer brands from their own country. A survey conducted by Yankelovich and Partners and its affiliates found that most European consumers' favourite commercials were for local brands even though they were more than likely to buy foreign brands. Respondents in Britain, France and Germany named Coca-Cola as the most often purchased soft drink. The French, however, selected the famous award-winning spot for France's Perrier bottled water as their favourite commercial. Similarly, in Germany, the favourite advertising was for a German brand of non-alcoholic beer, Clausthaler. In Britain, though, Coca-Cola was the favourite soft drink and also the favourite advertising. In the light of such findings, the important question was: does advertising help? Does it help increase the purchase probability of the brand or does it merely maintain the brand recognition rate high? One way of finding out was by running a regression where the dependent variable was the likelihood of brand purchase and the independent variables were brand attribute evaluations and advertising evaluations. Separate models with and without advertising could be run to assess any significant difference in the contribution. Individual *t* tests could also be examined to find out the significant contribution of both the brand attributes and advertising. The results could indicate the degree to which advertising plays an important part in brand purchase decisions. ■

Multiple regression model
An equation used to explain the results of multiple regression analysis.

The general form of the **multiple regression model** is as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + e$$

which is estimated by the following equation:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

As before, the coefficient *a* represents the intercept, but the *b*s are now the partial regression coefficients. The least squares criterion estimates the parameters in such a way as to minimise the total error, SS_{res}. This process also maximises the correlation between the actual values of Y and the predicted values of $\hat{Y}$. All the assumptions made in bivariate regression also apply in multiple regression. We define some associated statistics and then describe the procedure for multiple regression analysis.[14]

Most of the statistics and statistical terms described under bivariate regression also apply to multiple regression. In addition, the following statistics are used:

Adjusted R^2. R^2, the coefficient of multiple determination, is adjusted for the number of independent variables and the sample size to account for the diminishing returns. After the first few variables, the additional independent variables do not make much contribution.

Coefficient of multiple determination. The strength of association in multiple regression is measured by the square of the multiple correlation coefficient, R^2, which is also called the *coefficient of multiple determination*.

***F* test.** The F test is used to test the null hypothesis that the coefficient of multiple determination in the population, R^2_{pop}, is zero. This is equivalent to testing the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$. The test statistic has an F distribution with k and $(n - k - 1)$ degrees of freedom.

Partial *F* test. The significance of a partial regression coefficient, β_i, of X_i may be tested using an incremental F statistic. The incremental F statistic is based on the increment in the explained sum of squares resulting from the addition of the independent variable X_i to the regression equation after all the other independent variables have been included.

Partial regression coefficient. The partial regression coefficient, b_1, denotes the change in the predicted value, $\hat{Y}$, per unit change in X_1 when the other independent variables, X_2 to X_k, are held constant.

Conducting multiple regression analysis

The steps involved in conducting multiple regression analysis are similar to those for bivariate regression analysis. The discussion focuses on partial regression coefficients, strength of association, significance testing and examination of residuals.

Estimating the partial regression coefficients

To understand the meaning of a partial regression coefficient, let us consider a case in which there are two independent variables, so that

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

First, note that the relative magnitude of the partial regression coefficient of an independent variable is, in general, different from that of its bivariate regression coefficient. In other words, the partial regression coefficient, b_1, will be different from the regression coefficient, b, obtained by regressing Y on only X_1. This happens because X_1 and X_2 are usually correlated. In bivariate regression, X_2 was not considered, and any variation in Y that was shared by X_1 and X_2 was attributed to X_1. In the case of multiple independent variables, however, this is no longer true.

The interpretation of the partial regression coefficient, b_1, is that it represents the expected change in Y when X_1 is changed by one unit but X_2 is held constant or otherwise controlled. Likewise, b_2 represents the expected change in Y for a unit change in X_2 when X_1 is held constant. Thus, calling b_1 and b_2 partial regression coefficients is appropriate. It can also be seen that the combined effects of X_1 and X_2 on Y are additive. In other words, if X_1 and X_2 are each changed by one unit, the expected change in Y would be $(b_1 + b_2)$.

Conceptually, the relationship between the bivariate regression coefficient and the partial regression coefficient can be illustrated as follows. Suppose that one were to remove the effect of X_2 from X_1. This could be done by running a regression of X_1 on X_2. In other words, one would estimate the equation $\hat{Y}_1 = a + bX_2$ and calculate the residual $X_r = (X_1 - \hat{Y}_1)$. The partial regression coefficient, b_1, is then equal to the bivariate regression coefficient, b, obtained from the equation $\hat{Y} = a + bX_r$. In other words, the partial regression coefficient, b_1, is equal to the regression coefficient, b, between Y and the residuals of X_1 from which the effect of X_2 has been removed. The partial coefficient, b, can also be interpreted along similar lines.

Extension to the case of k variables is straightforward. The partial regression coefficient, b_1, represents the expected change in Y when X_1 is changed by one unit and X_2 to X_k are held constant. It can also be interpreted as the bivariate regression coefficient, b, for the regression of Y on the residuals of X_1, when the effect of X_2 to X_k has been removed from X_1.

The beta coefficients are the partial regression coefficients obtained when all the variables $(Y, X_1, X_2, \ldots X_k)$ have been standardised to a mean of 0 and a variance of 1 before estimating the regression equation. The relationship of the standardised to the non-standardised coefficients remains the same as before:

$$B_1 = b_1\left(\frac{S_{x1}}{S_y}\right)$$

$$\vdots$$

$$B_{\mathrm{k}} = b_{\mathrm{k}}\left(\frac{S_{xk}}{S_y}\right)$$

The intercept and the partial regression coefficients are estimated by solving a system of simultaneous equations derived by differentiating and equating the partial derivatives to 0. Since these coefficients are automatically estimated by the various computer programs, we will not present the details. Yet it is worth noting that the equations cannot be solved if (1) the sample size, n, is smaller than or equal to the number of independent variables, k, or (2) one independent variable is perfectly correlated with another.

Suppose that in explaining the attitude towards the city we now introduce a second variable, importance attached to the weather. The data for the 12 pre-test respondents on attitude towards the city, duration of residence and importance attached to the weather are given in Table 20.2. The results of multiple regression analysis are depicted in Table 20.4. The partial regression coefficient for duration (X_1) is now 0.4811, different from what it was in the bivariate case. The corresponding beta coefficient is 0.7636.

Table 20.4 Multiple regression

Multiple *R*	0.97210
R^2	0.94498
Adjusted R^2	0.93276
Standard error	0.85974

Analysis of variance			
	df	*Sum of squares*	*Mean square*
Regression	2	114.26425	57.13213
Residual	9	6.65241	0.73916
F = 77.29364	Significance of F = 0.0000		

Variables in the equation					
Variable	*b*	SE_b	β	*T*	*Sig. of T*
Importance	0.28865	0.08608	0.31382	3.353	0.0085
Duration	0.48108	0.05895	0.76363	8.160	0.0000
(Constant)	0.33732	0.56736		0.595	0.5668

The partial regression coefficient for importance attached to weather (X_2) is 0.2887, with a beta coefficient of 0.3138. The estimated regression equation is

$$(\hat{Y}) = 0.33732 + 0.48108X_1 + 0.28865X_2$$

or

$$\text{attitude} = 0.33732 + 0.48108\ (\text{duration}) + 0.28865\ (\text{importance})$$

This equation can be used for a variety of purposes, including predicting attitudes towards the city, given a knowledge of the respondents' duration of residence in the city and the importance they attach to weather.

Strength of association

The strength of the relationship stipulated by the regression equation can be determined by using appropriate measures of association. The total variation is decomposed as in the bivariate case:

$$SS_y = SS_{reg} + SS_{res}$$

where $SS_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

The strength of association is measured by the square of the multiple correlation coefficient, R^2, which is also called the coefficient of multiple determination:

$$R^2 = \frac{SS_{reg}}{SS_y}$$

The multiple correlation coefficient, R, can also be viewed as the simple correlation coefficient, r, between Y and $\hat{Y}$. Several points about the characteristics of R^2 are worth noting. The coefficient of multiple determination, R^2, cannot be less than the highest bivariate, r^2, of any individual independent variable with the dependent variable. R^2 will be larger when the correlations between the independent variables are low. If the independent variables are statistically independent (uncorrelated), then R^2 will be the sum of bivariate r^2 of each independent variable with the dependent variable. R^2 cannot decrease as more independent variables are added to the regression equation. Yet diminishing returns set in, so that after the first few variables, the additional independent variables do not make much of a contribution.[15] For this reason, R^2 is adjusted for the number of independent variables and the sample size by using the following formula:

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

For the regression results given in Table 20.4, the value of R^2 is

$$R^2 = \frac{114.2643}{114.2643 + 6.6524}$$

$$= 0.9450$$

This is higher than the r^2 value of 0.8762 obtained in the bivariate case. The r^2 in the bivariate case is the square of the simple (product moment) correlation between attitude toward the city and duration of residence. The R^2 obtained in multiple regression is also higher than the square of the simple correlation between attitude and importance attached to weather (which can be estimated as 0.5379). The adjusted R^2 is estimated as

$$\text{adjusted } R^2 = 0.9450 - \frac{2(1.0 - 0.9450)}{12 - 2 - 1}$$

$$= 0.9328$$

Note that the value of adjusted R^2 is close to R^2 and both are higher than r^2 for the bivariate case. This suggests that the addition of the second independent variable, importance attached to weather, makes a contribution in explaining the variation in attitude towards the city.

Test for significance

Significance testing involves testing the significance of the overall regression equation as well as specific partial regression coefficients. The null hypothesis for the overall test is that the coefficient of multiple determination in the population, R^2_{pop}, is zero:

$$H_0: R^2_{pop} = 0$$

This is equivalent to the following null hypothesis:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

The overall test can be conducted by using an F statistic

$$F = \frac{SS_{reg}/k}{SS_{reg}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

which has an F distribution with k and $n - k - 1$ degrees of freedom.[16] For the multiple regression results given in Table 20.4,

$$F = \frac{114.2642 / 2}{6.6524 / 9} = 77.2938$$

which is significant at $\alpha = 0.05$.

If the overall null hypothesis is rejected, one or more population partial regression coefficients have a value different from 0. To determine which specific coefficients (β_is) are non-zero, additional tests are necessary. Testing for the significance of the β_is can be done in a manner similar to that in the bivariate case by using t tests. The significance of the partial coefficient for importance attached to weather may be tested by the following equation:

$$t = \frac{b}{SE_b} = \frac{0.2887}{0.08608} = 3.353$$

which has a t distribution with $n - k - 1$ degrees of freedom. This coefficient is significant at $\alpha = 0.05$. The significance of the coefficient for duration of residence is tested in a similar way and found to be significant. Therefore, both the duration of residence and importance attached to weather are important in explaining attitude towards the city.

Some computer programs provide an equivalent F test, often called the partial F test, which involves a decomposition of the total regression sum of squares, SS_{reg}, into components related to each independent variable. In the standard approach, this is done by assuming that each independent variable has been added to the regression equation after all the other independent variables have been included. The increment in the explained sum of squares, resulting from the addition of an independent variable, X_i, is the component of the variation attributed to that variable and is denoted SS_{xi}.[17] The significance of the partial regression coefficient for this variable, β_i, is tested using an incremental F statistic

$$F = \frac{SS_{xi} / 1}{SS_{res} / (n-k-1)}$$

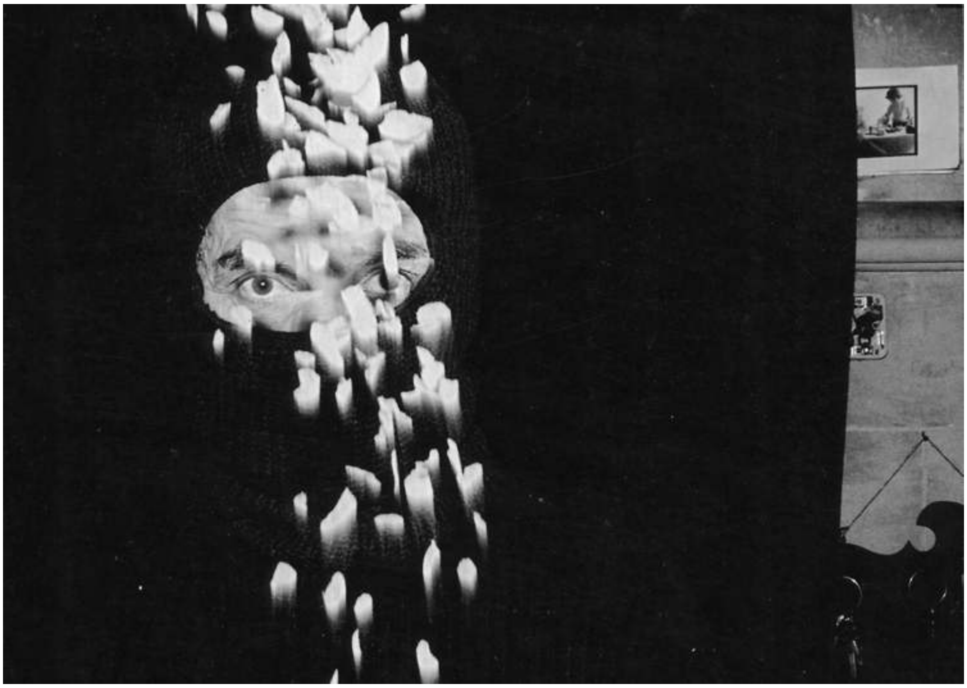
which has an F distribution with 1 and $(n - k - 1)$ degrees of freedom.

While high R^2 and significant partial regression coefficients are comforting, the efficacy of the regression model should be evaluated further by an examination of the residuals.

Examine the residuals

Residual
The difference between the observed value of Y_i and the value predicted by the regression equation $\hat{Y}_i$.

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. In addition, scattergrams of the residuals – in which the residuals are plotted against the predicted values, $\hat{Y}_i$, time, or predictor variables – provide useful insights in examining the appropriateness of the underlying assumptions and regression model fitted.[18]



Examining scattergrams of residuals provides useful insights.

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. Additional evidence can be obtained by determining the percentages of residuals falling within ±1 SE or ±2 SE. These percentages can be compared with what would be expected under the normal distribution (68% and 95%, respectively). More formal assessment can be made by running the K-S one-sample test.

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. Figure 20.6 shows a pattern whose variance is dependent on the $\hat{Y}_i$ values.

A plot of residuals against time, or the sequence of observations, will throw some light on the assumption that the error terms are uncorrelated. A random pattern should be seen if this assumption is true. A plot like the one in Figure 20.7 indicates a linear relationship between residuals and time. A more formal procedure for examining the correlations between the error terms is the Durbin-Watson test.[19]

Plotting the residuals against the independent variables provides evidence of the appropriateness or inappropriateness of using a linear model. Again, the plot should

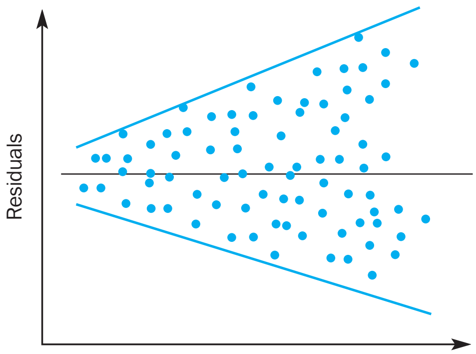


Figure 20.6
Residual plot indicating that variance is not constant

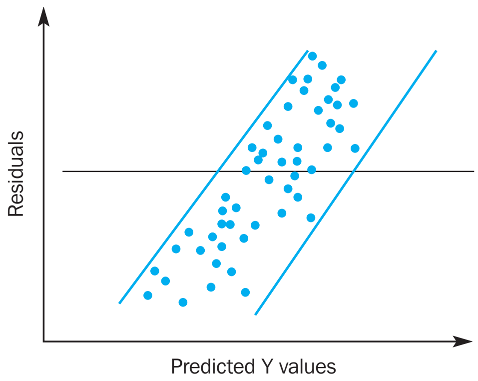


Figure 20.7
Plot indicating a linear relationship between residuals and time

result in a random pattern. The residuals should fall randomly, with relatively equal distribution dispersion about 0. They should not display any tendency to be either positive or negative.

To examine whether any additional variables should be included in the regression equation, one could run a regression of the residuals on the proposed variables. If any variable explains a significant proportion of the residual variation, it should be considered for inclusion. Inclusion of variables in the regression equation should be strongly guided by the researcher's theory. Thus, an examination of the residuals provides valuable insights into the appropriateness of the underlying assumptions and the model that is fitted. Figure 20.8 shows a plot that indicates that the underlying assumptions are met and that the linear model is appropriate.

If an examination of the residuals indicates that the assumptions underlying linear regression are not met, the researcher can transform the variables in an attempt to satisfy the assumptions. Transformations, such as taking logs, square roots or reciprocals, can stabilise the variance, make the distribution normal or make the relationship linear. We further illustrate the application of multiple regression with an example.

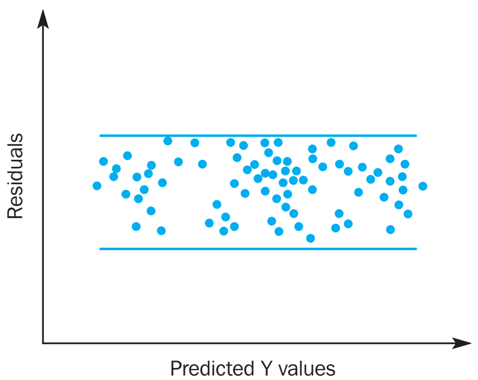


Figure 20.8
Plot of residuals indicating that a fitted model is appropriate

example

At no 'Ad' ditional cost[20]

It is widely believed that consumer magazines' prices are subsidised by the advertising carried within the magazines. A study examined the contribution of advertising to the price per copy of magazines.

Multiple regression analysis was used to examine the relationships among price per copy and editorial pages, circulation, percentage of news-stand circulation, promotional expenditures, percentage of colour pages, and per copy advertising revenues. The form of the analysis was

$$\text{PPC} = b_0 + b_1(\text{ed. pages}) + b_2(\text{circ.}) + b_3(\text{\% news circ.}) + b_4(\text{PE}) + b_5(\text{\% colour}) + b_6(\text{ad revs.})$$

where

PPC = price per copy (in €)
ed. pages = editorial pages per average issue
circ. = the log of average paid circulation (in thousands)
% news circ. = percentage news-stand circulation
PE = promotional expenditures (in €)
% colour = percentage of pages printed in colour
ad revs. = per copy advertising revenues (in €)

Table 1 shows the zero-order Pearson product moment correlations among the variables. The correlations provide directional support for the predicted relationships and show that collinearity among the independent variables is sufficiently low so as not to affect the stability of the regression analysis. The highest correlation among the independent variables was between promotional expenditures and circulation ($r = 0.42$).

Table 1 Zero-order correlation matrix of variables in analyses

	Price per copy	*Price per editorial page*	*Circulation*	*Editorial pages*	*Promotional expenditures*	*% colour pages*	*% news-stand circ.*
Price per editorial page	0.60[a]						
Circulation	–0.21[a]	–0.42[a]					
Editorial pages	0.52[a]	–0.30[a]	0.29[a]				
Promotional expenditures	–0.22[a]	–0.06	0.42[a]	–0.19			
% colour pages	0.01	–0.15	0.33[a]	0.19	–0.15		
% news-stand circ.	0.46[a]	0.17	0.09	0.31[a]	0.26[a]	0.02	
Ad. revenues per copy	0.29[a]	–0.04	–0.25[a]	0.30[a]	–0.14	0.15	0.08

[a] $p < 0.05$.

The results of the regression analysis using price per copy as the dependent variable are given in Table 2. Of the six independent variables, three were significant ($p < 0.05$): the number of editorial pages, average circulation, and percentage news-stand circulation. The three variables accounted for virtually all of the explained variance ($R^2 = 0.51$; adjusted $R^2 = 0.48$). The direction of the coefficients was consistent with prior expectations: the number of editorial pages was positive, circulation was negative, and percentage news-stand circulation was positive. This was expected, given the structure of the magazine publishing industry, and it confirmed the hypothesised relationship.

Table 2 Regression analysis using price per copy as dependent variable

	b	*SE*	*F*
Dependent variable: price per copy			
Independent variables:			
Editorial pages	0.0084	0.0017	23.04[a]
Circulation	–0.4180	0.1372	9.29[a]
Percentage news-stand circulation	0.0067	0.0016	18.46 [a]
Promotional expenditures	0.13–0.04[b]	0.0000	0.59
Percentage colour pages	0.0227	0.0092	0.01
Per copy ad. revenues	0.1070	0.0412	0.07
Overall $R^2 = 0.51$	$df = 6, 93$	Overall $F = 16.19$[a]	

[a] $p < 0.05$.
[b] Decimal moved in by four zeros.

Promotional expenditures, use of colour and per copy advertising revenues were found to have no relationship with price per copy, after the effects of circulation, percentage news-stand circulation and editorial pages were controlled in the regression analysis.

Because the effect of per copy advertising revenue was not significant, no support was found for the contention that advertising decreases the price per copy of consumer magazines. It was concluded that advertising in magazines is provided free to consumers, but does not subsidise prices. ■

In the preceding example, promotional expenditures, percentage of colour pages and per copy advertising revenues were not found to be significantly related to the price per copy of magazines. Some of the independent variables considered in a study often turn out to be non-significant. When there are a large number of independent variables and the researcher suspects that not all of them are significant, stepwise regression should be used.

Stepwise regression

Stepwise regression
A regression procedure in which the predictor variables enter or leave the regression equation one at a time.

The purpose of **stepwise regression** is to select, from a large number of predictor variables, a small subset of variables that account for most of the variation in the dependent or criterion variable. In this procedure, the predictor variables enter or are removed from the regression equation one at a time.[21] There are several approaches to stepwise regression.

1 *Forward inclusion*. Initially, there are no predictor variables in the regression equation. Predictor variables are entered one at a time, only if they meet certain criteria specified in terms of F ratio. The order in which the variables are included is based on the contribution to the explained variance.
2 *Backward elimination*. Initially, all the predictor variables are included in the regression equation. Predictors are then removed one at a time based on the F ratio for removal.
3 *Stepwise solution*. Forward inclusion is combined with the removal of predictors that no longer meet the specified criterion at each step.

Stepwise procedures do not result in regression equations which are optimal, in the sense of producing the largest R^2, for a given number of predictors.[22] Because of the correlations between predictors, an important variable may never be included or less important variables may enter the equation. To identify an optimal regression equation, one would have to compute combinatorial solutions in which all possible combinations are examined. Nevertheless, stepwise regression can be useful when the sample size is large in relation to the number of predictors, as shown in the following example.

example

Browsers step out[23]

A profile of browsers in regional shopping centres was constructed using three sets of independent variables: demographics, shopping behaviour, and psychological and attitudinal variables. The dependent variable consisted of a browsing index. In a stepwise regression including all three sets of variables, demographics were found to be the most powerful predictors of browsing behaviour. The final regression equation, which contained 20 of the possible 36 variables, included all the demographics. The table presents the regression coefficients, standard errors of the coefficients, and their significance levels.

In interpreting the coefficients, it should be recalled that the smaller the browsing index (the dependent variable), the greater the tendency to exhibit behaviours associated with browsing. The two predictors with the largest coefficients were gender and employment status. Browsers were more likely to be employed females. They also tend to be somewhat 'downscale', compared with other shopping centre patrons, exhibiting lower levels of education and income, after accounting for the effects of gender and employment status. Although browsers tend to be somewhat younger than non-browsers, they are not necessarily single; those who reported larger family sizes tended to be associated with smaller values of the browsing index.

The 'downscale' profile of browsers relative to other patrons indicates that speciality stores in shopping centres should emphasise moderately priced products. This may explain the historically low rate of failure in shopping centres among such stores and the tendency of high-priced speciality shops to be located in only the prestigious shopping centres or 'upscale' non-enclosed shopping centres. ■

Regression of browsing index on descriptive and attitudinal variables by order of entry into stepwise regression

Variable description	*Coefficient*	*SE*	*Significance*
Gender (0 = male, 1 = female)	–0.485	0.164	0.001
Employment status (0 = employed)	0.391	0.182	0.003
Self-confidence	–0.152	0.128	0.234
Education	0.079	0.072	0.271
Brand intention	–0.063	0.028	0.024
Watch daytime TV? (0 = yes)	0.232	0.144	0.107
Tension	–0.182	0.069	0.008
Income	0.089	0.061	0.144
Frequency of shopping centre visits	–0.130	0.059	0.028
Fewer friends than most	0.162	0.084	0.054
Good shopper	–0.122	0.090	0.174
Others' opinions important	–0.147	0.065	0.024
Control over life	–0.069	0.069	0.317
Family size	–0.086	0.062	0.165
Enthusiastic person	–0.143	0.099	0.150
Age	0.036	0.069	0.603
Number of purchases made	–0.068	0.043	0.150
Purchases per store	0.209	0.152	0.167
Shop on tight budget	–0.055	0.067	0.412
Excellent judge of quality	–0.070	0.089	0.435
Constant	3.250		
Overall $R^2 = 0.477$			

Multicollinearity

Multicollinearity
A state of high intercorrelations among independent variables.

Stepwise regression and multiple regression are complicated by the presence of **multicollinearity**. Virtually all multiple regression analyses done in marketing research involve predictors or independent variables that are related. Multicollinearity, however, arises when intercorrelations amongst the predictors are very high.[24] Multicollinearity can result in several problems, including the following:

1 The partial regression coefficients may not be estimated precisely. The standard errors are likely to be high.

2 The magnitudes as well as the signs of the partial regression coefficients may change from sample to sample.
3 It becomes difficult to assess the relative importance of the independent variables in explaining the variation in the dependent variable.
4 Predictor variables may be incorrectly included or removed in stepwise regression.

What constitutes serious multicollinearity is not always clear, although several rules of thumb and procedures have been suggested in the literature. Procedures of varying complexity have also been suggested to cope with multicollinearity.[25] A simple procedure consists of using only one of the variables in a highly correlated set of variables.

Alternatively, the set of independent variables can be transformed into a new set of predictors that are mutually independent by using techniques such as principal components analysis (see Chapter 22). More specialised techniques, such as ridge regression and latent root regression, can also be used.[26]

Relative importance of predictors

When multicollinearity is present, special care is required in assessing the relative importance of independent variables. In marketing research, it is valuable to determine the relative importance of the predictors. In other words, how important are the independent variables in accounting for the variation in the criterion or dependent variable?[27] Unfortunately, because the predictors are correlated, there is no unambiguous measure of relative importance of the predictors in regression analysis.[28] Several approaches, however, are commonly used to assess the relative importance of predictor variables.

1 *Statistical significance.* If the partial regression coefficient of a variable is not significant, as determined by an incremental F test, that variable is judged to be unimportant. An exception to this rule is made if there are strong theoretical reasons for believing that the variable is important.
2 *Square of the simple correlation coefficient.* This measure, r^2, represents the proportion of the variation in the dependent variable explained by the independent variable in a bivariate relationship.
3 *Square of the partial correlation coefficient.* This measure, $R^2 yx_i.x_j.x_k$, is the coefficient of determination between the dependent variable and the independent variable, controlling for the effects of the other independent variables.
4 *Square of the part correlation coefficient.* This coefficient represents an increase in R^2 when a variable is entered into a regression equation that already contains the other independent variables.
5 *Measures based on standardised coefficients or beta weights.* The most commonly used measures are the absolute values of the beta weights, $|\beta_i|$, or the squared values, β_i^2. Because they are partial coefficients, beta weights take into account the effect of the other independent variables. These measures become increasingly unreliable as the correlations among the predictor variables increase (multicollinearity increases).
6 *Stepwise regression.* The order in which the predictors enter or are removed from the regression equation is used to infer their relative importance.

Given that the predictors are correlated, at least to some extent, in virtually all regression situations, none of these measures is satisfactory. It is also possible that the different measures may indicate a different order of importance of the predictors.[29] Yet if all the measures are examined collectively, useful insights may be obtained into the relative importance of the predictors.

Cross-validation

Cross-validation
A test of validity that examines whether a model holds on comparable data not used in the original estimation.

Before assessing the relative importance of the predictors or drawing any other inferences, it is necessary to cross-validate the regression model. Regression and other multivariate procedures tend to capitalise on chance variations in the data. This could result in a regression model or equation that is unduly sensitive to the specific data used to estimate the model. One approach for evaluating the model for this and other problems associated with regression is cross-validation. **Cross-validation** examines whether the regression model continues to hold on comparable data not used in the estimation. The typical cross-validation procedure used in marketing research is as follows.

1. The regression model is estimated using the entire data set.
2. The available data are split into two parts, the *estimation sample* and the *validation sample*. The estimation sample generally contains 50 – 90% of the total sample.
3. The regression model is estimated using the data from the estimation sample only. This model is compared with the model estimated on the entire sample to determine the agreement in terms of the signs and magnitudes of the partial regression coefficients.
4. The estimated model is applied to the data in the validation sample to predict the values of the dependent variable, $\hat{Y}_i$, for the observations in the validation sample.
5. The observed values, Y_i, and the predicted values, $\hat{Y}_i$, in the validation sample are correlated to determine the simple r^2. This measure, r^2, is compared with R^2 for the total sample and with R^2 for the estimation sample to assess the degree of shrinkage.

Double cross-validation
A special form of validation in which the sample is split into halves. One half serves as the estimation sample and the other as a validation sample. The roles of the estimation and validation halves are then reversed and the cross-validation process is repeated.

A special form of validation is called **double cross-validation**. In double cross-validation the sample is split into halves. One half serves as the estimation sample, and the other is used as a validation sample in conducting cross-validation. The roles of the estimation and validation halves are then reversed, and the cross-validation is repeated.[30]

Regression with dummy variables

Cross-validation is a general procedure that can be applied even in some special applications of regression, such as regression with dummy variables. Nominal or categorical variables may be used as predictors or independent variables by coding them as dummy variables. The concept of dummy variables was introduced in Chapter 17. In that chapter, we explained how a categorical variable with four categories (heavy users, medium users, light users and non-users) can be coded in terms of three dummy variables, D_1, D_2 and D_3, as shown.

Suppose that the researcher was interested in running a regression analysis of the effect of attitude towards the brand on product use. The dummy variables D_1, D_2 and D_3 would be used as predictors. Regression with dummy variables would be modelled as

$$\hat{Y}_i = a + b_1D_1 + b_2D_2 + b_3D_3$$

Product usage category	*Original variable code*	*Dummy variable code*		
		D_1	D_2	D_3
Non-users	1	1	0	0
Light users	2	0	1	0
Medium users	3	0	0	1
Heavy users	4	0	0	0

In this case, 'heavy users' have been selected as a reference category and have not been directly included in the regression equation. Note that for heavy users, D_1, D_2 and D_3 assume a value of 0, and the regression equation becomes

$$\hat{Y}_i = a$$

For non-users, $D_1 = 1$, and $D_2 = D_3 = 0$, and the regression equation becomes

$$\hat{Y}_i = a + b_1$$

Thus, the coefficient b_1 is the difference in predicted Y_i for non-users, as compared with heavy users. The coefficients b_2 and b_3 have similar interpretations. Although heavy users was selected as a reference category, any of the other three categories could have been selected for this purpose.[31]

Analysis of variance and covariance with regression

Regression with dummy variables provides a framework for understanding the analysis of variance and covariance. Although multiple regression with dummy variables provides a general procedure for the analysis of variance and covariance, we show only the equivalence of regression with dummy variables to one-way analysis of variance. In regression with dummy variables, the predicted $\hat{Y}$ for each category is the mean of Y for each category. To illustrate using the dummy variable coding of product use we just considered, the predicted $\hat{Y}$ and mean values for each category are as follows:

Product usage category	*Predicted value*	*Mean value*
	$\hat{Y}$	$\bar{Y}$
Non-users	$a + b_1$	$a + b_1$
Light users	$a + b_2$	$a + b_2$
Medium users	$a + b_3$	$a + b_3$
Heavy users	a	a

Given this equivalence, it is easy to see further relationships between dummy variable regression and one-way ANOVA.[32]

Thus, we see that regression in which the single independent variable with c categories has been recoded into $c - 1$ dummy variables is equivalent to one-way analysis of variance. Using similar correspondences, one can also illustrate how n-way analysis of variance and analysis of covariance can be performed using regression with dummy variables.

Dummy variable regression	**One-way ANOVA**
$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$= SS_{within}$
$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	$= SS_{between}$
R^2	$= \eta^2$
Overall F test	$= F$ test

Internet and computer applications

The computer packages contain several programs to perform correlation analysis and regression analysis, calculating the associated statistics, performing tests for significance and plotting the residuals.

SPSS

CORRELATIONS can be used for computing Pearson product moment correlations, PARTIAL CORR for partial correlations, and NONPAR CORR for Spearman's ρs and Kendall's τ. The main program is REGRESSION which calculates bivariate and multiple regression equations, associated statistics and plots. It allows for easy examination of residuals. Stepwise regression can also be conducted. Regression statistics can be requested with PLOT, which produces simple scattergrams and some other types of plots.

SAS

The program CORR can be used for calculating Pearson, Spearman's, Kendall's and partial correlations. REG is a general-purpose regression procedure that fits bivariate and multiple regression models using the least-squares procedure. All the associated statistics are computed, and residuals can be plotted. Stepwise methods can be implemented. RSREG is a more specialised procedure that fits a quadratic response surface model using least squares regression. It is useful for determining factor levels that optimise a response. The ORTHOREG procedure is recommended for regression when the data are ill-conditioned. GLM uses the method of least squares to fit general linear models and can also be used for regression analysis. NLIN computes the parameters of a non-linear model using least squares or weighted least squares procedures.

Minitab

Correlation can be computed using the Stat>Basic statistics>Correlation function. It calculates Pearson's product moment. The Spearman's procedure ranks the columns first and then performs the correlation, on the ranked columns. To compute partial correlation, use the menu commands Stat>Basic Statistics>Correlation and Stat>Regression>Regression. Regression analysis, under the Stats>Regression function, can perform simple, polynomial and multiple analysis. The output includes a linear regression equation, table of coefficients, R^2, adjusted R^2, analysis of variance table, a table of fits and residuals that provide unusual observations. Other available features include stepwise, best subsets, fitted line plot and residual plots.

Excel

Correlations can be determined in Excel by using the Tools>Data analysis>Correlation function. Utilise the Correlation Worksheet function when a correlation coefficient for two cell ranges is needed. There is no separate function for partial correlations. Regression can be accessed from the Tools>Data analysis menu. Depending on the features selected, the output can consist of a summary output table, including an ANOVA table, a standard error of Y estimate, coefficients, standard error of coefficients, R^2 values and the number of observations. In addition, the function computes a residual output table, a residual plot, a line fit plot, a normal probability plot and a two-column probability data output table.

Summary

The product moment correlation coefficient, r, measures the linear association between two metric (interval or ratio scaled) variables. Its square, r^2, measures the proportion of variation in one variable explained by the other. The partial correlation coefficient measures the association between two variables after controlling, or adjusting for, the effects of one or more additional variables. The order of a partial correlation indicates how many variables are being adjusted or controlled. Partial correlations can be very helpful for detecting spurious relationships.

Bivariate regression derives a mathematical equation between a single metric criterion variable and a single metric predictor variable. The equation is derived in the form of a straight line by using the least squares procedure. When the regression is run on standardised data, the intercept assumes a value of 0, and the regression coefficients are called beta weights. The strength of association is measured by the coefficient of determination, r^2, which is obtained by computing a ratio of SS_{reg} to SS_y. The standard error of estimate is used to assess the accuracy of prediction and may be interpreted as a kind of average error made in predicting Y from the regression equation.

Multiple regression involves a single dependent variable and two or more independent variables. The partial regression coefficient, b_1, represents the expected change in Y when X_1 is changed by one unit and X_2 to X_k are held constant. The strength of association is measured by the coefficient of multiple determination, R^2. The significance of the overall regression equation may be tested by the overall F test. Individual partial regression coefficients may be tested for significance using the incremental F test. Scattergrams of the residuals, in which the residuals are plotted against the predicted values, $\hat{Y}_i$, time, or predictor variables, are useful for examining the appropriateness of the underlying assumptions and the regression model fitted.

In stepwise regression, the predictor variables are entered or removed from the regression equation one at a time for the purpose of selecting a smaller subset of predictors that account for most of the variation in the criterion variable. Multicollinearity, or very high intercorrelations among the predictor variables, can result in several problems. Because the predictors are correlated, regression analysis provides no unambiguous measure of relative importance of the predictors. Cross-validation examines whether the regression model continues to hold true for comparable data not used in estimation. It is a useful procedure for evaluating the regression model.

Nominal or categorical variables may be used as predictors by coding them as dummy variables. Multiple regression with dummy variables provides a general procedure for the analysis of variance and covariance.

Questions

1. What is the product moment correlation coefficient? Does a product moment correlation of 0 between two variables imply that the variables are not related to each other?
2. What are the main uses of regression analysis?
3. What is the least squares procedure?
4. Explain the meaning of standardised regression coefficients.
5. How is the strength of association measured in bivariate regression? In multiple regression?
6. What is meant by prediction accuracy?

7 What is the standard error of the estimate?

8 What is multiple regression? How is it different from bivariate regression?

9 Explain the meaning of a partial regression coefficient. Why is it called that?

10 State the null hypothesis in testing the significance of the overall multiple regression equation. How is this null hypothesis tested?

11 What is gained by an examination of residuals?

12 Explain the stepwise regression approach. What is its purpose?

13 What is multicollinearity? What problems can arise because of multicollinearity?

14 Describe the cross-validation procedure. Describe double cross-validation.

15 Demonstrate the equivalence of regression with dummy variables to one-way ANOVA.

Notes

1 Zajac, J., 'Avon's finally glowing thanks to global sales – and new lip-shtic', *Money* 26(9) (September 1997), 60; Miller, C., 'Computer modelling rings the right bell for Avon', *Marketing News* (9 May), 14.

2 Draper, N.R. and Smith, H., *Applied Regression Analysis*, 3rd edn (New York: Wiley, 1998); Doherty, M.E. and Sullivan, J.A., 'rho = p', *Organizational Behaviour and Human Decision Processes* 43(1) (February 1989), 136–44; Martin, W.S., 'Effects of scaling on the correlation coefficient: additional considerations', *Journal of Marketing Research* 15 (May 1978), 304–8; Bollen, K.A. and Barb, R.H., 'Pearson's *R* and coarsely categorized measures', *American Sociological Review* 46 (1981), 232–9.

3 Tacq, J., *Multivariate Analysis Techniques in Social Science Research* (Thousand Oaks, CA: Sage, 1997); Neter, J., Wasserman, W. and Kutner, M.J., *Applied Linear Statistical Models*, 3rd edn (Burr Ridge, IL: Irwin, 1990).

4 Although the topic is not discussed here, partial correlations can also be helpful in locating intervening variables and making certain types of causal inferences.

5 'Bates Saatchi & Saatchi, Budapest: accounting for change', *Accountancy* 116(224) (August 1995), 31; Kasriel, K., 'Hungary's million-dollar slap', *Advertising Age* (8 June 1992).

6 Another advantage to τ is that it can be generalised to a partial correlation coefficient. Pett, M.A., *Nonparametric Statistics for Health Care Research* (Thousand Oaks, CA: Sage, 1997); Siegel, S. and Castellan, N.J., *Nonparametric Statistics*, 2nd edn (New York: McGraw-Hill, 1988).

7 In a strict sense, the regression model requires that errors of measurement be associated only with the criterion variable and that the predictor variables be measured without error. For serially correlated errors, see Canjels, E. and Watson, M.W., 'Estimating deterministic trends in the presence of serially correlated errors', *Review of Economics and Statistics* 79(2) (May 1997), 184–200.

8 See any text on regression, such as Draper, N.R. and Smith, H., *Applied Regression Analysis*, 3rd edn (New York: Wiley, 1998); Neter, J., Wasserman, W. and Kutner, M.J., *Applied Linear Statistical Models*, 3rd edn (Burr Ridge, IL: Irwin, 1990).

9 Technically, the numerator is $b - \beta$. Since it has been hypothesised that $\beta = 0.0$, however, it can be omitted from the formula.

10 The larger the SEE, the poorer the fit of the regression.

11 The assumption of fixed levels of predictors applies to the 'classical' regression model. It is possible, if certain conditions are met, for the predictors to be random variables. Their distribution is not allowed to depend on the parameters of the regression equation. See Draper, N.R. and Smith, H., *Applied Regression Analysis*, 3rd edn (New York: Wiley, 1998).

12 For an approach to handling the violations of these assumptions, see Dispensa, G.S., 'Use logistic regression with customer satisfaction data', *Marketing News* 31(1) (6 January 1997), 13; Reddy, S.K., Holak, S.L. and Bhat, S., 'To extend or not to extend: success determinants of line extensions', *Journal of Marketing Research* 31 (May 1994), 243–62.

13 Rees, J., 'Tight ship keeps Coke on top of the world', *Marketing Week* 20(6) (8 May 1997), 28–9; Giges, N., 'Europeans buy outside goods, but like local ads', *Advertising Age International* (27 April 1992).

14 For other applications of multiple regression see Griffin, A., 'The effect of project and process characteristics on product development cycle time', *Journal of Marketing Research* 34 (February 1997), 24–35; Gatignon, H. and Xuereb, J.M., 'Strategic orientation of the firm and new product performance', *Journal of Marketing Research* 34 (February 1997), 77–90; Kumar, N., Scheer, L.K. and Steenkamp, J.B., 'The effects of supplier fairness on vulnerable resellers', *Journal of Marketing Research* 32 (February 1995), 54–65.

15 Yet another reason for adjusting R^2 is that as a result of the optimising properties of the least squares approach it is a maximum. Thus, to some extent, R^2 always overestimates the magnitude of a relationship. For applications of adjusted R^2 see Smith, N.C. and Cooper-Martin, E., 'Ethics and target marketing: the role of product harm and consumer vulnerability', *Journal of Marketing* 61(3) (January 1997), 1–20; Cohen, M.A. and Ho, T.H., 'An anatomy of a decision support system for developing and launching line extensions', *Journal of Marketing Research* 34 (February 1997), 117–29.

16 If R^2_{pop} is zero, then the sample R^2 reflects only sampling error, and the F ratio will tend to be equal to unity.

17 Another approach is the hierarchical method, in which the variables are added to the regression equation in an order specified by the researcher. For an application of multiple regression, see Corfman, K.P., 'Perceptions of relative influence: formation and measurement', *Journal of Marketing Research* 28 (May 1991), 125–36.
18 Atkinson, A.C., Koopman, S.J. and Shephard, N., 'Detecting shocks: outliers and breaks in time series', *Journal of Econometrics* 80(2) (October 1997), 387–422; Wang, G.C.S. and Akabay, C.K., 'Autocorrelation: problems and solutions in regression modeling', *Journal of Business Forecasting Methods and Systems* 13(4) (Winter 1994/95), 18–26; Belsley, D., Kuh, E. and Walsh, R.E., *Regression Diagnostics* (New York: Wiley, 1980).
19 The Durbin-Watson test is discussed in virtually all regression textbooks. See Draper, N.R. and Smith, H., *Applied Regression Analysis*, 3rd edn (New York: Wiley, 1998).
20 Berman, H., 'Selling the advertising/trade show partnership', *Folio: The Magazine for Magazine Management* (Special Sourcebook Issue for 1997 Supplement) 25(18) (1997), 214–15; Soley, L. and Krishnan, R., 'Does advertising subsidise consumer magazine prices?', *Journal of Advertising* 16 (Spring 1987), 4–9.
21 Fox, J., *Applied Regression Analysis, Linear Models and Related Methods* (Thousand Oaks, CA: Sage, 1997); McIntyre, S.H., Montgomery, D.B., Srinivasan, V. and Weitz, B.A., 'Evaluating the statistical significance of models developed by stepwise regression', *Journal of Marketing Research* 20 (February 1983), 1–11.
22 For applications of stepwise regression, see Ittner, D. and Larcker, D.F., 'Product development cycle time and organisational performance', *Journal of Marketing Research* 34 (February 1997), 13–23; Laroche, M. and Sadokierski, R., 'Role of confidence in a multi-brand model of intentions for a high-involvement service', *Journal of Business Research* 29 (January 1994), 1–12.
23 Crispell, D., 'Hispanics at the mall', *American Demographics* 19(10) (October 1997), 35–6; Jarboe, G.R. and McDaniel, C.D., 'A profile of browsers in regional shopping malls', *Journal of the Academy of Marketing Science* (Spring 1987), 46–53.
24 Greenberg, E. and Parks, R.P., 'A predictive approach to model selection and multicollinearity', *Journal of Applied Econometrics* 12(1) (January/February 1997), 67–75; Ofir, C. and Khuri, A., 'Multicollinearity in marketing models: diagnostics and remedial measures', *International Journal of Research in Marketing* 3 (1986), 181–205.
25 Possible procedures are given in Wang, G.C.S., 'How to handle multicollinearity in regression modelling', *Journal of Business Forecasting Methods and Systems* 15(1) (Spring 1996), 23–7; Mason, C.H. and Perreault Jr, W.D., 'Collinearity, power, and interpretation of multiple regression analysis', *Journal of Marketing Research* 28 (August 1991), 268–80; Hocking, R.R., 'Developments in linear regression methodology: 1959–1982', *Technometrics* 25 (August 1983), 219–30; Snee, R.D., 'Discussion', *Technometrics* 25 (August 1983), 230–7.
26 Holzworth, J.R., 'Policy capturing with ridge regression', *Organizational Behavior and Human Decision Processes* 68(2) (November 1996), 171–9; Wildt, A.R., 'Equity estimation and assessing market response', *Journal of Marketing Research* 31 (February 1994), 437–51; Sharma, S. and James, W.L., 'Latent root regression: an alternative procedure for estimating parameters in the presence of multicollinearity', *Journal of Marketing Research* (May 1981), 154–61.
27 Only relative importance can be determined, since the importance of an independent variable depends upon all the independent variables in the regression model.
28 Rugimbana, R., 'Predicting automated teller machine usage: the relative importance of perceptual and demographic factors', *International Journal of Bank Marketing* 13(4) (1995), 26–32; Green, P.E., Carroll, S.D. and DeSarbo, W.S., 'A new measure of predictor variable importance in multiple regression', *Journal of Marketing Research* (August 1978), 356–60; Jackson, B.B., 'Comment on "A new measure of predictor variable importance in multiple regression"', *Journal of Marketing Research* (February 1980), 116–18.
29 In the rare situation in which all the predictors are uncorrelated, simple correlations = partial correlations = part correlations = betas. Hence, the squares of these measures will yield the same rank order of the relative importance of the variables.
30 For more on cross-validation, see Litwin, M.S., *How to Measure Survey Reliability and Validity* (Thousand Oaks, CA: Sage, 1997); Song, X.M. and Perry, M.E., 'The determinants of Japanese new product success', *Journal of Marketing Research* 34 (February 1997), 64–76; Cooil, B., Winer, R.S. and Rados, D.L., 'Cross-validation for prediction', *Journal of Marketing Research* (August 1987), 271–9.
31 For an application of dummy variable regression, see Yavas, U., 'Demand forecasting in a service setting', *Journal of International Marketing and Marketing Research* 21(1) (February 1996), 3–11. For further discussion on dummy variable coding, see Cohen, J. and Cohen, P., *Applied Multiple Regression Correlation Analysis for the Behavioural Sciences*, 2nd edn (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 181–222.
32 Fox, J., *Applied Regression Analysis, Linear Models and Related Methods* (Thousand Oaks, CA: Sage, 1997). For an application of regression analysis to conduct analysis of covariance, see Dancin, P.A. and Smith, D.C., 'The effect of brand portfolio characteristics on consumer evaluations of brand extensions', *Journal of Marketing Research* 31 (May 1994), 229–42.