Block 14: Correlation and regression

(Activity solutions can be found at the end of the document.)

We consider **regression analysis** which is used for explaining the variation in market share, sales, brand preference etc. This may use explanatory variables such as advertising, price, distribution and product quality. Starting with **product moment correlation**, we proceed to the **bivariate regression** model followed by the **multiple regression model**. Special topics associated with regression are then considered.

Learning Objectives

- discuss the concepts of product moment correlation and the partial correlation coefficient and show how they provide a foundation for regression analysis
- 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
- explain the nature and methods of multiple regression analysis and the meaning of partial regression coefficients
- describe specialised techniques used in multiple regression analysis, particularly stepwise regression and regression with dummy variables.

Reading List

Malhotra, N.K., D. Nunan and D.F. Birks. Marketing Research: An Applied Approach. (Pearson, 2017) 5th edition [ISBN 9781292103129] Chapter 22.

14.1 Correlation and regression

For each section of *Correlation and regression,* use the LSE ELearning resources to test your knowledge with the Key terms and concepts flip cards.

Product moment correlation

The **product moment correlation**, *r*, summarises the *strength of the linear 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*. As it was originally proposed by Karl Pearson, it is also known as the *Pearson correlation coefficient*. It is also referred to as *simple correlation, bivariate correlation,* or merely the *correlation coefficient*.

From a sample of *n* observations on *X* and *Y*, the product moment correlation, r, can be calculated as:

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

Division of the numerator and denominator by n − 1 gives:

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

r varies between −1 and +1. The correlation coefficient between two variables will be the same *regardless of their underlying units of measurement*.

When it is computed for a population, rather than a sample, the product moment correlation is denoted by ρ, the Greek letter 'rho'. The coefficient *r* is an *estimator* of ρ. The statistical significance of the relationship between two variables measured by using *r* can be conveniently tested.

We test $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. The *test statistic* is:

$$t = r \times \sqrt{\frac{n-2}{1-r^2}} \sim t_2 - 2$$

Activity 14.1

What is the product moment correlation coefficient? Does a product moment correlation of zero between two variables imply that the variables are not related to each other?

Partial correlation

A **partial correlation coefficient** measures the association between two variables after *controlling for, or adjusting for, the effects of one or more additional variables*:

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

Partial correlations have an order associated with them. The order indicates how many variables are being adjusted or controlled for. The simple correlation coefficient, *r*, has a zero order, as it does not control for any additional variables while measuring the association between two variables.

The coefficient $r_{xy.z}$ is a first-order partial correlation coefficient, as 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 special case when a partial correlation is larger than its respective zero-order correlation involves a *suppressor effect*.

Regression analysis

Regression analysis examines associative relationships between a metric-dependent variable and one or more independent variables in the following ways.

- Determine whether the independent variables explain significant variation in the dependent variable: *whether a relationship exists*.
  Determine how much of the variation in the dependent variable can be explained by the independent variables: *strength of the relationship*.
- Determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables.
- *Predict* values of the dependent variable.
- *Control* for other independent variables when evaluating the contributions of a specific variable or set of variables.

Regression analysis is concerned with the nature and degree of association between variables and does not imply or assume any causality.

The statistics associated with bivariate regression analysis are the following.

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

Coefficient of determination: The strength of association is measured by the coefficient of determination, $R^2 = r^2$ (in bivariate regression). It varies between 0 and 1 and is the proportion of the total variation in *Y* which is accounted for by the variation in *X*.
Estimated or predicted value: The estimated or predicted value of $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ where $\hat{Y}_i$ is the predicted value of Y_i, and $\hat{\beta}_0 \; and \; \hat{\beta}_1$ and are estimators of $\hat{\beta}_0 \; and \; \hat{\beta}_1$, respectively.

Regression coefficient: The estimated parameter $\hat{\beta}_1$ is usually referred to as the unstandardised regression coefficient.

Scatterplot: A scatterplot is a plot of the values of two variables for all the cases or observations.

Standard error of estimate: This statistic is the standard deviation of the actual *Y* values from the predicted Y_i values.

Standard error: The standard deviation of $\hat{\beta}_1$, S. E. ($\hat{\beta}_1$), 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, i.e. $\sum e_i^2$.

***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 : β1 = 0, where:

$$\frac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)} \sim t_{n-2}$$

Activity 14.2

What are the main uses of regression analysis?

Conducting bivariate regression analysis

The process to conduct bivariate regression analysis is as follows:

Plot the scatter diagram

⇓⇓

Formulate the general model

⇓⇓

Estimate the parameters

⇓⇓

Estimate standardised regression coefficient

⇓⇓

Test for significance

⇓⇓

Determine strength and significance of association

⇓⇓

Check prediction accuracy

⇓⇓

Examine the residuals

⇓⇓

Cross-validate the model

Conducting bivariate regression analysis

A scatterplot (or scatter diagram) is a plot of the values of two variables for all the cases or observations. The most commonly-used technique for fitting a straight line to a scatterplot is the **least squares procedure**. In fitting the line, the least squares procedure *minimises the sum of squared errors*, $\sum e_1^2$. In the bivariate regression model, the general form of a straight line is:

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

The regression procedure adds an error term to account for the probabilistic or *stochastic nature* of the relationship such that:

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

where ε_i is the error term associated with the *i*th observation. We assume $\varepsilon_i \sim N(0, \sigma^2)$, where σ^2 is a parameter of the model denoting the (constant) variance of the error term for i = 1, 2, …, n. Figure 22.5 of the textbook shows examples of normally distributed error terms.

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

where $\hat{Y}_i$ is the estimated or predicted value of Y_i, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are *estimators* of β_0 and β_1, respectively, where:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2} \frac{\sum_{i=l}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=l}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=l}^{n} X_i^2 - n\bar{X}^2}$$

and:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Standardisation is the process by which the raw data are transformed into new variables which have a mean of 0 and a variance of 1. 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*:

$$\hat{B}_{yx} = B_{xy} = r_{xy}$$

There is a simple relationship between the standardised and unstandardised regression coefficients, given by:

$$B_{yx} = \hat{\beta}_1 \times \frac{S_X}{S_Y}$$

The *statistical significance* of the linear relationship between XX and YY may be tested by examining the hypotheses:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0.$$

A *t* statistic with $n - 2$ degrees of freedom can be used, such that:

$$\frac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)} \sim t_{n-2}$$

where $S.E.(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$, i.e. the standard deviation of $\hat{\beta}_1$.

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:

$$SS_Y = SS_{Reg} + SS_{Res}$$

where:

$$SS_Y = \sum_{i=l}^{n} (Y_i - \bar{Y})^2 \, , SS_{Reg} = \sum_{i=l}^{n} (\hat{Y}_i - \bar{Y})^2 \text{ and } SS_{Res} = \sum_{i=l}^{n} (Y_i - \hat{Y}_i)^2$$

Figure 22.6 of the textbook shows the decomposition of the total variation in bivariate regression.

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

$$R^2 = \frac{SS_{Reg}}{SS_Y} = \frac{SS_Y - SS_{Res}}{SS_Y}$$

Another *equivalent test* for examining the significance of the linear relationship between X and Y (i.e. the significance of $\hat{\beta}_1$) is the test for the significance of the coefficient of determination. The hypotheses in this case are:

$$H_0 : R^2_{Pop} = 0 \quad vs. \quad H_1 : R^2_{Pop} > 0.$$

The appropriate test statistic is the FF statistic:

$$\frac{SS_{Reg}}{SS_{Res}/(n-2)} \sim F_{1,n-2}$$

The *F* test is a generalised form of the *t* test. 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 in the numerator and denominator, respectively. Hence the *F* test for testing the significance of the coefficient of determination is *equivalent* to testing the following hypotheses:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0.$$

or:

$$H_0 : \rho = 0 \quad vs. \quad H_1 : \rho \neq 0.$$

To estimate the *accuracy of predicted values*, $\hat{Y}$, it is useful to calculate the standard error of estimate, S.E.E., given by:

$$S.E.E. = \sqrt{\frac{\sum_{i=l}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{SS_{Res}}{n-2}}$$

or, more generally, if there are k independent variables (in multiple regression):

$$S.E.E. = \sqrt{\frac{SS_{Res}}{n-k-1}}$$

Activity 14.3

What is the least squares procedure?

Activity 14.4

How is the strength of association measured in bivariate regression? How is it measured in multiple regression?

Activity 14.5

What is meant by prediction accuracy? What is the standard error of the estimate?

Assumptions of the bivariate regression model

Assumptions of the bivariate regression model are the following.

- The error term is normally distributed, so, for each fixed value of X, the distribution of Y is normal.
- The means of all these normal distributions of Y, given X, lie on a straight line with slope β_1
- The mean of the error term is 0.

- The variance of the error term is constant (homoskedasticity), i.e. the variance does not depend on the values assumed by X.
- The error terms are uncorrelated, i.e. the observations have been drawn independently.

Multiple linear regression

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

which is estimated by the following equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

As before, the coefficient $\hat{\beta}_0$ represents the intercept, but the $\hat{\beta}_i$s are now *the partial regression coefficients.*

Statistics associated with multiple linear regression are the following.

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**: This is used to test the null hypothesis that the coefficient of multiple determination in the population, R2Pop, is zero. The test statistic has an $F_{k,n-k-1}$ distribution.

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, $\hat{\beta}_1$ (say), 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*.

Consider a case in which there are two independent variables, such that:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

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. The interpretation of the partial regression coefficient, $\hat{\beta}_1$, is that it represents the expected change in Y when X_1 is changed by one unit while X_2 is held constant, or otherwise controlled. Likewise, $\hat{\beta}_2$ represents the expected change in Y for a unit change in X_2 when X_1 is held constant. Therefore, calling $\hat{\beta}_1$ and $\hat{\beta}_2$ partial regression coefficients is appropriate.

It can also be seen that the combined effects of X_1 andX_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 $\hat{\beta}_1 + \hat{\beta}_2$. Suppose one was to remove the effect of X_2 from X_1, which could be done by running a regression of X_1 on X_2. One

would estimate the equation $\hat{X}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_2$ and calculate the residual $e = X_1 - \hat{X}_1$. The partial regression coefficient, $\hat{\beta}_1$, is then equal to the bivariate regression coefficient, $\tilde{\beta}_1$, obtained from the equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 e$.

The extension to the case of *k* variables is straightforward. The partial regression coefficient, $\hat{\beta}_1$, represents the expected change in *Y* when X_1 is changed by one unit and X_2 through X_k are *held constant*. It can also be interpreted as the bivariate regression coefficient, $\tilde{\beta}_1$, for the regression of Y on the residuals of X_1, when the effects of X_2 through X_k have been removed from X_1. The relationship of the standardised to the unstandardised coefficients remains the same as before:

$$B1 = \hat{\beta}_1 \times \frac{SX_1}{SY} \quad and \quad Bk = \hat{\beta}_k \times \frac{SX_k}{S_Y}.$$

Recall that $SS_Y = SS_{Reg} + SS_{Res}$. 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*, given by:

$$R^2 = \frac{SS_{Reg}}{SS_Y}$$

R^2 is *adjusted for the number of independent variables and the sample size* by using the following formula:

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

We have seen H0:R2Pop=0, which is equivalent to the following null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \quad \beta_k = 0.$$

The overall test can be conducted by using an *F* statistic:

$$\frac{SS_{Reg/k}}{SS_{Res}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1.}$$

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 regression coefficient may be tested with:

$$\frac{\hat{\beta}_i}{S.E.(\hat{\beta}_i)} \sim t_{n-k-1}$$

Activity 14.6

Explain the meaning of a partial regression coefficient. Why is it called this name?

Activity 14.7

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

Examination of residuals

A **residual** is the difference between the observed value of Y_i and the value predicted by the regression equation, $\hat{Y}_i$. Scatterplots 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 fit.

The assumption of a normally-distributed error term can be examined by constructing a *histogram of the residuals*.

The assumption of a constant variance of the error term can be examined by plotting the residuals against the predicted values of the dependent variable, $\hat{Y}_i$.

A plot of residuals against time, or the sequence of observations, will shed some light on the assumption that the error terms are *uncorrelated*.

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

To examine if additional variables should be included in the regression, a regression of the residuals on the proposed variables could be run.

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.

Figure 22.7 of the textbook shows a residual plot indicating that the error term variance is not constant. Figure 22.8 of the textbook shows a plot indicating a linear relationship between residuals and time. Finally, Figure 22.9 of the textbook shows a plot of residuals indicating that a fitted model is appropriate.

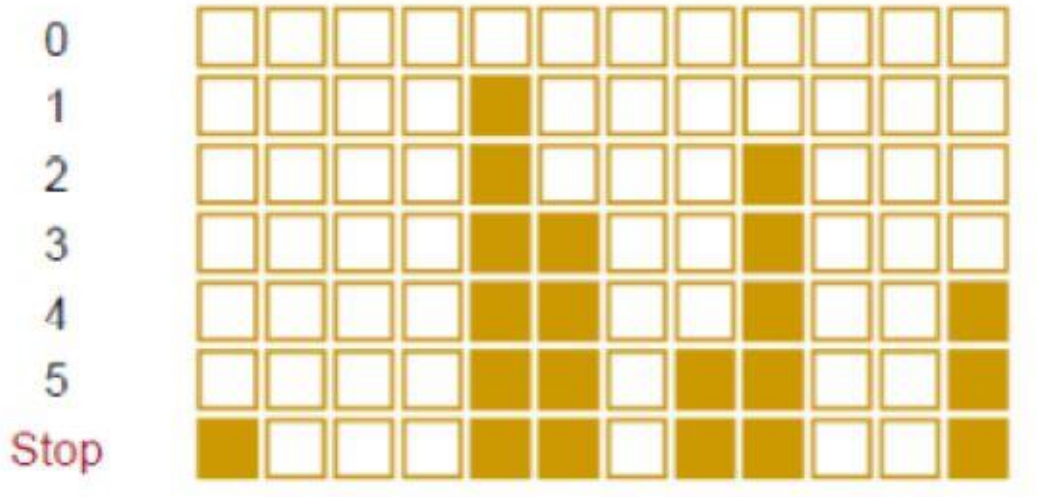
Activity 14.8

What is gained by an examination of residuals?

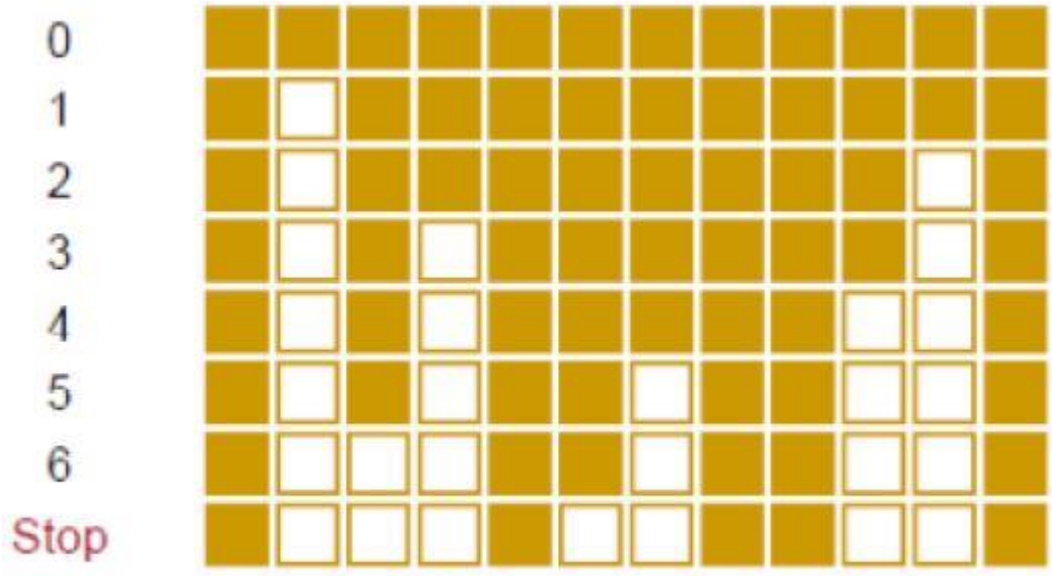
Stepwise regression

The purpose of **stepwise regression** is to select, from a large number of predictor variables, a small subset of variables which 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. There are several approaches to stepwise regression.

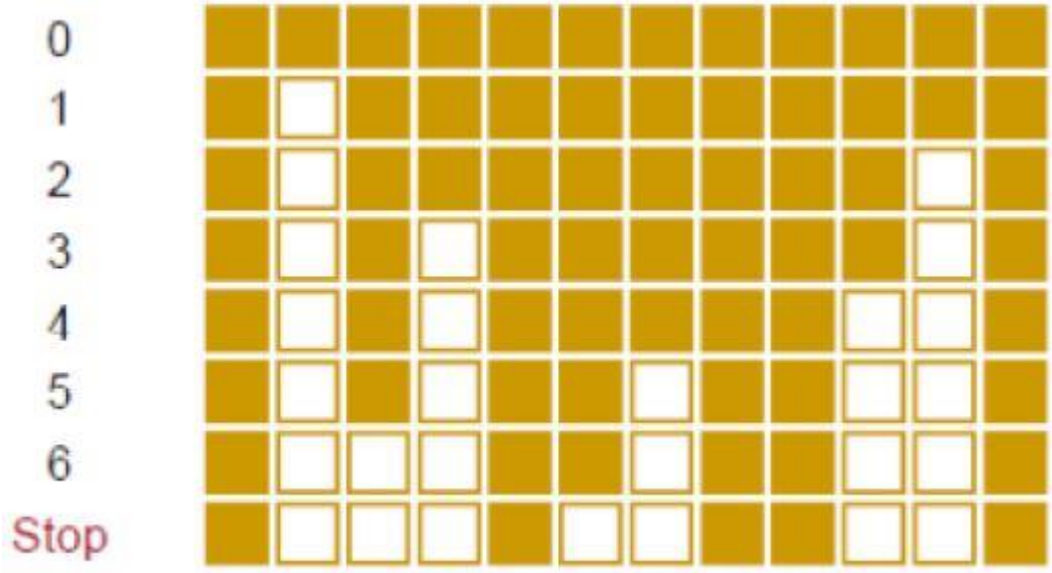
Forward selection starts with an empty model. The method computes an *F* statistic for each independent variable not in the model and examines the largest of these statistics. If it is significant at a specified significance level, the corresponding variable is added to the model. After a variable is entered in the model it is *never removed from the model*. The process is repeated until none of the remaining variables meets the specified level for entry.



Backward elimination starts off with the full model. Results of the FF tests for individual parameter estimates are examined, and the least significant variable which falls above the specified significance level is removed. After a variable is removed from the model *it remains excluded*. The process is repeated until no other variable in the model meets the specified significance level for removal.



Stepwise selection is similar to forward selection in that it starts with an empty model and incrementally builds a model one variable at a time. However, the method differs from forward selection in that variables already in the model *do not necessarily remain*. The backward component of the method removes variables from the model which do not meet the significance criteria specified to stay in the model. The stepwise selection process terminates if no further variable can be added to the model, or if the variable just entered into the model is the only variable removed in the subsequent backward elimination.



Activity 14.9

Explain the stepwise regression approach. What is its purpose?

Multicollinearity

Multicollinearity arises when intercorrelations among the predictors are very high. Multicollinearity can result in *several problems*, including the following.

- The partial regression coefficients may not be estimated precisely. The standard errors are likely to be high.
- The magnitudes, as well as the signs, of the partial regression coefficients may change from sample to sample.
- It becomes difficult to assess the relative importance of the independent variables in explaining the variation in the dependent variable.
- Predictor variables may be incorrectly included or removed in stepwise regression.

A simple procedure for adjusting for multicollinearity 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 which are mutually independent by using techniques such as **principal components analysis**. We will consider a closely related technique called **factor analysis** later in the course.

Activity 14.10

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

Regression with dummy variables

Sometimes we want to model *categorical explanatory variables*. For example, possession of a University of London degree *may* (no, *will*!) affect salary (positively, obviously!). Of course, this

leads to a problem - categorical variables are not measurable. So we use **dummy variables** which take only two values, 0 and 1. The value is 1 if a subject's value of a categorical variable is in a particular category (level), and 0 if it is not.

Golden rule: *k* categorical levels require *k*−1 dummy variables.

Suppose we wanted to differentiate between consumers in England, Scotland and Wales. You have *three* categories, so now what? You create *two* dummy variables - for example, D_1 = 1 if the country is Scotland; D_2 = 1 if the country is Wales. Why do you not need a third dummy variable?

- Because you have all the information you need here - the 'baseline' is when $D_1$and D_2 are both zero.
- The one you leave out becomes the *reference category* (in the above example, England).

Dummy variables can be included alongside interval- or ratio-level (measurable) variables as explanatory variables. However, the *interpretation* of the coefficients is different. We cannot talk about a one-unit increase in a dummy variable, rather the coefficient gives the *expected difference in the response variable of those in the '1' category with the reference category*. Tests of significance are as before.

In the case below, 'heavy users' has been selected as the reference category and has not been directly included in the regression equation. The coefficient $\hat{\beta}_1$is the difference in the predicted $\hat{Y}_i$ between non-users and heavy users.

Product usage category	**Original variable code**	**Dummy variable codes**		
		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

We have:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3.$$

Relative importance of predictor variables

Unfortunately, because the predictors are correlated, there is no unambiguous measure of the relative importance of the predictors in regression analysis. However, several approaches are commonly used to assess the relative importance of predictor variables.

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.

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 regression.

Square of the partial correlation coefficient: This measure, $R^2_{yxi,xj,xk}$, is the coefficient of determination between the dependent variable and the independent variable, controlling for the effects of the other independent variables.

Square of the part correlation coefficient: This coefficient represents an increase in R^2 when a variable is entered into a regression equation which already contains the other independent variables.

Measures based on standardised coefficients or beta weights: The most commonly-used measures are the absolute values of the beta weights, $|B_i|$, or the squared values, B_i^2

Stepwise regression: The order in which the predictors enter or are removed from the regression equation is used to infer their relative importance.

Cross-validation

The regression model is estimated using the entire dataset. The available data are then split into two parts, the **estimation sample** and the **validation sample**. The estimation sample generally contains 50-90% of the total sample. The regression model is estimated using the data from the estimation sample only. This model is compared to the model estimated using the entire sample to determine the *agreement in terms of the signs and magnitudes* of the partial regression coefficients.

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. 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 to R^2 for the total sample and to R^2 for the estimation sample to assess the degree of shrinkage.

Discussion forum and case studies

To access the solutions to these questions and case study, click here to access the printable Word document or click here to go to LSE's Elearning resources.

Activities on the block's topics

1. A supermarket chain wants to determine the effect of promotion on relative competitiveness. Data were obtained from 15 cities on the promotional expenses relative to a major competitor (competitor expenses = 100) and on sales relative to this competitor (competitor sales = 100). The data can be found in the file *Supermarket.sav*. (An Excel version of the dataset is *Supermarket.xlsx*.)

You are assigned the task of telling the manager whether there is any relationship between relative promotional expenses and relative sales.

   a. Plot the relative sales (on the y-axis) against the relative promotional expenses (on the x-axis), and interpret the diagram.

Video walkthrough of exercise 1a.

Which measure would you use to determine whether there is a relationship between the two variables? Why?

b. Run a bivariate regression analysis of relative sales on relative promotional expenses.

Video walkthrough of activity 1c.

c. Interpret the regression coefficients.

d. Is the regression relationship significant?

Video walkthrough of activity 1e.

e. If the company matched the competitor in terms of promotional expenses (if the relative expense was 100), what would the company's relative sales be?

f. Interpret the resulting r^2.

Video walkthrough of activity 1g.

2. To understand the role of quality and price in influencing the patronage of shoe shops, 14 major shoe shops in a large city were rated in terms of preference to shop, quality of shoes sold and price fairness. All the ratings were obtained on an 11-point scale, with higher numbers indicating more positive ratings. The data can be found in the file *Shoe_shops.sav*. (An Excel version of the dataset is *Shoe_shops.xlsx*.)

   a. Run a multiple regression analysis explaining shoe shop preference in terms of shoe quality and price fairness.

   b. Interpret the partial regression coefficients.

   c. Determine the significance of the overall regression.

   d. Determine the significance of the partial regression coefficients.

   e. Do you think that multicollinearity is a problem in this case? Why or why not?

Video walkthrough of activity 2.

3. You come across a magazine article reporting the following relationship between annual expenditure on prepared dinners in pounds, *PD*, and annual income in pounds, *INC*: [\widehat{PD} = 23.4 + 0.003 \times INC.\] The coefficient of the *INC* variable is reported as significant.

   a. Does this relationship seem plausible? Is it possible to have a coefficient which is small in magnitude and yet significant?

   b. From the information given, can you tell how good the estimated model is?

   c. What is the expected expenditure on prepared dinners for a family earning £30,000?

   d. If a family earning £40,000 spent £130 annually on prepared dinners, what is the residual?

   e. What is the meaning of a negative residual?

4. In a survey pretest, data were obtained from 20 participants on preference for boots on a seven-point scale (1 = not preferred, 7 = greatly preferred) (V_1). The participants also provided their evaluations of the boots on comfort (V_2), style (V_3) and durability (V_4), also on

seven-point scales (1 = poor, 7 = excellent). The resulting data are given in the file *Boots.sav*. (An Excel version of the dataset is *Boots.xlsx*.)

a. Calculate the simple correlations between V_1 to V_4 and interpret the results.

b. Run a bivariate regression with preference for boots, V_1, as the dependent variable and evaluation on comfort, V_2, as the independent variable. Interpret the results.

c. Run a bivariate regression with preference for boots, V_1, as the dependent variable and evaluation on style, V_3, as the independent variable. Interpret the results.

d. Run a bivariate regression with preference for boots, V_1, as the dependent variable and evaluation on durability, V_4, as the independent variable. Interpret the results.

e. Run a multiple regression with preference for boots, V_1, as the dependent variable and V_2 to V_4 as the independent variables. Interpret the results. Compare the coefficients for V_2, V_3 and V_4 obtained in the bivariate and the multiple regressions.

Video walkthrough of activity 4.

Discussion forum, exercises and discussion points

To access the solutions to these questions and case study, click here to access the printable Word document or click here to go to LSE's Elearning resources.

Discussion points

1. Regression is such a basic technique that it should always be used in analysing data.
2. What is the relationship between bivariate correlation, bivariate regression, multiple regression and analysis of variance?

Learning outcomes checklist

Use this to assess your own understanding of the chapter. You can always go back and amend the checklist when it comes to revision!

- Discuss the concepts of product moment correlation and the partial correlation coefficient and show how they provide a foundation for regression analysis
- 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
- Explain the nature and methods of multiple regression analysis and the meaning of partial regression coefficients
- Describe specialised techniques used in multiple regression analysis, particularly stepwise regression and regression with dummy variables.

Block 14: Correlation and regression

Solution to Exercise 14.1

It is a statistic used to determine whether a linear relationship exists between two metric variables. From a sample of *n* observations on variables *X* and *Y*, the product correlation coefficient, *r*, is given as:

$$r = \frac{S_{XY}}{S_X S_Y}$$

with $-1 \leq r \leq 1$, where S_{XY} denotes the covariance between *X* and *Y*, while S_X and S_Y are the standard deviations of X and Y, respectively.

The product correlation coefficient cannot reveal the presence of non-linear relationships, if any, between the variables. It can only indicate the degree to which variation in one variable is related to the variation in another, if the relationship is linear.

Solution to Exercise 14.2

Regression analysis is used to:

- Determine whether a relationship exists between the dependent and independent variables
- Determine the strength of the relationship
- Determine the mathematical equation relating the variables
- Predict values of the dependent variable
- Control other independent variables when evaluating the effect of a particular variable or set of variables.

Regression analysis, however, is concerned only with the nature and degree of association between variables and does not *by itself* imply or assume any causality.

Solution to Exercise 14.3

This method is the most popular technique for determining a line which approximates most accurately a given set of data points. Following the least squares method, the best-fitting line (regression line) is determined by minimising the vertical distances of all the points from the line. A point which does not fall on the regression line is not fully accounted for, and the vertical distance of the point from the line is called the error. The distances of all the points from the line are squared and added together to arrive at a measure of the total variation. In fitting the line, the least squares procedure minimises the sum of the squared errors.

Solution to Exercise 14.4

The strength of association between two variables, X and Y, in bivariate regression is measured by the coefficient of determination, R^2. The strength of the association is then calculated as:

$$R^2 = \frac{SS_{Reg}}{SS_Y} = \frac{SS_Y - SS_{Res}}{SS_Y}$$

Also, an *F* statistic, given as:

$$\frac{SS_{Reg}}{SS_{Res}/(n-2)} \sim F_{1,n-2}$$

is used to examine the significance of the linear relationship between X and Y.

In multiple regression, the strength of association is measured along similar lines. The strength of association is given by the multiple correlation coefficient:

$$R^2 = \frac{SS_{Reg}}{SS_Y} = \frac{SS_Y - SS_{Res}}{SS_Y}$$

The adjusted R^2 is used, which is given as:

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

where *k* = the number of independent variables, and *n* = the sample size.

Solution to Exercise 14.5

Prediction accuracy is a measure used when estimating the value of the dependent variable given a value of the independent variable.

The standard error of the estimate (S.E.E.) is used to test accuracy, where:

$$S.E.E. \sqrt{\frac{SS_{Res}}{n-k-1}}$$

where *n* = the number of observations, and *k* = the number of independent variables.

S.E.E. can be interpreted as an average error resulting from the regression equation. Therefore, the larger the S.E.E., the poorer the regression fit.

Solution to Exercise 14.6

A partial regression coefficient, $\hat{\beta}_i$, represents the expected change in Y when X_i is changed by one unit and the remaining independent variables remain unchanged. In effect, each $\hat{\beta}_i$ represents a part of the total change in the dependent variable caused by changes in the independent variables.

Solution to Exercise 14.7

The null hypothesis for testing the overall multiple regression equation is:

$$H_0 : R^2_{Pop} = 0.$$

As this implies that each of the regression coefficients equals 0, it may also be stated as:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$$

The null hypothesis is tested by an *F* statistic where:

$$\frac{SS_{Reg}/k}{SS_{Res}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

If the null hypothesis is rejected, *t* tests can be conducted on each coefficient to determine their significance in the model.

Solution to Exercise 14.8

Studying the residuals allows us to judge the appropriateness of both the underlying assumptions and the regression model used. A histogram of the residuals reveals whether the error term has a normal distribution. Also, the residuals can be plotted against the predicted values. The resulting graph should be random if the error terms have constant variances. Plotting the residuals against time, or the sequence of observations, reveals whether any correlation exists among error terms. Finally, a plot of the residuals against the independent variable reveals whether the model is appropriate, or whether additional variables are needed.

Solution to Exercise 14.9

In stepwise regression, predictor variables are entered or removed from the regression model one at a time based on their contribution toward explaining the variation inherent in the data. In forward selection, predictor variables are added to the model (initially starting with none) one at a time if they have significant FF ratios. In backward elimination, insignificant variables are removed from the model (initially all variables are included) until all remaining variables are significant. The stepwise procedure combines forward selection with backward elimination to arrive at a model.

The purpose of stepwise regression is to select a smaller subset of a larger number of predictor variables which will account for most of the variation in the data.

Solution to Exercise 14.10

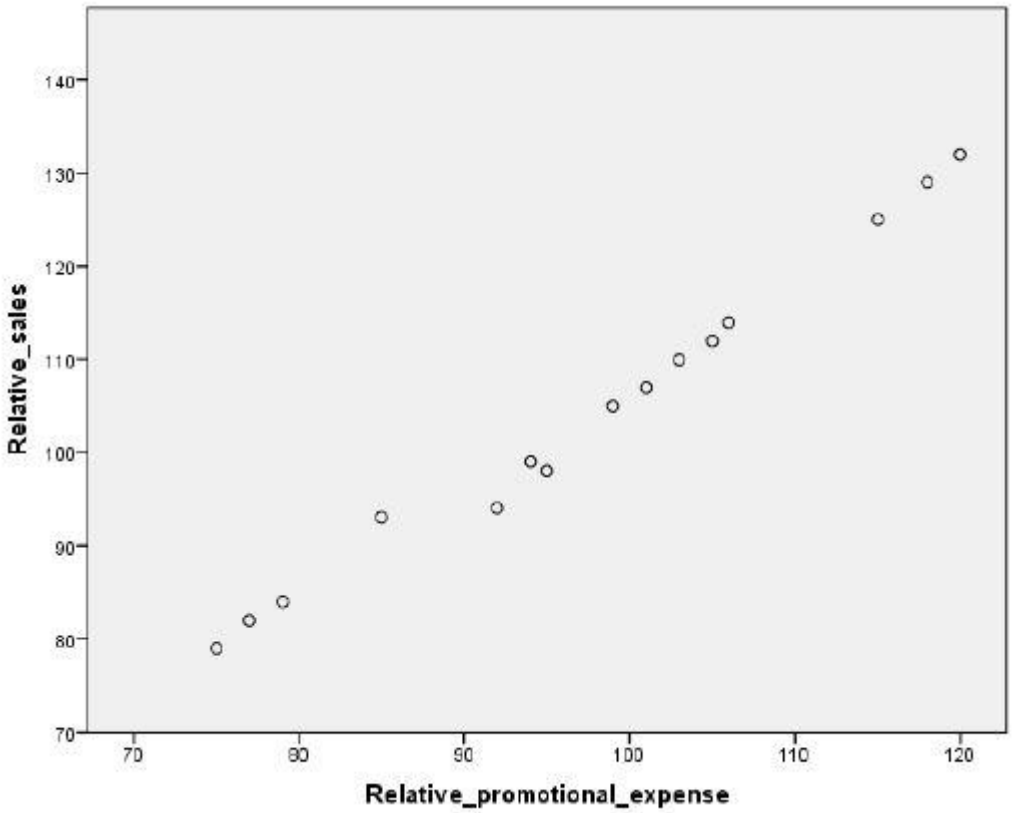
Multicollinearity refers to very high inter-correlations among the predictor variables. Multicollinearity can result in the following problems.

- Precise estimation of the partial regression coefficients may vary with the sample.
- The magnitudes and signs of the partial regression coefficients may vary with the sample.
- Unambiguous measurements of the relative importance of the independent variables in explaining the dependent variable become difficult.
- The incorrect removal or inclusion of predictor variables may occur in a stepwise procedure.

Solutions to exercises on the block's topics

1.

a. The scatterplot is:



There is a strong, positive linear relationship between relative sales and relative promotional expenses.

b. The product moment correlation should be used because an interval scale is used and no third variable can confound the interaction between X and Y.

c. We have:

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-7.927	3.596		-2.204	.046
	Relative promotional expense	1.149	0.036	.994	31.496	.000

a. Dependent Variable: Relative_sales

Hence:

$$\widehat{\text{Relative sales}} = -7.927 + 1.149 \times \text{Relative promotional expenses.}$$

d. $\hat{\beta}_0 = -7.927$ represents the y-intercept, i.e. the value of relative sales when there are no relative promotional expenses. Of course, relative sales cannot be negative, but from the scatterplot in (a) it is clear there were no instances of zero, or near-zero, relative promotional expenses ($x_{min} \approx 75$). $\hat{\beta}_1 = 1.149$ represents the slope of the regression line, so for a one-unit increase in relative promotional expenses there is a 1.149 unit increase in relative sales.

e. We have:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	3800.593	1	3800.593	991.981	.000[b]
	Residual	49.807	13	3.831		
	Total	3850.400	14			

a. Dependent Variable: Relative sales
b. Predictors: (Constant), Relative_promotional_expensive

f. Hence the model is statistically significant due to the pp-value of 0.000 corresponding to the *F* statistic value of 991.981.

g. We have:

$$\widehat{\text{Relative sales}} = -7.927 + 1.149 \times 100 = 106.973$$

We have:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.994[a]	.987	.986	1.957

a. Predictions: (Constant), Relative_promotional_expense

$R^2 = r^2$ = 0.987. This means that 98.7% of the variation in relative sales is explained by relative promotional expenses.

2.

a. We have:

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.535	.471		1.136	.280
	Quality	.976	.097	.798	10.096	.000
	Price	.251	.071	.278	3.522	0.005

a. Dependent Variable: Preference

Hence:

$$\widehat{\text{Preference}} = 0.535 + 0.976 \times Quality + 0.251 \times Price.$$

b. The partial regression coefficients indicate that a one-unit increase in quality rating will increase preference by 0.976 units, when price is held constant. A one-unit increase in price will increase preference by 0.251 units, when quality is held constant.

c. We have:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	122.831	2	61.415	105.826	.000[b]

	Residual	6.384	11	.580	
	Total	129.214	13		

a. Dependent Variable: Preference
b. Predictors: (Constant), Price, Quality

Hence the model is statistically significant due to the pp-value of 0.000 corresponding to the F statistic value of 105.826.

d. We use the regression output in (a). For quality, t=10.096 which is highly significant since its p-value is 0.000. Similarly for price, t=3.522 which is highly significant with a p-value of 0.005.
e. We have:

Correlations

		Quality	**Price**
Quality	Pearson Correlation	1	.531
	Sig. (2-tailed)		.051
	N	14	14
Price	Pearson Correlation	.531	1
	Sig. (2-tailed)	.051	
	N	14	14

Multicollinearity is unlikely to be a major problem because quality and price tend to be only moderately correlated, with r=0.531. This sample correlation coefficient is borderline insignificant (when testing H_0: ρ=0) with a pp-value of 0.051. Of course, if we assume a one-tailed test (if correlated at all, we would expect quality and price to be *positively* correlated) the p-value would be 0.051/2=0.0255, and hence significantly different from zero, although the true correlation is unlikely to be extremely high (which is typically required for multicollinearity to be problematic).

3.

a. The relationship seems believable since the higher a person's income, the more likely a person is to buy prepared dinners. A coefficient can be small in magnitude yet be significant if its standard error is small.
b. Without knowing the sums of squares, we cannot determine how good the model is.
c. We have:

$$\widehat{\text{Expenditure}} = 23.5 + 0.003 \times 30000 = £113.40$$

d. The residual is $130-(23.4+0.003\times40000)=£13.40$
e. A negative residual means that the model predicted a value which was greater than the actual value.

4.

a. We have:

Correlations

		Preference	Comfort	Style	Durability
Preference	Pearson Correlation	1	.573**	.642**	.559*
	Sig. (2-tailed)		.008	.002	.101
	N	20	20	20	20
Comfort	Pearson Correlation	.573**	1	.560*	.534*
	Sig. (2-tailed)	.008		.010	.015
	N	20	20	20	20
Style	Pearson Correlation	.642**	.560*	1	.364
	Sig. (2-tailed)	.002	.010		.114
	N	20	20	20	20
Durability	Pearson Correlation	.559*	.534*	.364	1
	Sig. (2-tailed)	.010	.015	.114	
	N	20	20	20	20

- **. Correlation is significant at the 0.01 level (2-tailed).
- *. Correlation is significant at the 0.01 level (2-tailed).

All correlations are (highly) statistically significant, except for the correlation between ‘Durability’ and ‘Style’.

b. We have:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.573[a]	.328	.291	1.599

. Predictors: (Constant), Comfort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	22.516	1	22.516	8.804	.008[b]
	Residual	46.034	18	2.557		
	Total	46.034	19			

a. Dependent Variable: Preference
b. Predictors: (Constant), Comfort

The overall regression is significant (the FF statistic is 8.804 with a pp-value of 0.008), and R2=0.328R2=0.328. Therefore, comfort explains 32.8% of the total variation in preference for boots. Also:

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.028	1.337		.021	.983
	Comfort	.921	.310	0.573	2.967	.008

c. Dependent Variable: Preference

Hence:

$$\widehat{\text{Preference}} = 0.028 + 0.921 \times \text{Comfort}$$

Comfort is significant at the 1% significance level (t=2.967, with a two-tailed p-value of 0.008).

c. We have:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.642[a]	.412	.380	1.496

. Predictions: (Constant), Style:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	28.272	1	28.272	12.634	.002[b]
	Residual	40.278	18	2.238		
	Total	68.550	19			

a. Dependent Variable: Preference
b. Predictors: (Constant), Style

The overall regression is significant (the FF statistic is 12.634 with a pp-value of 0.002), and R2=0.412R2=0.412. Therefore, style explains 41.2% of the total variation in preference for boots. Also:

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.078	.848		1.271	.220
	Style	.739	.208	.642	3.554	.002

c. Dependent Variable: Preference

Hence:

$$\widehat{Preference} = 1.078 + 0.739 \times \text{Style}$$

Style is significant at the 1% significance level (t=3.554, with a two-tailed p-value of 0.002).

d. We have:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.559[a]	.312	.274	1.619

. Predictions: (Constant), Durability

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	28.272	1	28.272	12.634	.002[b]

	Residual	40.278	18	2.238	
	Total	68.550	19		

a. Dependent Variable: Preference
b. Predictors: (Constant), Style

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	21.390	1	21.390	8.164	.010[b]
	Residual	47.160	18	2.620		
	Total	68.550	19			

a. Dependent Variable: Preference

b. Predictors: (Constant), Durability

The overall regression is significant (the F statistic is 8.164 with a p-value of 0.010), and R^2= 0.312. Therefore, durability explains 31.2% of the total variation in preference for boots. Also:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.307	.961		1.361	.190
	Durability	.598	.209	.559	2.857	.010

a. Dependent Variable: Preference

Hence:

$$\widehat{Preference} = 1.307 + 0.598 \times Durability$$

Durability is significant at the 1% significance level (t=2.857, with a two-tailed p-value of 0.010).

e. We have:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.741[a]	.548	.464	1.391

a. Predictors: (Constant), Durability, Style, Comfort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37.599	3	12.533	6.479	.004[b]
	Residual	30.951	16	1.934		
	Total	68.550	19			

a. Dependent Variable: Preference

b. Predictors: (Constant), Durability, Style, Comfort

The overall regression is significant (the F statistic is 6.479 with a p-value of 0.004), and R^2 = 0.548. Therefore, the model explains 54.8% of the total variation in preference for boots. Also:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.539	1.183		-.455	.655
	Comfort	.258	.360	.161	.717	.484
	Style	.504	.234	.438	2.152	.047
	Durability	.336	.214	.313	1.571	.136

a. Dependent Variable: Preference

Hence:

$$\widehat{Preference} = -0.539 + 0.258 \times Comfort + 0.504 \times Style + 0.336 \times Durability.$$

The coefficients for all three variables (comfort, style and durability) are significant in the individual bivariate regressions. However, only the coefficient for style is significant in the multiple regression (at the 5% significance level). In each case, the coefficient is smaller in the multiple regression than it is in the bivariate regression. In the bivariate case, all of the explained variation is attributed to the independent variable in the equation since it is the only independent variable. However, in multiple regression, each independent variable shares the explained variation with the other independent variables.

Commentary on Discussion points

1. Issues to discuss include the information gained from conducting regression and the data requirements for regression. Regression shows the relationship between a set of independent variables and the dependent variable. There are cases where this information is not needed by the decision-maker. Also, the data must be collected to correspond to the requirements of regression, i.e. the dependent variable must be metric.
2. Bivariate correlation, bivariate regression, multiple regression and ANOVA are all related in that they either build on one another or are variations of one another. Bivariate correlation is an index which is used to determine if a linear relationship exists between the dependent variable, Y, and the independent variable, X, of a

bivariate regression equation. (Note that the equation has only one X variable and one Y variable - two variables in total - hence the name bivariate regression.)