Simple linear regression and correlation

Contents

Introduction

In many instances, decision makers and researchers are interested in understanding a key variable, and, in particular, the variation in that variable. For example, why does productivity performance vary so much between employees—why don't they all achieve the same productivity score? Or, why do the heart rates of people vary—why don't we all have the same heart rate? Clearly, productivity performance can vary for a range of reasons—experience, job security, staff morale, family pressures, type of job, etc. And heart rates can vary for many reasons as well—fitness level, age, weight, race, diet, etc. If we can detect any relationship (or dependency) between our key variable and another variable, that second variable may be useful in predicting, explaining or controlling the variation in the key one.

In many areas of business we are interested in relationships between variables. A production manager, for example, may be interested in the relationship between units produced and costs of production and any findings may be useful in reducing costs.

A marketing manager may be interested in any potential dependency between gender and acceptance of brand or product, and a detected relationship may suggest different market segments and, hence, different marketing strategies for each gender group. If management found that productivity and morale were related, then it would suggest some company emphasis on staffing issues may improve morale and therefore productivity. It is clear that decisions and strategies can be based on the existence, or non-existence, of these relationships.

We include *descriptive techniques* such as scatter plots, summary measures and regression equations. However, we also need to cover *inferential techniques,* since much correlation and regression analysis is conducted on sample data.

We concentrate on regression techniques, as they allow us to determine the mathematical form of a relationship. That form is particularly useful from explanatory, predictive and decision-making points of view.

Note that we will often use the term 'equation' and 'model' interchangeably. Thus, a *regression model* and *regression equation* can be considered to be the same concept.

Objectives

- use scatter plots to determine the nature of relationships between numerical variables
- calculate a correlation coefficient for a set of data
- conduct a hypothesis test on the correlation coefficient
- understand the assumptions underlying the regression model

- generate a simple regression model and explain the regression equation
- analyse the results of simple linear regression and correlation analysis and carry out simple diagnostic checking on the validity of these results.

Examining relationships

We use *descriptive techniques* such as:

- the scatter plot as a graph for helping us explore for linear and non-linear relationships
- the correlation coefficient, r (and coefficient of determination, r^2)as measures of the strength of any linear relationship
- the regression equation as a mathematical representation of a possible relationship.

Correlation analysis provides an indication of the direction of a relationship as well as a numerical measure of the strength of *linear* association between two numerical variables or two groupings of numerical variables. It is very useful in cases where there is not a strict cause/effect relationship established between the variables. Job satisfaction v productivity is a good example. Does low productivity affect job satisfaction? Does job satisfaction affect productivity? In this situation we seek to establish an *association* between the variables. Thus, even though we can't say which variable 'causes' or 'affects' the other, we may be able to establish a significant link between the two, which points to the need for more in-depth analysis. Correlation can be determined for samples taken from whole populations and a hypothesis test can be used to make inferences about population associations between variables. Scatter plots are a useful measure for exploring for linear or non-linear relationships.

Regression analysis allows the user to quantify relationships between numerical variables (and between categorical variables which have been appropriately coded in numerical form). Regression enables us to determine the mathematical form of the relationship. This is called the *regression equation*.

Regression analysis mainly focuses on linear relationships between variables but can be modified to handle non-linear relationships.

Regression analysis is normally studied in two stages: *simple* regression or *multiple* regression analysis.

Simple regression analysis looks at the relationship between two variables only. As for correlation, regression also helps determine the direction of the relationship. One of the variables (independent, denoted by X) is thought to determine the other (dependent, denoted by Y). The cost of production and units of production provides a good example. It is clear that the number of

units of production (*X*) determine, to a large extent, the costs of production (*Y*).

Multiple regression analysis allows for more than one independent variable to be used in explaining variation in the dependent variable. It allows for relationships between many independent (*X*) variables and a solitary dependent (*Y)* variable of interest. For example, a company may wish to understand variation in productivity. This is the dependent *Y* variable. There are many candidate variables that could be considered as independent *X* variables: variables such as age, salary, gender and overtime hours. Used correctly, multiple regression analysis can determine which of the variables are important and which are unimportant. Further, it provides a quantitative representation of the effect of each variable on the dependent variable. Management can use this information to determine strategies to manage, control, monitor or predict productivity results, with subsequent impact on staff performance, efficiency and budgets.

Correlation and scatter diagrams

Scatter plots, also known as *scatter diagrams*, are plots which capture the joint position of pairs of observations of two variables in a two-dimensional diagram. They are useful in determining if a relationship exists between variables, and, if it does, what type of relationship is suggested (linear or non-linear; positive or negative).

When we construct a scatter plot, it is useful to draw the ‘line of best fit’ (think of ‘the straight line closest to all the points’) if a straight line seems to be the appropriate pattern, or to draw the ‘curve of best fit’ (think of ‘the curve that is closest to all the points’) if a straight line is not appropriate.

Correlation is a term used to describe certain aspects about the direction and relative strength of linear relationships between two numerical variables. It gives an indication of the direction of the relationship and also a quantified measure of the strength of linear association.

When you are examining bivariate relationships (just two variables) for a large data set with many variables, it is useful to begin with correlation analysis, followed by the analysis of scatter diagrams.

The advantage of conducting the correlation analysis first is that you can quickly detect potential linear relationships and check for potential multi-collinearity; but you will need to follow up with scatter diagrams to detect non-linear relationships.

The correlation coefficient is denoted by *r* and is a number between –1 and 1, with those extremes representing perfect linear relationships (negative and positive). A correlation of zero indicates no linear association. The closer *r* is to +1 or to –1, the stronger the linear relationship (everything else being equal); the closer to 0, the weaker the linear relationship (everything else being equal). A low value of *r* does not mean that variables are not related: they may in fact be related in a non-linear way.

Later in this topic you will learn how to conduct a hypothesis test for correlation. If you conclude that there appears to a linear relationship between two variables, we can proceed to generate a regression model relating the two variables.

Correlation is designed to detect the existence, direction and strength of *linear* relationships. Scatter plots are the primary tool we use for detecting *non-linear* relationships.

Scatter diagrams also indicate the direction of the relationship if one exists. This will be in terms of positive or negative (inverse) relationships. We can check what is suggested by our correlation work or scatter diagram work against our *a priori* reasoning. For example, we would expect productivity and morale to be positively related. A correlation coefficient or scatter diagram might confirm our expectation.

When you examine scatter diagrams, you should try to examine it from a viewpoint of whether or not a relationship exists between the variables. A random, horizontal, scatter of points may suggest no relationship, while a systematic scatter generally upwards, or downwards or in a curved form may suggest a form of relationship. The systematic relationships that are observed can then be defined as linear or non-linear, positive or negative.

Non-linear relationships are just as important to identify as linear ones. If the scatter diagram indicates a non-linear relationship, then adjustments to the modelling procedure are required, such as using a mathematical transformation of the variable (such as logarithms or squaring the variable). Not allowing for the non-linear effect may lead to a highly misleading model.

Remember that scatter diagram analysis may reveal that there is one or more very strong and important non-linear relationships that the correlation analysis doesn't detect.

Neither correlation analysis nor scatter diagrams should be taken as absolute proof of relationships or absence of relationships. In some cases, two variables that appear unrelated may be related via a third or other variables. Alternatively, two variables that appear related may not really be related once a third variable is introduced. For example, productivity may seem to be related to which work shift was operating. However, if there are differences in the age, experience etc. of the workers in the different shifts, then *these* may really be the variables that are related to productivity, not the shift itself. Identifying shift as a possible variable may still be useful; by that means we eventually realise that other variables are in fact the explanatory ones.

In cases where a regression model will be used, we need to define a dependent (or response) variable (Y) and an independent (explanatory or predictor variable (X). The main question you should ask yourself is 'Which variable am I trying to predict, explain, monitor or control?' With correlation analysis, choosing the dependent variable is not essential. However, with scatter diagrams, it is important to ensure the dependent variable, Y, is on the vertical axis.

You will see from your scatter diagram analysis, that you rarely (if ever) see a perfect relationship, that is, with all points falling on a straight line or

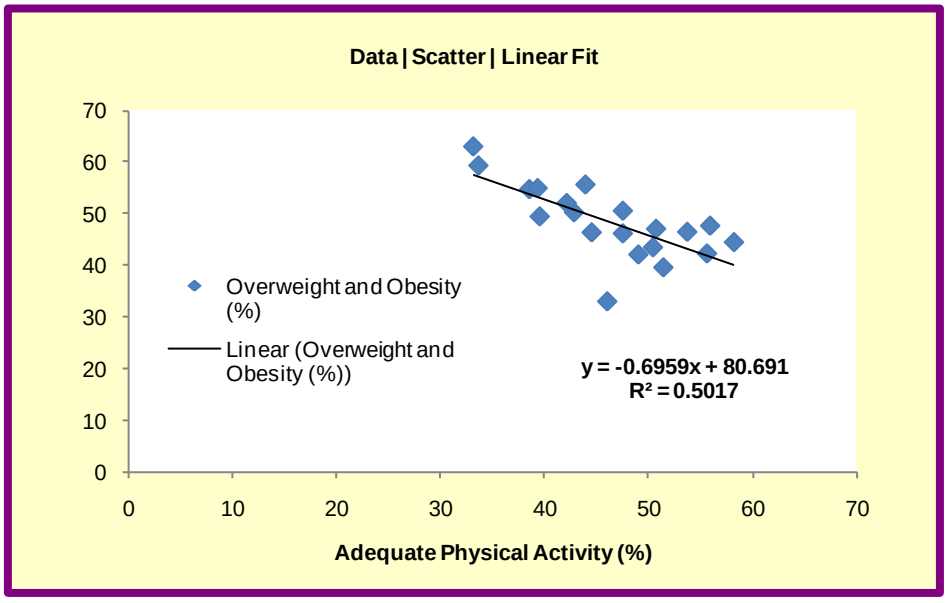
Before looking at simple regression and correlation in details, the following application summarises the key results for us.

EXAMPLE

This application examines the scatter diagram for the data on Adequate physical activity and Overweight and obesity. Note the following:

- The scatter of points.
- The line of best fit.
- The regression equation for that line, Y_{hat} = 80.691 - 0.6959X (with the order of terms rearranged compared to the graph).
- The R^2 value of 0.5017.
- *R,* (or *r*) the correlation coefficient, is not shown on the diagram but it has a value of -0.708.
- *R* is the square root of R^2. Or conversely, R^2 is the square of *R*.
- Normally we use just *r* for the correlation coefficient, but occasionally *R* is used instead. They mean the same thing.

Exhibit 1: Scatter diagram of Adequate physical activity and Overweight and obesity



In this instance, Overweight and obesity is the *Y* (or dependent) variable. Adequate physical activity is the *X* (independent, explanatory or predictor) variable.

While we cannot be absolutely sure of the direct effect of Adequate physical activity on Overweight and obesity, we can note the following:

- The relationship does appear to be linear and negative which accords with our *a priori* expectations or theory: 'the greater the Adequate physical activity rate the more likely the Overweight and obesity rate will be lower'.
- There appears to be neither a strong or weak association between the two variables as indicated by the high R^2 close to 0.50. This might suggest that, with caution, Adequate physical activity could be used as a predictor of Overweight and obesity in a regression model. In fact, it seems from our regression equation that with every percentage point increase in the Physical activity rate, Overweight and obesity tends to decrease on average by about 0.696%.
- Note that the *R* value of -0.708 is a misleading indicator of the absolute strength – ensure you always find R^2 and use it as your measure of strength. Here we estimate that 50.17% of variation in the Overweight and obesity rate can be explained by or attributed to variation in the Adequate physical activity rate. (Note, that for lower R values, the R^2 value can be very low:, an R value of 0.6 has an R^2 of just 0.36 or 36%, while an R value of 0.3 has an R^2 of just 0.09 or 9%.)
- By substituting values for Adequate physical activity into the equation we can predict Overweight and obesity. For example, if the Adequate physical activity rate is 40, then Y_{hat} = 80.691 - 0.6959x40 = 52.857%. Thus, as a point estimate we predict an Adequate physical activity rate of 40% will result in, on average, Overweight and obesity of 52.857%.

This problem is an instance of simple regression, since we only have two numerical variables involved: one dependent and one independent. Notice how we have been able to construct a simple graph for the problem, as well as gain a measure of the strength of the relationship and derive the mathematical form of this relationship, which itself has useful interpretations and can be used for predictions.

Simple regression

Simple regression analysis and correlation analysis are ways in which a linear relationship between two variables may be quantified.

You should note that the sample regression is an estimate of the population regression equation, that is, the equation we would obtain if we could look at *all* pairs of variables, rather than a relatively small sample. The sample equation is written as:

$$\hat{y} = b_0 + b_1 x$$

One of your key objectives should be to understand the equation, its components and associated results. The estimated intercept term (b_0), in

many cases, represents an estimate of the average value of the dependent variable when the independent variable effect is zero. In some circumstances it has meaning, but in many cases it is meaningless. For our Adequate physical activity/Overweight and obesity application, this intercept does not seem to have a useful meaning as it does not make sense to have a zero physical activity rate! Thus, the regression model will calculate an intercept but it does not necessarily have a valid interpretation. This is due to the intercept being an extrapolation from the sample values in the data. In other cases, the intercept can be meaningful. You should examine the data. If the *X* (independent) data values are near zero, then the intercept probably has some interpretation, otherwise it may be meaningless.

By far the most important interpretation is for the estimated slope b_1. It is the estimate of the average change predicted for the dependent variable when the independent variable changes by one unit. In the Adequate physical activity/Overweight and obesity application it would represent the expected change in the Overweight and obesity rate when there was a change of one unit of *X*, namely 1% in the Adequate physical activity rate.

This equation would have very useful implications for strategy, control and prediction. We have at our disposal a tool that not only is able to predict Overweight and obesity for a given level of Adequate physical activity, it is also able to predict changes in the Overweight and obesity rate brought about by changes in the Adequate physical activity rate.

Note that you should understand that you should only use your equation for predicting *Y* for values of *X* within the range of *X* values used to generate the model. The reason is that the straight line relationship might not hold outside those values. A good example is height v age. If we fitted a regression line through the data for 20 children and teenagers aged from 3 to 18, we might find a noticeable positive relationship: that is, the older the person, the taller they are. But would it make sense to use that equation to make predictions beyond 18: for example, to predict the height of a 22-year-old or a 42-year-old, or indeed the height of an 82-year-old? Thus, use your regression results carefully.

Measures of variation and goodness of fit

Since the underlying regression model is likely to incorporate some random error component, this implies that the predicted regression line will not fit the sample data exactly. This can be seen quite clearly from the scatter plot in exhibit 1, where the points scatter around the line and there only a few points actually right on the line. Thus, there will be error in using our equation for providing point estimates for each value of *X*. The error for a given data point is measured as the difference between the prediction and the actual data values.

This section deals with ways of measuring that error over all data points. An analysis of variation of the data values and corresponding error variation can provide measures of model effectiveness. We introduce two new but

very important terms: the *coefficient of determination* (which we have already introduced as r^2 or R^2) and the *standard error of the estimate*.

The differences between the points on the line and the data points are called the residuals.

The size of the residuals over all data points gives us an indication of how well or poorly the model predicts the actual dependent variable values. Larger residuals suggest that the model is not a good fit and not a good predictor of the dependent variable. There are two ways of using the residuals for analysing goodness of fit.

The first is through the *coefficient of determination* which provides us with a quantitative measure of goodness of fit and hence the predictive ability and overall usefulness of the model. Since it is expressed as a percentage, the r^2 is bound by 0 and 1 (0 and 100%), with 1 (or 100%) indicating a perfect explanation.

The *standard error of the estimate* is another measure for analysing goodness of fit. It might make some sense to calculate an 'average' residual which can be used as an indication of the level of prediction error. The regression model is set up in such a way so that the residuals will be both positive and negative and will sum to zero. Hence, their average will be zero. The standard error of estimate is a way of calculating an 'average' residual for the regression model.

Assumptions of the regression model

The preceding analysis in simple regression was undertaken with scant mention of the underlying assumptions of the model. The assumptions of the model are extremely important because inferences or generalisations of sample relationships (estimated by the model) to population relationships of interest depends crucially on the validity of the assumptions. Before and after applying the regression model, the user should make some elementary checks on assumption validity.

From this section onwards we are discussing how the regression model estimated from the sample of data can be applied to the much wider population. We are primarily concerned with the population relationship between the variables of interest, even though we spend most of our efforts dealing with the sample data and the results derived there from. The techniques discussed in previous sections were measures that described sample relationships, that is, we determined the regression equations etc. from sample data. Thus, if we cannot take a census, but can only take a sample, how can the sample results be utilised? The answer is that we have to generalise from the sample relationships to a broader population level. As a first step, we should determine if the results of the sample regression or correlation model are valid. Are the assumptions on which the model is based valid, or is there some violation of those assumptions? Generalisations from samples to populations will crucially depend upon the validity of the assumptions.

Depending on the severity of the violation, the results can be invalidated. We will not be able to discuss assumption violations in detail in this course but it is a crucial part of regression model building. We will use some simple diagnostic checks in the next section to help detect assumption violations. They are based around the calculation and plotting of the residuals.

Regression diagnostics—Residual analysis

The residuals can be useful in two main ways.

First, they indicate goodness of fit—as discussed earlier when we discussed the coefficient of determination and the standard error of the estimate. If the model is good, then the residuals should be relatively small and should truly reflect the random variation. Thus, R^2 should be relatively large and S_{yx} should be relatively small.

The second main use of residuals is as a simple way to check violations of the assumptions about the regression model.

The residuals should be randomly distributed if our regression assumptions are to hold. Any non-randomness in the residual pattern would suggest that there has been an assumption violation of some description. The validity of the results may then be questionable.

The examination of residual plots is only a first step in examining violation of assumptions. For the non-expert, however, it is some indication.

Non-random residual plots indicate that there are assumption violations. You will see that residuals can be calculated and plotted in different ways, in particular, in absolute form or in some standardised form.

Standardised residuals will give some indication of points at which the model is not predicting well. If we assume normality, then the rule for the standardised normal distribution is ± 3 (as one rule of thumb we could choose). Standardised residuals within this bound are no real cause for concern. However, residuals outside these bounds may be an indication of an outlier (or extreme value), though these can also be detected from a scatter plot. In some circumstances the extreme values can radically alter the estimates of the slope and intercept of the regression model and provide misleading results. (Genuine mistakes must be corrected; otherwise a choice has to be made as to whether or not to exclude a potential outlier from the data.)

Inference in regression and correlation

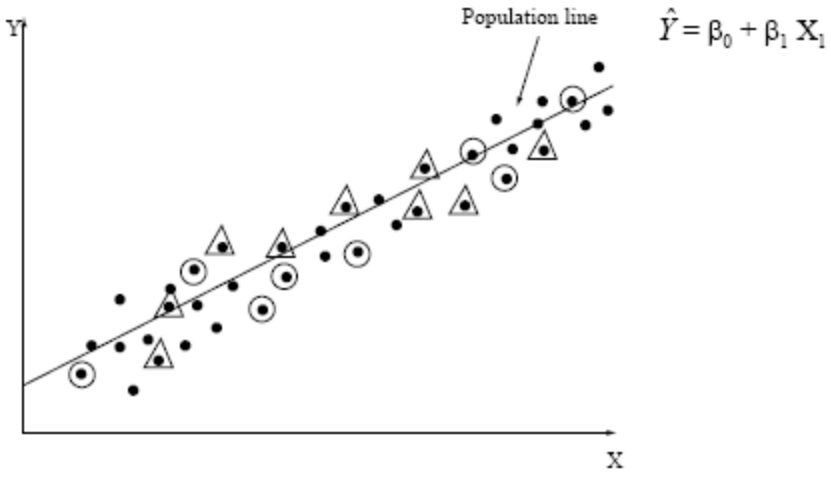
The need for inferential techniques in regression and correlation

Typically, we wish to have results that are indicative of the population relationship between the variables, not just from a particular sample. We desire results that can be generalised to the wider population. We therefore

need inferential procedures to make this link. There are actually several very important areas in regression and correlation that require inferential techniques. In this section we cover some of the most important ones.

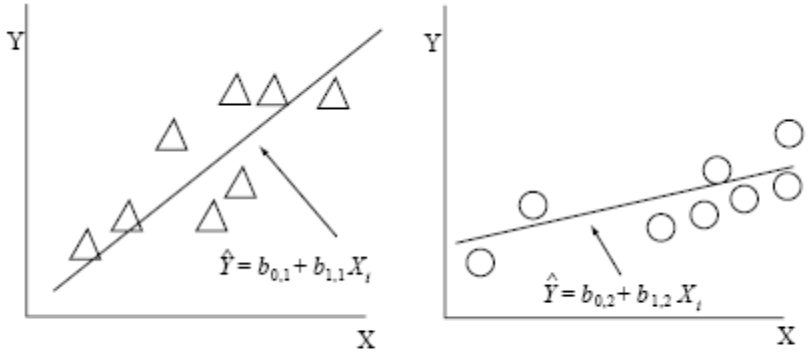
The inferential methods we use in regression and correlation analysis can be thought of in a similar way to the discussion of sampling distributions and the use of confidence intervals and hypothesis tests. The estimated statistics from the regression—in particular the estimated intercept (b_0) and the estimated slope (b_1), and the sample population correlation coefficient in the correlation model—are sample statistics that will vary with different samples taken from the population. Each of the above statistics has a particular sampling distribution.

Exhibit 2: **Population regression line for *Y*, *X***



Consider two separate randomly chosen samples from the population denoted by Δ and O respectively. The sample values (for the two samples) and the (two) fitted least squares regression lines appear in exhibit 9.3. These are two extreme examples of two different samples that could be drawn from a population, but you should see from them that a sample regression line is likely to give different (hopefully not great) results compared to the population line.

Exhibit 3: **Two samples from the population and the estimated sample regression lines**



Sample 1 Sample 2

Applying the least-squares method to each sample, yields a different sample regression line with different estimates of b_0 and b_1. Clearly, the sample repression line will vary depending on the sample. Hence, there must be sampling distributions for both b_0 and b_1. (We will see later that similar analysis can be applied to the sampling distribution of the sample correlation coefficient, r.)

The sampling distribution of the statistics will almost always be normal or follow the *t* distribution if we our regression assumptions are true, and/or we are dealing with large samples.

In the next two sub-sections you learn how to make inferences about the coefficients in the regression equation and how to make inferences about the simple correlation coefficient.

t test for the slope in simple regression

We can make inferences about the slope coefficient in a number of ways. One is to test whether or not the equivalent population parameter slope could be 0 using a *t test for the slope*. Another approach is to construct a confidence interval estimate of the population slope: if that interval contains 0 then we cannot reject the hypothesis of the population slope being zero; if the two sides of the interval are both the same sign (either both positive or both negative), then the interval indicates that the population slope is non-zero and also provides us with a useful estimate of the true value of the true slope.

The main test for the slope is generally a hypothesis test with the null hypothesis that the true slope equals zero, as opposed to the alternative that it is different from zero. There is a special significance to this hypothesis test and, in particular, to the implications of the null and alternative hypotheses.

If we test that the slope equals zero, then we are suggesting that if the null hypothesis is supported, there is no linear relationship between the variables (*Y*, *X*). A rejection of the null hypothesis suggests that a linear relationship *does* exist between the variables. Other hypotheses can be tested but the above hypothesis is generally the most informative. Further, most software packages produce regression models which provide the statistic necessary to conduct the hypothesis test that the slope is equal to zero.

The relevant standardised statistic of the hypothesis test as given by software. The denominator is the *standard error* for the equation coefficients. Software output includes the *p*-value, which provides us with a quick method of assessing the significance of individual slope coefficients. If the *p*-value is less than 5% (or alpha, the level of significance we wish to use) then we would conclude significance (that the population slope was not 0). On the other hand, if the *p*-value is large, we could conclude the variable is not significant in our model, as it seems the slope coefficient might be 0.

t test for the correlation coefficient

We can also make inferences about the correlation coefficient, *r*. The test for a single correlation coefficient is set up with the null hypothesis which assumes there is no linear correlation between the pair of variables, that is, that the true population correlation coefficient (denoted by the Greek symbol ρ, and pronounced 'rho') is zero. As the alternative hypothesis, we test that the true correlation coefficient is not zero, that is, we test for $\rho \neq 0$.

Again the hypothesis test can be interpreted in the same way as in other cases: we would observe that 'high *t* statistics and low *p*-values' indicate that there is correlation while 'low *t* statistics and high *p*-values' indicate no correlation.

Estimation of predicted values

One of the main uses for a regression model is for prediction. You will see that we can actually make two types of predictions, and we give them different terms. If we wish to *estimate the mean value of Y for a given value of X*, we construct a *confidence interval estimate.* If we wish to *predict the value of Y for a given value of X*, we construct a *prediction interval for Y.*

To explain the difference, say we are interested in investigating the annual incomes of full-time males and their ages. Income is the *Y* variable, and age is the *X* variable. If we are particularly interested in 28-year-olds, there are two types of predictions or estimates we might want to do:

- Estimate the mean income for all 28-year-olds. In this case, you would construct a *confidence interval estimate for the mean,* using the regression model.

- Predict the mean income for a given 28-year-old. In this case, you would construct a *prediction interval* for the income for that individual, using the regression model.

In neither case can we use just the value given by the regression line, since it is only a point estimate. We need to work with a margin of error either side of the line. The margin of error is somewhat less for the *confidence interval for the mean*, compared to the *prediction interval for individual Y*.

Putting it all together

APPLICATION

In this application, the marketing manager of a supermarket chain would like to determine the effect of shelf space on weekly sales of pet food. A random sample of 12 equalised stores was selected. Assume the marketing manager has estimated the simple regression model to predict weekly sales based on shelf space. The equation and associated results are given below to be discussed in general terms to give an overview of how to interpret the results. Later, in the next exercise, you are asked to generate the computer output using software.

The estimated equation is:

Sales = 1.45 + 0.074 (Space)

The intercept has no viable interpretation, as sales would be zero if the amount of shelf space allocated was zero. The slope suggests that every extra unit of shelf space will increase sales on average by .074 units. In other words, an increase of shelf space of 1 foot would increase sales by $7.4. A confidence interval for the true value of the slope parameter is provided. It suggests that a one foot increase in shelf space will have an effect on sales somewhere between $3.9 and $10.9 (with 95% confidence).

The r^2 value (R-squared) for this regression is reasonable at 68.39%, indicating reasonable explanatory power.

The inferential procedures that were discussed in the previous sections can now be used to suggest if the sample results can be applied to the population.

We check the t statistics on the slope and intercept. Since the intercept has no meaning here, we will concentrate on the t statistic for the slope. Recall that the t statistic is for the hypothesis test that the true slope parameter is zero against the true slope parameter that is not zero. In this case, the t statistic for the slope is 4.6517. The t statistic is large (and out of general interest, exceeds the rule of thumb value of ± 2). It suggests that we should reject H_0 which implies that there is a relationship in the population between sales and shelf space. Do not use ± 2 as the overriding rule, use it as a quick check only. The p-value is the preferred way—is it bigger or smaller

than your level of significance, for example, 5%? In this case, the *p*-value is small at only 0.00091 or 0.091%.

Overall, the inferential procedures suggest that there is a relationship between sales and shelf space. In the absence of further evidence, the estimated equation represents that relationship. However, you should be aware that the sample is small and conclusions from small samples should be qualified.

As a matter of course you should check the residuals from the regression. A non-random scatter of residuals may invalidate the above results. When we plotted the residuals, we obtained an acceptable random pattern, with general scattering either side of the zero line. One slight concern is the reduction in variability of the residuals and, hence, around the regression line as our *X* variable, Space, increases. There is a distortion of the plot because the variable shelf only takes on values of 5, 10, 15 and 20. Also, the number of observations is small and caution should be taken when making conclusions. There does not seem to be a definite reason from the plot to invalidate the regression results.

Hence, we could use these results with confidence.

Summary

This topic introduced regression and correlation analysis. These techniques are used to examine relationships between numeric variables. Correlation is a procedure to determine the linear association between two variables. A correlation approaching 1 (–1) indicates a strong positive (negative) relationship while a correlation approaching 0 suggests no linear association.

We also discussed scatter plots in detail as important tools in assessing whether a linear relationship is appropriate and, if not, in helping detect if a non-linear relationship is appropriate. In a sense, residual plots are also scatter plots.

Regression analysis differs from correlation in that it attempts to quantify the relationship between variables with one variable being the response or dependant variable. Simple regression involved a single dependent variable and a single independent variable. Multiple regression is for a single dependent variable and more than one independent variable.

Estimation of these models is generally from sample data and hence inferential techniques come into play when trying to use the sample regression and correlation results to draw conclusions about the equivalent population parameters.

The mathematics and calculations involved in regression and correlation are quite demanding, and we rely on computer software to determine results. Results studied so far include the estimated equation and its two coefficients, r^2 (or R^2), the correlation coefficient, *r,* the standard error of the estimate, residuals and prediction intervals.

Inferential procedures for the regression coefficients include confidence intervals and hypothesis tests for the slope and tests for explanatory power of the model. The main statistics for inferential procedures are the *t* and *F* statistics.

Further resources

Black, K 2008, *Business statistics for contemporary decision making*, 5th edn, Wiley, NJ.

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati.

Selvanathan, A, Selvanathan, S, Keller, G & Warrack, B 2006, *Australian business statistics*, 4th edn, Nelson Thomson Learning, Melbourne.