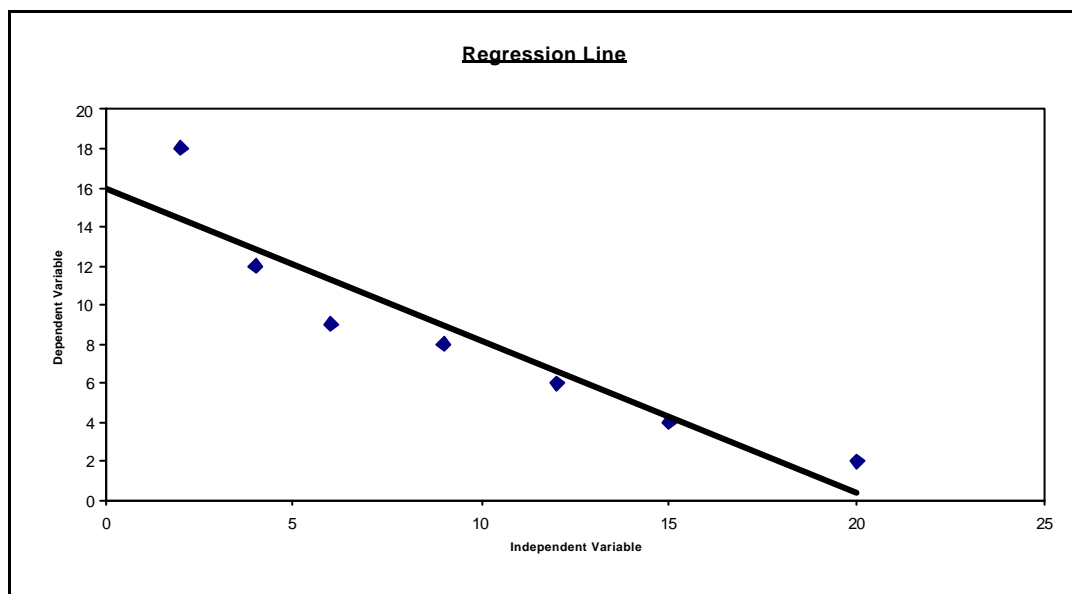


Chapter 10: Exploring Bivariate Analysis

Progress Questions

1. **Correlation** is a measure of the strength of association between two variables. If a line graph were plotted of one variable against the other, there would be perfect correlation if all the plotted points fell along a straight line. In reality, though, nothing is likely to be so perfect – even only taking into account sampling errors – so the closeness of fit to a straight line is a good indicator (see the illustrations in the chapter). The **correlation coefficient** is used to indicate the closeness of fit. This will range from +1 to –1, with both extremes indicating perfect relationship between the values of the two variables. The + and – labels indicate the direction of fit, i.e. a direct relationship (in which both increase or decrease together) or an inverse relationship (where one increases as the other decreases). A correlation of 0 means no closeness of fit at all.

Linear regression is related to correlation in the sense that the process attempts to make mathematical sense of the pairs of values. It determines the formula of the best straight line graph that is the best fit for the points on the scatter graph. An example is shown below.



Linear regression has calculated the best straight line as being represented by the formula:

$$y = mx + c$$

$$y = -0.8x + 15.9$$

In other words, if a value of the *independent variable* is known the value of the *dependent variable* may be predicted (or vice versa). However, the degree to which we might wish to trust this calculation will depend on a number of factors, and one of these will be the degree of correlation, because the greater the *correlation* the

better the regression line will fit the observations. In the example above, the correlation coefficient is -0.93 . This is a very high degree of correlation and would give more confidence in prediction.

2. **Pearson's r** is the method for calculating the *correlation coefficient* when the data is *interval*. **Spearman's ρ** is the method used with *ordinal* data. **Phi** is the method used when the data is *dichotomous*. The example quoted in the tutorial of this chapter presents a frequent complication – that of different data types being used. In this case salary is *interval* but gender will be a *dichotomy*. In such circumstances, the rule to be applied is that of using the method appropriate for the lowest order of data type. The hierarchy of data type is:

Interval - Ordinal - Dichotomy - Nominal

This means that **Phi** would be used as the form of *correlation coefficient*.

3. A *correlation coefficient* of -0.3 is quite low. If you recall from chapter 10, the *coefficient of determination*, the proportion of variance in the *dependent variable* accounted for by the *independent variable*, is calculated by squaring the *correlation coefficient*. This means that in this example, the *coefficient of determination* is 0.09 , or 9% . In other words, while there is a relationship between these two variables (assuming it is statistically significant) it is not a strong or near exclusive one and I would not be keen to use linear regression alone as a means of predicting likely *dependent variable* values. Chapter 11 describes a related technique called *multiple regression* that assumes a number of variables will interact to influence the dependent variable. It is this technique I would look towards.
4. A scatter graph is not obligatory, but it does produce a very powerful visual impression of the strength of a relationship between two variables. It may even help me decide not to bother with further analysis by *correlation* or *regression*.