2012-10

Correlation and Regression

Donal O'Brien
Technological University Dublin, donal.obrien@tudublin.ie

Pamela Sharkey Scott
Technological University Dublin

CHAPTER 6

CORRELATION AND REGRESSION

Donal O'Brien & Pamela Sharkey Scott

INTRODUCTION

A correlation is a measure of the linear relationship between two variables. It is used when a researcher wishes to describe the strength and direction of the relationship between two normally continuous variables. The statistic obtained is Pearson's product-moment correlation (r), and SPSS also provides the statistical significance of r. In addition, if the researcher needs to explore the relationship between two variables while statistically controlling for a third variable, partial correlation can be used. This is useful when it is suspected that the relationship between two variables may be influenced, or confounded, by the impact of a third variable.

Correlations are a very useful research tool but they do not address the predictive power of variables. This task is left to regression. Regression is based on the idea that the researcher must first have some valid reasons for believing that there is a causal relationship between two or more variables. A well known example is the consumer demand for products and the level of income of consumers. If income increases then demand for normal goods such as cars, foreign travel will increase. In regression analysis, a predictive model needs to fit to both the data and the model. And then we can use the result to predict values of the dependent variable (DV) from one or more independent variables (IVs). In straight forward

terms, simple regression seeks to predict an outcome from a single predictor; whereas multiple regression seeks to predict an outcome from several predictors.

CORRELATION

Correlation analysis is useful when researchers are attempting to establish if a relationship exists between two variables. For example, a researcher might be interested in whether there is a relationship between the IQ level of students and their academic performance. It would be expected that high levels of IQ will be linked to high levels of academic performance. Such a relationship indicates a high degree of positive correlation between the IQ level of students and their academic performance.

It is also possible to find correlation where the direction is inverse or negative. For example, a company making computer chips records a high level of defective products. They decide to invest considerable capital in improving their machines and processes. Over time it is seen that as the investment was increased the number of defective products declined. Therefore there is a negative correlation between investment and defective products.

It is important to note that correlation provides evidence that there is a relationship between two variables. It does not, however, indicate that one variable causes the other. In other words, the correlation between variable A and B could be a result of A causing B, or B causing A, or there could be a third variable that causes both A and B. It is for this reason that researchers must always take into account the possibility of a third variable impacting on the observed variables. By using partial correlation it is possible to statistically control for these additional variables, which allows the possibility of a clearer, less contaminated, indication of the relationship between the two variables of interest.

It is also important to understand the difference between a statistically significant correlation coefficient between the variables and what is of practical significance for the

sample. When using large samples, even quite small correlation coefficients can reach statistical significance. For example, although it is statistically significant, the practical significance of a correlation of 0.09 is quite limited. In a case like this, the researcher should focus on the actual size of Pearson's r and the amount of shared variance between the two variables. To interpret the strength of the correlation coefficient, it is advisable to take into account other research that has been conducted in that particular area.

SPSS can calculate two types of correlation. First it will give a simple bivariate correlation between two variables, also known as zero order correlation. Secondly, SPSS can explore the relationship between two variables, while controlling for another variable. This is called partial correlation. In this text, only zero order correlation will be discussed. The two most popular correlations are: Spearman's and Pearson's product-moment correlation coefficients. The difference between them is that Pearson's product-moment correlation deals with interval or ratio data while Spearman rank-order can deal with ordinal data apart from interval and ratio data.

REGRESSION

Regression is particularly useful to understand the predictive power of the independent variables on the dependent variable once a causal relationship has been confirmed. To be precise, regression helps a researcher understand to what extent the change of the value of the dependent variable causes the change in the value of the independent variables, while other independent variables are held unchanged.

Simple and Multiple Regressions

In simple linear regression, the outcome or dependent variable Y is predicted by only one independent or predictive variable. Their relationship can be expressed in a math equation as follows:

$$Y = \alpha + \beta X + e \quad (6.1)$$

Where:

Y is the dependent variable;

α is a constant amount;

β is the coefficient;

X is the independent variable;

e is the error or the 'noise' term that reflect other variables to have an effect on Y.

It should be stressed that in very rare cases, the dependent variable can only be explained by one independent variable. To avoid omitted variable bias, multiple regression is applied. Its math equation is as follows:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + e \quad (6.2)$$

where Y, α and e remain to be the dependent variable, a constant amount and the error respectively. But as you can see, the number of the independent variables is now more than one.

Multiple regression is not just a technique on its own. It is, in fact, a family of techniques that can be used to explore the relationship between one continuous dependent variable and a number of independent variables or predictors. Although multiple regression is based on correlation, it enables a more sophisticated exploration of the interrelationships among variables. This makes it suitable for investigating real life rather than laboratory-based research questions. An important point must be made here. There is a temptation to see regression as the shortcut through the forest of quantitative analysis. Just stick the variables

into regression, wait for the answers, and off we go. Job Done! This unfortunately is not the case. There must be a sound theoretical and conceptual reason for the analysis, and in particular, the order of the variables entering the equation. It is therefore vital to spend the time on the research process prior to undertaking regression analysis.

Imagine that the college in which you are enrolled is attempting to analyse the reasons behind students choosing a particular course. The administration team already knows that advertising accounts for 33% of the variation in student enrolment but a much larger 67% remains unexplained. The college could add a new predictor to the model in an attempt to explain some of the unexplained variation in student numbers. They decide to measure whether prospective students attended college open days in the year prior to enrolment. The existing model can now be extended to include this new variable, and conclusions drawn from the model are now based on two predictors. This analysis will be fulfilled by multiple regression.

Having said that, approached correctly, multiple regression has the potential to address a variety of research questions. It can tell you how well a set of variables is able to predict a particular outcome. For example, you may be interested in exploring how well a set of organisational variables are able to predict organisation performance. Multiple regression will provide you with information about the model as a whole and the relative contribution of the variables that make up the model. It will also allow you to measure whether including an additional variable makes a difference, and to control for other variables when exploring the predictive ability of the model.

Some of the main types of research questions that multiple regression can be used to address include:

i) how well a set of variables can predict a particular outcome;

ii) identification of the best predictor of an outcome amongst a set of variables;

iii) whether a particular variable is still able to predict an outcome when the effects of another variable are controlled for.

Assumptions behind Multiple Regression

Multiple regression makes a number of assumptions about the data, and is important that these are met. The assumptions are:

a. Sample Size
b. Multicollinearity of IVs
c. Linearity
d. Absence of outliers
e. Homoscedasticity
f. Normality

Tests of these assumptions are numerous so we will only look at a few of the more important ones.

a. Sample size

You will encounter a number of recommendations for a suitable sample size for multiple regression analysis (Tabachinick & Fidell, 2007). As a simple rule, you can calculate the following two values:

$$104 + m$$

$$50 + 8m$$

where *m* is the number of independent variables,

and take whichever is the largest as the minimum number of cases required.

For example, with 4 independent variables, we would require at least 108 cases:

[104+4=108] [50+8*4=82]

With 8 independent variables we would require at least 114 cases:

[104+8=112] [50+8*8=114]

With Stepwise regression, we need at least 40 cases for every independent variable (Pallant, 2007).

However, when any of the following assumptions is violated, larger samples are required.

b. Multicollinearity of Independent Variables

Any two independent variables with a Pearson correlation coefficient greater than .9 between them will cause problems. Remove independent variables with a tolerance value less than 0.1. A tolerance value is calculated as $1 - R_i^2$, which is reported in SPSS.

c. Linearity

Standard multiple regression only looks at linear relationships. You can check this roughly using bivariate scatterplots of the dependent variable and each of the independent variables[1].

d. Absence of outliers

Outliers, such as extreme cases can have a very strong effect on a regression equation. They can be spotted on scatterplots in early stages of your analysis. There are also a number of more advanced techniques for identifying problematic points. These are very important in multiple regression analysis where you are not only interested in extreme values but in unusual combinations of independent values.

e. Homoscedasticity

This assumption is similar to the assumption of homogeneity of variance with ANOVAs.

[1] More advanced methods include examining residuals.

It requires that there be equality of variance in the independent variables for each value of the dependent variable. We can do this in a crude way with the scatterplots for each independent variable against the dependent variable. If there is equality of variance, then the points of the scatterplot should form an evenly balanced cylinder around the regression line.

f. Normality

The dependent and independent variables should be normally distributed. Please refer to Chapter 3 on how to examine univariate and multivariate normality.

Types of Multiple Regression

Standard Multiple Regression: All of the independent (or predictor) variables are entered into the equation simultaneously.

Hierarchical Multiple Regression: The independent variables are entered into the equation in the order specified by the researcher based on their theoretical approach.

Stepwise Multiple Regression: The researcher provides SPSS with a list of independent variables and then allows the program to select which variables it will use and in which order they can go into the equation, based on statistical criteria. This approach has received a lot of criticism and will not be covered in any detail in this chapter.

WORKED EXAMPLE

Research Question: What is the predictive power of '**backgdmusic**', '**quality**' and '**waitingtime**' on customer perceived satisfaction ('**satisfaction**') of a clothes store?

The Steps for Standard Multiple Regression in SPSS

1. From the menu at the top of the screen, click on '**Analyze**', then '**Regression**', then '**Linear**'.

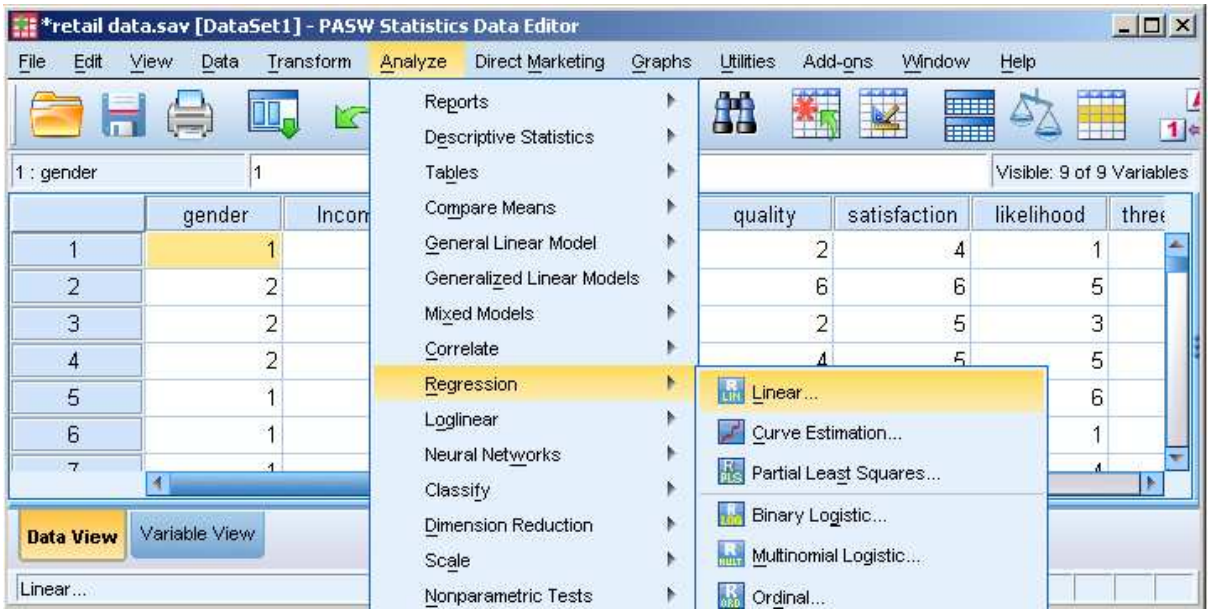

Figure 6.1 Analyse Menu in SPSS with Regression and Linear Submenus

1. Once you are on the window as shown in Figure 6.2, click on the dependent variable, '**satisfaction**' and move it into the box of '**Dependent**'. Move the variables you want to control into the box of '**Independent(s)**' under Block 1 of 1. The variables we normally want to control are demographic variables such as age, gender, income. In this case, move '**gender**' and '**income**' into the box of '**Independent(s)**'.

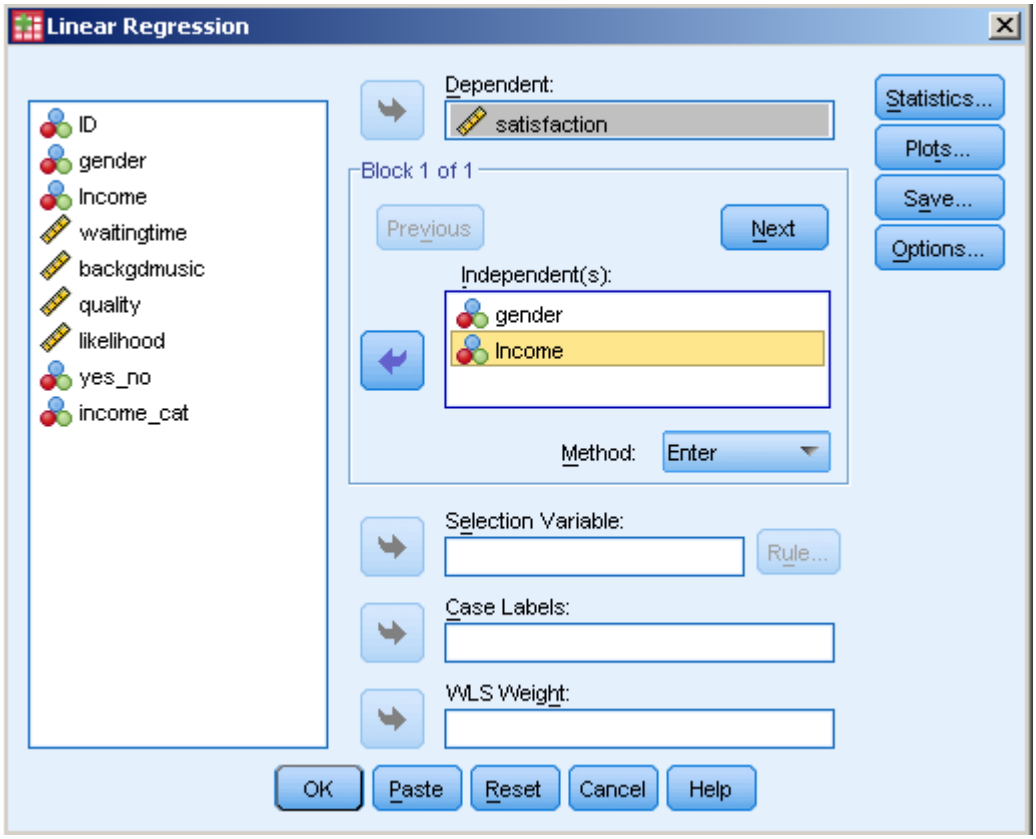

Figure 6.2 Linear Regression Dialogue Box (1)

2. Then click on the button marked '**Next**'**.** This will give you a second independent variables box to enter the second block of variables under '**Block 2 of 2**'[2]. Move the independent variables, '**waitingtime**', '**backgdmusic**' and '**quality**' into the box of '**Independent(s)**'.
3. For '**Method**', make sure '**Enter**' is selected. This will give you standard multiple regression.

[2] If you do not want to control the variables of '**gender**' and '**income**', you can move all independent variables, in the case, '**gender**', '**income**', '**waitingtime**', **backdgmusic**', and '**quality**' into the box of '**Independent(s)**'at one go.

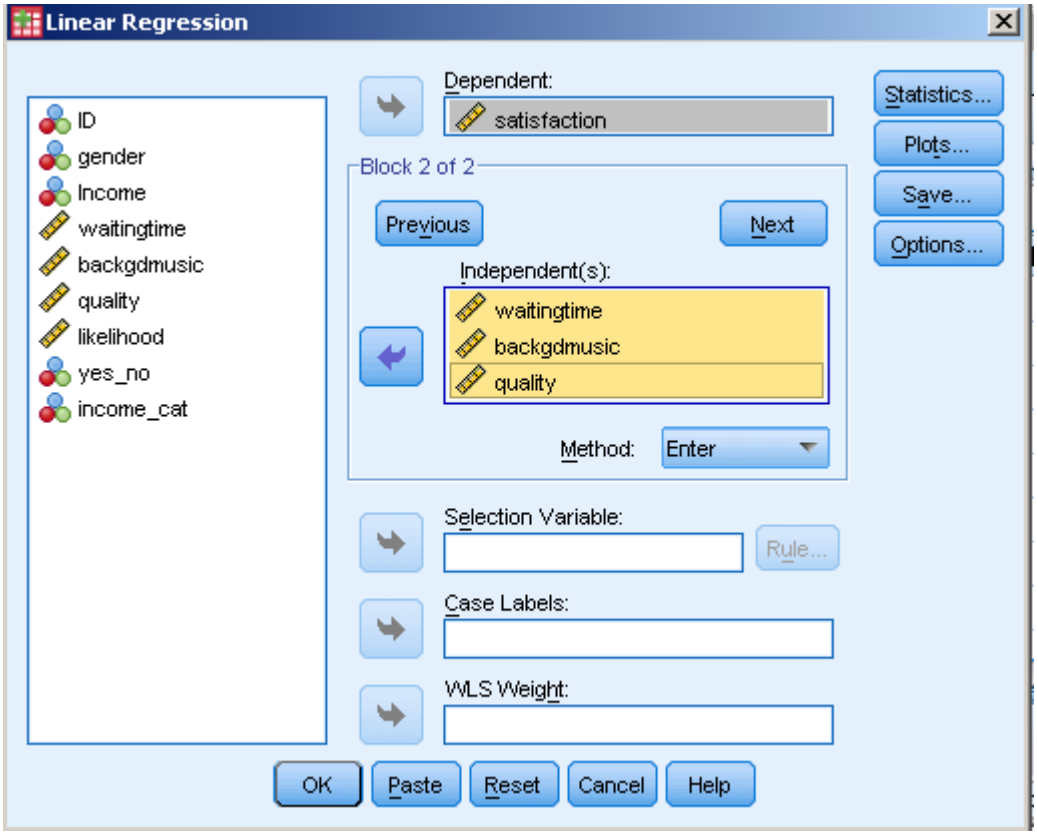

Figure 6.3 Linear Regression Dialogue Box (2)

4. Click on the '**Statistics**' button**.**

   - Tick the boxes marked '**Estimates**', '**Confidence Intervals**', '**Model fit**', '**Part and partial correlations**' and '**Collinearity diagnostics**'.
   - In the '**Residuals**' section, tick the '**Casewise diagnostics**' and '**Outliers outside 3 standard deviations**'. Click on '**Continue**'.

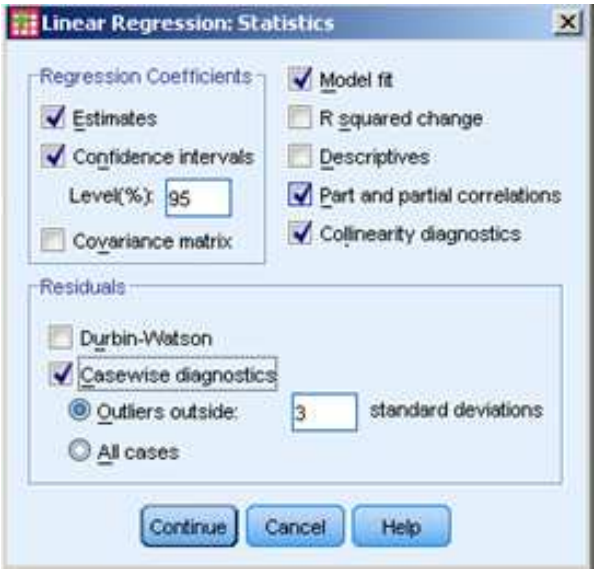

Figure 6.4 Linear Regression Statistics Dialogue Box

5. If there is missing data, click on the '**Options**' button. In the '**Missing Values**' section, click on '**Exclude cases pairwise**'. Click on '**Continue**'.

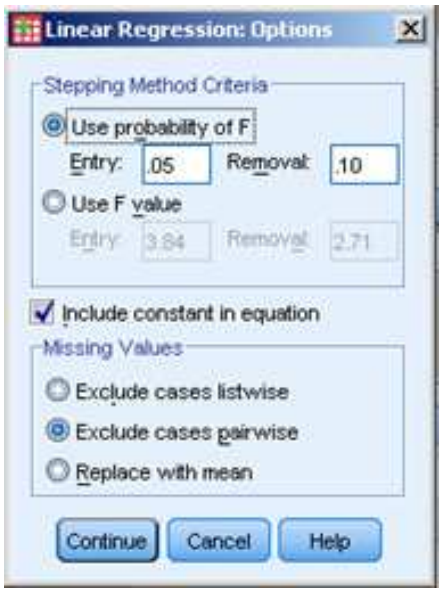

Figure 6.5 Linear Regression Options Dialogue Box

6. Click on the '**Save'** button.
   - In the next section, labelled '**Distances**'**,** tick the '**Mahalanobis**' box and '**Cook's**'.
   - Click on '**Continue**' and then '**OK**'.

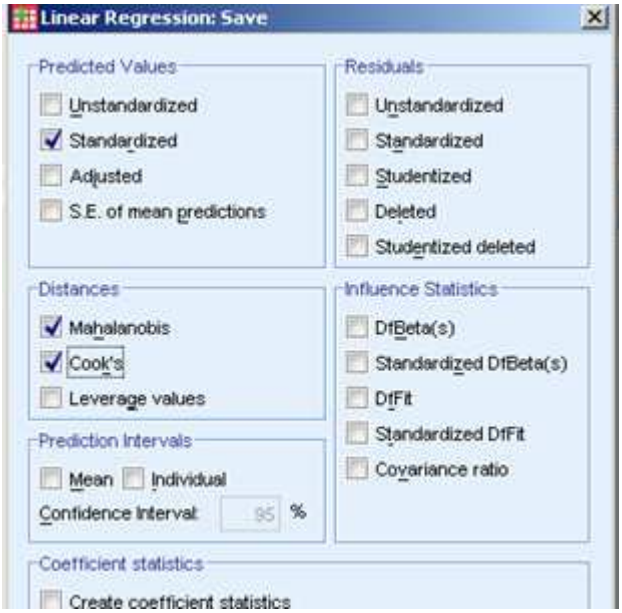

Figure 6.6 Linear Regression Save Dialogue Box

A selected view of the output is as follows:

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.224[a]	.050	.026	1.609	.050	2.037	2	77	.137
2	.484[b]	.234	.183	1.473	.184	5.930	3	74	.001

a. Predictors: (Constant), gender. income

b. Predictors: (Constant), gender, income, waitingtime, backgdmusic, quality

c. Dependent Variable: satisfaction

Table 6.1 Model Summary

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.541	2	5.271	2.037	.137[a]
	Residual	199.259	77	2.588		
	Total	209.800	79			
2	Regression	49.161	5	9.832	4.529	.001[b]
	Residual	160.639	74	2.171		
	Total	209.800	79			

a. Predictors: (Constant), Income, gender

b. Predictors: (Constant), Income, gender, quality, waitingtime, backgdmusic

c. Dependent Variable: satisfaction

Table 6.2 ANOVA

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2.792	.613		4.556	.000	1.572	4.012					
	gender	.565	.377	.166	1.498	.138	-.186	1.316	.168	.168	.166	1.000	1.000
	Income	.193	.145	.148	1.333	.186	-.095	.481	.150	.150	.148	1.000	1.000
2	(Constant)	.048	.861		.056	.956	-1.667	1.763					
	gender	.463	.352	.136	1.316	.192	-.238	1.164	.168	.151	.134	.963	1.038
	Income	.106	.145	.082	.731	.467	-.183	.395	.150	.085	.074	.831	1.203
	waitingtime	.277	.108	.264	2.555	.013	.061	.493	.287	.285	.260	.967	1.034
	backgdmusic	.175	.143	.139	1.223	.225	-.110	.461	.150	.141	.124	.804	1.245
	quality	.332	.105	.332	3.150	.002	.122	.542	.308	.344	.320	.931	1.074

a. Dependent Variable: satisfaction

Table 6.3 Coefficients

Interpretation of the Output from Hierarchical Multiple Regression

In Table 6.1, there are two models listed. **Model 1** refers to the control variables that were entered in Block 1 of 1 ('gender' and 'income'), while **Model 2** includes all the variables that were entered in both blocks ('**gender**', '**income**', '**waitingtime**', '**backgdmusic**', and '**quality**').

1. *Evaluate the Model*

Firstly check the **R Square** values in the third column**.** After the variables in Block 1 have been entered the overall model explains 5% of the variance (.050 x 100%). After Block 2 variables have also been entered, the model in its entirety explains 23.4% (.234 x 100%). Do not forget that the R squared value in model 2 includes all of the five variables, not just those three from the second block.

Now we want to establish how much of this overall variance is explained by our variables of interest after the effects of gender and income are removed. To do this, we must look at the column labelled **R Squared change.** In the output that was produced in Table 6.1, on the marked line **Model 2,** the **R square change** value is .184. This means that '**waitingtime**', '**backgdmusic**', and 'quality' explain an additional 18.4% of '**satisfaction**', even when statistically controlling for the effects of gender and income. This is statistically significant contribution, as indicated by the **Sig. F Change** value for this line (.001).

2. *Evaluate each of the independent variables*

To find out how well each of the variables predicts the dependent variable, we must now look in Table 6.3 in the row of Model 2. This information contains a summary of the results, with all the variables entered into the equation. Upon reviewing the **Standardise Coefficient Beta (β)** column there are two variables that make a statistically significant contribution. The t of each coefficient **β** needs to be greater than 2 or less than -2; and the sig. level less than .05. In

the example, '**waitingtime**' has a **β** of .264 at a sig. level of .013, and t=2.555 and '**quality**' has a **β** of .332 at a sig. level of .002, and t=3.150. Therefore, we can conclude that the more customers are pleased with the waiting time at the check-out and the quality of the merchandises, the higher satisfaction is perceived by customers. Neither '**gender**', '**income**' and '**backgdmusic**' can predict customers perceived satisfaction significantly.

REFERENCES

Cohen, J.W. (1988). *Statistical Power Analysis for The Behavioural Sciences*. 2nd ed., Hillsdale, NJ: Lawrence Erlbaum Associates.

Field, A. (2002). *Discovering Statistics Using SPSS for Windows*. London: Sage.

Gravetter, F.J. & Wallau, L.B. (2004). *Statistics for The Behavioural Sciences*. 6th ed., CA: Wadsworth.

Pallant, J. (2007). *SPSS Survival Manual*. Berkshire: Open University Press.

Stevens, J. (1996). *Applied Multivariate Statistics For The Social Sciences*. NJ: Laurence Erlbaum.

Tabachnick, B.G. & Fidell, L.S. (2007). *Using Multivariate Statistics*. 5th ed., Boston: Pearson.