Appendix F
Technical Appendix

Technical Appendix

The following gives more technical information regarding two key areas of statistical analysis – significance testing, and regression analysis.

Measures of Confidence and Significance Testing

Where it is not possible to survey the entire target population a sample of this population is used. In this case, a sample of Australian and New Zealand consumers was surveyed as conducting a census of the entire population would have been a costly exercise. Using a random sample, we assume that the statistics gathered are representative of the total population. We can make inferences about the actual population statistic by creating confidence intervals around the sample statistic.

A **confidence interval** assumes that the statistics gathered are distributed on an approximately normal distribution, and is used to describe the precision around a statistic, and to give a range of reasonable values for the population parameter.

The width of the confidence interval for a proportion depends on:

- sample size (n);
- level of confidence (95% in this case); and
- size of the proportion (p).

The sample size required to assume a normal distribution is:

$$n*p \geq 5 \text{ and } n*(1-p) \geq 5$$

As a rule of thumb, an acceptable confidence interval is $\pm$5% at a 95% confidence level. That is, if a sample proportion is 50%, we can be 95% confident that the population proportion is between 45% and 55%.

The width of the confidence interval for a mean depends on:

- sample size;
- level of confidence; and
- standard deviation of sample.

If this is not the case, the tests are not valid and another test would have to be used. The sample sizes used for this survey meet this requirement, so these tests can be used.

A **significance test** is used to determine whether a particular estimate of the population parameter is reasonable.

Factors impacting on statistically significant differences between two means are:

- sample size;
- standard deviations of the samples; and
- confidence level.

In order to conduct this test we require that:

- the 2 sample sizes are both at least 30; and
- the 2 samples are independent. The same people can not be in both samples (e.g. Males versus Total, multiple response questions).

Factors impacting on tests of significant difference between proportions are:

- sample size;
- confidence level; and
- size of proportions.

Again, this test assumes that the difference in proportions has an approximately normal distribution, and that the two samples are independent:

- the sample size required to assume a normal distribution is
  $n1*p1 \geq 5$ and $n1*(1-p1) \geq 5$ and $n2*p2 \geq 5$ and $n2*(1-p2) \geq 5$

In the case of weighted data (as is the case for the 2007 Consumer Attitudes Survey), unweighted base sizes are used in all tests of difference, to ensure that differences observed are actual differences, and not due to the change in sample sizes as a result of weighting.

Multivariate Analysis

The analysis for the 2007 Consumer Attitudes Survey utilised regression analysis, a form of multivariate analysis, to understand the interplay between individual variables and overall confidence. Regression seeks to explain the relationship between independent variables and a dependent variable; that is, if one or more independent variables change, how will the dependent variable change.

Regression analysis is also a measure of association, but with the added features of:

- implying causality: that is, variable X causes variable Y to change; and
- the ability to consider relationships beyond 2 variables.

The statistical objective of regression analysis is to explain as much of the variation in the dependent variable with as few independent variables as possible. However, the managerial objective sometimes differs in a business sense, and many of the variables outside of the control of the business are not measured, because these variables cannot be impacted by the work of the business.

Multiple regression is used to understand the inter-relationships between a group of independent variables (e.g. performance issues) and a dependent variable (e.g. overall satisfaction). The objective of multiple regression is to determine which performance issues have the most significant and unique impact on overall satisfaction and which in combination, explain the most about overall performance.

Regression analysis generates two important pieces of information:

- **The relative importance of particular drivers.** These are the percentages shown as 'drivers' or 'importance scores'. The driver percentages are derived from a linear regression model. Linear regression is conducted to determine which service level issues have the most significant impact on satisfaction with the area. The output from the linear regression which indicates impact, or importance, is called a standardised beta coefficient. These beta coefficients are converted into percentages which total 100% and indicate the importance of each service level issue.

- **The strength of the model, or how well the combination of independent variables explain the dependent variable.** Model strength is expressed as a percentage and is called an Adjusted R-squared:
    - the Adjusted R-squared figure is interpreted as the amount of variance that two or more independent variables explain in a dependent variable;
    - an Adjusted R-squared figure of 80% indicates that 80% of satisfaction is explained by the independent variables. The remaining 20% consists of things that were not measured and would probably not be significant enough to be included in the model; and
    - in customer satisfaction research, Adjusted R-squared figures ranging from 60% to 80% are typical and are indicative of strong models, however in other types of research, particularly where there are a large number of variables not explored in the research, a model which explains 40% or more is acceptable.

There are several types of regression:

- independent variables: simple (one independent) vs. multiple (2 or more independents).
- dependent variables: standard (scale) vs. logistic (binary); and
- 'method' (stepwise, backwards, etc), 'enter' (linear).

It is best to have scale (interval or ratio) independent variables. This is the case in most of the questions included in the Consumer Attitudes Survey. Binary variables are possible but difficult to interpret beyond 2 binary variables.

The equation for a **simple regression** is:

$$Y = a + bx + e$$

Where:

- Y = Dependent variable
- x = Independent variable
- a = point the regression line intercepts the y axis
- b = beta coefficient
- e = residual error

The equation for a **multiple regression** is:

$$Y = a + b_1x_1 + b_2x_2 + \ldots + e$$

Where:

- b_1 = beta for independent variable 1
- x_1 = independent variable 1
- b_2 = beta for independent variable 2
- x_2 = independent variable 2

Managing missing cases

Cases with completed responses were taken into the regression models. Thus, missing cases and missing values were excluded in the analysis.

Tests for multicollinearity

Multicollinearity is the undesirable situation when one independent variable is a linear function of other independent variables. Eigenvalues of the scaled and uncentred cross-products matrix, condition indices, and variance-decomposition proportions are displayed along with variance inflation factors (VIF) and tolerances for individual variables. Particularly with this project, VIF was used to indicate whether the independent variable is highly correlated with the dependent measure. Factors with VIF over 3 would have been removed from the regression models. However no factors were removed on this basis.

Rationale for excluding socio-demo variables from the regression models

In general, socio-demo variables are moderating variables rather than independent. Factors such as age, gender, income, etc tend to heighten the impact of independent variables on the dependent variable rather than having a direct impact.

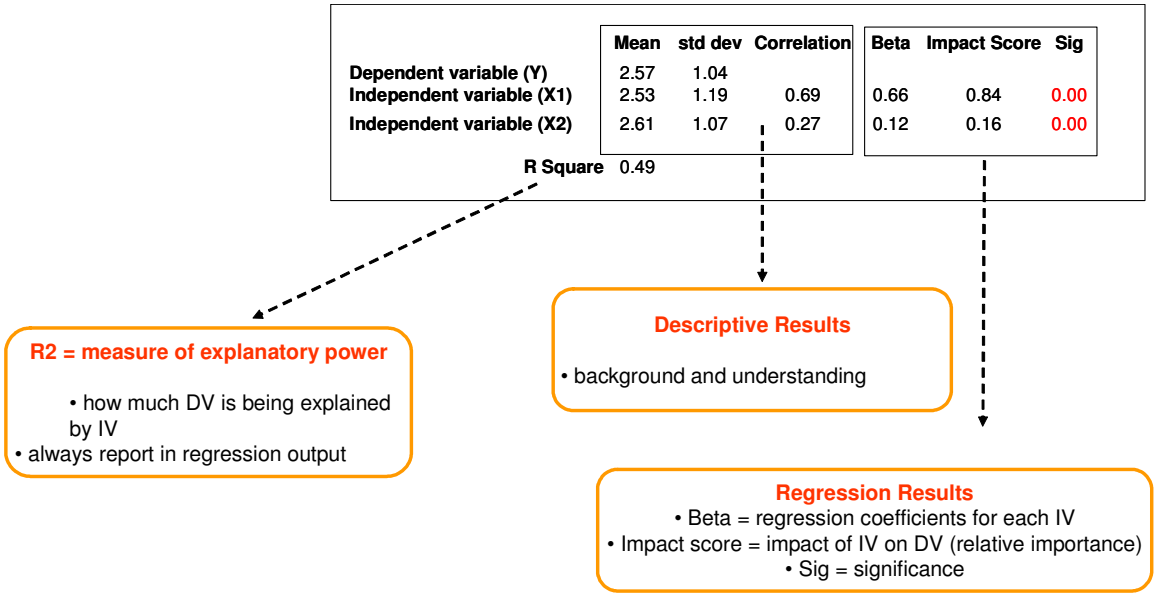
Regression models would also be more useful when independent variables are the factors that businesses can 'do something about' (i.e. increasing or reducing) to have an impact on the dependent variable. Socio-demographic variables are uncontrollable factors, therefore have limitations in terms of business implications.

One-factor regression model

All of the hypothesized independent variables had been entered into the regression model, however the result showed only one factor with significant impact on the dependent measure. A linear regression model consists of one dependent variable and at least two independent variables. However, to determine whether the regression model is plausible is also up to the researcher as well.

Understanding the results

The following diagram provides a guide to interpreting the results.



There is one beta coefficient for each independent variable:

- betas may be positive or negative and also ranges between -1 and +1;
- simple regression: the beta coefficient = correlation coefficient; and
- multiple regression: the beta coefficient can also be thought of as a weighting reflective of the magnitude of the relationship between an IV and the DV.

For example: Y = 1 + 0.59 x1 - 0.37 x2 + e

- for 1 unit increase in x1 , Y increases by 0.59
- for 1 unit increase in x2, Y decreases by 0.37.

Impact scores = beta value/sum of all betas

- % reflecting the contribution of each IV to explaining the DV
- e.g.,30% is explained by x1 and 70% explained by x2.