We draw upon our experience to examine the following key topics within the context of customer experience analysis:

- How regression analysis works;
- What confidence intervals are;
- How regression analysis should be interpreted;
- How to interpret changes over time.

White Paper written by John Bee, Director, White Space Insight
June 2011
Using Regression in Customer Experience Analysis

Regression analysis is used as standard by White Space in most customer experience and consumer strategy projects. When used properly, it can help predict future trends or identify key drivers of customer behaviour, but if used incorrectly it can lead to dangerously inaccurate findings and recommendations.

This white paper explores some of the key themes and challenges consultants and their clients need to address when using and interpreting regression analysis.

In particular, we draw upon our experience to examine the following key topics:

- How regression analysis works;
- What confidence intervals are;
- How regression analysis should be interpreted;
- How to interpret changes over time.

1. How Does Regression Work?

The main aim of using regression analysis in customer experience work is to estimate which factors affect overall satisfaction or spend. This will allow a company to focus their resources on the key drivers of satisfaction/spend, and not waste these resources on areas which might seem important but in fact are not.

Regression works by calculating a line of best fit between survey respondents’ answers to one or more ‘input’ questions and a single ‘output’ question. By using this line of best fit, you can then predict what any given respondent’s response to the ‘output’ question will be, based on the answers they give to the ‘input’ question(s).

As with most statistics (including the most basic measures such as averages), regression is only ever able to provide an estimation of the real relationships which exist within a population. It is therefore subject to a degree of error, with results subject to revision as more data becomes available.

Regression’s accuracy depends on several factors, including:

- **The sample size:**
  - Larger sample size = more accurate regression

- **The sample structure:**
  - Tighter sample structure = more accurate regression

- **The level of natural spread in the data** (i.e. how closely the data points are scattered around the line of best fit):
  - Less spread = more accurate regression

Introduction

Regression’s accuracy depends on several factors, including:

- **The sample size:**
  - Larger sample size = more accurate regression

- **The sample structure:**
  - Tighter sample structure = more accurate regression

- **The level of natural spread in the data** (i.e. how closely the data points are scattered around the line of best fit):
  - Less spread = more accurate regression
For example, the ‘output’ measure might be ‘Overall satisfaction’, and the ‘input’ factors might be ‘Satisfaction with product’ and ‘Satisfaction with service’. Taking the example of just one ‘input’ factor, ‘Satisfaction with product’, we can compare how customers rate satisfaction with product versus satisfaction overall by plotting all survey responses:

As can be seen in the diagram above, there is a possible pattern in the data, with a higher score in ‘Satisfaction with product’ coinciding with a higher score in ‘Overall satisfaction’. Using Regression we can fit a line of best fit to this data to help us predict how an increase in ‘Satisfaction with product’ might affect ‘Overall satisfaction’.

The gradient (slope) of this line tells us how much of an increase in ‘Overall satisfaction’ there is, on average, if ‘Satisfaction with product’ goes up by one unit. This value is called the regression coefficient. In the case above, the value is 0.69, which means that for an increase in ‘Satisfaction with product’ of 1, on average there will be an increase in ‘Overall satisfaction’ of 0.69.
Testing for Significance

In addition to creating a best fit model, the regression process also calculates whether this line is statistically significant. This is important because you can plot a line of best fit through any randomly plotted points on a graph – the key is to know whether they have a meaningful pattern, or whether the variables are actually linked.

Significance tests are performed using differences between the regression model’s predicted values (the line) and the actual values. If the distances are too large, the model or coefficients will be rejected. In the model above, the model and the coefficient were judged to be statistically significant at a 95% level of confidence. This means that, 19 times out of 20, the relationship between ‘Satisfaction with product’ and ‘Overall satisfaction’ would not be caused by chance (i.e. it is very likely that a change in one variable is causing a change in the other).

Multiple Regression

Although interesting, regression on a single factor is of limited use in a real world situation in which multiple factors influence behaviour. The true power of regression becomes apparent when the influence of more than one factor is assessed, e.g. satisfaction with product, service, customer services etc.

Using multiple regression techniques it is possible to model how a large number of factors affect overall satisfaction and their relative influences. It is also possible to identify those factors that do not have a statistically significant effect and remove them.

2. What are Confidence Intervals?

Without an extremely large sample size, regression analysis is only ever able to approximately estimate to what extent specified ‘input’ variables affect an ‘output’ variable. In reality, the true scale of this effect may be bigger or smaller. It is therefore useful to know what the likely maximum and minimum scale of the effect might be (i.e. the largest and smallest values the regression coefficients are likely to have).

To do this, regression analysis can provide a confidence interval, expressed as a maximum value and a minimum value, which, in 95% of cases, will contain the true value of the coefficient. This might look as follows, assessing the impact of satisfaction with product and service on overall satisfaction:

The conclusions we would draw from this chart are as follows:

• The most likely true value of the coefficient for ‘Product’ is 0.28. To a 95% level of certainty, the true value lies between 0.20 and 0.36

• The most likely true value of the coefficient for ‘Service’ is 0.21. To a 95% level of certainty, the true value lies between 0.12 and 0.30

• ‘Satisfaction With Product’ and ‘Satisfaction With Service’ are both highly likely to be significant drivers of customer satisfaction

• It is most likely that ‘Satisfaction With Product’ is a more important cause than ‘Satisfaction with Service’. However, because the confidence intervals overlap we cannot say this to a high level of certainty – there is a good chance if the survey was repeated with a different sample, the values (and order) would be different. However, 19 times out of 20 (95% of the time) they would lie within the confidence intervals reported.
What determines the width of a confidence interval, and how should confidence intervals be used?

If the sample size is small, the confidence interval will usually be wide, as there is little data on which to base calculations. If the sample size is big, the confidence interval will usually be narrower, as there is more data to work with, to make a more accurate prediction.

This does not mean that it is not useful (and usual) to report the single estimated coefficient values. On the basis of the sample data available, this is the best estimate possible of what the true value is. This is therefore the value White Space usually report to clients.

However, it can sometimes be useful to also consider and report the confidence interval. This is for 2 reasons:

1) It provides an indication of the highest and lowest values of the true value

2) It provides a useful way of working out whether the differences between two different regression coefficients should be considered significant. This is useful when either comparing two different regression coefficients in the same survey wave, or when comparing old and new regression versions of the same coefficient in different survey waves

When comparing regression coefficients, if the confidence intervals of two regression coefficients overlap, there is a possibility that the true values may actually be the same. When confidence intervals overlap, White Space therefore generally consider both factors to be of similar importance, whilst noting that the value that is higher is still more likely to be the most important.

3. How Should Regression Analysis Be Interpreted?

Regression analysis, though a very useful tool, can be difficult to interpret partly due to the complexity of the outputs, and partly due to the probabilistic reasoning upon which it is based. Making a regression relevant to a business requires understanding both its strengths and limitations. The key points to interpreting the results of a regression are listed below.

1. The results must make practical sense

As with all analysis, the results of a regression should be sense checked against the reality of a business. For example, it is possible for a regression to show a factor as significant, because it is very important for a small proportion of the customer base, but does not affect the rest of the customer base. In this case, it makes little business sense to make such a factor a business priority even if it is the most important factor in the regression. Alternatively, regression may indicate that a certain factor is very important, but the client may have no practical ability to change in this area. This would reduce the overall significance of the result, and the extent to which we would draw upon it in decision making.

2. Confidence intervals should be considered, as well as actual coefficients

The actual coefficient provides a very useful precise number, which can be used as a basis for modelling and prediction. However, it should be considered as part of its confidence interval (even if, ultimately, only the actual coefficient is reported). This allows us to assess the maximum and minimum amount a given 'input' factor is likely to have on the 'output' measure, and provides a solid basis to compare movement of coefficients over time.

For this reason, it is often difficult to identify a single most important priority, as overlap of confidence intervals means that two could be interpreted as being of similar importance (even if the actual coefficient values are clearly different).

3. When comparing regression analyses over time, what remains constant is as important as what changes

When looking at a series of regression analyses of the same customer base, it is key from a business perspective
to identify what remains constant, to gain increased certainty that this is (or is not) truly a key driver of satisfaction. Repeated independent surveys with regression analysis are a good way to validate choice of business priorities from a previous wave. There will often be changes in a regression from one wave to the next, but often the most significant factors do not change. If there is a change in the most significant factors (beyond the boundaries of the confidence intervals) further analysis should be performed to try to identify the reason why.

Changes will naturally occur from one wave to the next due to a variety of factors. Some of these may be 'natural' and not indicate a change in customer needs (such as variation in sample), some may be due to temporary changes in customer needs (e.g. seasonality) and some may be due to longer term changes in customer needs (e.g. unemployment increasing price sensitivity). However, the key is in identifying significant changes, outside of the confidence intervals, that have real business relevance.

4. How Should Changes Over Time Be Interpreted?

Because of the complex nature of regression analysis, requiring straight lines to be fitted to often highly spread data points, the amount of natural variation in coefficients can be quite large over time (much more so than for means, where the maths involved is much simpler).

However, coefficients are unlikely to change to be smaller or larger than the boundaries of their confidence intervals, unless the underlying causes of satisfaction have changed. The only other potential cause of this would be a change in sampling method and/or sample structure (which would normally lead to a change in coefficients because the nature of the survey data would be fundamentally different).

White Space therefore expect to see a level of change in coefficients over time naturally, due to the nature of sampling and the complexity of the analysis involved. However, we usually only pay close attention to these changes when they are beyond the boundaries of the confidence intervals.

5. Conclusion: A Best Practice Approach

This white paper has demonstrated that regression analysis is both extremely powerful but also complex to set up and interpret accurately. This is in contrast to how easy it is to run the actual test procedure on most software packages such as SPSS. This creates a serious risk of it being run inaccurately by people who can run the test procedure but don’t know how to set it up properly. In turn, this can lead to false conclusions which seem highly compelling at a surface level, but are actually highly dangerous.

In order to avoid this pitfall, consultants and clients should ask the following 5 key questions when using or interpreting regression:

1) What process has been used to run the regression, from data preparation to development of the final model?
2) What are the confidence intervals for the coefficients in the final model?
3) Is the original sample size and structure statistically valid?
4) Does the person running the regression analysis understand the statistical principles involved?
5) Are the final outputs clearly presented and explained so that senior decision makers with no background in statistics can understand/use them?

In summary, it is critically important to ensure regression analysis is set up, run and interpreted correctly, based around these key questions. When this is done, the full potential of regression to drive improvement of the customer experience can be safely unlocked.

For further information on the themes contained within this white paper, please contact
John Bee, Director
john.bee@whitespaceinsight.com

ABOUT WHITE SPACE

White Space is one of the UK’s leading providers of research consultancy services, bridging the gap between traditional market research and management consultancy. We use cutting edge research methods and consultancy tools to help blue chip clients find answers to strategic questions. Our clients include many of the world’s biggest companies, global private equity houses, top tier strategy consultancies and smaller organisations ambitious to succeed. Over 85% of them repeat purchase – to find out why, go to www.whitespaceinsight.com