KEY DRIVER ANALYSIS USING LATENT CLASS REGRESSION

Steven M. LaLonde, Eastman Kodak Company
343 State Street, Rochester, NY 14650-1015

KEY WORDS: regression, segmentation, clustering, satisfaction.

1. INTRODUCTION

Regression analysis is typically applied to customer satisfaction surveys to determine which variables are important to overall customer satisfaction. This is often referred to as a "key driver analysis". In addition to regression analysis, respondents are sometimes grouped using a post-hoc clustering algorithm in order to explore segmentation.

Latent class regression combines the two analysis objectives, key driver analysis and segmentation, into one step. Latent class regression fits regression equations to classes of respondents exhibiting similar response patterns. The result is a number of customer segments, each with its own key drivers.

This paper compares the results from the usual key driver analysis with a latent class analysis of customer satisfaction data. A principal components regression analog to latent class analysis is also illustrated.

Other looks at the data will also be discussed to lend insight into other possible approaches.

THE PROBLEM DESCRIPTION

This paper will focus on the analysis of a sample of one hundred sixty customers surveyed in the second half of 1995. The sample was chosen for illustrative purposes and may not represent the total population of customers. The data was collected in twenty-five minute telephone interviews with individual dealers.

Customers were asked about their overall satisfaction with as well as their satisfaction on a number of component categories including:

Pricing Policies
Order and Delivery
Sales Rep Support
Advertising & Promotions
Billing & Credit
Product Quality
Range and Variety
Technical Support

The objective of the analysis is to determine which of these components of overall satisfaction are most important. This information, along with a measure of current performance, can be used to focus subsequent efforts on those processes that have the greatest impact on overall satisfaction.

Do all customers fit the same model? Interest in segmenting the sample comes from the belief that different customers want different things, or, that there are different drivers of satisfaction for different retailers. Understanding the particular needs of heterogenous customer segments would allow a more tailored marketing program (product/ pricing/ channel/ communications).

ISSUES TO BE CONSIDERED

For the purpose of this investigation we will assume that the sample we are working with has no obvious a priori segmentation. In the more general situation, there may be some a priori segmentation; based on the size of the account, or the type of channel, or some other demographic variable.

We will conduct a multiple regression of the component satisfaction categories on the measure of overall satisfaction and use the regression coefficients as a measure of importance for the baseline key driver analysis. The mean of each component category will be used as a measure of current performance.

The performance on a given component category must be taken into account when interpreting the importance. Indeed, some component categories may appear to be less important merely because there is smaller range of responses (e.g., a ceiling effect). Also, component categories exhibit varying amounts of inter-correlation that can lead to biased estimates of importance. These coefficients can lead to very problematic interpretations of component category effects.

2. BASELINE KEY DRIVER ANALYSIS

The results of the baseline regression on the entire sample of one-hundred sixty respondents is displayed in Table 1. The component categories appear in the table sorted by descending size of the parameter estimate. The first four component categories: PRICING, ORDERDEL, SALESUPT, and ADVPROMO are statistically significant at the 0.05 level. Of those categories that are significant, PRICING also appears to exhibit the lowest average performance. As a whole, the averages on each component category, and the overall measure of satisfaction, are quite high (all greater than a 4.5 on a 7-point scale).

The tolerances in the table would indicate that there is some intercorrelation in the predictors, although there is nothing that looks too far out of the ordinary for this type of data.

Table 1: The Baseline Regression:

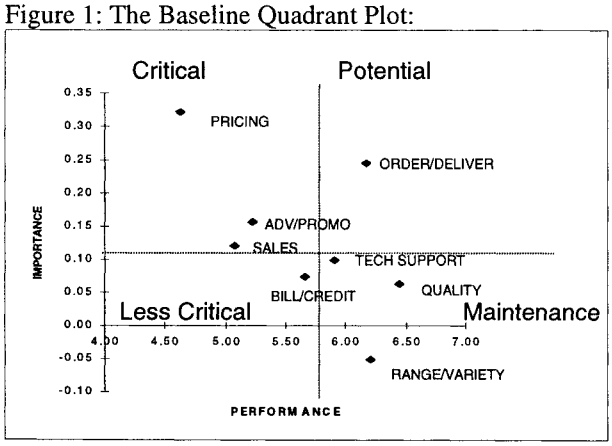
Variable	Parameter	Standard			
Label	Estimate	Error	Prob > T	Tolerance	Mean
OVERALL					5.29
PRICING	0.32	0.06	0.00	0.65	4.62
ORDERDEL	0.25	0.08	0.00	0.81	6.17
SALESUPT	0.12	0.05	0.02	0.85	5.07
ADVPROMO	0.16	0.08	0.04	0.63	5.23
TECHSUPT	0.10	0.07	0.14	0.75	5.91
BILLCRED	0.07	0.06	0.23	0.78	5.66
PRODQUAL	0.06	0.15	0.67	0.67	6.45
RANGEVAR	-0.05	0.12	0.68	0.56	6.21

The importance and performance measures are often displayed as a scatterplot (see Figure 1). In addition to plotting the bivariate points, lines are added to separate the plotting region in four distinct areas. Each area is labeled and the plot is referred to as a quadrant plot.

The placement of the lines is somewhat arbitrary. In these plots I have placed the horizontal line near the point of statistical significance of the regression coefficients and the vertical lines all at a performance of 5.75.

Given this view of the data, one would concentrate efforts on improving the performance of the PRICING, SALESSUPT, and ADV/PROMO component categories. The ORDER/DEL category, although important, appears to be performing at a high level of satisfaction currently.

Figure 1: The Baseline Quadrant Plot:



This view of the data represents the typical approach to a key driver analysis and will form the basis of comparison of other techniques.

3. LATENT CLASS REGRESSION

In latent class regression one assumes that the respondents come from an unknown number of latent classes (DeSarbo and Cron, 1988). Each latent class has its own regression parameters and each respondent has a probability of belonging to each of the latent classes. A variety of information heuristics are employed to aid in the choice of the number of latent classes. A maximum likelihood solution is arrived at via the EM algorithm.

A two-class solution was chosen here for illustrative purposes. Repondents were assigned to the class with the highest probability, resulting in one-hundred-sixteen (116) in latent class 1 and forty-four (44) in latent class 2. The importance and performance measures are displayed in Table 2 and Table 3, while the quadrant plots can be found in Figure 2 and Figure 3.

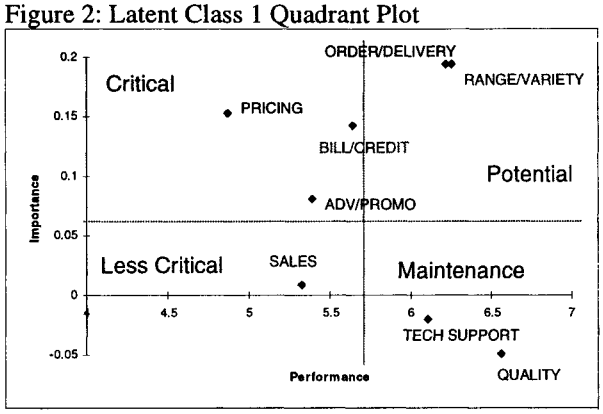
In latent class 1 respondents appear to put much more importance on RANGEVAR than did the baseline analysis indicated. The satisfaction with performance of RANGEVAR is relatively high, so no immediate action would be required. However, in this class of respondents, BILLCRED has taken on more importance than in the total sample, and probably would warrant some attention. SALES appears to have less importance in this latent class.

The tolerances appear to be reasonable, making the interpretation of the regression coefficients relatively straightforward.

Table 2: Latent Class 1 Statistics

	Parameter	Standard			
Variable	Estimate	Error	Prob > ITI	Tolerance	Mean
OVERALL					5.93
PRICING	0.15	0.03	0.00	0.65	4.87
ORDERDEL	0.19	0.04	0.00	0.72	6.22
BILLCRED	0.14	0.04	0.00	0.59	5.64
RANGEVAR	0.19	0.07	0.01	0.58	6.25
ADVPROMO	0.08	0.04	0.05	0.70	5.39
PRODQUAL	-0.05	0.08	0.54	0.70	6.56
TECHSUPT	-0.02	0.04	0.65	0.69	6.10
SALESUPT	0.01	0.03	0.78	0.82	5.33

Figure 2: Latent Class 1 Quadrant Plot

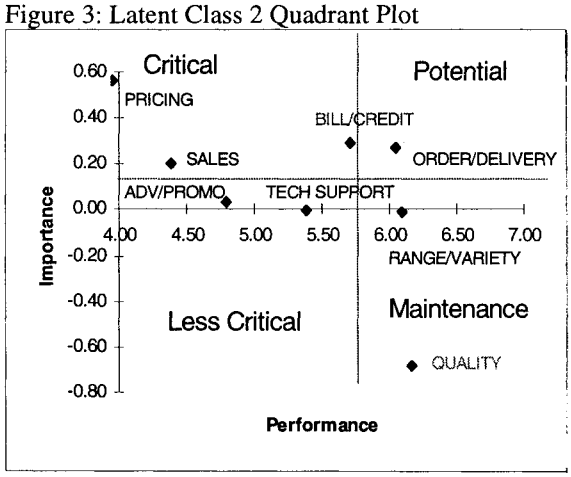


In latent class 2 ADV/PROMO appears to have taken on less importance and BILLCRED taken on more importance than in the total sample. Table 3 shows some alarmingly low tolerances, an indication of multicollinearity of the predictors. Indeed, the very high negative for PRODQUAL raises a great deal of concern. Further analysis shows that the correlation between OVERALL and PRODQUAL to be positive.

Table 3: Latent Class 2 Statistics

Variable	Parameter Estimate	Standard Error	Prob > ITI	Tolerance	Mean
OVERALL					3.61
PRICING	0.56	0.09	0.00	0.49	3.95
BILLCRED	0.29	0.09	0.00	0.67	5.70
SALESUPT	0.20	0.06	0.00	0.75	4.39
PRODQUAL	-0.68	0.23	0.01	0.49	6.16
ORDERDEL	0.27	0.10	0.01	0.67	6.05
ADVPROMO	0.03	0.11	0.76	0.45	4.80
TECHSUPT	-0.01	0.07	0.94	0.78	5.39
RANGEVAR	-0.01	0.18	0.96	0.37	6.09

Figure 3: Latent Class 2 Quadrant Plot



Clearly, the regression coefficients cannot be interpreted as is. Although there was some indication of multicollinearity in the total sample, it was not nearly as severe as that which has manifested itself in latent class 2. The next step might be to start dropping predictors from the model until the multicollinearity problem is resolved. This solution is not desirable, because the process owner of each component category is expecting a customer satisfaction report and it will be difficult to explain why we chose to drop his/her component category from the model rather than some other.

We could alternatively perform a factor analysis of the predictors in latent class 2 and regress the factors on overall satisfaction. This line of reasoning led the author to the idea of doing the factor analysis on the data before doing the latent class analysis. This is described in the next section of the paper.

4. PRINCIPAL COMPONENT LATENT CLASS REGRESSION

It is not a new idea to use principal components to reduce multicollinearity in a regression problem. Jackson (1991) provides a good discussion, and criticism, of the technique described as principal component regression.

In this case, the data was factor analyzed and rotated to a varimax solution keeping all eight factors. This results in uncorrelated factors which, in total, explain all of the original variance. The loading matrix for the rotated factor solution appears in Table 4. Each of the original variables loads highly on one and only one of the factors and visa-versa. This need not be the case, but leads to a easily interpreted solution here.

Table 4: The Rotated Factor Solution

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7	FACTOR8
PRICING	0.15	0.12	0.10	0.08	0.09	**0.93**	0.24	0.14
BILLCRED	0.02	**0.96**	0.15	0.04	0.08	0.11	0.10	0.14
SALESUPT	0.05	0.04	0.04	**0.98**	0.16	0.07	0.08	0.05
PRODQUAL	**0.94**	0.02	0.07	0.06	0.12	0.15	0.16	0.21
ORDERDEL	0.06	0.15	**0.97**	0.04	0.09	0.09	0.06	0.14
ADVPROMO	0.17	0.11	0.07	0.10	0.09	0.25	**0.92**	0.16
TECHSUPT	0.12	0.09	0.09	0.18	**0.95**	0.08	0.09	0.14
RANGEVAR	0.25	0.18	0.17	0.06	0.16	0.16	0.18	**0.89**

These eight factors were then submitted to a latent class regression analysis, predicting OVERALL satisfaction, and a two-class solution obtained. The importance and performance statistics for each latent class appear in Table 5 and Table 6 while the corresponding quadrant plot appear in Figure 4 and Figure 5.

The solution converged to classes of exactly the same proportions as in the orginal latent class analysis. Indeed, upon further inspection it was realized that the

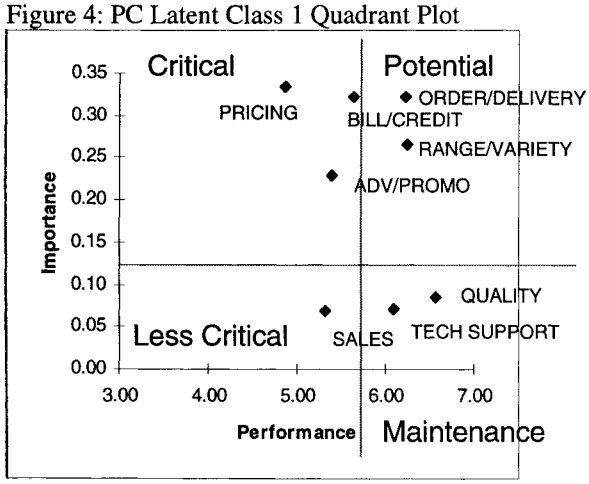
class memberships were exactly as it was in the first latent class analysis. The mean performances are the same as they were in the previous analysis. The regression coefficients, and hence importances, have changed somewhat. The tolerances have also been reduced.

In latent class 1 the relative value of the importances remained the same and the points are all in the same quadrant as they were in the latent class analysis.

Table 5: PC Latent Class 1 Statistics

Variable	Parameter Estimate	Standard Error	Prob > ITI	Tolerance	Mean
OVERALL					5.93
FPRICING	0.33	0.05	0.00	0.93	4.87
FORDERDEL	0.32	0.05	0.00	0.92	6.22
FBILLCRED	0.32	0.05	0.00	0.88	5.64
FRANGEVAR	0.27	0.05	0.00	0.98	6.25
FADVPROMO	0.23	0.05	0.00	0.96	5.39
FPRODQUAL	0.09	0.05	0.07	0.95	6.56
FTECHSUPT	0.07	0.06	0.22	0.95	6.10
FSALESUPT	0.07	0.05	0.18	0.94	5.33

Figure 4: PC Latent Class 1 Quadrant Plot



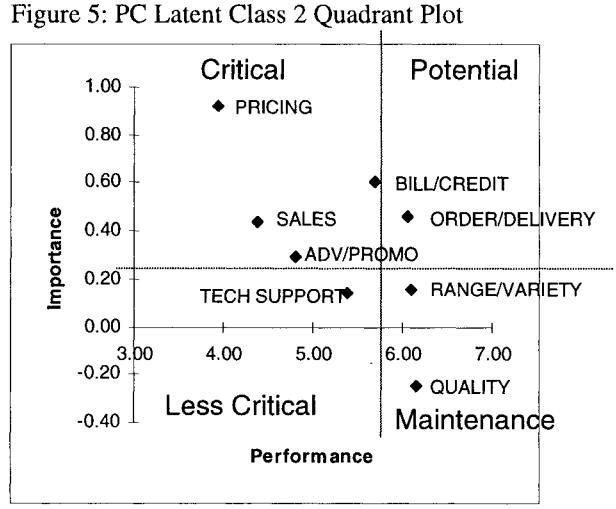
In the second latent class the relative value of the importances is the same as in the first latent class analysis. The points on the quadrant plot all land in the same quadrant as they did in the original latent class analysis.

The tolerances have been reduced substantially although the bothersome negative coefficient for PRODQUAL remains. The PRODQUAL category persists as a problem with all the methods of analysis. This is proably due to the fact that ratings of 5, 6, and 7 only were given to PRODQUAL, resulting in a highly compacted and discretized scale.

Table 6: PC Latent Class 2 Statistics

Variable	Parameter Estimate	Standard Error	Prob > ITI	Tolerance	Mean
OVERALL					3.61
FPRICING	0.92	0.12	0.00	0.76	3.95
FBILLCRED	0.60	0.12	0.00	0.69	5.70
FORDERDEL	0.46	0.11	0.00	0.76	6.05
FSALESUPT	0.44	0.11	0.00	0.80	4.39
FADVPROMO	0.29	0.11	0.01	0.91	4.80
FRANGEVAR	0.16	0.11	0.16	0.87	6.09
FTECHSUPT	0.14	0.09	0.11	0.90	5.39
FPRODQUAL	-0.25	0.13	0.06	0.84	6.16

Figure 5: PC Latent Class 2 Quadrant Plot



The analysis based on the rotated principal components mirrors the original latent class analysis. The same latent classes were derived from the data, and the relative size of the regression coefficients remained the same. The was some apparent relief from the multicollinearity seen in the first analysis, although the negative coefficient for PRODQUAL was not eliminated.

5. OTHER LOOKS AT THE DATA

Some other approaches were also used to look at the data. A dual scaling analysis was done to examine what component categories were important in moving

respondents up on the OVERALL satisfaction scale. Dual scaling found:

1) SALES and TECHSUPT moved OVERALL from less than a 4 to a 5 rating
2) SALES, ORDERDEL, and PRICING moved OVERALL from a 5 to a 6 rating
3) all component categories were elements of moving OVERALL from a 6 to a 7 rating.

A MARS analysis was also run on the data. Nonlinear main effects of PRICING, as well as weak interaction effects involving BILLCRED, SALESUPT, and ORDERDEL were found in the data.

6. SUMMARY AND RECOMMENDATIONS

Latent class analysis has fundamental appeal in the analysis of customer satisfaction data. If we could successfully identify customer segments whose overall satisfaction was a function of more specific components, products and services could be tailored to each customer segment.

The results of the application of latent class regression to this customer satisfaction data was relatively disappointing. Little new insight was gained from the splitting the population into two latent classes. Indeed, it only exasperated the multicollinearity problem.

The author suspects that the failure to get interesting results was due in large part to "problems" with the data. These problems include:

1) a great deal of multicollinearity
2) restricted ranges of responses, and
3) highly discretized scales of measurement.

The procedure may do better after removing problematic variables (e.g. PRODQUAL, ADVPROMO, and RANGEVAR).

7. ACKNOWLEDGMENTS

I would like to thank Chuck Heckler for the MARS analysis and a custom program for mixture regression program, Dan Lawrence for the dual scaling analysis, and Bill Novik for many insightful discussions.

8. REFERENCES

DeSarbo, W. S. and Cron, W. L. (1988), "A Maximum Likelihood Methodology for Clusterwise Linear Regression", *Journal of Classification*, *5*, 249-282.

Jackson, J. E. (1991), *A User's Guide to Principal Components*, John Wiley and Sons.