An Investigation to Assess the use of Multiple-outcome Logistic Regression Techniques in Credit Scoring

Christopher Tyers BSc

Abstract

Traditionally credit scoring systems have aimed to reduce the risk in a lender's portfolio by predicting whether a customer would default on their account. However, the current models offer limited depth, as they only use binary outcomes. It has been suggested that by utilising multiple outcomes more predictive models could be produced. This paper investigates the use of multiple-outcome logistic regression models in the context of credit scoring. These multiple-outcome models are then compared to a binary logistic model; a widely used model in the credit industry, to offer a comparison in discriminatory power. Finally the paper discusses the impact that multiple outcome models would have if implemented into a credit scoring solution.

1. Introduction

Mester (1997) describes credit scoring as "a statistical method used to predict the probability that a [credit] applicant or existing borrower will default or become delinquent". By utilising credit scoring lenders are able to reject applicants who are likely to default or become delinquent. Prior to the advent of automated systems, lenders would make decisions on an application-by-application basis. A small team of credit analysts would process all of the applications manually leading to inconsistencies in the lending decisions. To resolve this problem some mail-order companies introduced numerical scoring systems. With the start of World War II all of the credit lenders began to experience difficulties with credit management, as credit analysts were being drafted into military service. As there was a shortage of experienced credit analysts, lenders had their analysts write down the rules that they used to decide upon an applicant's credit worthiness. Non-experts could then use these rules to help make lending decisions. After the war people began to connect the classification techniques being developed by statisticians and the automation of lending decisions (Wonderlic, 1952). The connections being made at this time led to the formation of the first credit consultancy: Fair Isaac, in 1956.

Traditionally, credit scoring has used statistical methods such as discriminant analysis, and more latterly the more robust logistic regression (Thomas et al, 2002). Hand & Henley (1997) discuss the merits of various statistical methods that can be used to model credit data. When looking at the use of linear regression Hand & Henley make the point that although logistic regression appears to be more suitable for use in a two class situation such as modelling goods and bads it is no better than linear regression as a large proportion of the applicants have estimated probabilities of good between 0.2 and 0.8. In this case the logistic curve is approximated by a straight line making linear regression suitable. More recently research into the field of credit scoring has focused more on the use of data mining techniques to classify customers. Chye, Chin & Peng (2004) detail the use of a variety of data mining techniques that can be used for the classification of customers, such as decision trees and neural networks. Although many of the techniques used to classify credit applicants are capable of modelling more than two categories, credit scoring is normally conducted using only two categories. The aim of this paper is to investigate the use of multiple outcome logistic regression techniques, and appraise their use in the credit scoring industry.

2. Data

The data used for the analysis comprised 12000 applicants for a fixed term loan and contained 10 explanatory variables and an outcome flag.

Table 1, below, shows the number of applicants that fall into each of the 3 classifications.

	Frequency	*Percent*
Good	10000	83.33
Bad	1000	8.33
Early Payer	1000	8.33

Table 1: Frequency of classification

The data were a mixture of application and performance characteristics. The characteristics supplied were as follows: Time at Bank (Months), Time at Address (Months), Marital Status, Total Number of Accounts (Last 6 Months), Number of Credit Searches (Last 3 Months), Age of Applicant, Number of Defaults, Time Since Most Recent County Court Judgement (CCJ), Demographic Index and Credit Risk Score. Where possible the variables were used as continuous variables. However due to the nature of some of the variables they needed to be banded into discrete categories. Firstly, Number of defaults was recoded into three groups: 0, 1 and 2+, this was because there were so few applicants with high numbers of defaulted accounts. Secondly, Time Since Most Recent CCJ was recoded as a binary flag; CCJ vs No CCJ. This was again because of the low number of applicants who had this derogatory data.

3. Binary Logistic Regression

a. Methodology

The binary logistic regression model will be used as a control model so that any uplift produced by the use of multiple outcomes can be seen. As logistic regression is used to model a binary response (Jones and Kilner, 2007), we would anticipate that the number of customers that are deemed to be 'bad' customers, as detailed in section 2, would follow a binomial distribution,

$$y_i \sim B(n_i, \pi_i).$$

Where y_i represents the observed number of subjects, who have a value of x_i for the factor of interest, who have been classified as 'bad' customers. n_i is the number of customers in the sample with a value of x_i for the factor. π_i represents the underlying true probability that a customer in the sample is a 'bad' customer. If we were to fit a model with π as the response, the model would need to be constrained to values of π between 0 and 1. To achieve this, we first consider the odds ratio, $\pi/(1 - \pi)$. It can be seen that as π approaches zero, the odds ratio also approaches zero; whilst as π approaches 1, the odds ratio approaches infinity. We now consider the log odds ratio, also known as the logistic function:

$$\eta = \log\left(\frac{\pi}{1-\pi}\right). \qquad (3.1)$$

As π approaches zero the odds ratio also approaches zero, so η approaches negative infinity. As π approaches 1 the odds ratio approaches infinity such that η also approaches infinity.

By exponentiating equation 3.1 we get:

$$e^{\eta} = \frac{\pi}{1-\pi}, \qquad (3.2)$$

by rearrangement it can be shown that:

$$\pi = \frac{e^{\eta}}{1+e^{\eta}}. \qquad (3.3)$$

Equation 3.3 is known as the inverse logistic function. Notice that as η approaches negative infinity e^{η} approaches zero, and as a result, π approaches zero. As η approaches infinity e^{η} approaches infinity, such that π approaches 1. Therefore, regardless of the value that the logistic function, η, takes, the corresponding probability, π, must lie between 0 and 1. Collett (2003) shows that the logistic function can be modelled by a linear function of x_i, known as the linear predictor.

$$\eta_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \qquad (3.4)$$

The β terms in equation 3.4 can be estimated by maximum likelihood and can be interpreted as the additive effect on the log odds ratio for a unit change in the *j*th explanatory variable.

b. Analysis

Prior to beginning the analysis the response variable was dichotomised to a binary response of 'Good' and 'Bad or Early Payer'. The first stage of the modelling process was to build single term models using each of the explanatory variables. This would allow for any insignificant variables to be removed from the latter stages of the analysis. All of the variables were found to be significant at the 5% level. All of the terms were than passed to stepwise modelling procedures, with entry and exit parameters of 0.05. Interaction terms between each of the remaining variables were then added to the model, and stepwise regression was repeated. The final model inclused the terms: Time at bank, Time at address, Marital status, Number of credit searches, Age, Credit risk score and the interaction term between marital status and number of credit searches.

4. Proportional Odds Model

a. Methodology

Agresti (2002) details that the proportional odds model utilises the ordinality of a response variable, and in doing so improves model parsimony and power, over binary models. Cumulative probabilities can be used to express the order of the categories,

$$P(Y \leq j \mid x) = \pi_1(x) + \ldots + \pi_j(x), \qquad (4.1)$$

where j is the response category, and $j = 1, \ldots, J$, and x is a factor of interest.

The cumulative probabilities shown in expression 4.1 can then be expressed as cumulative logits, defined as:

$$\begin{aligned}\text{logit}[P(Y \le j \mid x)] &= \log \frac{P(Y \le j \mid x)}{1 - P(Y \le j \mid x)} \\ &= \log \frac{\pi_1(x) + \ldots + \pi_j(x)}{\pi_{j+1}(x) + \ldots + \pi_J(x)},\end{aligned} \tag{4.2}$$

where $j = 1, \ldots, J-1$ and each of the cumulative logits uses all J response categories. The proportional odds model is a model that uses all of the cumulative logits simultaneously. The model can be given by:

$$\text{logit}[P(Y \le j \mid x)] = \alpha_j + \beta' x, \tag{4.3}$$

where $j = 1, \ldots, J-1$.

Each of the cumulative logits has its own intercept, α_j. As the value of j increases, so does the value of α as $P(Y \le j \mid x)$ increases in j for a fixed value of x. Each of the logits include the term β` to represent the parameters for each of the explanatory variables. The inclusion of the β` them in expression 4.3 constrains the $J - 1$ response curves to have the same shape. The fit of this model is not the same as fitting separate logit
models for each j.

$$\begin{aligned}&\text{logit}[P(Y \le j \mid x_1)] - \text{logit}[P(Y \le j \mid x_2)] \\ &= \log \frac{P(Y \le j \mid x_1)/P(Y > j \mid x_1)}{P(Y \le j \mid x_2)/P(Y > j \mid x_2)} \\ &= \beta'(x_1 - x_2).\end{aligned} \tag{4.4}$$

The cumulative odds ratio; the odds ratio for cumulative probabilities, for a response $\le j$ at $x = x_1$ is $\exp[\beta'(x_1 - x_2)]$ times the odds at $x = x_2$. This property gave the model its name, as the log cumulative odds ratio is proportional to the distance between x_1 and x_2; shown in expression 4.4. If each of the explanatory variables cannot satisfy this property, the model is said to violate the proportional odds assumption thus invalidating the model.

b. Analysis

Before any analysis could be conducted the score test for the proportional odds assumption needed to be carried out for each of the explanatory variables. This test would show which of the variables met the assumptions imposed be the model. After conducting the test for each of the variables it emerged that seven of the 10 variables violated the proportional odds assumption. This demonstrated that the use of the proportional odds model would not be viable for these data.

5. Partial Proportional Odds Model

a. Methodology

The partial proportional odds model is similar to the proportional odds model, in that it uses $J - 1$ cumulative logits simultaneously to model J outcomes. However, rather than constraining all of the parameters to be the same across all logits, it allows the variables that violate the proportional odds assumption to have differing parameters. The variables that do meet the proportional odds assumption are still constrained as in the proportional odds model. As stated above, the partial proportional odds model uses cumulative probabilities;

$$P(Y \le j \mid x) = \pi_1(x) + \ldots + \pi_j(x),$$

to express the ordinality of the data. As with the full proportional odds model these cumulative probabilities can be expressed as a set of $J - 1$ cumulative logits,

$$\begin{aligned}\text{logit}[P(Y \le j \mid x)] &= \log \frac{P(Y \le j \mid x)}{1 - P(Y \le j \mid x)} \\ &= \log \frac{\pi_1(x) + \ldots + \pi_j(x)}{\pi_{j+1}(x) + \ldots + \pi_J(x)}.\end{aligned}$$

The partial proportional odds model differs from the full proportional odds model in that it allows the parameters of the covariates that do not meet the proportional odds assumption to vary across the different logits, while constraining those that do meet the proportional odds assumption to be the same. Where the full proportional odds model takes the form:

$$\text{logit}[P(Y \le j \mid x)] = \alpha_j + \beta' x,$$

Stokes, Davis and Koch (2000) show that the partial proportional odds model, instead, takes the form:

$$\text{logit}[P(Y \leq j \mid x)] = \alpha_j + \beta_1' x_{i1} + \beta_{2j}' x_{i2}. \qquad (5.1)$$

Where x_{i1} are the explanatory variables that meet the proportional odds assumption, and β'_1 are the parameters associated with these variables. x_{i2} are the explanatory variables that do not meet the proportional odds assumption, and β'_{2j} are the parameters associated with these variables. Unlike the β'_1s, the β'_{2j}s can be different for each value of j.

b. Analysis

Stokes, Davis and Koch (2000) detail the process for fitting the partial proportional odds model in SAS, as it is not directly available. The method approximates the score test for the proportional odds assumption by modelling interaction terms between each logit and each of the explanatory variables. The results of this model showed that five of the 10 explanatory variables do not meet the proportional odds assumption. The logit interactions for these variables are retained in the model for the remaining analysis. A pseudo-stepwise procedure was then conducted to remove the insignificant terms from the model. The variables remaining in the final model were Time at bank, Time at address, Number of credit searches, Marital status, Total number of accounts, Age, Demographic Index and Credit risk score.

6. Baseline-Category Logit Model

a. Methodology

If Y is a categorical response with J categories, a baseline-category logit model will simultaneously describe the log odds for all $\binom{J}{2}$ pairs of categories. However, all but $J-1$ of these pairs are redundant.

Let $\pi_j(x) = P(Y \leq j \mid x)$ at a fixed value of x for explanatory variables, with $\Sigma_j \pi_j(x) = 1$. For observations where $X=x$ we can treat the counts for each of the J categories of Y as multinomial with probabilities $\{\pi_1(x), \ldots, \pi_J(x)\}$

A baseline category is often paired with each response category in logit models. The model;

$$\log \frac{\pi_j(x)}{\pi_J(x)} = \alpha_j + \beta_j' x, \qquad (6.1)$$

simultaneously describes the effects of x on the $J-1$ logits, where $j = 1, \ldots, J-1$. The effects vary according to the response level paired with the baseline. These $J-1$ equations determine the parameters for the logits of the other pairs of response categories, as;

$$\log \frac{\pi_a(x)}{\pi_b(x)} = \log \frac{\pi_a(x)}{\pi_J(x)} - \log \frac{\pi_b(x)}{\pi_J(x)}. \qquad (6.2)$$

The equation used to express a multinomial logit model, such as the baseline-category logit model, in terms of the response probability $\pi_j(x)$ is;

$$\pi_j(x) = \frac{\exp(\alpha_j + \beta_j' x)}{1 + \Sigma_{h=1}^{J-1} \exp(\alpha_h + \beta_h' x)}, \qquad (6.3)$$

with $\alpha_J = 0$ and $\beta_J = 0$. This follows from equation 6.1, in the fact that equation 6.1 also holds with $j = J$ by setting $\alpha_J = 0$ and $\beta_J = 0$.

In situations where $J = 2$ equation 6.3 simplifies to the equation for binary logistic regression,

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}, \qquad (6.4)$$

which is equivalent to equation 3.3.

b. Analysis

The modelling process for the baseline-category logit model followed a similar process to the binary logistic regression.

Firstly, single term models were used to identify which of the variables were likely to be significant in any further modelling. All of the variables were found to be significant, at the 5% level, in at least one of the logits at this stage so none of the variables were removed from further analysis. Stepwise selection was then performed on the variables with entry and exit parameters of 0.05. Interaction terms between each of the variables were then added to the model, and the stepwise selection was repeated. The final model consisted of the terms: Time at bank, Marital Status, Total number of accounts, Age and Credit risk score.

7. Measures of Association

The measures of association that have been calculated from the concordance analysis are: the concordance index (*c*), Somers' D, Godman-Kruskal Gamma and Kendall's Tau-a. The journal *Nature* (Concordance Index, 2008) defines the concordance index: *c*, as "the proportion of subject pairs in which the subject with the higher true response also has the higher predicted response", and also states that *c* has a range from 0.5; representing no discriminating ability, to 1; perfect discrimination. For a binary outcome, the concordance index is an approximation to the area under the ROC curve. The SAS support article, *Rank Correlation of Observed Responses and Predicted Probabilities (2008)*, defines Somers' D as the difference between the number of concordant and discordant pairs divided by the total number of pairs, and as such penalises for the presence of ties. Somers' D has a range from 0 to 1, 1 being perfect discrimination.

Goodman-Kruskal Gamma is similar to Somers' D, in that it is a ratio of the difference between the number of concordant and discordant pairs, and the total number of pairs. However, Goodman-Kruskal Gamma does not penalise for tied pairs in the computation. When there are no ties Goodman-Kruskal Gamma is synonymous with Somers' D. Preston (2006) details the use of Kendall's Tau-a. The measure is described as a ratio of the difference between the number of concordant and discordant pairs. Although, rather than being divided by the total number of concordant, discordant or tied pairs, is divided by the number of all pairs in the sample.

8. Results

After conducting the analysis using each of the modelling techniques, it was found that all of the models had a poor predictive power. This was due to the fact that most of the applicants had predicted probabilities of good over 0.5. However, if the resultant models were to be used as a credit risk tool, the score would not be used to classify the applicants directly, being used instead to rank the customers; any applicant achieving a score over a given cut-off would be accepted; being deemed a probable good. Because of the way that the final model would be used, to rank the customers in term of risk, a concordance analysis can be performed on the predicted probabilities of each of the models. The concordance analysis will allow for the derivation of several measures of association, which show the models ability to rank the applicants according to their known classifications.

	Binary Logistic Regression	*Partial Proportional Odds Model*	*Baseline-Category Logit Model*
c	0.6957	0.7156	0.7210
Somers' D	0.3914	0.4313	0.4421
Gamma	0.3914	0.4313	0.4421
Tau-a	0.1087	0.1257	0.1289

Table 2: Measures of Association

Table 2 shows the measures of association for each of the models. The concordance index shows that there is at least a 0.2 increase in the number of concordant pairs, after adjusting for ties, between the binary logistic model and the multiple outcome models. The values of Somers' D and Gamma confirm that the multiple outcome models have an increased discriminatory power over the binary logistic regression model. Agresti (2002) details that utilising the ordinality of a response variable can improve model parsimony and power when compared to binary models. Due to the limitations in the data this research cannot confirm Agresti's research, however, it does show that multiple outcome logistic regression models have a greater discriminatory power than binary logistic models when used in the context of credit scoring.

9. Conclusions

By utilising a model that has a higher discriminatory power a lender will be able to reduce the bad rate of their book: the number of accepted customers that become bad for every customer accepted. The simplest interpretation of this understanding is that fewer potentially bad customers are accepted and as bad customers lead to a financial loss for the lender, the lenders losses are reduced. In the current financial climate it is important for lenders to show that their books do not represent too large a risk. Additionally, by introducing the concept of a multiple outcome solution to the lending decision, lenders would be able to tailor the terms of acceptance for each of the different predicted outcomes; for instance, ensuring that customers who are likely to pay off their credit early face higher penalty charges for this early repayment. Due to the fact that all of the models investigated use the same set of explanatory variables, there will only ever be a limited scope to improving the discriminatory power of the model. Although this research has shown that the use of multiple outcome logistic regression models lead to a more discriminatory model, there are several shortcomings to the analysis that prevent a concrete recommendation from being made. Firstly, the Proportional Odds model, and less so the Partial Proportional Odds Model, are only valid if the outcome is truly ordinal. Secondly, all of the models assume that the outcomes are mutually exclusive. Although in this case it is reasonable to assume that the outcomes, at a fixed point in time, are mutually exclusive, in other cases, however, this may not be the case.

To fully investigate the use of multiple outcome logistic regression models it is recommended that more analysis be conducted on a range of datasets, each containing a different set of outcomes. Also, the use of survival analysis, in particular competing risks models, could be used to assess how the risk of one of the outcomes occurring changes over time.

10. Acknowledgements

As well as cited references I would like to acknowledge the input that has been made by Keith Jones & Karen Kilner, Sheffield Hallam University, and Alex Gillespie, Experian Decision Analytics.

11. References

AGRESTI, Alan (2002). *Categorical data analysis.* 2nd ed., New Jersey, Wiley-Interscience.

CHYE, Koh Hian, CHIN, Tan Wei and PENG, Goh Chwee (2004). Credit Scoring Using Data Mining Techniques. *Singapore Management Review*, **26** (2), 25-47.

COLLETT, David (2003). *Modelling binary data*. 2nd ed., Boca Raton, Chapman and Hall/CRC.

Concordance Index (2008). [online]. Last accessed 17 April 2008 at: http://www.nature.com/glossary/clinicalpractice/defDetails.do?uid=ncp_164

DURAND, D (1941). *Risk elements in consumer instalment financing*. National bureau of economic research.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. **7**, 179-188.

HAND, D. J. and HENLEY, W. E. (1996). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society.* Series A (Statistics in Society), **160** (3), 523-541.

JONES, K and KILNER, K (2007). *Logistic Regression for Epidemiological Studies,* v1.1. Sheffield Hallam University.

LEWIS, E. M (1992). *An Introduction to Credit Scoring*. Athena Press.

MESTER, Loretta (1997). What's the point of credit scoring? *Federal Reserve Bank of Philadelphia, Business Review.* **1997** (September), 3-16.

PRESTON, S (2006). Assessing the U.S. Senate Vote on the Corporate Average Fuel Economy (CAFE). *Standard Journal of Statistics Education*, **14** (2).

Rank Correlation of Observed Responses and Predicted Probabilities (2008). [online]. Last accessed 16 April 2008 at: http://support.sas.com/ documentation/cdl/en/statug/59654/HTML/default/statug_logistic_sect035.htm

STOKES, DAVIS and KOCH (2000), Partial proportional odds model. [online]. Last accessed 1 April 2008 at: http://www.bios.unc.edu/~jpreisse/bios765/ notesS.pdf

THOMAS, Lyn, EDELMAN, David and CROOK, Jonathon (eds.) (2002). *Credit scoring and its applications.* Philadelphia, Society for Industrial and Applied Mathematics.

WONDERLIC, E. F. (1952). An analysis of factors in granting credit. *Indiana University Bulletin.* **50** (27), 163-176.