Logistic regression and its application in credit scoring

Christine Bolton

2009

Abstract

Credit scoring is a mechanism used to quantify the risk factors relevant for an obligor's ability and willingness to pay. Credit scoring has become the norm in modern banking, due to the large number of applications received on a daily basis and the increased regulatory requirements for banks. In this study, the concept and application of credit scoring in a South African banking environment is explained, with reference to the International Bank of Settlement's regulations and requirements. The steps necessary to develop a credit scoring model is looked at with focus on the credit risk context, but not restricted to it. Applications of the concept for the whole life cycle of a product are mentioned. The statistics behind credit scoring is also explained, with particular emphasis on logistic regression. Linear regression and its assumptions are first shown, to demonstrate why it cannot be used for a credit scoring model. Simple logistic regression is first shown before it is expanded to a multivariate view. Due to the large number of variables available for credit scoring models provided by credit bureaus, techniques for reducing the number of variables included for modeling purposes is shown, with reference to specific credit scoring notions. Stepwise and best subset logistic regression methodologies are also discussed with mention to a study on determining the best significance level for forward stepwise logistic regression. Multinomial and ordinal logistic regression is briefly looked at to illustrate how binary logistic regression can be expanded to model scenarios with more than two possible outcomes, whether on a nominal or ordinal scale. As logistic regression is not the only method used in credit scoring, other methods will also be noted, but not in extensive detail. The study ends with a practical application of logistic regression for a credit scoring model on data from a South African bank.

Contents

1 Introduction **4**
1.1 The New Basel Capital Accord (Basel II) . . . 5
1.2 Multivariate analysis . . . 7
1.3 Credit scoring . . . 8
1.4 Brief outline of the study . . . 11

2 Steps in credit scoring model development **12**
2.1 Step 1: Understanding the business problem . . . 12
2.2 Step 2: Defining the dependent variable . . . 12
2.3 Step 3: Data, segmentation and sampling . . . 14
2.3.1 Historical lending experience . . . 14
2.3.2 Data retention . . . 14
2.3.3 Known outcomes of past decisions . . . 14
2.3.4 Age of decision . . . 14
2.3.5 Sample size . . . 15
2.4 Step 4: Fitting of a model and optimization of selected criteria . . . 17
2.5 Step 5: Generalization . . . 18
2.6 Step 6: Ongoing monitoring . . . 18

3 Credit scoring methods **19**
3.1 Linear discriminant analysis . . . 20
3.2 Linear regression . . . 21
3.3 k-Nearest Neighbour Classification . . . 21
3.4 Classification and Regression Trees (CART) . . . 22
3.5 CHAID Analysis . . . 23
3.5.1 Preparing predictors . . . 24
3.5.2 Merging categories . . . 24
3.5.3 Selecting the split variable . . . 24
3.5.4 Exhaustive CHAID algorithms . . . 24
3.6 Neural networks . . . 25
3.7 Which method is best? . . . 26

4 Linear regression and its assumptions **28**
4.1 Simple linear regression . . . 28
4.2 Assumptions of linear regression . . . 31
4.2.1 Constant variance of the error terms . . . 33
4.2.2 Independence of error terms . . . 35
4.2.3 Normality of the error term distribution . . . 37

5 Simple logistic regression **38**
5.1 Deviance . . . 44
5.2 Likelihood-ratio test . . . 45
5.3 Wald Test . . . 45
5.4 Score Test . . . 45

6 Multivariate logistic regression **48**
6.1 Testing significance of the model . . . 50
6.2 Interpretation of the fitted model . . . 56

7 Variable reduction and analysis in credit scoring **66**
7.1 Introduction . . . 66
7.1.1 Logical . . . 67
7.1.2 Predictive . . . 67
7.1.3 Multicollinearity . . . 67
7.1.4 Available and stable . . . 67
7.1.5 Compliant . . . 68
7.1.6 Customer related . . . 68
7.1.7 Minimum information loss . . . 68
7.2 Bivariate analysis . . . 69
7.2.1 Likelihood ratio test . . . 69
7.2.2 Pearson chi-square test . . . 69
7.2.3 Spearman rank-order correlation . . . 70
7.2.4 Weight of Evidence (WOE) . . . 71
7.2.5 Information Value . . . 72
7.2.6 Gini coefficient . . . 72
7.3 Variable cluster analysis . . . 80
7.3.1 Interpreting VARCLUS procedure output . . . 83

8 Different modeling techniques for logistic regression **84**
8.1 Stepwise logistic regression . . . 84
8.1.1 Step 0 . . . 85
8.1.2 Step 1 . . . 86
8.1.3 Step 2 . . . 86
8.1.4 Step 3 . . . 86
8.1.5 Step S . . . 87
8.2 Best subsets logistic regression . . . 88
8.3 Testing the importance of variables in the model . . . 90
8.4 Analysis of the variables in the model . . . 91

9 Best significance level in forward stepwise logistic regression **92**
9.1 Performance criterion . . . 92
9.2 Selection criteria . . . 92
9.3 Monte Carlo Experimental Design . . . 93
9.3.1 Step 1: . . . 93
9.3.2 Step 2: . . . 93
9.3.3 Step 3: . . . 94
9.3.4 Step 4: . . . 94
9.4 Multivariate normal case . . . 94
9.5 Multivariate binary case . . . 96
9.6 Results of the simulations experiments . . . 97
9.6.1 Multivariate Normal Case . . . 97
9.6.2 Multivariate Binary Case . . . 97
9.7 Conclusion . . . 98

10 Assessing the fit and predictive power of the model **99**
10.1 Summary measures of Goodness-of-fit . . . 99
10.1.1 Pearson Chi-square and Deviance . . . 100
10.1.2 The Hosmer-Lemeshow Tests . . . 102
10.1.3 Classification tables . . . 105
10.2 Separation statistics . . . 106
10.2.1 Divergence statistic . . . 107
10.2.2 Area under the ROC curve . . . 107

10.2.3 KS statistic . . . 108
10.3 Logistic regression diagnostics . . . 110
10.4 Assessment of fit via external validation . . . 118

11 Multinomial and Ordinal Regression **120**
11.1 The Multinomial Logistic Regression model . . . 120
11.1.1 Interpreting and Assessing the Significance of the Estimated Coefficients . . . 123
11.1.2 Model building strategies for multinomial logistic regression . . . 131
11.1.3 Assessing the fit and diagnostics for the multinomial logistic regression model . . . 131
11.2 Ordinal logistic regression models . . . 132
11.2.1 Model building strategies for Ordinal Logistic Regression Models . . . 139

12 Practical example **140**
12.1 Initial bivariate analysis . . . 141
12.2 Secondary bivariate analysis . . . 143
12.3 Variable clustering . . . 144
12.3.1 Cluster 1 . . . 145
12.3.2 Cluster 2 . . . 146
12.3.3 Cluster 3 . . . 146
12.3.4 Cluster 4 . . . 146
12.3.5 Cluster 5 . . . 146
12.3.6 Cluster 6 . . . 146
12.3.7 Cluster 7 . . . 146
12.3.8 Cluster 8 . . . 147
12.3.9 Cluster 9 . . . 147
12.3.10 Cluster 10 . . . 147
12.4 Stepwise logistic regression . . . 147
12.4.1 Significance levels . . . 147
12.4.2 Design/dummy variables . . . 147
12.4.3 Stepwise summary . . . 148
12.4.4 Model coefficients . . . 149
12.4.5 Model refinement . . . 150
12.5 Final model . . . 167
12.6 Next steps . . . 169
12.7 Closing thoughts . . . 169
12.8 Appendix 1: Initial bivariate analysis . . . 170
12.9 Appendix 2: Secondary bucketing of variables . . . 186
12.10 Appendix 3: Cluster analysis SAS code and output . . . 202
12.11 Appendix 4: Stepwise logistic regression SAS code and output . . . 207
12.12 Appendix 5: Default rate sloping and model coefficients . . . 217
12.13 Appendix 6: Default rate sloping and model coefficients version 2 . . . 223
12.14 Appendix 7: Default rate sloping and model coefficients version 3 . . . 228

13 Glossary **232**

14 References **236**

Chapter 1

Introduction

In 2004, at the start of my career, I was seconded to a leading international company in the USA and Europe, which had a joint venture with a South African bank, with the specific purpose of obtaining experience in credit scoring in a banking environment. The ultimate goal of this secondment was to implement the methodology of credit scoring in the South African situation.

The objectives of this study are:

1. To research and study applications of categorical data analysis with specific reference to best practices in credit scoring,
2. To make an academic contribution in the field of credit scoring, and
3. To clarify the practical application of the methodology of credit scoring for the general banking industry.

Banks and the overall banking system are critical components of any country's economy and the world economy at large. Banks primarily rely on taking in deposits from clients, consumers, businesses and large companies, to fund the lending they then provide to a variety of clients and sectors in their economy and globally.

The safety and soundness of the banking system is of paramount importance. The collapse of a large bank could result in diminished confidence in the banking system, and this could have dire consequences for a country and have possible global impacts. This is due to the potential knock-on impact to other banks. An example of this is the global financial crisis of September-October 2008. It began with failures of large financial institutions in the United States, and then rapidly evolved into a global crisis resulting in a number of European bank failures and sharp reductions in the value of stocks and commodities worldwide. Clients panicked and abnormally withdrew their deposits. This lead to a liquidity problem that further accelerated this crisis. This crisis has its roots in the subprime mortgage crisis in the US and affected the financial systems internationally.

It is for this reason that banks are highly regulated (in South Africa by the South African Reserve Bank) and that the Bank of International Settlements (BIS) first issued a comprehensive set of principles and practices to help minimize this risk in 1988 (Basel I[1]). These mostly involve risk management, capital management and corporate governance.

The New Basel Capital accord (Basel II), replaces Basel I and is a comprehensive response to the significant developments in banking over the past 20 years.

[1] BIS has its capital in Basel, Switserland.

Some of the key criticisms of Basel I was that:

- It was risk insensitive as there was little differentiation between high, medium and low risk taking in a bank's business activities.
- It didn't reward good and penalized poor risk and capital management.
- It was generally out of touch with technological advancements in banking, financial markets and risk management.

The South African Reserve Bank confirmed an implementation date of 1 January 2008 for Basel II in South Africa.

1.1 The New Basel Capital Accord (Basel II)

The New Basel Capital Accord, (known as Basel II) that was released in June 2004 is the latest initiative by the BIS to regulate the global financial services industry. It tries to achieve this by more appropriate aligning of bank capital requirements with underlying risks - credit risk, market risk and operational risk. In short, the key objective of Basel II is to enhance the safety and soundness of the banking system through vastly improved risk and capital management, tailored to each individual bank and banking group.

Basel II is based on three mutually reinforcing pillars covering:

- Minimum capital requirements
- Supervisory review, and
- Market discipline.

Pillar 1 covers sophisticated risk measurement and management for large banks using internally developed models. Different approaches are available that varies in sophistication and complexity. It covers different types of risk, including credit risk, investment risk, operational risk, market risk and more.

Implementation of Basel II resulted in a significant increase in the regulatory role of the South African Reserve Bank (SARB) in South Africa. The SARB reviews and evaluates each bank's risk and capital management in great detail and might require higher capital levels based on the quality thereof. Pillar 2 also requires assessment of all other major risks not covered in Pillar 1.

Under Pillar 3, banks are now required to release information publicly about their respective risk profiles. This includes increased public disclosure of risk measurement and management practices, capital management and capital adequacy. This allows financial markets and investors to differentiate a bank's investment potential, based on the quality of their risk and capital management.

It is expected that disclosure of risks measured based on the Basel II requirements will strengthen prudential aspects of financial stability. This study will be confined to the estimation of default probability as a major component of credit risk provision for compliance with the Basel II standard. Estimation of default probability is also more commonly known as credit scoring.

Credit risk is the risk that the borrower may be unable or unwilling to honor his obligations under the terms of the contract for credit. A major part of the assets of a bank consists of a loan portfolio. Banks suffer maximum loss due to non-performing assets. Credit risk is thus a dominant concern to manage the asset portfolio of any bank.

The first pillar (Basel II) proposes the following two approaches of credit risk and risk weights:

- Standardized approach
- Internal Rating Based (IRB) approach

The standardized approach is designed to be applicable for every bank. In this approach a portfolio of bank loans will be characterized by a relatively small number of risk categories, and the risk weight associated with a given category is based on an external rating institution's evaluation of counterpart risk.

The underlying idea of the IRB approach is to make further use of the information collected and processed in the bank's counterpart rating operation. Since banks make it a business to evaluate risks, these evaluations ought to be reasonable basis for risk-contingent capital adequacy information (Carling,Jacobson, Lindè, and Roszbach (2002)).

The major South African banks are using the IRB approach, as specified by the Accord. The basic requirement for this is to have a reliable estimate of the probability distribution of the loss for each type of loan asset.

As per the Basel II requirements, expected loss is estimated using four components:

- Probability of default (PD), which will be the focus of this study
- Loss given default (LGD)
- Exposure at default (EAD) and
- Maturity of exposures (M)

Probability of default is the likelihood that a loan will not be repaid. Loss given default is the fraction of the exposure at default that will not be recovered in the case of a default event. The exposure at default is an estimation of the extent to which a bank may be exposed to a counterparty in the event of, and at the time of, that counterparty's default. These components are calculated for the minimum of a year or the maturity of the loan.

Expected loss (EL) is not the risk, but the cost of providing credit and is calculated as follows:

$$EL = PD \times LGD \times EAD$$

Unexpected loss (UL) represents the volatility of actual loss rates that occur around EL. It is the presence of UL that creates the requirement for a capital "cushion" to ensure the viability of the bank during the year when losses are unexpectedly high. Credit risk is not the risk from expected losses from the credit but the risk of unexpected losses from credit. The entire focus of credit risk management under Basel II is to predict and manage unexpected loss (Bhatia (2006)).

Under the IRB-approach, banks have the choice of application at either of two levels of sophistication. The more advanced approach requires bank internally generated inputs on PD, LGD and EAD, whereas the simpler, foundation approach only requires the bank to provide estimates of PD.

As the basic purpose of analysis of credit risk as part of Basel II is to provide for adequate capital as a safety net against possible default, it is a method of quantifying the chance of default. Thus, it is the frequency of default and the regularity with which it occurs that matters.

If a frequency function approach for estimation of probability is assumed, the probability density function needs to be estimated based on the data to be in a position to estimate the probability of default. This involves statistical science to find an empirically valid estimate of default probabilities representative of the population under consideration (Barman (2005)). Several techniques can be used to find the estimate of the default probabilities, such as discriminant analysis, neural networks and regression techniques. This study will focus on the use of regression techniques to estimate this.

1.2 Multivariate analysis

Successful modeling of a complex data set is part science, part statistical methods, and part experience and common sense.

Multivariate analysis is all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation. It is any appropriate method of analysis when the research problem involves a single dependent variable that is presumed to be related to one or more independent variables. The objective is to predict the changes in the dependent variable, using the changes in the independent variables.

1.3 Credit scoring

Credit scoring is a method to assess risk. The meaning of credit scoring is to assign scores to the characteristics of debt and borrowers and historical default and other loss experienced as an indication of the risk level of the borrower. The aim of the credit score model is to build a single aggregated risk indicator for a set of risk factors.

Prior to the application of formal methods in banking, decisions were made on judgmental basis: the bank manager would assess the creditworthiness of an individual on the basis of personal knowledge of the applicant. This had several shortcomings, in that it was unreliable (it can change from day to day with the bank manager's mood), not replicable (another manager may make a different decision, and the reasoning behind the decisions may not be reproducible), difficult to teach, unable to handle large numbers of applicants, and, in general, subjective, with all the risks of irrational personal prejudice that that implies.

Credit scoring as a method of credit evaluation has been used for more than 50 years. The first successful application of credit scoring was in the area of credit cards. According to Anderson (2007), the first retail credit scoring model for credit cards in the US was proposed in around 1941, based on the following parameters for scoring credit card applications:

- The applicant's job/position
- The number of years spent in the current position
- The number of years spent at the current address
- Details on bank accounts and life insurance policies
- Gender
- The amount of the monthly installment

The increase in the US credit card business mandated a reduction in the decision time. In 1956, Fair, Isaac & Co. (FICO) was established to help consumer credit evaluation and in the 1960's computers were bought to process credit card applications. Anderson (2007) also notes that in 1963, Myers and Forgy proposed the application of multivariate discriminant analysis for credit scoring. In 1975, with the passing of the "US Equal Credit Opportunity Act I", credit scoring received complete acceptance.

The first UK credit card, Barclaycard, was launched in 1966. The dramatic growth in the number of credit decisions needing to be made encouraged the development and growth of automated and objective methods of making such decisions.

Legislation, like the South African Constitution, forbids the use of certain factors in the decision making process. The Constitution of the Republic of South Africa (No. 108 of 1996) states in article 2, section 9 the following:

> (3) The state may not unfairly discriminate directly or indirectly against anyone on one or more grounds, including race, gender, sex, pregnancy, marital status, ethnic or social origin, colour, sexual orientation, age, disability, religion, conscience, belief, culture, language and birth.
>
> (4) No person may unfairly discriminate directly or indirectly against anyone on one or more grounds in terms of subsection (3). National legistlation must be enacted to prevent or prohibit unfair discrimination.

This makes the use of formal methods inevitable, how else, other than having some decision-making strategy explicitly articulated, can one ensure that some prescribed factor is not being used?

The modern age of computers provided the necessary tools to implement automated, objective procedures. These procedures are based on formal statistical models of customer behaviour.

Apart from the sheer impossibility of making decisions personally today, it has also become clear that formal and objective methods yield superior decisions. At the very least, these procedures permit the decision to take into account more potential factors than a human could. Of course, automated decision making can go wrong and allowance is generally made for the system to be overridden in such cases, although overrides need to be used with caution.

Although the first application of statistical methods was mainly for application processing, it can be used in all stages of a product life cycle. It is also not just confined to credit risk, but is used to measure, understand, predict, monitor and detect virtually all aspects of customer behaviour. Statistical models used to predict these different behaviours are often referred to as scorecards.

The Credit Risk Management Cycle is split into five stages: marketing, application processing, account management, collections and recoveries (Anderson (2007)).

In the marketing stage, scorecards can be used in several ways. For direct marketing campaigns, response models are often used to predict which of the potential customers in the prospect pool is most likely to respond. Those that are least likely to respond are not contacted, and the overall marketing cost is lowered. Unfortunately, people who are most likely to respond are mostly those that are the riskiest. Therefore response models are used in conjuction with a type of marketing stage credit risk model, to only target those potential customers that are likely to be accepted for the marketed product.

Application processing involves making the accept/reject decision, as well as determining the amount of money to be loaned to the customer, the interest rate that should be charged and the repayment terms of the loan. As this process is now automated, decisions can be made almost instantly.

Behaviour scores used in the account management cycle are distinguished from
application scores because they include characteristics representing the borrower's own payment pattern on the loan. The number of times the borrower has gone delinquent, the seriousness of the delinquency, and even the point during the month when payments are typically received are all very predictive of future behaviour. They include variables related to the borrower's demonstrated willingness and ability to pay on the loan under consideration and tend to be more predictive than application scores, which, of course, are only based on data available when the loan is originated.

Behaviour scores are often used to change limits or lines of credit. If a customer has a bad behaviour score on its credit card, the bank can lower the credit limit and thereby reduce its exposure to the customer.

Behaviour scores are also used for streamlined residential mortgage refinancing programs. Customers who have shown a pattern of paying on time and have excellent behaviour scores may be eligible to refinance without having to provide the income and asset documentation that is typically required.

Behaviour scores are also used to cross-sell products to existing customers. For example, a lender may use a behaviour score generated for its mortgage portfolio to select customers for a favourable offer on a credit card. On the flip side, a bad behaviour score on one product might prevent the bank from lending to the customer on another product.

Scorecards are also used in the account management stages to help with customer retention. It is not as expensive for a bank to retain clients as it is to acquire new clients. Models are developed to predict the likelihood of attrition for clients and pro-active retention strategies can be implemented.

One of the most important uses of scorecards is in the collections stage to assist in collecting delinquent accounts (early-stage collections). Customers with poor collection scores are contacted earlier in the month and the method of contact (phone calls versus letters, etc.) may be varied with the score. Often collections strategies uses collection scores in combination with outstanding balance (or loan amount minus expected collateral values if the loans are secured) to decide whom to contact and how often.

Collection scores are also used to determine strategies for handling seriously delinquent accounts (late-stage collections). Scores that predict the likelihood a customer can recover from serious delinquency may be used to select loans for modification programs or other special treatment.

Another application of scorecards, recovery scores, is used to estimate the likelihood that all or some portion of a bad debt will be recovered. After an account has gone seriously delinquent or even been charged off, recovery scores can be generated to rank accounts by the likelihood that some of the debt will be collected. Recovery scores help lenders make sound decisions about which accounts to retain and attempt to recover on themselves and which to sell to debt collection agencies or third parties.

It is clear that credit scoring, or more particular, scorecards can affect every decision made on an account at any stage in the credit life cycle. Credit scoring therefore forms part of the decision sciences.

1.4 Brief outline of the study

In the second chapter, the steps in developing a credit scoring model will be outlined. It mainly focuses on the application of scorecards for credit risk, but as explained above, can be generalized to almost any decision in the life cycle of an account.

Chapter three will briefly outline methods other than logistic regression that can be and is used for credit scoring.

Chapter four briefly recaps normal linear regression and its basic assumptions

Chapter five is an introduction to logistic regression. It shows why linear regression cannot be used for binary outcomes as well as explain the basics of logistic regression.

Chapter six discusses multivariate/multiple logistic regression. The strength of a modeling technique lies in its ability to model many variables, of which some are on a different measurement scales. The logistic model will be expanded to the case of more than one independent variable.

For a typical credit scoring model, the possible factors may vary from as little as 20 to over 200! Different strategies and associated methods for reducing the number of predictor variables are discussed in chapter seven.

Chapter eight discusses methods and strategies that can be used to model a logistic regression.

Chapter nine presents a study that was conducted to find the best significance level for a stepwise logistic regression.

Chapter ten assumes that a model has been fit and shows procedures and methods to see how well the model describes the dependent variable.

In the previous chapters the focus was on modeling where the outcome variable is binary. In chapter eleven the model is extended to handle the cases where the outcome variable is nominal with more than two levels. It will focus on the multinomial model and only give a brief introduction to the ordinal models.

In the last chapter, an example of an actual credit scoring model fitting will be shown.

Chapter 2

Steps in credit scoring model development

Credit scoring is a mechanism used to quantify the risk factors relevant for an obligor's ability and willingness to pay. The aim of the credit score model is to build a single aggregate risk indicator for a set of risk factors. The risk indicator indicates the ordinal or cardinal credit risk level of the obligor. To obtain this, several issues needs to be addressed, and is explained in the following steps.

2.1 Step 1: Understanding the business problem

The aim of the model should be determined in this step. It should be clear what this model will be used for as this influences the decisions of which technique to use and what independent variables will be appropriate. It will also influence the choice of the dependent variable.

2.2 Step 2: Defining the dependent variable

The definition identifies events vs. non-events (0-1 dependent variable). In the credit scoring environment one will mostly focus on the prediction of default. Note that an event (default) is normally referred to as a "bad" and a non-event as a "good".

Note that the dependent variable will also be referred to as either the outcome or in traditional credit scoring the "bad" or default variable. In credit scoring, the default definition is used to describe the dependent (outcome) variable.

In order to get a default definition, it is important to define default. Default risk is the uncertainty regarding a borrower's ability to service its debts or obligations. It is quantified by measuring the probability of default (PD). PD reflects the probabilistic assessment of the likelihood that an obligor or counterparty will default on its contractual obligation within a certain period of time.

Thus, a definition of a dependent variable in a credit scoring problem is two-fold: a delinquency definition and a time period in which to reach that level of delinquency, which is generally known as the outcome period. The outcome period is thus the period of time over which the loans in the sample is observed to classify them as good or bad.

Under paragraph 452 in the Basel Accord (Basel Committee on Banking Supervision (2004), default is defined as:

> A default is considered to have occurred with regards to a particular obligor when either or both of the following two events have taken place:

- The bank considers that the obligor is unlikely to pay its credit obligations to the banking group in full, without recourse by the bank to actions such as realizing security (if held).
- The obligor is past due more than 90 days on any material credit obligation to the banking group. Overdrafts will be considered as being past due once the customer has breached an advised limit or has been advised of a limit smaller than the current outstanding.

Paragraph 453 has defined unlikely to pay as follows:

- The bank puts the credit obligation on a non-accrued status.
- The bank makes a charge-off or account specific provision, resulting from a significant perceived decline in credit quality subsequent to the bank taking on the exposure.
- The bank sells the credit obligation at a material credit-related economic loss.
- The bank consents to a distressed restructuring of the credit obligation where this is likely to result in a diminished financial obligation, caused by the material forgiveness, or postponement of principal, interest or fees (where relevant).
- The bank has filed for the obligor's bankruptcy or a similar order in respect of the obligor's credit obligation ot the banking group.
- The obligor has sought or has been placed in bankruptcy or some similar protection where this would avoid or delay payment of the credit obligation to the banking group.

Business and other restrictions should be considered here. As stated in the previous chapter, the components for the EL are calculated for a one year time frame. The Basel II definition to be more than 90 days delinquent in a year might not be applicable for a certain model. Often, the choice of outcome definition is limited by data constraints. The outcome period window should ideally be long enough to cover the period of the peak of default for the product in question. Observing performance only over the first 12 months of life, for example, for a product whose default rates do not peak until, say, 3 or 4 years, may produce a development sample that does not reflect the bad loans the scorecard is being designed to identify. Because early defaulters may have characteristics different from late defaulters, the resulting scorecard could misinterpret the importance of the characteristics exhibited by the late defaulters, who make up the largest portion of total defaulted loans. Also, a lot of accounts can cure (move out of default scenario) after being 90 days delinquent, which might not make this the most appropriate definition of an "event". Note that if the model was developed on a different default definition and the outputs of the model will be used for Basel II capital calculations, one will have to calibrate the model to give the Basel II default definition PD equivalent.

The amount of data available will also influence the definition. If data is an issue in terms of very few events, a slightly more relaxed delinquency definition might render more events and can make the model development possible.

A practice which has been common in the credit industry, is to define three classes of risk (good and bad, as stated above and indeterminate), to design the scorecard using only the extreme two classes. Indeterminants are loans that perform worse than a good loan but better than a bad loan. For example, if goods are defined as loans that have been at most one cycle delinquent and bads as those that have been three cycles delinquent, indeterminants would be those that went a maximum of two cycles delinquent. Developers delete indeterminants from a sample with the hope that eliminating gray loans will produce a scorecard that can better distinguish between goods and bads. To some, this practice seems curious and difficult to justify and speculated that the practice arose because 'default' is a woolly concept. (Hand & Henley (1997)). Other developers find this practice quite useful. Although these observations are removed from the modeling, it is again considered in the post development analysis. The use of indeterminants will mostly be the choice of the developer, dependent on the size of the percentage of the population that will fall into the particular classification. If the indeterminant part is more than 15% of the population, it is common practice to rather change the indeterminant definition, or not use indeterminants at all.

2.3 Step 3: Data, segmentation and sampling

Few credit situations are absolutely perfect for modeling. Therefore trade-offs exist between what would be ideal and what can be done. There are few data requirements that need to be considered before a model can be developed:

2.3.1 Historical lending experience

Since development of a scoring system requires the analysis of past decisions, the creditor must have offered credit in the past. Therefore, no historical data equals no scoring system.

2.3.2 Data retention

Information used to support past decisions must have been retained in a usable form in order to build a custom model. For example, the credit application and credit bureau report existing when a new applicant was evaluated would be relevant as a database for model development, but not a more recent credit report or updated application.

2.3.3 Known outcomes of past decisions

The outcomes of past decisions must be available in a quantifiable form. Account payment histories can be used to classify outcomes as good or bad loans. The level of detail of historical payment records must be examined, and data archiving and purging procedures are important. For instance, when creditors purge charge-off accounts from the records, efforts must be made to recover information on these accounts.

2.3.4 Age of decision

The decisions must have aged enough to allow appropriate measurement and classification of the outcomes. For example, accounts approved three months previously are not old enough to be accurately classified, whereas accounts approved two years ago probably are. The appropriate time will vary with the product and type of decision. At the other extreme, accounts approved 10 years ago are too old, since the relationships between their historical credit applications and credit bureau reports and their outcomes would not likely reflect current relationships. Model developers will specify a sample time frame in which decisions must have occurred if they are to be included in the development.

2.3.5 Sample size

The number of credit decisions made must have been large enough to allow an appropriate sample size. The least frequent outcome that must be predicted will often determine if a large enough sample can be obtained. Since bad accounts should be the least frequent outcome, the number of available bad accounts would be the limiting factor. In fact, sample availability may influence the sample time frame - a creditor with fewer accounts might sample decisions made from two to four years ago, while a larger creditor might only sample from two to three years ago.

Data selection is the most important step in the development process. This is also typically the step that will require the most time and effort. Having clean, accurate and appropriate data is extremely important.

The type of model being developed as well as the position in the life cycle will influence the availability of the data. This will also play a role in segmentation.

At acquisition stage, mostly external data is available and typically these models will not be very predictive but is still used for the value that it can add. These types of models are mostly used for direct marketing which is very expensive. The responsiveness to an offer is modeled to exclude the least responsive people in the prospect pool. Note that a "wrong" classification by the model is not so expensive to the company. It might mean that a responsive prospect wasn't contacted, which results in lost opportunity, but credit wasn't granted to a risky applicant that wouldn't pay the company back and thus resulting in a real monetary loss.

When applying for a credit product, an application form is filled out. This is then typically combined with data from the credit bureaus as well as information on other products held by the applicant at the financial institution. Application models are more predictive than models at the acquisition stage, but a more important decision rests on the output of these models. Firstly the decision needs to be made whether to accept or reject the applicant and then how much credit to grant.

Another problem of particular relevance in credit scoring is that, in general, only those who are accepted for credit will be followed up to find out if they really do turn out to be good or bad risks according to the default definition adopted. This implies that the data available for future models will be a biased (truncated) sample from the overall population of applicants. The design sample is the set of applicants that were classified as good risks by an earlier scorecard or a judgmental decision. Those in the "reject" region were not granted credit and hence were not followed up to determine their true risk status. This distortion of the distribution of applicants clearly has implications for the accuracy and general applicability of any new scorecard that is constructed. Attempts to compensate for this distortion in distribution, model developers use what information there is on the rejected applicants (their value on the characteristics, but not their true classes) and this is called reject inference. It describes the practice of attempting to infer the likely true class of the rejected applicants and using this information to yield a new scorecard superior to the one built on only the accepts. The higher the rejection rate, the more important the problem and the less effective the compensation.

Instead of obtaining the actual performance on a set of what normally would be rejected loans or using a surrogate for that performance based on credit accounts, other techniques for dealing with sample selection bias use statistical methods. Two such methods are the augmentation and extrapolation methods.

With the augmentation method, only accepted accounts are include in the score development sample but each is weighted by the inverse of its probability of being accepted. This probability is derived by building a second logistic regression model that includes both accepted and rejected loans, but instead of predicting which loans will be good or bad, it predicts whether an applicant will be approved or rejected. The reciprocal of the probability derived from this regression model is used as the weight in the credit scoring model. In this way the accepted applicants with a high probability of rejection (which are presumably are more like those of actual rejects) are given more weight when the application credit-scoring model is built.

The extrapolation method actually uses the rejected applicants as part of the development sample for the credit-scoring model. Typically a preliminary regression is estimated using only the accepted applicants (for which the outcome is known), the rejects are scored with that model, and the model is used to derive a good and bad probability for each. Finally, the regression is estimated again, this time using both accepted and rejected applicants. Rejects are duplicated and then weighted based on their estimated bad rate. If, for example, a rejected applicant has a 40 percent probability of being bad, it is included as a bad risk with a weight of 0.4 and as a good risk with a weight of 0.6. The theory is that this second model that includes rejects and their inferred performance as part of the development sample should be free of sample selection bias.

Improved results could be produced if information was available in the reject region- if some applicants who would normally be rejected were accepted. This would be a commercially sensible thing to do if the loss due to the increased number of delinquent accounts was compensated for by the increased accuracy in classification. A related practice, increasingly common, is to obtain information on rejected applicants from other credit suppliers (via the credit bureaus) who did grant them credit.

To reach the account management stage, an account is typically left for three to six months to build up some sort of behaviour. This behaviour on the account is then used to predict an outcome, usually over the next twelve months. Normally internal account behaviour data is enough to build a powerful model, but in some instances external data can also be used. Accounts in this stage, will be scored frequently through the model (typically monthly) and not just once as in the application stage.

Segmentation divides the population into groups and builds a separate scorecard for each. Three types of reasons for segmenting scorecards have been identified: (1) strategic, (2) operational, and (3) statistical.

A lender may want to segment its scorecard strategically to target certain customer segments, such as borrowers who already have a loan with that lender. It wants a separate scorecard for this group because it wants to treat them differently.

An operational reason for segmenting would arise where different data are available for different customer segments. For example, different loan applications may be used for applicants applying for credit in a bank branch, those phoning into a call center, or those applying through a website. This could mean that certain predictive characteristics are available for some applicant segments but not for others, necessitating segmented scorecards. Developing different models for applicants with information at the credit bureaus versus those that don't might also be a useful operational segmentation.

Finally, a statistical reason for segmentation arises when characteristics affect the outcome variable differently for some subpopulations of the general applicant population than for others. For example, applicants with other products already at the institution are less risky than applicants new to the organization, as they already had to pass a credit screening. The behaviour between these two groups is normally different, as well as the time period needed for maturity.

One way to handle this is to build a single scorecard for all applicants but incorporate an interactive effect into the regression model. An interactive effect is present when the effect of a predictor on the outcome variable varies depending on the effect of a second predictor. This is accounted for in a regression by creating a combination variable – an "interaction" – from the two predictors. To continue with the example, instead of including delinquency and the number of other products at the financial institution in the regression as predictors, a new (interactive) variable would be created combining delinquency level with number of products. Adding this variable to the regression model accounts for the fact that delinquency on a score varies with the total number of products with the institution.

There may be many other variables in the score whose affects differ depending on the total number of other accounts the applicant has. Rather than incorporating several interactive effects into a single scorecard, the applicant population might be segmented into those with other products with the financial institution and those that don't, building separate scorecards for each. Separate scorecards for these two populations are more straightforward and easier to understand than a single scorecard with several interactive variables.

Segmentation may or may not lead to a set of scorecards that is more predictive than a single scorecard, depending on the situation. The first consideration in deciding whether or not to segment is where there is a strategic, operational or statistical need to do so. If there is, the performance of the segment cards should be compared to that of a single model to see which would be most beneficial for the business.

The performance of segmented models should be significantly better than a single model before segmentation is adopted because there are drawbacks to using segmented scorecards. First, the incremental costs of using multiple scorecards can be significant- reviewing 10 sets of monitoring reports on the scorecards is considerably more time consuming than reviewing a single set. Second, there may not be enough bad loans within each segment for reliable scorecard validation. Finally, scorecard policies should be set taking into consideration the characteristics that comprise the score. Drafting and maintaining policies for 10 different scores within the same business can be both time consuming and confusing. The gain from segmentation should be substantial before a developer opts for it.

Once the bad definition and the outcome period have been specified, the relevant data collected and the population segmented, the data set for scorecard development can be created. It is standard practice to create both a development and a holdout sample. As the names imply, the development sample is used to build the scorecards while the holdout sample is used to check the accuracy of the completed scorecard on a set of loans that had no influence on generation of the point weights.

Scorecard building is a combination of art and science: The science lies in the statistical methods at the core of scorecard development. The art lies in the many choices the scorecard developer must make throughout the model building process. These choices have a major effect on the final scorecard. Uninformed or incorrect decisions can result in an important variable being excluded or an improper variable being included. Some of the choices that must be mare are how to treat data errors, missing values, and outliers (extreme values of the characteristics); whether to use continuous or transformed variables or to make categorical (binned) variables out of continuous variables; and whether to include variable interactions. Perhaps the most important, the modeler must choose which characteristics to incorporate in the scorecard.

It is extremely important for the scorecard developer to have a good grasp of the business for which the scorecard is being built, and a firm command of the business's data. Otherwise, he can make mistakes in working with the data and interpreting the results of test runs.

2.4 Step 4: Fitting of a model and optimization of selected criteria

Several techniques can be used to fit a model. Most scorecards are built by estimating a regression model. Regression models examine how a particular variable (the outcome variable) is explained by another variable- or, more typically, by a whole set of other variables. The output from a regression model is a set of factors called regression coefficients. Each of these can be interpreted as the correlation between the outcome variable one is trying to predict and the explanatory variable or characteristic, holding constant all other influences on the outcome variable.

Irrespective of the technique used to fit the model, certain criteria should be set to determine the goodness-of-fit and predictive ability of the model. These criteria indicate the scorecard's ability to perform the tasks it is intended to perform and permit comparisons of different model specifications, pointing up the strength of the scorecard when it contains one set of characteristics rather than another set. The most commonly used is the Gini index, calculated from the Lorenz curve. The different criteria will be discussed later in greater detail.

2.5 Step 5: Generalization

In order for a credit scoring model to be useful, it must be applicable to the larger population and not just on the development sample. Care must be taken not to over-fit the model to the development data. Normally, a holdout sample of the same time period is used when developing the model. After development, a completely independent sample from a different period can also be used to test if the model is predictive. Thus, the intent of the validation sample is to insure that the score is robust across different time periods.

2.6 Step 6: Ongoing monitoring

Once the model is developed and implemented, it is important that the model is monitored at regular intervals. In a developing economy, it is especially important to monitor that the model still predicts and the population hasn't changed. If the population has changed, it doesn't mean that the model doesn't predict anymore, it might just require a few changes. Monitoring the model also indicates when the predictive power of the model is below acceptable levels and when the model should be redeveloped.

Chapter 3

Credit scoring methods

The methods generally used for credit scoring are based on statistical pattern-recognition techniques. Historically, discriminant analysis and linear regression were the most widely used techniques for building scorecards. Both have the merits of being conceptually straightforward and widely available in statistical software packages. Typically the coefficients and the numerical scores of the attributes were combined to give single contributions which are added to give an overall score. Logistic regression is now probably the most used technique for credit scoring.

Other techniques which have been used in the industry include probit analysis, nonparametric smoothing methods, mathematical programming, Markov Chain Models, recursive partitioning, expert systems, genetic algorithms, neural networks and conditional independent models. If only a few characteristics are involved, with a sufficiently small number of attributes, an explicit classification table can be drawn up, showing classification to be given to each combination of attributes. In this chapter, a short review of some of the techniques/methods mentioned here will be given.

The concept of incurred cost associated with the probabilities of repayment of loans will be used to illustrate some of the methods. For simplicity, assume that the population of loans consist of two groups or classes G and B that denote loans that (after being granted) will turn out to be good or bad in the future, respectively. Good loans are repaid in full and on time. Bad loans are subject to the default definition chosen as explained in chapter 2.

Usually the class/group sizes are very different, so that for the probability that a randomly chosen customer belongs to group G, denoted as p_G, one has $p_G > p_B$. Let $\mathbf{x}$ be a vector of independent variables (also called the measurement vector) used in the process of deciding whether an applicant belongs to group G or B. Let the probability that an applicant with measurement vector $\mathbf{x}$ belongs to group G be $p(G|\mathbf{x})$, and that of B be $p(B|\mathbf{x})$. Let the probability $p(\mathbf{x}|G)$ indicate that a good applicant has measurement vector $\mathbf{x}$. Similarly, for bad applicants the probability is $p(\mathbf{x}|B)$. The task is to estimate probabilities $p(.|\mathbf{x})$ from the set of given data about applicants which turn out to be good or bad and to find a rule for how to partition the space $\mathbf{X}$ of all measurement vectors into the two groups A_G and A_B based on these probabilities, so that in A_G would be the measurement vectors of applicants who turned out to be good and vice versa.

It is usually not possible to find perfect classification as it may happen that the same vector is given by two applicants where one is good and the other is bad. Therefore it is necessary to find a rule that will minimize the cost of the bank providing credit connected with the misclassification of applicants. Let c_G denote the costs connected with misclassifying a good applicant as bad and c_B the costs connected with classifying a bad applicant as good. Usually $c_B > c_G$, because costs incurred due to misclassifying a bad customer are financially more damaging than cost associated with the former kind of error. If applicants with $\mathbf{x}$ are assigned to class G, the expected costs are $c_B p(B|\mathbf{x})$ and the expected loss for the whole sample is $c_B \sum_{\mathbf{x}\epsilon A_G} p(B|\mathbf{x})p(\mathbf{x}) + c_G \sum_{\mathbf{x}\epsilon A_B} p(G|\mathbf{x})p(\mathbf{x})$, where $p(\mathbf{x})$ is a probability that the measurement vector is equal to $\mathbf{x}$. This is minimized when, into group G, such applicants are assigned who have their group of measurement vectors

$$A_G = \{\mathbf{x}|c_B p(B|\mathbf{x}) <= c_G p(G|\mathbf{X})\}$$

which is equivalent to

$$A_G = \{\mathbf{x}|p(G|\mathbf{x}) >= \frac{c_B}{c_G + c_B}\}.$$

Without loss of generality, the misclassification costs can be normalized to $c_G + c_B = 1$. In this case, the rule for classification is to assign an applicant with $\mathbf{x}$ to class G of $p(G|\mathbf{x}) > c_B$ and otherwise to class B.

An important task is to specify the cost of lending errors and to accurately as possible specify the optimal cutoff-score for credit scoring, as banks have to choose the optimal trade-off between profitability and risk. Credit policies that are too restrictive may ensure minimal costs in terms of defaulted loans, but the opportunity costs of rejected loans may exceed potential bad debt costs and thus profit is not maxmized. Conversely, policies that are too liberal may result in high losses from bad debt.

3.1 Linear discriminant analysis

The aim of Linear Discriminant Analysis (LDA) is to classify a heterogeneous population into homogenous subsets and further the decision process on these subjects. One can assume that for each applicant there are a specific number of explanatory variables available. The idea is to look for such a linear combination of explanatory variables, which separates most subsets from each other. In a simple case of two subsets, the goal is to find the linear combination of explanatory variables, which leaves the maximum distance between means of the two subsets.

In a general case, consider the distributions $p(\mathbf{x}|G)$ and $p(\mathbf{x}|B)$ which are multivariate normal distributions with common variance. Then the above equation reduces to

$$A_G = \mathbf{x}|\sum w_i x_i > c$$

as follows from econometrics theory. Here x_i are explanatory variables and w_i are associated coefficients (weights) in the linear combination of explanatory variables. If one takes $s(\mathbf{x}) = \sum w_i x_i$ then it is possible to discriminate according to this "score" and thus to reduce the problem to only one dimension.

The need for multivariate normality is a common misconception. If the variables follow a multivariate ellipsoidal distribution (of which the normal distribution is a special case), then the linear discriminant rule is optimal (ignoring sampling variation). However, if discriminant analysis is regarded as yielding that linear combination of the variables which maximizes a particular separation criterion, then clearly it is widely applicable. The normality assumption only becomes important if significance tests are to be undertaken.

The advantages of the LDA method are that it is simple, it can be very easily estimated and it actually works very well. The disadvantage is that LDA requires normally distributed data, but the credit data are often non-normal (and categorized).

3.2 Linear regression

Ordinary linear regression has also been used for the two-class problem in credit scoring. Since regression using dummy variables for the class labels yields a linear combination of the predicting characteristics which is parallel to the discriminant function, one might also expect this method to perform reasonably. Linear regression is briefly discussed in the next chapter, and it will be explained why logistic regression is a better method to employ.

3.3 k-Nearest Neighbour Classification

The k-nearest neighbour classifier serves as an example of a non-parametric statistical approach. This technique assessess the similarities between the pattern identified in the training set and the input pattern. One chooses a metric on the space of applicants and takes the k-nearest neighbour (k-NN) of the input pattern that is nearest in some metric sense. A new applicant will be classified in the class to which the majority of the neighbours belong (in the case when the cost of misclassification is equal) or according to the rule expressed by the above equation This means that this method estimates the $p(G|\mathbf{x})$ or $p(B|\mathbf{x})$ probability by the proportion of G or B class points among the k-nearest neighbours to the point $\mathbf{x}$ to be classified.

When performing the k-NN methodology, a very important step is the choice of metric used. A commonly used metric is the standard Euclidean norm given by

$$\rho_1(\mathbf{x}, \mathbf{y}) = \left[(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})\right]^{1/2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are measurement vectors.

However, when the variables are in different units or categorized, it is necessary to use some appropriate standardization of variables as well as to select some data-dependent version of the Euclidean metric such as:

$$\rho_2(\mathbf{x}, \mathbf{y}) = \left[(\mathbf{x} - \mathbf{y})'\mathbf{A}(\mathbf{x} - \mathbf{y})\right]^{1/2}$$

where $\mathbf{A}$ is a $n \times n$ matrix with n number of variables. As matrix $\mathbf{A}$ can depend on $\mathbf{x}$, two types of metrics can be defined according to how $\mathbf{A}$ is selected: local metrics are those where $\mathbf{A}$ depends on $\mathbf{x}$; global metrics are those where $\mathbf{A}$ is independent of $\mathbf{x}$.

The choice of the number of nearest neighbours to be chosen (k) determines the bias/variance trade-off in the estimator. The k has to be smaller than the smallest class. In problems where there are two unbalanced classes, the fact that k is finite (and thus asymptotic properties do not hold) results in a non-monotonic relationship between the k and the proportion of each class correctly classified. That means, in general, that a larger k may not yield better performance than a smaller k. For example, if the number of points from the smallest class is less than $(1 - c_B)^{-1}$, the best classification rule for predicting class G membership is to use $k = 1$.

There is a lack of a formal framework for choosing the k and the method can only make discrete predictions by reporting the relative frequencies which have no probabilistic interpretation. These difficulties can possibly be overcome by using a Bayesian approach, which integrates over the choice of k. Such approach leads to the conclusion that marginal predictions are given as proper probabilities.

An advantage of this method is that the non-parametric nature of this method enables modeling of irregularities in the risk function over the feature space. The k-NN method has been found to perform better than other non-parametric methods such as kernel methods when the data are multidimensional. It is a fairly intuitive procedure and as such it could be easily explained to business managers who would need to approve its implementation. It can also be used dynamically by adding applicants when their class becomes known and deleting old applicants to overcome problems with changes in the population over time. Despite this, nearest neighbour models have not been widely adopted in the credit scoring industry. One reason for this is the perceived computational demand: not only must the design set be stored, but also the nearest few cases among maybe $100,000$ design set elements must be found to classify each applicant.

3.4 Classification and Regression Trees (CART)

Classification and regression trees (CART) (Lewis (2000)) method is a non-parametric method. It is a flexible and potent technique; however, it is used in credit scoring practice chiefly only as a supporting tool to accompany parametric estimation methods. It serves, for example, in the process to select characteristics with the highest explanatory power. The CART method employs binary trees and classifies a data set into a finute number of classes. It was originally developed as an instrument for dealing with binary responses and as such it is suitable for use in credit scoring where the default and non-default responses are contained in the data. In most general terms, the purpose of analysis using tree-building methods is to determine a set of if-then logical split conditions that permit accurate prediction or classification of cases.

CART analysis is a form of binary recursive partitioning. The term "binary" implies that each group of observations, represented by a node in a decision tree, can be only split into two groups. Thus, each node can be split into two child nodes, in which case the original node is called the parent node. The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes themselves can be split, forming additional children. The term "partitioning" refers to the fact that the dataset is split into sections or partitioned.

Similarly as with the methodologies reviewed earlier, one makes the assumption of having a training set of measurement vectors $\mathbf{x}_T = \{\mathbf{x}^i\}$ along with information whether an individual j defaulted or not (y_j is coded 1 or 0 respectively). The CART tree consists of several layers of nodes: the first layer consists of a root node; the last layer consists of leaf nodes. Because it is a binary tree, each node (except the leaves) is connected to two nodes in the next layer. The root node contains the entire training set; the other nodes contain subsets of this set. At each node, the subset is divided into 2 disjoint groups based on one specific characteristic x_i from the measurement vector. If x_i is ordinal, the split results from the fact, related to a particular individual, as to whether $x_i > c$, for some constant c. If the previous statement is true, an individual j is classified into the right node; if not, an individual is classified into the left node. A similar rule applies, if x_i is a categorized variable.

The characteristic x_i is chosen among all possible characteristics and the constant c is chosen so that the resulting sub-samples are as homogenous in y as possible. In other words: x_i and c are chosen to minimize the diversity of resulting sub-samples. The classification process is a recursive procedure that starts at the root node and at each further node (with exception of the leaves) one single characteristic and a splitting rule (or constant c) are selected. First, the best split is found for each characteristic. Then, among these characteristics the one with the best split is chosen.

The result of this procedure is that sub-samples are more homogenous than the parent sample. The procedure ends when the node contains only individuals with the same y_j or it is not possible to decrease diversity further. For illustration purposes, let $p(.|t)$ be the proportion of G and B groups present at node t. The Gini index function can be used as an example of the diversity functions, which is defined as $d(t) = p(G|t)p(B|t)$. The value of constant c that is instrumental to splitting between nodes (i.e. nodes to which the node t is parental). Formally, the aim is to choose c minimizes $p_L d(t_L) + p_R d(t_R)$, where p_L and p_R are the proportions of individuals going into nodes t_L and t_R respectively. The complete tree is usually very large but algorithms exist for pruning it into a simpler, final tree. The most common and efficient pruning methods are based on the fact that if one tries to select a sub-tree of the maximal tree that minimizes the misclassification costs, a large number of trees yield approximately the same estimated misclassification costs. Therefore it is reasonable to stop the search for the best pruned tree once a sub-tree with similar misclassification costs to the maximal tree is found. Non-linearities and characteristics can be included in what is superficially a linear model.

The advantages of the CART method in credit scoring are that it is very intuitive, easy to explain to management and it is able to deal with missing observations. The interpretation of results (in most cases) summarized in a tree is very simple. This simplicity is useful not only for purposes of rapid classification of new observations, but can also often yield a much simpler "model" for explaining why observations are classified in a particular manner. As the final results of using tree methods for classification can be summarized in a series of logical if-then conditions, there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear, follow some specific non-linear function or that they are even monotonic in nature. CART can handle numerical data that are highly skewed or multi-modal, as well as catergorical predictors with either ordinal or non-ordinal structures. Tree methods are particularly well suited for data mining tasks, where there are often little prior knowledge or any coherent set of theories or predictions regarding which variables are related and how. Tree methods can often reveal simple relationships between just a few variables that could have easily gone unnoticed using other analytic techniques.

The major disadvantage is the computational burden in case of large datasets since at each node every characteristic has to be examined. Very often the resulting tree is quite large so that the process of model-learning becomes too time consuming. Some emperical studies also note that often the trees are not stable since small changes in a training set may considerably alter the structure of the whole tree. A significant problem is also the fact that CART optimizes only locally on a single variable at a time and thus may not minimize the overall costs of misclassification.

3.5 CHAID Analysis

CHAID (Hoare (2004)), in one or other of its many forms, is a great way to sift certain kinds of data to find out where interesting relationships are buried, especially when the relationships are more complex than the linear or at least monotonic ones usually sought.

The acronym CHAID stands for Chi-squared Automated Interaction Detector. Although it can be used for regression problems, it is mostly used to build classification trees. The Chi- squared part of the name arises because the technique essentially involves automatically constructing many cross-tabulations and working out statistical significance in the proportions. The most significant relationships are used to control the structure of a tree diagram.

Because the goal of classification trees is to predict or explain responses on a categorical dependent variable, the technique has much in common with the techniques used in the more traditional methods such as discriminant analysis. The flexibility of classification trees makes them a very attractive analysis option, but it is not to say that their use is recommended to the exclusion of more traditional methods. When the typically more stringent theoretical and distributional assumptions of more traditional methods are met, the traditional methods may be preferable. As an exploratory technique, or as a technique of last resort when traditional methods fail, classification trees are, in the opinion of many, unsurpassed.

CHAID will "build" non-binary trees (i.e. trees where more than two branches can attach to a single root or node), based on a relatively simple algorithm that is particularly well suited for the analysis of larger datasets. Both CHAID and CART techniques construct trees, where each (non-terminal) node identifies a split condition, to yield optimum prediction or classification.

The basic algorithm used to contruct the tree relies on the Chi-square test to determine the best next split at each step. Specifically, the algorithm proceeds as follows:

3.5.1 Preparing predictors

The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distribution into a number of categories with an approximately equal number of observations. For categorical predictors, the categories (classes) are "naturally" defined.

3.5.2 Merging categories

The next step is to cycle through the predictors to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable by calculating a Pearson chi-square. If the test for a given pair of predictor categories is not statistically significant as defined by an alpha-to-merge value, then it will merge the respective predictor categories and repeat this step (i.e. find the next pair of categories, which may now include previously merged categories). If the statistical significance for the respective pair of predictor categories is significant (less than the respective alpha-to-merge value), then (optionally) it will compute a Bonferroni adjusted p-value for the set of categories for the respective predictor.

3.5.3 Selecting the split variable

The next step is to choose the split predictor variable with the smallest adjusted p-value, i.e. the predictor variable that will yield the most significant split. If the smallest Bonferroni adjusted p-value for any predictor is greated than the alpha-to-split value, no further splits will be performend and the respective node is a terminal node.

Continue this process until no further splits can be performed given the alpha-to-merge and alpha-to-split values.

3.5.4 Exhaustive CHAID algorithms

A modification to the basic CHAID algorithm, called Exhaustive CHAID, performs a more thorough merging and testing of predictor variables, and hence requires more computing time. Specifically, the merging of categories continues (without reference to any alpha-to-merge value) until only two categories remain for each predictor. The algorithm then proceeds as explained above and selects among the predictors the one that yields the most significant split.

The advantages of the CHAID method is that it is simple to use and easy to explain. A general issue that arises is that the final trees can become very large. When the input data are complex and contain many different categories for classification and many possible predictors for performing the classification, the resulting trees can become very large and presenting the trees in an easily accessible method becomes a problem. Another major critism of CHAID is that it supresses the real underlying structure of the data. Variables entered at an early stage may become insignificant, but no test is done to remove variables already in the tree.

3.6 Neural networks

A neural network (NNW) is a mathematical representation inspired by the human brain and its ability to adapt on the basis of the inflow of new information. Mathematically, NNW is a non-linear optimization tool.

The NNW design called multilayer perceptron (MLP) is especially suitable for classification. The network consists of one input layer, one or more hidden layers and one output layer, each consisting of several neurons. Each neuron processes its inputs and generates one output value that is transmitted to the neurons in the subsequent layer. Each neuron in the input layer (indexed $i = 1, 2, ..., n$) delivers the value of one predictor (or the characteristics) from vector $\mathbf{x}$. When considering default/non-default discrimination, one output neuron is satisfactory.

In each layer, the signal propagation is accomplished as follows. First, a weighted sum of inputs is calculated at each neuron: the output value of each neuron in the proceeding network layer times the respective weight of the connection with that neuron. A transfer function $g(\mathbf{x})$ is then applied to this weighted sum to determine the neuron's output value. So, each neuron in the hidden layer (indexed $j = 1, ..., g$) produces the so-called activation

$$a_j = g(\sum_i w_{ij} x_i).$$

The neurons in the output layer (indexed $k = 1, ..., m$) behave in a manner similar to the neurons of the hidden layer to produce the output of the network:

$$y_k = f(\sum_j w_{jk} a_j) = f(\sum_j w_{jk} g(\sum_i w_{ij} x_i))$$

where w_{ij} and w_{jk} are weights.

The Sigmoid (or logistic) function $f(x) = \frac{1}{1+e^x}$ or hyperbolic tangent function $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is usually employed in the above network output for functions f and g. The logistic function is appropriate in the output layer if one has a binary classification problem, as in credit scoring, so that the output can be considered as default probability. According to the theory, the NNW structure with a single hidden layer is able to approximate any continuous bounded integrable function arbitrarily accurately.

There are two stages of optimization. First, weights have to be initialized, and second, a nonlinear optimization scheme is implemented. In the first stage, the weights are usually initialized with a small random number. The second stage is called the learning or training of the NNW. The most popular algorithm for training multilayer perceptrons is the back-propagation algorithm. As the name suggests, the error computed from the output layer is back-propagated through the network, and the weights are modified according to their contribution to the error function. Essentially, back-propagation performs a local gradient search, and hence its implementation; although not computationally demanding, it does not guarantee reaching a global minimum. For each individual, weights are modified in such a way that the error computed from the output layer is minimized.

NNWs were described in the 1960s but their first use in the credit-scoring-related literature appears only at the beginning of the 1990s. Research suggested that logistic regression is a good alternative to neural models.

The major drawback of NNWs is their lack of explanation capability. While they can achieve a high prediction accuracy rate, the reasoning behind why and how the decision was reached is not available. For example, in a case of a denied loan it is not possible to determine which characteristic(s) was exactly the key one(s) to prompt rejection of the application.

The National Credit Act of 2005 states in article 62 the following:

"62. (1) On request from a consumer, a credit provider must advise that consumer in writing of the dominant reason for –

(a) refusing to enter into a credit agreement with that consumer;

(b) offering that consumer a lower credit limit under a credit facility than applied for by the consumer, or reducing the credit limit under an existing credit facility;

(c) refusing a request from the consumer to increase a credit limit under an existing credit facility; or

(d) refusing to renew an expiring credit card or similar renewable credit facility with that consumer"

Consequently, it is almost impossible to use a neural network for a credit scoring model for application decisioning and limit management and still comply with the National Credit Act.

3.7 Which method is best?

In general there is no overall "best" method. What is the best will depend on the details of the problem, the data structure, the characteristics used, the extent to which it is possible to separate the classes by using those characteristics and the objective of the classification (overall misclassification rate, cost-weighted misclassification rate, bad risk rate among those accepted, some measure of profitability, etc.) The various methods are often very comparable in results. This fact can be partly explained by the mathematical relationships between these models: for example, the NNW can be seen as a generalization of the logit method. Often, there is no superior method for diverse data sets.

If the classes are not well separated, then $p(G|\mathbf{x})$ is a rather flat function, so that the decision surface separating the classes will not be accurately estimated. In such circumstances, highly flexible methods such as neural networks and nearest neighbour methods are vulnerable to over-fitting the design data and considerable smoothing must be used (e.g. a very large value for k, the number of nearest neighbours).

Classification accuracy, however measured, is only one aspect of performance. Others include the speed of classification, the speed with which a scorecard can be revised and the ease of understanding of the classification method and why it has reached its conclusion. As far as the speed of classification goes, an instant decision is much more appealing to a potential borrower than is having to wait for several days. Instant offers can substantially reduce the attrition rate. Robustness to population drift is attractive-and, when it failes, an ability to revise a scorecard rapidly (and cheaply) is important.

Classification methods which are easy to understand (such as regression, nearest neighbours and tree-based methods) are much more appealing, both to users and to clients, than are methods which are essentially black boxes (as neural networks). They also permit more ready explanations of the sorts of reasons why the methods have reached their decisions.

Neural networks are well suited to situations where there is a poor understanding of the data structure. In fact, neural networks can be regarded as systems which combine automatic feature extraction with the classification process, i.e. they decide how to combine and transform the raw characteristics in the data, as well as yielding estimates of the parameters of the decision surface. This means that such methods can be used immediately, without a deep grasp of the problem. In general, however, if one has a good understanding of the data and the problem, then methods which makes use of this understanding might be expected to perform better. In credit scoring, where people have been constructing scorecards on similar data for several decades, there is a solid understanding. This might go some way towards explaining why neural networks have not been adopted as regular production systems in this sector, despite the fact that banks have been experimenting with them for several years.

Because there is such a good understanding of the problem domain, it is very unlikely that new classification methodologies will lead to other than a tiny improvement in classification accuracy. Significant improvements are more likely to come from including new, more predictive characteristics.

The logit method is the most favoured method in practice, mainly due to (almost) no assumptions imposed on variables, with the exception of missing values and multicollinearity among variables. Contrary to this, non-parametric methods can deal with missing values and multicollinearity (or correlations) among variables, but often are computationally demanding. The rules that are constructed on the basis of some of these methods can be hard to explain to a business manager as well as a clients however. Logistic regression produce models that are easy to explain and implement and ha s been widely accepted in the banking industry as the method of choice. The focus of the rest of the study will therefore be logistic regression's application in credit scoring.

Chapter 4

Linear regression and its assumptions

Linear regression is a statistical technique that can be used to analyze the relationship between a single dependant variable and one or more independent (predictor) variables. This chapter provides an overview of the basics of linear regression and in the next chapter it will be shown why it is not applicable for a binary outcome variable, and thus for credit scoring.

4.1 Simple linear regression

Let there be a data sample, of which one wants to predict the value of one of the variables (the dependent variable). If no independent variables were taken into account, there are several descriptive measures that one can use to predict the value of the dependent variable, such as the average, median or mode. The only question is then to determine how accurate each of these measures is in predicting. The sum of squares of errors (SSE) can be used as an indication of the accuracy. The objective is to obtain the smallest SSE, because this would mean that the predictions are most accurate.

For a single set of observations, the mean will produce the smallest sum of squared errors than any other measure of centrality, e.g. the median, mode or any other more sophisticated statistical measure. Thus, prediction using the mean is used as a baseline for comparison as it represents the best prediction without using independent variables.

Example 1 *Use an example of the number of credit cards a family holds:*

Observation	*Number of credit cards*
1	3
2	6
3	6
4	5
5	8
6	7
7	9
8	10

Using these 8 observations, the mean, the median and the mode is calculated:

Mean	6.75
Median	6.50
Mode	6.00

Now, calculate the squared prediction errors, using each of these measures:

Observation	*Num of CC*	*Mean (err)*^{2}	*Median (err)*^{2}	*Mode (err)*^{2}
1	3	14.0625	12.25	9
2	6	0.5625	0.25	0
3	6	0.5925	0.25	0
4	5	3.0625	2.25	1
5	8	1.5625	2.25	4
6	7	0.0625	0.25	1
7	9	5.0625	6.25	9
8	10	10.5623	12.25	16
	SUM	**35.5**	**36**	**40**

The mean has the lowest prediction error of these three measures.

In simple linear regression, the aim is to predict the dependent variable, using a single independent variable.

Without taking into account the independent variable, one can predict with the highest level of accuracy that:

$$\widehat{y} = \overline{y}.$$

By using other information, one could try and improve the predictions by reducing the prediction errors, i.e. the sum of squared errors. To do so, a variable is needed that can be associated with (is correlated to) the independent variable. Note that the conditional mean will now be estimated, given the correlated variable. The correlation coefficient is therefore fundamental to regression analysis and describes the relationship between two variables.

Two variables are said to be correlated if changes in the one variable are associated with changes in the other variable. In this way, as one variable changes, one would know how the other variable changes. It is also often found that the prediction can be improved by adding a constant value.

Example 2 *Following the previous example, a predictor variable, family size is now available. Now:*
Predicted number of credit cards = (Change in the number of credit cards held associated with family size)×family size.

$$\widehat{y} = b_1 x_1$$

If a constant value is added in this example, the constant will represent the number of credit cards held by a family, irrespective of its size. The equation now becomes:

$$\widehat{y} = b_0 + b_1 x_1.$$

The regression equation as above, can also be expressed in terms of the conditional mean. The key quantity in any regression problem is the mean value of the outcome variable, given the value of the independent variable. Therefore, the above equation can also be expressed in terms of the conditional mean of y, given x:

$$E(y|x) = \beta_0 + \beta_1 x_1$$

where b_0 and b_1 are estimates of β_0 and β_1.

If the constant term (intercept) does not help to make a better prediction, the constant term will be removed.

The terms β_0 and β_1 are called the regression coefficients. Their values are found by minimizing the SSE:

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

This gives the following results for $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

and

$$\widehat{\beta}_1 = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}.$$

This can also be written as:

$$\widehat{\beta}_1 = \frac{\sum (x_i - x)(y_i - y)}{\sum (x_i - \overline{x})^2}.$$

Example 3 *Returning to the earlier example around credit cards, assume that family size is available as a predictor (independent variable).*

Observation	*Number of credit cards*	*Family size*
1	3	2
2	6	3
3	6	4
4	5	4
5	8	5
6	7	5
7	9	6
8	10	6

Now, solving β_0 and β_1 with the equations derived above, the following is obtained:
Number of credit cards=0.2072+1.4955×Family size

This equation results in the following prediction error (SSE):

Obs	*Number of credit cards*	*Family size*	*Prediction*	*Squared error*
1	3	2	3.1982	0.0392832
2	6	3	4.6937	1.70641969
3	6	4	6.1892	0.03579664
4	5	4	6.1892	1.41419664
5	8	5	7.6847	0.09941409
6	7	5	7.6847	0.46881409
7	9	6	9.1802	0.03247204
8	10	6	9.1802	0.67207204

Summing over the squared prediction errors for all the observations gives:

$$SSE = 4.46846847$$

The same criterion was used as in the earlier example, minimizing the squared prediction error. It is clear that knowledge of family size improved the predicted number of credit cards held, compared to when only the arithmetic mean was used. The SSE has decreased from 35.5 to 4.5, indicating that the regression is better than just the average.

When interpreting the regression equation, it is clear that for each additional family member, the number of credit cards held increases by 1.5.

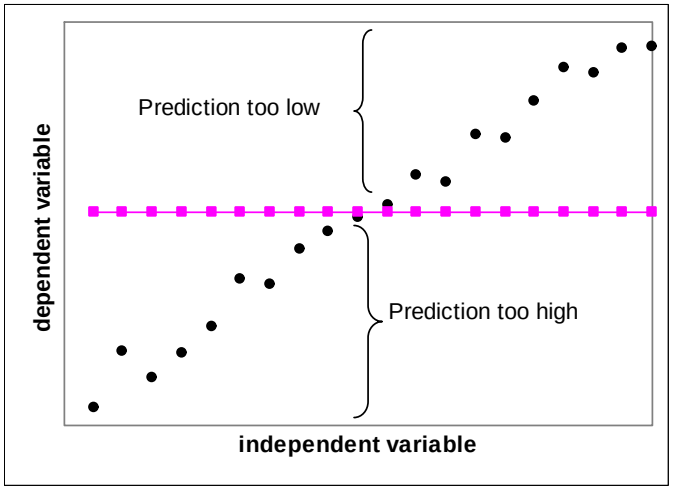
4.2 Assumptions of linear regression

The validity of the results in the previous section is dependent on whether the assumptions of regression analysis have been met. The question one asks is whether the error in prediction is a result of an actual absence of a relationship among the variables or are they caused by some characteristics of the data not accommodated by the regression model. The most used measure of prediction error is the residual, which is the difference between observed and predicted value.

The first and foremost assumption is that of correct functional form- is there a linear relationship between the independent variable(s) and the dependent variable? If the functional form is incorrect, the residual plots constructed by using the model will often display a pattern. This pattern can then give an indication of what a more appropriate model would be.

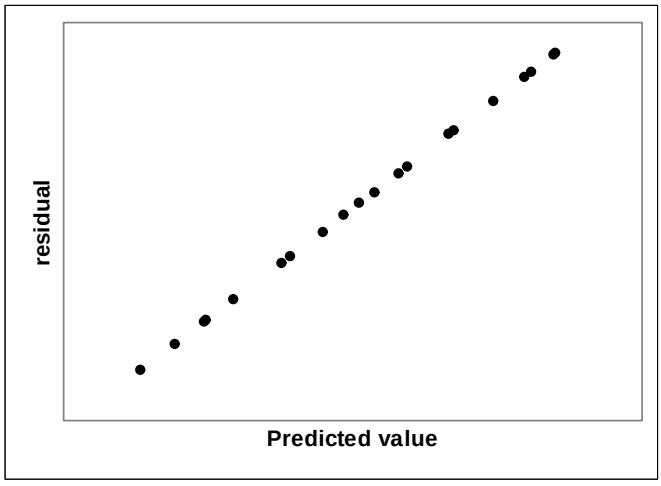
The concept of correlation is based on a linear relationship, making it a critical issue in regression analysis. Say there is dependent variable y and an independent variable x and a positive linear relationship exists between them. Suppose that, instead of using a simple linear regression model, the mean is used as the predictor.

This can be shown by the following graph:



Estimation using mean

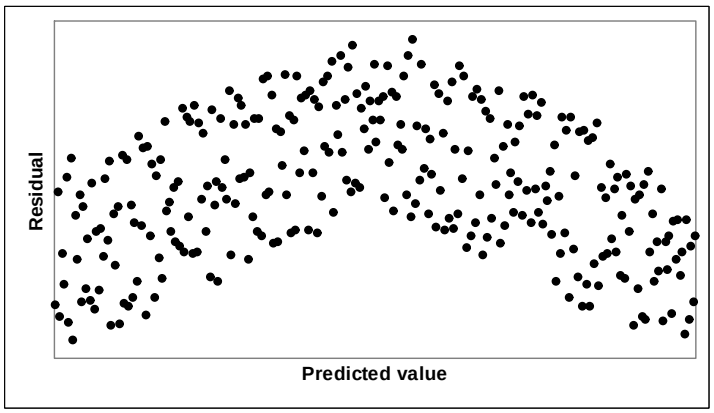
This situation will lead to the following residual plot:



Residual using mean

This clearly indicates that using the mean will not account for the linear relationship between the dependent and independent variables.

Residual plots can also be indicative of a nonlinear relationship that is not represented by the fitted linear model, for example:



Non-linear relationship

Any consistent non-linear pattern in the residuals indicates that corrective action will increase both the predictive accuracy of the model and the validity of the estimated coefficients.

Note that in multivariate regression with more than one independent variable, the residual plot shows the combined effects of all the independent variables. To examine a single predictor variable, partial regression plots are constructed, which show the relationship of a single independent variable to the dependent variable.

The scatter plot of points depicts the partial correlation between the two variables, with the effects of the other independent variables held constant. This is particularly helpful in assessing the form of a relationship (linear vs. nonlinear). Curvilinear patterns of the residuals are looked for, indicating a nonlinear relationship between a specific independent variable and the dependent variable. This is the most useful method when several independent variables are used as one can tell which specific variables violates the assumption of linearity and apply corrective measures to them. Also, the outliers and/or influential values can also be identified, one independent variable at a time.

In addition to the above assumption of the correct functional form, the validity of the inference around the regression (confidence intervals, prediction intervals, hypothesis tests etc.) are dependent on a further 3 assumptions, called the inference assumptions.

The inference assumptions of a regression model:

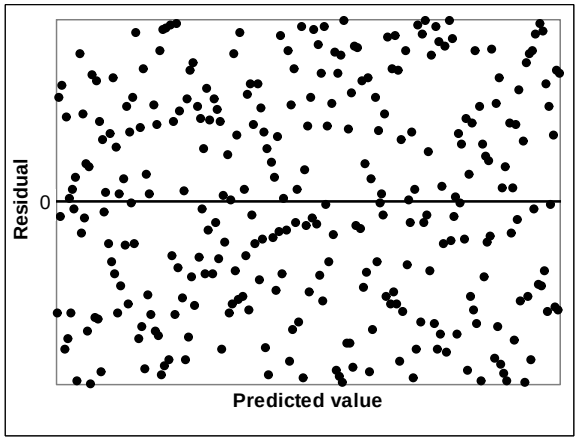
1. Constant variance of the error terms
2. Independence of the error terms
3. Normality of the error term distribution

4.2.1 Constant variance of the error terms

For any value of the independent variable, the corresponding population of potential values of the dependent variable has a variance σ^2 which is not dependent on the value of the independent variable. Equivalently, the different populations of possible error terms corresponding to the different values of the independent variable have equal variances.

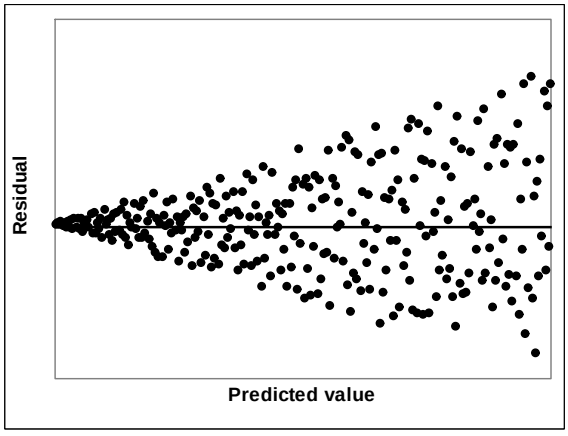
The presence of unequal variances (heteroscedasticity) is one of the most common assumption violations. Residual plots are used to assess the validity of the constant variance assumption. Residuals are plotted against the predicted values and clear patterns are looked for.

The null plot is preferable, as it indicates that the assumption of constant variance is met. There seems to be no pattern in the plot- the residuals are distributed randomly across the predicted values.



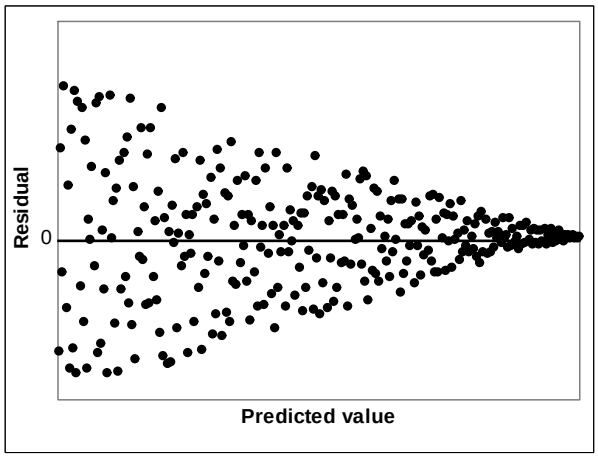
Null plot

Perhaps the most common shape is the triangle-shape, either fanning out or funneling in.



Residuals fanning out

This graph clearly indicates that the error variance increases with increasing values of the criterion.

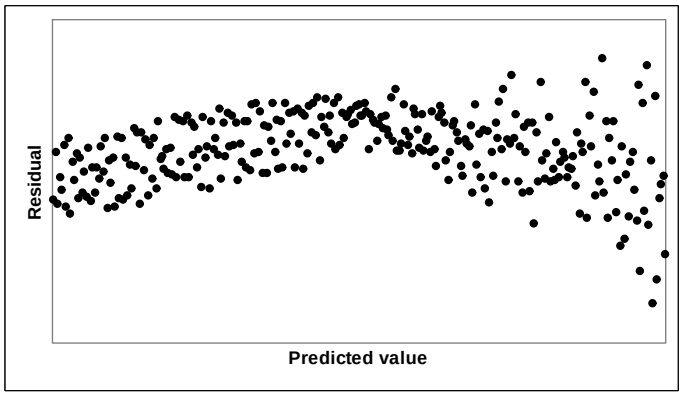


Residuals funneling in

This graph also shows non-constant variance of the error term as the error variance decreases with increasing values of the predicted value.

If heteroscedasticity is present remedies are available. Once the independent variable(s) that causes the heteroscedasticity is identified, the easiest and most direct corrective measure is to apply a variance-stabilizing transformation that allows the transformed variable(s) to be used directly in the regression model.

Many times, more than one violation can occur at the same time, such as non-linearity and heteroscedasticity.



Non-linear and heteroscedasticity

Remedies for one of the violations often correct problems in other areas as well.

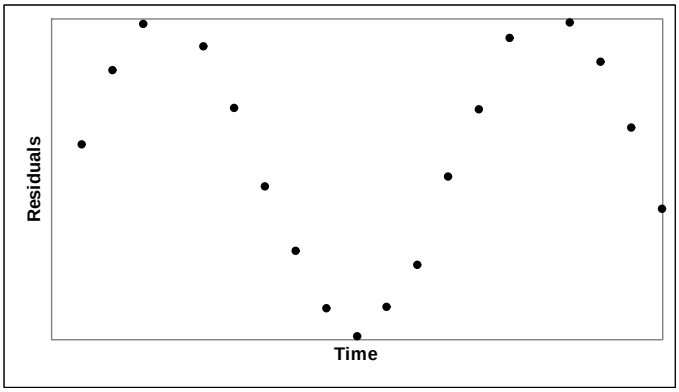
4.2.2 Independence of error terms

Any one value of the dependent variable is statistically independent of any other value of the dependent variable. Equivalently, any one value of the error term is statistically independent of any other value of the error term.

In regression, the assumption is that the predicted values are independent, in other words not related to any other prediction and not sequenced by any variable. This assumption is therefore most likely violated when the data being used in a regression problem is time series data (data collected in a time sequence).

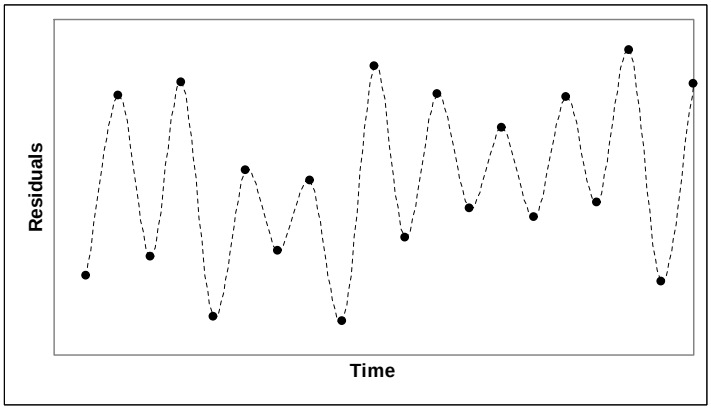
One of the ways to identify if this assumption is violated is to plot the residuals against any possible sequencing variable (for example, time). If the assumption is not violated, the pattern should appear random and similar to the null plot of residuals. Violations will be identified by consistent patterns in the residuals.

If the data has been collected in a time sequence, the time-ordered error terms might be autocorrelated. If a positive (negative) error term in time period t is followed by another positive (negative) error term in time $t + k$, the error terms have positive autocorrelation. The residuals will have a cyclical pattern over time.



Positive autocorrelation

Negative autocorrelation exists when positive error terms tend to be followed by negative error terms and negative error terms tend to be followed by positive error terms. The residuals will have an alternating pattern over time.



Negative autocorrelation

One type of positive or negative autocorrelation in the error term is called first-order autocorrelation, which says that the error term in time period t is related to the error term in time period $t - 1$. The Durban-Watson test can be used to test for first-order positive or negative autocorrelation.

4.2.3 Normality of the error term distribution

For any value of the independent variable, the corresponding population of potential values of the dependent variable has a normal distribution. Equivalently, for any value of the independent variable, the corresponding population of potential error terms has a normal distribution.

Important to note before checking this assumption, is that violations of the other two inference assumptions as well as an incorrect functional form, can often cause that the error term distribution to not be normal. It is therefore usually a good idea to use residual plots to check for incorrect functional form, non-constant error variance, and positive or negative autocorrelation before attempting to validate the normality assumption.

Several methods can be employed to validate the normality assumption. One can construct a bar chart or histogram of the residuals. If the assumption holds, the histogram of the residuals should look relatively bell-shaped and symmetric around zero. Note that one is looking for pronounced, rather than subtle deviations from the normality assumption, as small deviations do not hinder the statistical inferences. If one wishes to perform a formal test for normality, the chi-square goodness-of-fit test as well as the Kolmogorov-Smirnov test can be used.

Another method is to use standardized residuals, which is calculated by dividing the residuals by the standard error. One of the properties of the normal distribution is that 68.26% of the values in a normal distribution are within 1 standard deviation from the mean and 95.44% of the values in a normal distribution are within 2 standard deviations from the mean. Now, if the constant variance assumption holds (and remember the mean for the error terms is zero) one can say the normality assumption is met if about 68% of the standardized residuals lie between -1 and 1 and about 95% of the standardized residuals lie between -2 and 2.

Another graphical technique is to construct a normal plot. The method requires the residuals to be arranged from smallest to largest. Let $e_{(i)}$ be the i-th residual in the ordered list. Plot $e_{(i)}$ on the vertical axis against the point $z_{(i)}$ on the horizontal axis. Here $z_{(i)}$ is the point on the standard normal curve so that the area under this curve to the left of $z_{(i)}$ is$\frac{(3i-1)}{(3n+1)}$. If the normality assumption holds, this plot should have a straight line appearance.

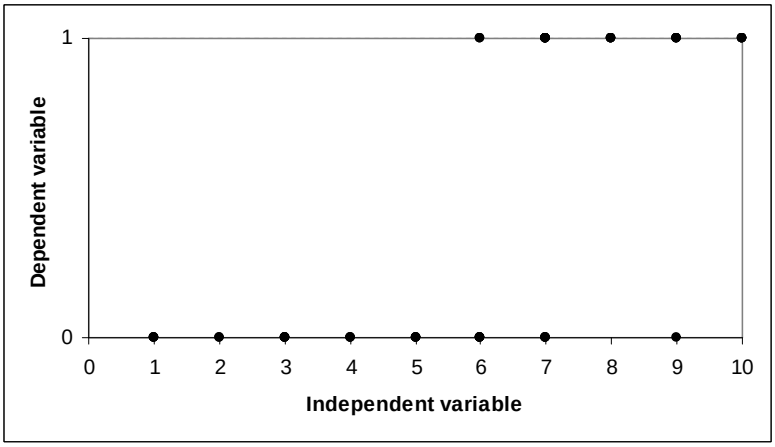
If any of the assumptions are seriously violated, remedies should be employed. These remedies are beyond the scope of this study.

Chapter 5

Simple logistic regression

The goal of an analysis using logistic regression is the same as that of any model-building technique used in statistics - to find the best fitting and most parsimonious, yet reasonable model to describe a relationship between a dependent and one or more independent variables. The primary difference between normal regression and logistic regression is the use of a binary or dichotomous dependent variable. This type of variable is called a Bernoulli variable. The use of normal linear regression for this type of dependent variable would violate several of the assumptions.

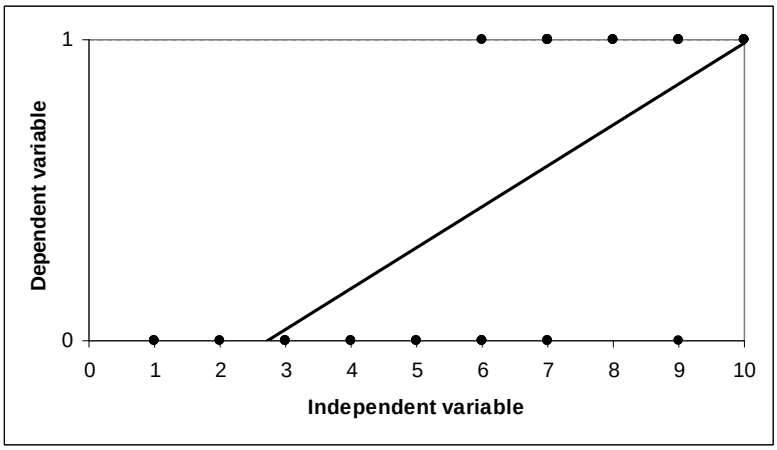
An extreme case is the following described. Note that each dot on the graph represents a data point.



Binary outcome variable

If the dependent variable can only be a zero or a 1, the mean of the distribution of the dependent variable is equal to the proportion of 1's in the distribution. This is also the probability of having outcome of 1. The mean of a binary distribution is denoted as p, which is the proportion of ones. The proportion of zeros is then $(1-p)$, which is also denoted as q. The variance of such distribution is pq.

Fitting the least squares line to this data gives the following:



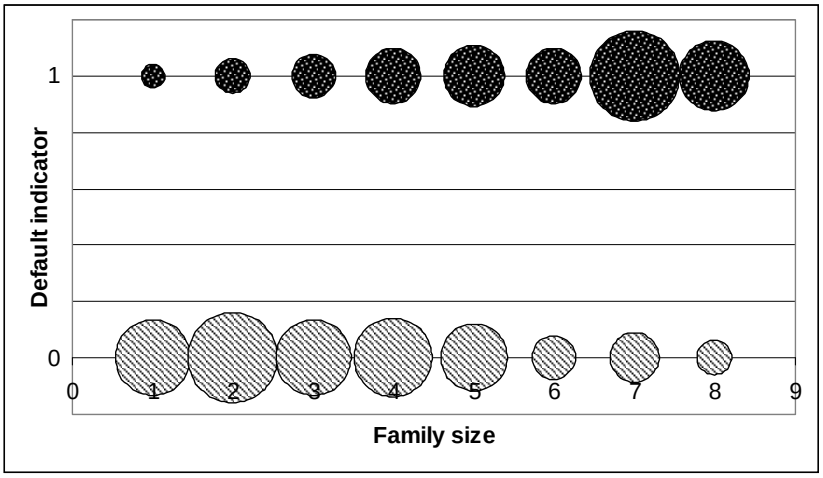
Fitting a straight line

It is quite clear from this illustration that the fundamental assumption of linear regression (i.e. correct functional form) is already violated. The following are also reasons why linear regression cannot be used:

- If one uses linear regression, the predicted values will become greater than one or less than zero, if moved far enough on the x-axis. These values are theoretically inadmissable.
- Another assumption that is violated is that of constant variance of the dependent variable across the independent variables. If the dependent variable is binary, the variance is pq. Now, if 50% of the observed sample has 1 as dependent variable, the variance is at its maximum of $pq = 0.25$. As one moves to the extremes, the variance decreases. If $p = 0.1$, the variance is $0.1 \times 0.9 = 0.09$ so as p approaches 0 or 1, the variance approaches zero.
- The significance tests of the parameters rest upon the assumption that the errors of the prediction (residuals) are normally distributed. As the dependent variable takes the values of 1 and 0 only, this assumption is very hard to justify, even just approximately.

Example 4 *Using the credit card concept again, let the dependent variable be a 1 if the family defaults on its credit card (event occurred) and 0 if the family does not default on its credit card. As an independent variable, family size is used again.*

Plotting this on a graph gives the following (where the size of the circle is an indication of the number of observations per family size). Note now that, since there is more than one observation per family size, the dependent variable is now a binomial variable.



Family size and default

As discussed earlier, fitting a simple linear regression line using least squares will violate several of the assumptions of linear regression. Yet, looking at the above graph, there seems to be a relationship between the family size and whether the client defaults or not. As the family size increases, it appears that the proportion of clients that default on their credit cards increases.

In linear regression it is assumed that the conditional mean of the dependent variable (y), given the independent variable (x) can be expressed as an equation linear in x.

$$E(y|x) = \beta_0 + \beta_1 x.$$

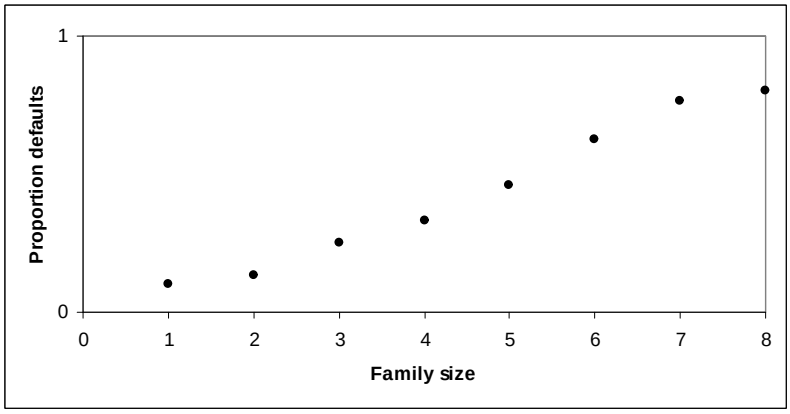
This expression implies that it is possible for $E(y|x)$ to take any value as x ranges between $-\infty$ and ∞. With a binary dependent variable, this conditional mean must be greater or equal to 0 and less than or equal to 1:

$$0 \leq E(y|x) \leq 1.$$

Example 5 *In order to have the conditional mean bounded, the proportion of defaults at each family size is taken. The proportion is now an estimate of this conditional mean.*

Thus, for the example:

	Default indicator		
Family size	*0*	*1*	*Proportion*
1	9	1	0.1
2	13	2	0.13333
3	9	3	0.25
4	10	5	0.33333
5	7	6	0.4615
6	3	5	0.625
7	4	13	0.7647
8	2	8	0.8



S- curve

It can be seen from the graph that this proportion approaches 0 and 1 gradually. This type of curve is said to be S-shaped.

The curve as in the above example, is defined by the equation

$$p = \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

where $\pi(x)$ represents $E(y|x)$, the proportion of 1s or the probability of a 1.

Rearranging this becomes:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x}.$$

Therefore:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x,$$

$$G(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x.$$

This transformation, $G(x)$ is called the logit transformation. Now, $G(x)$ have many of the desirable properties of a linear regression model. It is linear in its parameters, may be continuous and may range from $-\infty$ to ∞, depending on the range of x.

In a linear regression model, it is assumed that the dependent variable can be expressed as $y = \beta_0 + \beta_1 x + \varepsilon$. As seen in the previous chapter, the assumption in linear regression is that this ε is distributed with a mean of zero and a constant variance. With a binary variable, the dependent variable given x is now expressed as $y = \pi(x) + \varepsilon$. The error ε can now assume only one of two possible values. If $y = 1$, then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$. Thus, ε is distributed with a mean of zero and a variance of $\frac{\pi(x)}{1-\pi(x)}$. Now, since the error term has this distribution, if follows that the conditional distribution of the dependent variable follows a binomial distribution with a probability given by the conditional mean $\pi(x)$.

To fit a logistic regression model $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ to a set of data requires that the values for the unknown parameters β_0 and β_1 be estimated. Recall that in the previous chapter, the least squares approach was used to estimate the unknown parameters. The values of β_0 and β_1 were chosen that minimized the sum of squared errors (SSE), in other words the sum of squared deviations of the observed values and the predicted values based on the model. Now with some models, like the logistic curve, there is no mathematical solution that will produce explicit expressions for least square estimates of the parameters. The approach that will be followed here is called maximum likelihood. This method yields values for the unknown parameters that maximize the probability of obtaining the observed set of data. To apply this method, a likelihood function must be constructed. This function expressed the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators of these parameters are chosen that this function is maximized, hence the resulting estimators will agree most closely with the observed data.

If y is coded as 0 or 1, the expression for $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ provides the conditional probability that $y = 1$ given x, $P(y = 1|x)$. It follows that $1 - \pi(x)$ gives the conditional probability that $y = 0$ given x, $P(y = 0|x)$. For an observation (y_i, x_i) where $y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$ and where $y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$. Now this can be expressed for the observation (y_i, x_i) as:

$$\pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}.$$

The assumption is that the observations are independent, thus the likelihood function is obtained as a product of the terms given by the above expression.

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

where $\boldsymbol{\beta}$ is the vector of unknown parameters.

Now, $\boldsymbol{\beta}$ has to be estimated so that $l(\boldsymbol{\beta})$ is maximized. It is mathematically easier to work with the natural logarithm of this equation. The log likelihood function is defined as:

$$L(\boldsymbol{\beta}) = \ln(l(\boldsymbol{\beta})) = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\}.$$

In linear regression, the normal equations obtained by minimizing the SSE, was linear in the unknown parameters that are easily solved. In logistic regression, minimizing the log likelihood yields equations that are nonlinear in the unknowns, so numerical methods are used to obtain their solutions.

Example 6 *How does this all relate to the example? Where did all of this come from? In the example, the aim is to predic default, using family size. Assume that the probability of defaulting for a family of size 8 is 0.8.*

The odds of defaulting for a family of size 8 are then obtained by:

$$\begin{aligned} Odds &= \frac{p}{(1-p)} \\ &= 0.8/0.2 \\ &= 4 \end{aligned}$$

or odds of 4 to 1.

Alternatively, odds are calculated by dividing the number of defaults by the number of non-defaults:

$$\begin{aligned} Odds &= 8/2 \\ &= 4 \end{aligned}$$

or odds of 4 to 1.

The odds of not defaulting are $0.2/0.8$ *or* $2/8$ *or* 0.25.

The asymmetry is not appealing as one would want the odds of defaulting to be the opposite of the odds of not defaulting. Using the natural log takes care of this. The natural log of 4 is 0.602 and the natural log of 0.25 is -0.602. Now the log odds of defaulting are exactly the opposite to the log odds of not defaulting.

Now, the logit transformation is used for logistic regression. This is just the natural log of the odds:

$$G(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

$$G(x) = \log(odds)$$

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x}$$

$$p = \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

which is the same formula said to define the specific s-shape curve earlier.

After estimating the coefficients for the model, the first look at the model concerns an assessment of the significance of the variables in the model. This involves the formulation and testing of statistical hypotheses to determine whether the independent variables in the model are "significantly" related to the dependent variable. Note that the question of whether the predicted values are an accurate representation of the observed values - this is called goodness-of-fit - is not considered here. Goodness-of-fit will be discussed at a later stage. The question here is more: Does the model that includes the variable in question tell one more about the dependent variable than the model that does not include the variable? Several techniques can be used to test for the significance of the coefficients and is discussed below.

5.1 Deviance

In linear regression, the assessment of the significance of the slope coefficient is approached by forming an analysis of variance table. This table partitions the total sum of squared deviations of observations about their mean into two parts: the sum of squares of observations about the regression line SSE, and the sum of squares of predicted values, based on the regression model, about the mean dependent variable SSR. This is just a convenient way of displaying the comparison of observed and predicted values under two models, one containing the independent variable and one using the mean as prediction. In linear regression, interest focuses in the size of SSR. A large value of SSR suggests that the independent variable is important, whereas a small value suggests that the independent variable is not helpful in predicting the response variable.

The guiding principle with logistic regression is the same: Compare the observed values of the response variable to predicted values obtained from models with and without the variable in question. In logistic regression, comparison of observed to predicted values is based on the log likelihood function:

$$L(\boldsymbol{\beta}) = \ln(l(\boldsymbol{\beta})) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}.$$

To better understand this comparison, it is helpful conceptually to think of an observed value of the response variable as also being a predicted value resulting from a saturated model. A saturated model is one that contains as many parameters as there are data points.

The comparison of the observed to predicted values using the likelihood function is based on the following expression:

$$D = -2 \ln \left[\frac{likelihood(fitted)}{likelihood(saturated)} \right].$$

The quantity inside the large brackets in the above expression is called the likelihood ratio. Using minus 2 times the log is necessary to obtain a quantity whose distribution is known and can therefore be used for hypothesis testing. Such a test is called the likelihood ratio test and is described fully below.

Substituting the likelihood function gives us the deviance statistic:

$$D = -2 \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\widehat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \widehat{\pi}_i}{1 - y_i} \right) \right].$$

This statistic will be used in some approaches to assessing goodness-of-fit, which will be discussed in a later chapter. The deviance for logistic regression plays the same role that the residual sum of squares plays in linear regression.

Where the values of the outcome variable are either 0 or 1, the likelihood of the saturated model is 1. Specifically it follows from the definition of a saturated model that $\widehat{\pi}_i = y_i$ and the likelihood is:

$$l(saturated\ model) = \prod_{i=1}^{n} y_i^{y_i} \times (1 - y_i)^{(1-y_i)} = 1.$$

Thus it follows that the deviance is:

$$D = -2 \ln(likelihood(fitted)).$$

5.2 Likelihood-ratio test

The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_o). The full model has all the parameters of interest in it. The simpler model is said to be a nested, reduced model, where an independent variable is dropped from the overall model. The likelihood-ratio test tests if the logistic regression coefficient for the dropped variable can be treated as zero, thereby justifying the dropping of the variable from the model. A non-significant likelihood-ratio test indicates no difference between the full model and the reduced model, hence justifying dropping the given variable so as to have a more parsimonious model that works just as well. The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2\left[\log(L_0) - \log(L_1)\right] = -2(L_0 - L_1).$$

This log transformation of the likelihood function yields an appproximate chi-square statistic.

5.3 Wald Test

The Wald test is used to test the statistical significance of each coefficient (β) in the model. A Wald test calculates a Z statistic which is:

$$W = \frac{\widehat{\beta}}{SE(\widehat{\beta})}.$$

This value is squared which yields a chi-square distribution and is used as the Wald test statistic. (Alternatively the value can be directly compared to a normal distribution.)

Several statisticians have identified problems with this statistic. For large logit coefficients, the standard error is inflated, lowering the Wald statistic and leading to Type II errors (false negatives: thinking the effect is not significant, when it is).

5.4 Score Test

A test for significance of a variable, which does not require the computation of the maximum likelihood estimates for the coefficients, is the Score test. The Score test is based on the distribution of the derivatives of the log likelihood.

Let L be the likelihood function which depends on a univariate parameter θ and let x be the data. The score is $U(\theta)$ where

$$U(\theta) = \frac{\partial \log L(\theta|x)}{\partial\theta}.$$

The observed Fisher information is:

$$I(\theta) = \frac{\partial^2 \log L(\theta|x)}{\partial\theta^2}.$$

The statistic to test $H_0 : \theta = \theta_0$ is:

$$S(\theta) = \frac{U(\theta_0)^2}{I(\theta_0)}$$

which take a $\chi^2(1)$ distribution asymptotically when H_0 is true.

Note that an alternative notation, in which the statistic $S^*(\theta) = \sqrt{S(\theta)}$ is one-sided tested against a normal distribution can also be used. These approaches are equivalent and gives identical results.

Example 7 *Expanding the earlier credit card example, the dependent variable is still whether the client will default or not. Default is denoted by a 1. A different independent variable is now available: whether the client has other products with the bank, where 1 denotes yes.*

The data:

Client	*Default*	*Other products*
1	1	1
2	1	1
3	1	1
4	1	0
5	1	0
6	1	0
7	1	0
8	1	0
9	1	0
10	1	0
11	0	1
12	0	1
13	0	1
14	0	1
15	0	1
16	0	1
17	0	0
18	0	0
19	0	0
201	0	0

Note that the data table can be summarized to the following:

	Other products		
Default	***1***	***0***	***Total***
1	3	7	10
0	6	4	10
Total	9	11	20

Now, the odds of a client that has other products with the bank, to default is $3/6 = 0.5$. *Logistic regression will now be used to test whether its prediction will be close to this.*

The correlation matrix and some descriptive statistics:

	Default	***Other products***
Default	1	
Other products	-0.3	1
Mean	0.5	0.45
Standard deviation	0.51	0.51

Note that half the clients have defaulted. If no other knowledge was available on the clients and a prediction was to be made whether the client would default or not, one could just flip a fair coin to make the prediction. In order to make a better prediction, the use of independent variables becomes a necessity. According to the above correlation matrix, the presence of other products with the bank is negatively correlated with the default indicator. This shows that those clients with other products with the bank are less likely to default.

The following is run in SAS:

```
proc logistic data=simple.one descending;
  model default=Other_products;
run;
```

*Note that the **descending** option is used so that the log(odds) of the event (1) is modeled.*

SAS output:

Criterion	Intercept Only	Intercept and Covariates
-2 Log L	27.726	25.878

The first log-likelihood is that of the model without the independent variable. The second is the log-likelihood with the independent variable.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1.8480	1	0.1740

The likelihood ratio is the difference between the two above likelihoods. The test statistic has a chi-square distribution with 1 degree of freedom. The p-value for the hypothesis is quite large so the null hypothesis of no significance of the variable is rejected.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.5596	0.6268	0.7971	0.3720
Other_products	1	-1.2526	0.9449	1.7574	0.1850

The log odds are:

$$Log(odds) = 0.5596 - 1.2528 \times otherproducts$$

$$Odds = e^{0.5596 - 1.2528 \times other\ products}$$

Hence, if a client has other products with the bank, the predicted odds are:

$$Odds = 0.50007$$

which is close to the calculated odds of 0.5 above.

This implies that the probability of default for a client with other products at the bank is 33.34%.

Chapter 6

Multivariate logistic regression

In the previous chapter, univariate logistic regression was dicussed. As with linear regression, the strength of a modeling technique lies in its ability to model many variables, of which some are on different measurement scales. The logistic model will now be used where more than one independent variable is available and this is called multivariate logistic regression.

Some variables used in models are discrete, categorical variables. It is inappropriate to include them in the model as if they were interval scale variables if they are represented by numbers. The number merely indicates the category and doesn't have any numerical significance. This necessitates the use of dummy or design variables, a concept which is explained below.

As an illustration, suppose that a variable, education level, which have three categories; grade 12, diploma and degree, is available. One option is to create a dummy variable for each of the categories:

- $D_1 = 1$ if grade 12, else 0
- $D_2 = 1$ if diploma, else 0
- $D_3 = 1$ if degree, else 0

A problem arises due to the fact that D_3 is a linear combination of D_1 and D_2 and will create multicollinearity in the model. The existence of multicollinearity inflates the variance of the parameter estimates. That may result, particularly for small and moderate sample sizes, in lack of statistical significance of individual independent variables while the overall model may be strongly significant. It may also result in wrong signs and magnitudes of regression coefficient estimates, and consequently in incorrect conclusions about relationships between independent and the dependant variables. Note that if D_1 and D_2 are both zero,this indicates that the category is degree, which means that an extra variable for this category is not needed. The example can be summarized as follows:

	Design/dummy variable	
Education level	D_1	D_2
Grade 12	1	0
Diploma	0	1
Degree	0	0

So in general, if a nominal scaled variable has k possible values, then $k-1$ dummy variables will be needed. Most logistic regression software, like SAS, will generate dummy variables and in SAS the reference category can be specified.

Consider a collection of p independent variables denoted by the vector $\mathbf{x} = (x_1, x_2, ..., x_p)$. Denote the conditional probability that the event is observed by: $P(y = 1|\mathbf{x}) = \pi(\mathbf{x})$.

The logit of the multivariate logistic regression is then given by the equation:

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p.$$

This means the logistic regression is given by

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}.$$

Now, if the j^{th} independent variable has k_j levels, $k_j - 1$ dummy variables will be needed. These variables will be denoted by D_{jl} and the coefficients for these dummy variables will be denoted by β_{jl}, where $l = 1, 2, ..., k_j - 1$. Thus the logit for a model with p variables and the j^{th} variable being discrete would be:

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + ... + \beta_p x_p.$$

Now, how is a multiple/multivariate logistic model fitted?

Assume that a sample of n independent observations $(\mathbf{x}_i, y_i)$, $i = 1, 2, ..., n$. The estimates of the following vector need to be obtained:

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, ...\beta_p).$$

The method of estimation in the multivariate case is also maximum likelihood. The likelihood function will now be:

$$l(\beta) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i} \left[1 - \pi(\mathbf{x}_i)\right]^{1 - y_i}$$

with $\pi(\mathbf{x}_i)$ defined as:

$$\pi(\mathbf{x}_i) = \frac{e^{g(\mathbf{x}_i)}}{1 + e^{g(\mathbf{x}_i)}}.$$

The $p + 1$ likelihood equations will be obtained by differentiating the log likelihood function with respect to the $p + 1$ coefficients. As with the univariate case, there is no easy solution for these equations and solving them requires special software packages and numerical methods.

Let $\widehat{\boldsymbol{\beta}}$ denote the solution to these equations. In the previous chapter, the standard error of the estimate was used. It will now be considered in more detail.

The method of estimating the variances and covariances of the estimated coefficients follows from the theory that estimators are obtained from the matrix of second partial derivatives of the log likelihood function.

Let the $(p + 1) \times (p + 1)$ matrix containing the negative of these partial derivatives be denoted by $\mathbf{I}(\boldsymbol{\beta})$. This matrix is called the observed information matrix. The variances and covariances are obtained from the inverse of the matrix, which is denoted by:

$$\mathbf{var}(\widehat{\boldsymbol{\beta}}) = \mathbf{I}^{-1}(\boldsymbol{\beta}).$$

In most cases, if not always, it is not possible to write explicit expressions for the elements in this matrix. $var(\widehat{\beta}_j)$ will be used to denote the j^{th} diagonal element of the matrix, which is the variance of $\widehat{\beta}_j$ and $Cov(\widehat{\beta}_j, \widehat{\beta}_i)$ to denote the covariance of $\widehat{\beta}_j$ and $\widehat{\beta}_i$. The estimators of the variances and covariances are obtained by evaluating $\mathbf{var}(\boldsymbol{\beta})$ at $\widehat{\boldsymbol{\beta}}$.

The estimated standard errors of the estimated coefficients will mostly be used, which are:

$$SE(\widehat{\beta}_j) = var(\widehat{\beta}_j)^{\frac{1}{2}}$$

for $j = 1, 2, ..., p$.

A useful formulation of the information matrix is:

$$\widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{V}\mathbf{X}$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ . & . & . & ... & . \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix}$$

and

$$\mathbf{V} = \begin{pmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & 0 & ... & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & 0 & ... & 0 \\ . & . & . & ... & . \\ . & . & . & ... & . \\ 0 & 0 & 0 & ... & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{pmatrix}.$$

6.1 Testing significance of the model

Once the multivariate logistic regression model has been fitted, the model assessment begins. The first step is to assess the overall significance of the p independent variables in the model, using the likelihood ratio as in the univariate case. The likelihood of the fitted model is compared to the likelihood of a constant only model.

The hypothesis test:
$H_0 : \beta_i = 0, i = 1, 2, ..., p$
or
H_0 : There is no difference between the fitted and full (/intercept only) model

The test statistic G :

$$\begin{aligned} G &= -2\ln(\text{likelihood constant only model/likelihood fitted model}) \\ &= -2\ln(L_0/L_1) \\ &= -2\left[\ln(L_0) - \ln(L_1)\right] \\ &= -2(L_0 - L_1) \end{aligned}$$

Under H_0, G will have a chi-square distribution with p degrees of freedom.

Note that if H_0 is rejected, the conclusion is that at least one or perhaps all p coefficients are significantly different from zero.

Example 8 *Building again on the concept of the credit card example, the dependent variable is still whether the client will default or not. Default is denoted by a* 1*. Two independent variables are now avaible, whether the client has other products with the bank, where* 1 *denotes yes, and family size.*

The data

Client	***Default***	***Other products***	***Family size***
1	1	1	4
2	1	1	5
3	1	1	1
4	1	0	5
5	1	0	5
6	1	0	4
7	1	0	4
8	1	0	3
9	1	0	3
10	1	0	2
11	0	1	4
12	0	1	2
13	0	1	3
14	0	1	3
15	0	1	5
16	0	1	1
17	0	0	4
18	0	0	3
19	0	0	3
20	0	0	2

The correlation matrix and some descriptive statistics:

	Default	***Other products***	***Family size***
Default	1		
Other products	-0.3	1	
Family size	0.24	-0.14	1
Mean	0.5	0.45	3.30
Standard deviation	0.51	0.51	1.26

According to the above correlation matrix, those clients with other products in the bank are less likely to default. Also, family size is positively correlated with default, indicating that the higher the family size, the more likely the client is to default.

SAS code given below:

```
proc logistic data=simple.multi descending;
  model default=Other_products Family_size;
run;
```

SAS output:

Criterion	Intercept Only	Intercept and Covariates
-2 Log L	27.726	24.927

The first log-likelihood is that of the model without the independent variables. The second is the log-likelihood with the independent variables.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2.7987	2	0.2468

The likelihood ratio is the difference between the two above likelihoods. This test statistic has a chi-square distribution with 2 *degrees of freedom, as two independent variables as well as the intercept are in the model. The p-value for the hypothesis* $P(\chi^2 > 2.7987) = 0.2468$*, is quite large so the null hypothesis of no significance of the variables is rejected.*

Note that this test indicates that either one or both of the variables are significant, but does not give an indication of their individual significance. Before concluding that any or all of the coefficients are nonzero, one needs to look at the univariate Wald test statistics:

$$W_j = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}.$$

Under the null hypothesis that the coefficient is equal to zero, this statistic follows a standard normal distribution. Alternatively, this statistic can be squared and will then follow a chi-square distribution with 1 degree of freedom. Either way, it gives equivalent results.

Example 9 *Following on the previous example, the Wald statistics are also given as standard output in SAS:*

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7737	1.5313	0.2553	0.6134
Other_products	1	-1.1936	0.9697	1.5153	0.2183
Family_size	1	0.3919	0.4157	0.8888	0.3458

If significance of 0.05 *is used, the conclusion is that both independent variables are significant.*

If a variable is found to not be significant, a reduced model can be fitted and compared to the full model. The aim is to get the best fitting model, while keeping the number of variables to a minimum. The likelihood ratio test can again be used. If the null hypothesis is not rejected, the conclusion is that the reduced model is just as good as the full model. Note that statistical significance isn't the only consideration when variables are included or excluded in the model in practice. Sometimes business might require a variable to be included or legally a variable must be excluded. These considerations might prove to be more important than statistical significance.

Example 10 *Expanding the example again, to include a categorical independent variable, which has more than 2 categories and therefore will require more than one dummy variable.*

Education level is added, which has the following categories:

Education level	***Description***
1	*Grade 12*
2	*Diploma*
3	*Degree*

The data:

Client	***Default***	***Other products***	***Family size***	***Education level***
1	1	1	4	1
2	1	1	5	1
3	1	1	1	2
4	1	0	5	1
5	1	0	5	1
6	1	0	4	3
7	1	0	4	1
8	1	0	3	1
9	1	0	3	2
10	1	0	2	2
11	0	1	4	1
12	0	1	2	3
13	0	1	3	3
14	0	1	3	2
15	0	1	5	1
16	0	1	1	2
17	0	0	4	2
18	0	0	3	3
19	0	0	3	3
20	0	0	2	3

The SAS code:

```
proc logistic data=simple.edumulti descending;
  class Education_level (ref='3') /param=ref;
  model default=Other_products Family_size Education_level;
run;
```

*The **class** statement allows use of the categorical variables in **proc logistic**. Note that the **class** statement must appear before the **model** statement. The **ref=** allows the model developer to indicate the reference category for the class variable. This is the category that will be left out; the dummy variables for the remaining categories would be included in the **model** statement as independent variables. The quotation marks are used for both numerical and character variables. The **param** option specifies the parameterization method for the classification variable or variables. Design matrix columns are created from class variables according to the different schemes. **param=ref** specifies reference cell coding. This results in the same analysis as including dummy variables for each category, except the reference category. The default coding is **param=effect** which results in the coding of -1, 0 and 1.*

*SAS gives the following output for the class variable to indicate how the dummy variables were created, using the **param=ref** option:*

Class Level Information

Class	Value	Design Variables	
Education_level	1	1	0
	2	0	1
	3	0	0

*For interest, note the difference of the design variable coding, when the **param=effect** option (default design variable coding option in SAS) is used:*

Class Level Information

Class	*Value*	*Design Variables*	
Education_level	*1*	*1*	*0*
	2	*0*	*1*
	3	*-1*	*-1*

Whenever a categorical independent variable is included in a model, all of its design variables should be included. To do otherwise, only means that one has recoded the variable. If only D_1 is included, then education level is only included as Grade 12 or not Grade 12. Now, if education level is included in the model, the number of degrees-of-freedom added for the likelihood ratio test, is not 1, but $k-1$, where k is the number of categories in education level.

SAS output:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
-2 Log L	27.726	18.045

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.6811	4	0.0462

Note that adding education level increased the degrees-of-freedom from 2 to 4, as 2 dummy variables were used to code it. Using the likelihood ratio test, the significance of the coefficients in the model might be rejected under H_0. In the previous example, the null hypothesis was clearly rejected. This indicates that education level might not add to the significance of the overall model.

Because of the multiple degrees-of-freedom, the Wald statistics must be used with caution to assess the significance of the coefficients. If the W-statistics for both coefficients (of the dummy variable) exceed 2, the conclusion is that design variables are significant. Alternatively, if one coefficient has a W statistic of 3.0 and the other of 0.1, one cannot be sure about the contribution of the variable to the model.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6675	2.5327	0.0695	0.7921
Other_products	1	-2.9948	1.7572	2.9044	0.0883
Family_size	1	-0.6243	0.7813	0.6385	0.4242
Education_level 1	1	5.2975	2.8366	3.4876	0.0618
Education_level 2	1	2.2749	1.6365	1.9323	0.1645

Now, given that one of the coefficients is barely significant, and using the likelihood ratio test that indicates that education level might not be significant, education level is not a strong predictor in this model.

The Wald test is obtained from the following matrix calculation:

$$\begin{aligned} M &= \widehat{\boldsymbol{\beta}}' \left[\widehat{var}(\widehat{\boldsymbol{\beta}})\right]^{-1} \widehat{\boldsymbol{\beta}} \\ &= \widehat{\boldsymbol{\beta}}' (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1} \widehat{\boldsymbol{\beta}} \end{aligned}$$

which will be distributed as chi-square with $p+1$ degrees-of-freedom under the hypothesis that each of the $p+1$ coefficients is equal to zero. Tests for just the p slope coefficients are obtained by eliminating $\widehat{\beta}_0$ from $\widehat{\boldsymbol{\beta}}$ and the relevant row and column from $(\mathbf{X}'\mathbf{V}\mathbf{X})$. Since evaluation of this test requires the capability to perform vector-matrix operations and to obtain $\widehat{\boldsymbol{\beta}}$, there is no gain over the likelihood ratio test of the significance of the model.

The multivariate analog of the Score test for the significance of the model is based on the distribution of the p derivatives of $L(\boldsymbol{\beta})$ with respect to $\widehat{\boldsymbol{\beta}}$.

Suppose that $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ under the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Then:

$$U'\left(\widehat{\boldsymbol{\beta}}_0\right) I^{-1}\left(\widehat{\boldsymbol{\beta}}_0\right) U\left(\widehat{\boldsymbol{\beta}}_0\right) \sim \chi^2(k)$$

asymptotically under H_0, where k is the number of constraints imposed by the null hypothesis and

$$U\left(\widehat{\boldsymbol{\beta}}_0\right) = \frac{\partial \log L\left(\widehat{\boldsymbol{\beta}}_0|\mathbf{x}\right)}{\partial \boldsymbol{\beta}}$$

and

$$I\left(\widehat{\boldsymbol{\beta}}_0\right) = -\frac{\partial^2 \log L\left(\widehat{\boldsymbol{\beta}}_0|\mathbf{x}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}`}.$$

The computation of this test has the same level of complexity as the calculation of the Wald test.

6.2 Interpretation of the fitted model

Even though the interpretation of the coefficient of the model in the univariate case was briefly looked at, it will now be formally discussed here.

After fitting a model the emphasis shifts from the computation and assessment of significance of the estimated coefficients to the interpretation of their values. It will now be assumed that the logistic regression model has been fit, that the variables in the model are significant in either a statistical or business sense, and that the model fits according to some statistical measure of fit.

The answer to the following question is now needed: "What do the estimated coefficients in the model tell one about the question that motivated the model?" Interpretation involves two issues: determining the functional relationship between the dependent variable and the independent variable(s), and appropriately defining the unit(s) of change for the independent variable(s).

The first step is to determine the link function, i.e. what function of the dependent variable yields a linear function of the independent variables. In linear regression, it is the identity function, as the dependent variable is by definition linear in its parameters. In logistic regression, it is the logit transformation:

$$G(x) = \ln\left(\frac{\pi(x)}{1+\pi(x)}\right) = \beta_0 + \beta_1 x.$$

In linear regression, the slope coefficient is equal to the difference between the value of the dependent variable at $x+1$ and the value of the dependent variable at x, for any value of x. Thus, the interpretation of the coefficient is relatively straightforward as it expresses the resulting change in the measurement scale of the dependent variable for a unit change in the independent variable.

In the logistic regression model, the slope coefficient represents the change in the logit corresponding to a unit change in the independent variable, $\beta_1 = G(x+1) - G(x)$. A meaning needs to be placed on the difference between the two logits, to interpret the coefficient.

A basic example will first be used, as this will provide the conceptual foundation for all other situations. Assume that the independent variable x, is coded as either 1 or 0. The difference in the logit for a subject for $x = 1$ and $x = 0$ is:

$$G(1) - G(0) = [\beta_0 + \beta_1] - \beta_0 = \beta_1.$$

In this case the logit difference is β_1. In order to interpret this, a measure of association, called the odds ratio needs to be introduced

The possible values of the logistic regression are displayed in the following table:

	Independent variable (x)	
Outcome variable (y)	$x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y = 0$	$1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$
Total	1.0	1.0

The odds of the outcome being present among individuals with $x = 1$ is defined as:

$$\frac{\pi(1)}{1 - \pi(1)}.$$

Similarly, the odds of the outcome being present among individuals with $x = 0$ is defined as:

$$\frac{\pi(0)}{1 - \pi(0)}.$$

The odds ratio, OR, is defined as the ratio of the odds for $x = 1$ and the odds for $x = 0$ and is given by the equation:

$$OR = \frac{\pi(1)/\left[1 - \pi(1)\right]}{\pi(0)/\left[1 - \pi(0)\right]}.$$

Substituting the expressions in the above table gives:

$$\begin{aligned} OR &= \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} / \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} / \frac{1}{1+e^{\beta_0}}} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{(\beta_0+\beta_1)-\beta_0} \\ &= e^{\beta_1} \end{aligned}$$

Hence, for a logistic regression with a dichotomous independent variable, coded zero and one, the relationship between the odds ratio and the regression coefficient is:

$$OR = e^{\beta_1}.$$

The odds ratio is a measure of association which approximates how much more likely it is for the outcome to be present among those with $x = 1$ than those with $x = 0$. For instance, if y indicates whether a person defaults or not and x indicates whether the person has other products with the bank, then $\widehat{OR} = 0.5$ indicates that the occurrence of default is half as likely to occur for among those clients who have other products with the bank and those who don't.

Example 11 *Revisiting the earlier example used:*

	Other products		
Default	**1**	**0**	***Total***
1	3	7	10
0	6	4	10
Total	9	11	20

SAS was used to obtain the estimates of β_0 *and* β_1*:*

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.5596	0.6268	0.7971	0.3720
Other_products	1	-1.2526	0.9449	1.7574	0.1850

The estimate for the odds ratio is:

$$\widehat{OR} = e^{\widehat{\beta}_1} = e^{-1.2526} = 0.2858.$$

This odds ratio could have been obtained directly from the cross-product ratio from the data table, namely:

$$OR = \frac{3/6}{7/4} = 0.2857.$$

Thus

$$\widehat{\beta}_1 = \ln\left[(3/6)/(7/4)\right] = -1.2528.$$

The odds ratio (OR) is usually the parameter of interest in a logistic regression due to its ease of interpretation. Its estimate $\widehat{OR}$ tends to have a distribution that is skewed, due to the fact that the possible values range between 0 and ∞. Inferences are usually based on the sampling distribution of $\ln(\widehat{OR}) = \widehat{\beta}_1$. A $100(1-\alpha)\%$ confidence interval estimate for the odds ratio is obtained by first calculating the interval for the coefficient, β_1, then taking the exponent of these values. In general, the endpoints are given by:

$$\exp\left[\widehat{\beta}_1 \pm z_{1-\alpha/2} SE(\widehat{\beta}_1)\right].$$

Because of the importance of the odds ratio as a measure of association, software packages, like SAS, automatically provide point and confidence interval estimates based on the exponent of each coefficient in a fitted logistic regression model.

Example 12 *SAS OR output for the previous example:*

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Other_products	0.286	0.045	1.821

It is also important to consider the effect that coding of the variable has on the computation of the odds ratio. It was noted that the estimate of the odds ratio was $\widehat{OR} = e^{\widehat{\beta}_1}$. This is correct when the independent variable is coded as 0 or 1. Other coding requires that the value of the logit difference for the specific coding used is calculated and then this difference exponent is taken to estimate the odds ratio.

The estimate of the log of the odds ratio for any independent variable at two levels, say $x = a$ versus $x = b$ is the difference between the estimated logits computed at these two values:

$$\begin{aligned} \ln\left[\widehat{OR}(a,b)\right] &= \widehat{G}(a) - \widehat{G}(b) \\ &= \left(\widehat{\beta}_0 + \widehat{\beta}_1 \times a\right) - \left(\widehat{\beta}_0 + \widehat{\beta}_1 \times b\right) \\ &= \widehat{\beta}_1 \times (a - b) \end{aligned}$$

The estimate for the odds ratio is obtained by taking the exponent of the logit difference:

$$\widehat{OR}(a,b) = \exp\left[\widehat{\beta}_1 \times (a - b)\right].$$

In this chapter, the concept of design/dummy variables was introduced. This is also referred to as reference cell coding. Another coding method that is frequently used is referred to as deviation from means coding (effect in SAS). This method assigns -1 to the lower code and 1 to the higher code. If the variable, other products, is recoded:

Other products	**Design variable**
Yes	1
No	-1

Estimate the odds ratio where this coding is used:

$$\begin{aligned} \ln\left[\widehat{OR}(yes,no)\right] &= \widehat{G}(D = 1) - \widehat{G}(D = -1) \\ &= \left(\widehat{\beta}_0 + \widehat{\beta}_1 \times 1\right) - \left(\widehat{\beta}_0 + \widehat{\beta}_1 \times -1\right) \\ &= 2\widehat{\beta}_1 \end{aligned}$$

If the only the exponent of the coefficient from the computer output is taken, one would have obtained the wrong estimate of the odds ratio. The method of coding also influences the calculation of the endpoints of the confidence interval. In general it is given by:

$$\exp\left[\widehat{\beta}_1(a - b) \pm z_{1-\alpha/2}|a - b| \times SE(\widehat{\beta}_1)\right].$$

This relationship between the logistic regression coefficient and the odds ratio provides the foundation of the interpretation of all logistic regression results.

Now, how would one treat an independent variable with more than 2 categories, like education level in the example?

Suppose the independent variable has $k > 2$ distinct values. The variables have a fixed number of discrete values and a nominal scale. It cannot be treated as an interval scale and therefore a set of design variables must be created.

Example 13 *Referring back to the earlier example:*

	Education level			
Default	*Grade 12*	*Diploma*	*Degree*	*Total*
Yes	6	3	1	10
No	2	3	5	10
Total	8	6	6	20

Two design variables are needed to represent this variable:

	Design/dummy variable	
Education level	D_1	D_2
Grade 12	1	0
Diploma	0	1
Degree	0	0

The extension to variables with more than 3 categories is not conceptually different.

For this example, $k = 3$, the category "Degree" was used as the reference group. The estimated odds for Grade 12 is therefore:

$$(6/2)/(1/5) = (6/2) \times (5/1) = 15.$$

	Education level		
	Grade 12	***Diploma***	***Degree***
Odds ratio	15	5	1
Ln(odds ratio)	2.708	1.609	0

The estimated odds ratio for the reference group is indicated by 1. The output is typically found in literature and can be obtained from SAS by specifying the appropriate choice of design variables (as shown earlier).

SAS output:

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Education_level 1 vs 3	14.998	1.031	218.245
Education_level 2 vs 3	4.999	0.344	72.748

The estimated coefficients from SAS:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6093	1.0954	2.1584	0.1418
Education_level 1	1	2.7079	1.3662	3.9285	0.0475
Education_level 2	1	1.6093	1.3662	1.3875	0.2388

A comparison of the estimated coefficients to the log odds ratios shows that:

$$\ln\left[\widehat{OR}(Grade12, Degree)\right] = \widehat{\beta}_1 = 2.708$$

and

$$\ln\left[\widehat{OR}(Diploma, Degree)\right] = \widehat{\beta}_2 = 1.609.$$

This is not by chance, calculation of the logit difference shows that it is by design.

$$\begin{aligned}
\ln\left[\widehat{OR}(Grade12, Degree)\right] &= \widehat{G}(Grade12) - \widehat{G}(Degree) \\
&= \left[\widehat{\beta}_0 + \widehat{\beta}_1 \times (D_1 = 1) + \widehat{\beta}_2 \times (D_2 = 0)\right] - \\
&\quad \left[\widehat{\beta}_0 + \widehat{\beta}_1 \times (D_1 = 0) + \widehat{\beta}_2 \times (D_2 = 0)\right] \\
&= \left[\widehat{\beta}_0 + \widehat{\beta}_1\right] - \widehat{\beta}_0 \\
&= \widehat{\beta}_1
\end{aligned}$$

A similar calculation would demonstrate that the other coefficient estimated using logistic regression is also equal to the log of odds ratios computed from the data.

Confidence limits for odds ratios are obtained using the same approach as for a dichotomous independent variable. One starts by computing the confidence limits for the log odds ratio (the logistic regression coefficient) and then taking the exponent of these limits to obtain limits for the odds ratio. In general, the limits for a $100(1-\alpha)\%$ confidence interval for the coefficients are of the form:

$$\widehat{\beta}_j \pm z_{1-\alpha/2} SE(\widehat{\beta}_j).$$

Taking the exponent provides the corresponding limits for the odds ratios.

Reference cell coding (like used in the above example) is the most commonly employed coding method. It has widespread use as it estimates the risk of an "experimental" group relative to a "control" group.

There are other methods of coding the variables, like deviation from means coding. The interpretation of the estimated coefficients is not as easy or as clear as in the situation where a reference group is used. Taking the exponent of the estimated coefficients yields the ratio of the odds for the particular group to the geometric mean of the odds. However, the estimated coefficients obtained using deviation from means coding may be used to estimate the odds ratio for one category relative to a reference category. The equation for the logit difference is just more complicated than the one obtained using the reference cell coding. If the objective is to obtain odds ratios, the use of deviation from means coding is computationally much more complex than reference cell coding.

When a logistic regression model contains a continuous independent variable, interpretation of the estimated coefficients depends on how it is entered into the model and the particular units of the variable. For simplicity, assume that the logit is linear in the variable. Under this assumption, the equation for the logit is $G(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

The slope coefficient β_1 gives the change in the log odds for an increase of 1 unit in x: $\beta_1 = G(x + 1) - G(x)$ for any value of x.

Sometimes a unit change is not particularly interesting for the business. If household income is taken in rands, a R1 difference might not be interesting enough. A R1,000 difference might be considered more useful. To provide a useful interpretation for continuous scale covariates one needs to find a method for point and interval estimation for an arbitrary change of c in the covariate. The log odds ratio for a change of c units in x is obtained from the logit difference $G(x + c) - G(x) = c\beta_1$ and the associated odds ratio is obtained by taking the exponent of this logit difference, $OR(c) = OR(x + c, x) = \exp(c\beta_1)$. An estimate is then obtained by replacing β_1 with $\widehat{\beta}_1$. An estimate of the standard error needed for confidence interval estimation is obtained by multiplying the estimated standard error of $\widehat{\beta}_1$ by c. The endpoints of the $100(1 - \alpha)\%$ confidence interval estimate of $OR(c)$ are:

$$\exp\left[c\widehat{\beta}_1 \pm z_{1-\alpha/2} cSE(\widehat{\beta}_1)\right].$$

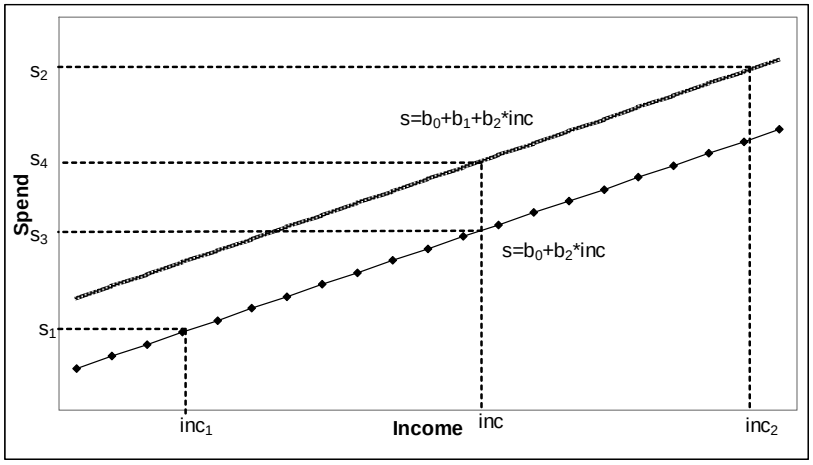
If it is believed that the logit is not linear in the covariate, then grouping and the use of dummy variables should be considered. Alternatively, use of higher order terms (e.g. x^2, x^3,...) or other nonlinear scaling in the covariate (e.g. $\log(x)$) could be considered. The scale in the logit is an important modeling consideration for continuous covariates. The interpretation of the estimated coefficients for a continuous variable is similar to that of a nominal scale variable, using an estimated log odds ratio. The primary difference is that a meaningful scale must be defined for the continuous variable.

Fitting a series of univariate models does not provide an adequate analysis of the data and is not useful in practice. The independent variables are usually associated with one another and may have different distributions within levels of the outcome variable. One generally considers a multivariate analysis for a more comprehensive modeling of data. One of the goals of such analysis is to statistically adjust the estimated effect of each variable in the model for differences in distributions of and associations among the independent variables. Each estimated coefficient now provides an estimate of the log odds adjusting for all other variables included in the model.

What does adjusting statistically for other variables mean? For example, there is a situation where two independent variables are available – one binary and the other continuous. The primary focus is on the effect of the binary variable. In linear regression the analogous situation is called analysis of covariance.

Suppose one wants to compare the average spend on a credit card for two different products (normal and platinum). Spend is associated with many characteristics, one which might be income. Assume that on all characteristics except income, the two groups have nearly identical distributions. If the income distribution is also the same for the two products, then a univariate analysis would suffice and the mean spend of the two groups can be compared. This comparison would provide one with the correct estimate of the difference in spend between the two groups. However, platinum products are mostly associated with more wealthy individuals. A comparison of the two groups would now be meaningless as at least a portion of the difference would be likely due to which product they have. Hence, one is looking for the partial effect of product type given a certain income, in other words, income is kept constant. This is different from the marginal effect of product type, i.e.the effect when income is ignored.

The situation is described graphically in the graph below:



Income vs. spend

In this figure it is assumed that the relationship between income and spend is linear, with the same significant non-zero slope in each. Both these assumptions are usually tested in an analysis of covariance before making inferences about group differences. It is assumed that this has been checked and supported by the data.

The statistical model that describes the situation in the above figure states that the value of spend, s, may be expressed as $s = b_0 + b_1x + b_2inc$, where $x = 0$ for group 1 and $x = 1$ for group 2 and inc denotes income. In this model the parameter b_1 represents the true difference in spend between the two groups and b_2 is the rate of change in spend per unit change in income. Suppose the mean income is inc_1 for group 1 and inc_2 for group 2. Comparison of the mean spend of group 1 and the mean spend of group 2 amounts to a comparison of s_1 and s_2. In terms of the model this difference is $(s_2 - s_1) = b_1 + b_2(inc_2 - inc_1)$. Thus, the comparison involves not only the true difference between the groups, b_1, but a component, $b_2(inc_2 - inc_1)$ which reflects the differences between the income of the groups.

The process of statistically adjusting for income, involves comparing the two groups at some common value of income. The value used is the mean of the two groups which is denoted by inc in the above graph. In model terms, this involves a comparison of s_3 and s_4, $(s_4 - s_3) = b_1 + b_2(inc - inc) = b_1$, the true difference between the two groups. The choice of the overall mean makes sense, it is biologically reasonable and lies within the range for which one believes that the association between income and spend is linear and constant within each group.

Now consider the same situation as shown in the graph, but instead of spend being the dependent variable, assume it is a dichotomous variable and that the vertical axis denotes the logit. Under the model the logit is given by the equation:

$$g(x, inc) = b_0 + b_1 x + b_2 inc.$$

A univariate comparison obtained from 2×2 table cross-classifying outcome and group would yield a log odds ratio approximately equal to $b_2(inc_2 - inc_1)$. This would incorrectly estimate the effect of group due to the difference in the distribution of income. To account or adjust for this difference, income is included as a variable in the model and calculate the logit difference at a common value of income, such as the combined mean, inc. This logit difference is $g(x = 1, inc) - g(x = 0, inc) = b_1$. The coefficient b_1 is the log odds ratio that one would expect to obtain from a univariate comparison if the two groups had the same distribution of income.

Exercise 14 *Returning to the earlier example with dependent variable default (1=yes, 0=no). Also used is the example of having other products with the bank (group1=no, group2=yes) and adding income as a variable. The following table will be used as a basis example for interpreting the logistic regression coefficient for a binary variable when the coefficient is adjusted for a continuous variable.*

	Other products=no		*Other products=yes*	
Variable	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*
Default	0.64	0.504	0.33	0.5
Income	5850.00	2235.73	7922.22	1136.64

Using this table if follows that the univariate log odds ratio for group 2 vs. group 1 is:

$$\ln(\widehat{OR}) = \ln(0.33/0.67) - \ln(0.64/0.36) = 0.1135$$

and the unadjusted estimated odds ratio is $\widehat{OR} = 1.1429$.

There is a considerable difference in income distribution of the two groups. The people in group two's income is on average R2000 higher that those in group 1.

The data is analyzed with a bivariate model using "other_products"=0 for group 1 and "other_products"=1 for group 2. The model is fitted using the following code in SAS:

```
proc logistic data=simple.income descending;
  model default=Other_products income;
run;
```

The output for the coefficients:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	14.0079	6.6209	4.4762	0.0344
Other_products	1	0.6060	1.4966	0.1639	0.6856
Income	1	-0.00199	0.000923	4.6676	0.0307

The income adjusted log odds ratio is given by the estimated coefficient for "other_products" and is 0.6060. *The income adjusted odds ratio is* $\widehat{OR} = \exp(0.6060) = 1.833$. *A lot of the apparent difference between the two groups is due to the difference in income.*

Look at it now in terms of the previous graph. An approximation to the unadjusted odds ratio is obtained by taking the exponent of the difference $s_2 - s_1$. *In terms of the fitted model, this is:*

$[14.0079 + 0.6060 - 0.00199(7922.22)] - [14.0079 - 0.00199(5850.00)]$
$= 0.6060 - 0.00199(7922.22 - 5850.00)$
$= -3.5177$

The value of this odds ratio is:

$$e^{-3.5177} = 0.02967.$$

The discrepancy between 0.02967 *and the actual unadjusted odds ratio,* 1.1429 *is due to the fact that the above comparison is based on the difference in the average logit while the crude odds ratio is approximately equal to the calculation based on the average estimated logistic probability for the two groups. The income adjusted odds ratio is obtained by taking the exponent of the difference* $s_4 - s_3$, *which is equal to the estimated coefficient for "other_products".*

The difference in the example is:
$[14.009 + 0.6060 - 0.00199(6782.5)] - [14.009 - 0.00199(6782.5)]$
$= 0.6060$

The method of adjustment when there is a mixture of variables (dichotomous, categorical, and continuous) is identical to described above. For example, suppose that, instead of using income as continuous, it was dichotomized into low and high income. To obtain the income adjusted effect of "other_products", a bivariate model is fitted containing the two binary variables and the logit difference at the two levels of "other_products" and a common value of the dichotomous variable for income is calculated. The procedure is similar for any number and mix of variables. Adjusted odds ratios are obtained by comparing individuals who differ only in the characteristic of interest and have the values of all other variables constant. The adjustment is statistical as it only estimates what might be expected to be observed had the subjects indeed differed only on the particular characteristic being examined, while all other variables have identical distributions within two levels of outcome.

One important issue to keep in mind when interpreting statistically adjusted log odds ratios and odds ratios is that the effectiveness of the adjustment is entirely dependent on the adequacy of the assumptions of the model: linearity and constant slope. Violations of these assumptions may render the adjustment useless.

Chapter 7

Variable reduction and analysis in credit scoring

7.1 Introduction

In the previous chapters the focus was on the estimation and interpretation of the coefficients in a logistic regression model. The examples used had only a few independent variables and there was perceived to be only one possible model. In practice there are typically many more independent variables that could potentially be included in the model. (For a typical credit scoring model, the possible factors may vary from as little as 20 to over 200!). A strategy and associated methods for handling these complex situations is thus needed.

In this chapter two important issues in scorecard development, variable analysis and variable reduction is discussed. Variable analysis investigates the relationship between variables one wants to test for predictive ability and the outcome variable one hopes to predict. Variable reduction narrows a large number of variables that are candidates for inclusion in the scorecard to a smaller number that can be more easily analyzed with the multivariate statistical techniques used to build scorecards. Not only does the modeler want a smaller number of variables, but also retain only those that are most predictive of the outcome variable.

The goal of any method is to select the variables that results in a "best" model within the context of the problem. Irrespective of the method used, statistical model development involves seeking the most parsimonious model that still explains the data. Minimizing the number of variables in the model has most likely as a result a more stable and easily generalized model. The more variables in the model, the greater the estimated standard errors become and it is more likely that there is over fitting on the data, which means that the model will only be applicable to the specific data set.

The methods discussed in this chapter are not to be used as a substitute for, but rather as an addition to, clear and careful thought.

Scorecard building is also a combination of art and science: The science lies in the statistical methods at the core of scorecard development. The art lies in the many choices the scorecard developer must make throughout the model building process. These choices have a major effect on the final scorecard. Uninformed or incorrect decisions can result in an important variable being excluded or an improper variable being included. Some of the choices that must be made are how to treat data errors, missing values, and outliers (extreme values of the characteristics); whether to use continuous or transformed variables or to make categorical (binned) variables out of continuous variables; and whether to include variable interactions. Perhaps the most important, the modeler must choose which characteristics to incorporate in the scorecard.

It is extremely important for the scorecard developer to have a good grasp of the business for which the scorecard is being built, and a firm command of the business's data. Otherwise, he/she can make mistakes in working with the data and interpreting the results.

There are a few inviolable rules on how characteristics should be chosen for inclusion in scorecards. Here is a list of the primary factors to be considered.

Characteristics should be

- Logical.
- Predictive.
- If included, not create multicollinearity.
- Available and stable.
- Compliant.
- Customer related.
- If excluded, result in unacceptable levels of information loss.

7.1.1 Logical

It is important to remember that simpler is better. The ultimate goal is to provide a robust model that works not only when implemented, but also for a significant time period thereafter. This is aided if the variables make logical sense, which also make the results easier to explain to the business.

7.1.2 Predictive

The variables of interest are those with a significant degree of predictive power. Determining the predictive power of the variables is explained later in this chapter.

7.1.3 Multicollinearity

In many instances, a lot of variables will be highly correlated with each other, especially those calculated using the same, or similar, base inputs. This gives rise to potential multicollinearity, which can lead to poor out-of-sample performance as the model is over-fitted to the data. Over-fitting is typically characterized by extremely large coefficients, high standard errors, or changes in coefficient for some variables that do not make logic sense. This issue can be dealt with by doing variable cluster analysis, which will be briefly discussed in this chapter. In instances where coefficients in the model have the "wrong" sign, a careful investigation of the relationships between the variables are necessary.

7.1.4 Available and stable

Variables should only be used if they will be available in future, have been stable since the sample was taken and are expected to be stable in the future. They should be excluded if they will not be captured in the future, are new and poorly populated, unstable due to infrastructure changes or problems, sensitive to inflation like income, or if they are easily manipulated by either the clients or staff.

7.1.5 Compliant

If the models are to support decision-making, the developer must ensure that the variables used are compliant with any legal, policy or ethical restrictions on their use. This was briefly discussed in chapter one.

7.1.6 Customer related

When rating clients' risk, the variables should relate to them and not the lender's strategy. Lenders' interest is in customer risk, independent of the decision made, so that the decision can be made. For example, in application scoring, the customer demographics, indebtedness, loan purpose and payment behaviour are acceptable variables, but the product offered, loan term and the credit limit granted are not.

7.1.7 Minimum information loss

When reducing the number of variables, it should be done with the minimum information loss. There may be variables that are seemingly weak and highly correlated with other variables, whose exclusion reduces the final model's power.

Another issue to consider here is also the sources of inaccuracy that can influence the model and ultimately the correctness of the prediction of one's credit score.

A number of factors can lead to inaccuracy in the regression coefficients. Inaccuracy means that if one could observe the true and exact effect of each predictor on the outcome, it would not be what the regression coefficients tell one it is. A predictor variable could have a greater or smaller effect than the coefficient indicates. Even worse, a predictor variable that may not appear predictive could be excluded, when it actually is predictive.

Some of the most significant sources of inaccuracy in building scorecards are:

- Inaccuracy from omitting important predictive variables from the score model.
- Inaccuracy due to errors in the predictor variables.
- Sample selection inaccuracy.

Omitted-variables inaccuracy occurs when an important predictor is left out of the regression model. Because many of the variables used to build scoring models are correlated with one another, excluding one or more variables that actually do affect the outcome can have an effect on the coefficients of the predictors that are incorporated into the model. In leaving an important variable out, one may find that the coefficients for other variables are biased or some highly predictive variables may not appear predictive at all.

The second source of possible inaccuracy in scorecard coefficients arises when one or more of the predictor variables are measured wrong. This can happen in a number of ways. The reliability of the data depends both on the method used to collect the information and on the type of documentation, if any, to evaluate its veracity. Carefully cleaning the model development sample may help avoid this type of bias. Data is cleaned by looking at distributions of variables and determining how to deal with outliers, data errors, and missing values.

Sample selection inaccuracy arises when a regression model is built using a subset of the population that is not entirely representative of the full population. Sample selection inaccuracy was already dealt with in chapter two and will not be repeated here.

Which of the three inaccuracy issues described is the most serious or deserving of the scorecard developer's attention is likely to vary case by case. The best one can do is to deal with those that can be dealt with and recognize that one may have to live with some degree of scorecard bias. Even though the scorecard may not capture the exact relationship between the predictor variables and the outcome variable, scorecards have proven themselves over and over again as reliable predictors. Even with some inaccuracy, businesses are almost always better off relying on a scorecard to making decisions than to rely purely on a subjective method.

7.2 Bivariate analysis

The selection process should begin with careful bivariate analysis of each possible variable and the outcome variable. For nominal, ordinal and continuous variables with few integer values, this can be done with a contingency table of outcome ($y = 0, 1$) versus the k levels of the independent variable. The likelihood ratio chi-square test with $k-1$ degrees of freedom is equal to the value of the likelihood ratio test for the significance of the coefficients for the $k-1$ design variables in a logistic regression model that contains that single independent variable.

7.2.1 Likelihood ratio test

The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the model containing the variable (L_1) over the maximize value of the likelihood function for the intercept-only model (L_0). The likelihood-ratio test tests if the logistic regression coefficient for the added variable can be treated as zero, and thereby can be excluded from the model development. A non-significant likelihood ratio test indicates no difference between the model containing the variable and the intercept only model, hence justifying dropping the given variable so as to have a more parsimonious model that works just as well.

The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2\left[\log(L_0) - \log(L_1)\right] = -2(L_0 - L_1).$$

This log transformation of the likelihood function yields a chi-square statistic.

The Pearson chi-square test is asymptotically equivalent to the likelihood ratio chi-square test and can also be used.

7.2.2 Pearson chi-square test

The chi-square test looks for a linear relationship between two characteristics and the resulting p-value provides a measure of reliability – the probability that the similarity (goodness-of- fit) or difference (independence) between them is not a chance occurrence. It is usually used to evaluate a theory or hypothesis, by comparing observed (actual) and expected (estimated) distributions.

There are several variations of the chi-square calculation, but the original and most commonly used is Pearson's chi-square test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency and E_i is the expected frequency for each class. A χ^2 value of zero indicates perfect fit and χ^2 increases as the distributions become dissimilar, eventually becoming so high that one can only conclude that the two distributions bear no relationship to each other (independent).

The associated p-value can be used– a percentage that indicates whether or not the fit is a random occurrence. As χ^2 approaches zero, the p-value approaches 100% and as χ^2 increases, the p-value approaches zero. The conversion depends on the degrees of freedom (the number of independent pieces of information contained in a statistic). The degrees of freedom is calculated as $(n - 1 - a)$ where n is the number of classes and a is the number of assumptions, if any, made in the null hypothesis. It is recommended that variables with a p-value of greater than 0.05 be eliminated from further analysis. Thus, the variable set would be reduced by excluding such variables from the multivariate analysis that would follow.

Chi-squared statistics can be compared across variables for insight into which are most closely associated with the outcome variable.. This statistic allows one to compare predictive variables and learn which are most closely associated with the outcome variable. Note that the value of the chi-squared statistic depends in part on the population default rate, the percentage in the sample experiencing the event. Thus, while one may compare the chi-squared statistic across variables for a given model and data set, one should not compare their values for different models or data sets.

Particular attention should be paid to any contingency table with a zero cell. This yields a point estimate for one of the odds ratios of either zero or infinity. Including such a variable in any logistic regression program causes undesirable numerical outcomes. Strategies for handling the zero cells includes: collapsing the categories of the independent variable in a sensible fashion, eliminating the category completely or, if the variable is ordinal scaled, modeling the variable as if it were continuous.

Note that this chi-squared statistic is a test for a linear association between the candidate predictive variable and the log-odds of the outcome variable. If there is a non-linear relationship, the chi-square may lead one to believe the variables are not associated when in fact they are – though not in a linear relationship.

7.2.3 Spearman rank-order correlation

The Spearman correlation statistic measures the correlation between the rankings of the predictive variable and the outcome variable. That is, instead of looking at the association between the actual values of the predictive and the binary outcome variables, in this calculation the rank assigned to each value of the predictive variable replaces the actual value. It is recommended to use this correlation statistic for binary analysis because it is less sensitive to outliers and nonlinear relationships between outcome and the input variables than some other statistics.

The Spearman correlation coefficient is calculated as:

$$S = 1 - 6\frac{\sum (x_R - y_R)^2}{n^3 - n}$$

where $(x_R - y_R)$ refers to the difference in the respective ranks of the independent variable (x) and the dependent variable (y) for the same observation, and n refers to the total number of cases that are being ranked.

Though the Spearman correlation, unlike the chi-squared statistic, does not require that a variable have a linear relationship to the outcome variable to detect the relationship, the relationship must be monotonic- that is, increasing values of a variable must generally be associated with a higher incidence of being bad (or vice versa) in order for the Spearman correlation to be large. If the Spearman correlation and the chi-square statistic give different indications of the strength of a variable, it is probably because the variable does have a monotonic relationship to the outcome, though not a linear one and a variable transformation may be in order.

An advantage of the Spearman correlation is that it shows the direction of the relationship of a predictive variable with the outcome variable. If the correlation is negative, it indicates that higher values of the predictive variable are associated with lower values of the outcome variable. One should always evaluate whether the effect of each variable on the outcome is consistent with the beliefs about how that variable should affect the outcome.

In credit scoring, variables are seldom used in their original form. Even continuous variables are bucketed, i.e. variables are categorized into logical intervals, for ease of interpretation and implementation. If continuous variables are used in its raw form, truncation is usually used, to reduce the effect and influence of outliers/ extreme values. Categorical variables are also seldom used with all the categories. The weight of evidence (WOE) of the different categories are examined and grouped together if they have similar relative risk (in this case, WOE is similar).

7.2.4 Weight of Evidence (WOE)

Each decision made is based on the probability of some event occurring. One assesses the circumstances and determine a weight of evidence. The weight of evidence converts the risk associated with a particular choice into a linear scale that is easier for the human mind to assess:

$$WOE_i = \ln\left(\left(\frac{N_i}{\sum_{i=1}^{n} N_i}\right) / \left(\frac{P_i}{\sum_{i=1}^{n} P_i}\right)\right)$$

where P is the number of occurrences, N is the number of non-occurrences and i the index of the attribute being evaluated. The precondition is non-zero values for all N_i and P_i.

The WOE is used to assess the relative risk of different attributes for a characteristic and as a means to transform characteristics into variables. It is also a very useful tool for binning.

The WOE formula discussed above is the one most often used. It can be restated as:

$$WOE_i = \ln(N_i/P_i) - \ln(\sum_{i=1}^{n} N_i / \sum_{i=1}^{n} P_i)$$

which illustrates two components: a variable portion for the odds of that group and a constant portion for the sample or population odds. The WOE for any group with average odds is zero. A negative WOE indicates that the proportion of defaults is higher for that attribute than the overall proportion and indicates higher risk.

For a characteristic transformation, the WOE variable has a linear relationship with the logistic function, making it well suited for representing the characteristic when using logistic regression.

The WOE does not consider the proportion of observations with that attribute, only the relative risk. Other tools are used to determine the relative contribution of each attribute and the total information value.

7.2.5 Information Value

The information value is used to rank order variables in terms of their predictive power. It is extended in the credit scoring context by looking at the non-occurrence (N) as non-defaults (goods) and the occurrence (P) as defaults (bads). A high information value indicates a high ability to discriminate. Values for the information value will always be positive and may be above 3 when assessing highly predictive characteristics and is typically seen in behavioural scorecards. Characteristics with information values less than 0.10 are typically viewed as weak, while values over 0.30 are sought after, and are likely to feature in scoring models. Please note that weak characteristics may provide value in combination with others or have individual attributes that could provide value as dummy variables. They should thus not be discarded indiscriminately. It can be difficult to interpret the information value, because there are no associated statistical tests. As a general rule, it is best to use the information value in combination with other measures for final selection.

The information value, as it is known in credit scoring, is technically referred to as the Kullback divergence measure. It measures the difference between two distributions. It is expressed as:

$$IV = \sum_{i=1}^{n} \left[\left(\frac{N_i}{\sum_{i=1}^{n} N_i} - \frac{P_i}{\sum_{i=1}^{n} P_i} \right) \times WOE_i \right]$$

where N is the number of non-occurrences, P is the number of occurrences, WOE is the weight of evidence and i is the index of the attribute being evaluated and n is the total number of attributes.

Note that the information value is sensitive to how the variable is grouped and the number of groups.

Another bivariate measure that is used in credit scoring is to calculate the Gini coefficient for each of the variables.

7.2.6 Gini coefficient

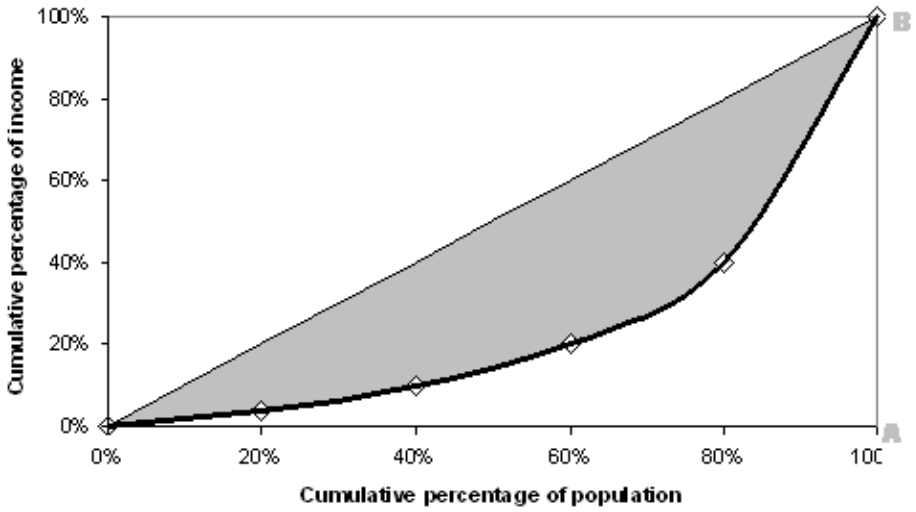
The Gini coefficient had its first application in economics measuring the degree of inequality in income distribution. It is explained here in the economical context and then extended it to its application in credit scoring. It is calculated using the Lorenz curve (Mohr (1998).

The Lorenz curve (named after the American statistician Lorenz, who developed it in 1905) is a simple graphic device which illustrates the degree of inequality in the distribution of the variable concerned. To construct a Lorenz curve that reflects the distribution of income among the individuals or households in an economy, the latter first have to be ranked from poorest to richest. This is done on a cumulative percentage basis. In other words, one starts with the poorest percent of the population, the second poorest percent of the population and so on until one comes to the richest percent of the population. The cumulative percentages of the population are plotted along the horizontal axis. The vertical axis shows the cumulative percentage of total income. If the poorest percent of the population earns 0.1% of the total income in the economy, that number will be plotted vertically above the first percent of the population. If the second poorest percent of the population earns 0.2% of the total income in the economy, it means that the first 2% earned a cumulative share of 0.3% (0.1% + 0.2%) of the income. This number 0.3% will be plotted vertically above the 2% on the horizontal axis.

The following table shows a hypothetical distribution of income. To keep things simple for this illustration, only income shares of each successive 20% of the population are shown:

Percentage of		**Cumulative percentage of**	
Population	**Income**	**Population**	**Income**
Poorest 20%	4%	20%	4%
Next 20%	6%	40%	10%
Next 20%	10%	60%	20%
Next 20%	20%	80%	40%
Richest 20%	60%	100%	100%

The first two columns in this table contain the basic data. The last two columns are the cumulative totals. These two columns show that the poorest 60% of the population earn 20% of the total income in the economy. The last two columns are then plotted.



Lorenz curve and Gini

The figure shows that the poorest 20% of the population earn 4% of the income; the poorest 60% of the population earn 20% of the income and so on.

Note two other features of the diagram. The first is that the axes have been joined to form a square. The second feature is the diagonal running from the origin 0 to the opposite point **B** (top right) of the rectangle. The diagonal serves as a reference point. It indicates perfectly equal distribution of income. Along the diagonal the first 20% of the population receives 20% of the income; the first 40% receives 40% and so on. Like the diagonal, the Lorenz curve must start at the origin 0 (since 0% of the population will earn 0% of the income) and end at **B** (since 100% of the population will earn 100% of the income). The degree of inequality is shown by the deviation form the diagonal. The greater the distance between the diagonal and the Lorenz curve, the greater the degree of inequality. The area between the diagonal and the Lorenz curve has been shaded on the graph and is called the area of inequality. The greatest possible inequality will be where one person earns the total income. If that is the case, the Lorenz curve will run along the axes from **A** to **B**.

Since only the bottom triangle in the above figure has any real significance, the Lorenz curve can be presented simply in a triangle, instead of a square. The meaning and interpretation of the curve remains unchanged. The nearer the position of the Lorenz curve to the diagonal of the triangle, the more even the distribution. If the distribution is absolutely equal, the Lorenz curve will coincide with the diagonal. The further away the curve lies from the diagonal, the less even the distribution. To compare different distributions, different Lorenz curves have to be constructed and compared. Provided the curves do not intersect, such a comparison will reveal whether one distribution is more (or less) equal than the other. However, if they intersect, other criteria have to be considered. One of these is the Gini coefficient.

The Gini coefficient (or Gini ratio) is named after the Italian demographer Corrodo Gini, who invented it in 1912. It is a useful quantitative measure of the degree of inequality, obtained by dividing the area of inequality shown by the Lorenz curve by the area of the right-triangle formed by the axes and the diagonal. (The triangle formed by points 0, **A** and **B** in the above graph).

The Gini coefficient can vary between 0 and 1. If the incomes are distributed perfectly equally, the Gini coefficient is zero. In this case the Lorenz curve coincides with the line of perfect equality (the diagonal) and the area of inequality is therefore zero. At the other extreme, if the total income goes to one individual or household (i.e. if the incomes are distributed with perfect inequality), the Gini coefficient is equal to one. In this case the area of inequality will be the same as the triangle 0**AB**. The Gini coefficient can also be expressed as a percentage, in which case it is called the Gini index.

The Gini coefficient is calculated using the following formula:

$$Gini = \sum_{i=1}^{n} \left(\left(cpY_i - cpY_{i-1} \right) \left(cpX_i + cpX_{i-1} \right) \right) - 1$$

where cpY is the cumulative percentage of ranked income and cpX is the cumulative percentage of population. Note that the result is a rank correlation coefficient which is exactly the same as the Somer's D statistic provided by SAS. The Gini coefficient is not used for hypothesis testing, but does provide a powerful measure of separation, or lack of it.

Example 15 *The Gini index is calculated for the example:*

Cumulative percentage of				
Population	***Income***	$cpX_i + cpX_{i-1}(A)$	$cpY_i - cpY_{i-1}(B)$	$Z_i(= A \times B)$
20%	4%	20%	4%	0.8%
40%	10%	60%	6%	3.6%
60%	20%	100%	10%	10%
80%	40%	140%	20%	28%
100%	100%	180%	60%	108%
Gini $= \sum_i \mathbf{Z}_i \mathbf{-1} = \mathbf{50.4}\%$				150.4%

Note that the calculation formula of the Gini is dependent on the ranking. In the above example, the ranking was done from the poorest to the richest and resulted in the above formula for the Gini calculation. If the ranking was from richest to poorest, the Gini calculation formula would be:

$$Gini = 1 - \sum_{i=1}^{n} \left(\left(cpY_i - cpY_{i-1} \right) \left(cpX_i + cpX_{i-1} \right) \right).$$

The calculation of the Gini index has been co-opted into a lot of other disciplines, including credit scoring, where it is often referred to as an accuracy ratio or power ratio. The Gini coefficient is used as a measure of how well a scorecard or variable is able to distinguish between goods and bads.

The cumulative distribution function $F(x)$ for a logistic model can be expressed in the form:

$$H\{F(x)\} = \lambda_0 + \lambda_1 \log(x) + \lambda_2 (\log(x))^2$$

where $H\{F(x)\}$ is a known function of $F(x)$. Then the frequencies $\mathbf{f}_i$ of the specific values or class intervals of $\mathbf{x}$ follows a multinomial distribution that , in terms of n, the total number of observations, have an asymptotically multinomial normal distribution. The estimation methods used for categorical data, can be used to determine the λ-coefficients in the above equation. The expected frequency of the i^{th} class interval is m_i and the probability that an observation falls in the i^{th} class interval is π_i, with the observed probability $\mathbf{p}_i = \frac{\mathbf{f}_i}{n}$.

The generalized linear model of the above equation is:

$$H\{F(x)\} = A\boldsymbol{\lambda}.$$

As a measurement of the discrepancy of the model in terms of the observed frequencies, the statistic

$$D\chi^2 = \chi^2/n$$

is used, where n is the total number of observations and χ^2 is the Pearson goodness-of-fit test statistic.

If Δ_1 is the coefficient of the average difference, it follows that

$$\begin{aligned}\Delta_1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|x-y|f(x)f(y)dxdy \\ &= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x)f(y)dxdy - 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x)f(y)dxdy \\ &= 2\mu - 4\int_{-\infty}^{\infty}\{\int_{-\infty}^{\infty}yf(y)dy\}f(x)dx \\ &= 2\mu\{1 - \int_{-\infty}^{\infty}\Phi(x)fxdx\}\end{aligned}$$

With a sample of size 2 from the distribution of $\mathbf{x}$, it follows that the expected value of the smallest order statistic is given by

$$\mu^{(1)} = 2\int_{-\infty}^{\infty}\{\int_{-\infty}^{\infty}yf(y)dy\}f(x)dx.$$

It then follows that the Gini -index is given by

$$G = 1 - \frac{\mu^{(1)}}{\mu}.$$

The Gini index is also defined by Kendall and Stuart (1958) as:

$$G = \frac{\Delta}{2\mu}.$$

The Gini index can also be calculated using the odds distribution (goods:bads)

	Outcome			**Cumulative %**	
Score	**Goods**	**Bads**	**G/B odds**	**Goods**	**Bads**
Low	5,000	2,000	2.5	2.0	33.3
Middle	45,000	2,000	22.5	20.0	66.7
High	200,000	2,000	100.0	100.0	100.0
	250,000	6,000	41.7		

Score	$\mathbf{cpN}_i+\mathbf{cpN}_{i-1}(\%)$	$\mathbf{cpP}_i-\mathbf{cpP}_{i-1}(\%)$	$\mathbf{Z}_i(\%)$
Low	2.0	33.3	0.7
Middle	22.0	33.3	7.3
High	120.0	33.3	40.0
	Gini index=$(\mathbf{1}-\sum_i \mathbf{Z}_i)$		52.0

The odds distribution for the above table is as follows:

$$\begin{array}{lll} f(x) & = \frac{1}{3} \quad \text{when} & x = 2.5 \\ & = \frac{1}{3} \quad \text{when} & x = 22.5 \\ & = \frac{1}{3} \quad \text{when} & x = 100 \end{array}$$

For the distribution:
$\mu = 41.67$
$\Delta = \frac{2}{3^2}\{20 + 97.5 + 77.5\} = 43.33333$

$Gini = \frac{\Delta}{2\mu} = 0.52$

The cumulative distribution function of the smallest order statistic for a sample of size 2 is:

$$\begin{array}{lll} F_{X^{(1)}}(x) & =0 & \text{when} \quad x < 2.5 \\ & =1 - (\frac{2}{3})^2 = \frac{5}{9} & \text{when} \quad 2.5 \leq x \leq 22.5 \\ & =1 - (\frac{1}{3})^2 = \frac{8}{9} & \text{when} \quad 22.5 \leq x < 100 \\ & =1 & \text{when} \quad x \geq 100 \end{array}$$

with the probability function

$$\begin{array}{lll} f_{X^{(1)}}(x) & = \frac{5}{9} \quad \text{where} & x = 2.5 \\ & = \frac{3}{9} \quad \text{where} & x = 22.5 \\ & = \frac{1}{9} \quad \text{where} & x = 100 \end{array}$$

with $\mu^{(1)} = 20$

Then the Gini index is:

$$Gini = 1 - \frac{\mu^{(1)}}{\mu} = 0.52.$$

The Gini index can also be used to determine the power of an individual variable. The attributes of the variable is grouped from the best to the worst risk. The cumulative distribution is still plotted on the x-axis, but the cumulative percentage of defaults (events) is plotted on the y-axis. A certain minimum cut-off can be set, depending on the model to be built, and variables can be chosen for inclusion in the modeling process based on this cut-off. Note that the Gini-coefficient is equivalent to the test statistic in the Mann-Whitney-Wilcoxon nonparametric test, and also to the area under the Receiver Operating Characteristic cuve. The latter will be explained later in more detail.

Example 16 *The WOE, Information Value and Gini index is now illustrated using an example.*

The same data is used as in the previous examples, with the explanations of the categories also the same and is given below:

Client	***Default***	***Other Products***	***Family Size***	***Education Level***
1	1	1	4	1
2	1	1	5	1
3	1	1	1	2
4	1	0	5	1
5	1	0	5	1
6	1	0	4	3
7	1	0	4	1
8	1	0	3	1
9	1	0	3	2
10	1	0	2	2
11	0	1	4	1
12	0	1	2	3
13	0	1	3	3
14	0	1	3	2
15	0	1	5	1
16	0	1	1	2
17	0	0	4	2
18	0	0	3	3
19	0	0	3	3
20	0	0	2	3

Weights of evidence

First, the WOE is calculated for each attribute for each variable:

	Default		
Other products	***1***	***0***	***WOE***
0	7	4	−0.5596158
1	3	6	0.6931472

	Default		
Family size	***1***	***0***	***WOE***
1	1	1	0.0000000
2	1	2	0.6931472
3	2	4	0.6931472
4	3	2	−0.4054651
5	3	1	−1.0986123

Note that the WOE for family size of 2 and 3 is identical. This indicates that there is benefit in grouping them together.

	Default		
Education level	***1***	***0***	***WOE***
1	6	2	−1.0986123
2	3	3	0.0000000
3	1	5	1.6094379

Information value

As the WOE does not give an indication of the overall power of a variable, the Information Value is calculated.

Other products	***WOE***	***Calculation***
0	−0.5596158	0.1678847
1	0.6931472	0.2079442
Information Value		**0.3758289**

Family Size	***WOE***	***Calculation***
1	0.0000000	0.0000000
2	0.6931472	0.0693147
3	0.6931472	0.1386294
4	−0.4054651	0.0405465
5	−1.0986123	0.2197225
Information Value		**0.4682131**

What would happen to the Information Value if family size of 2 and 3 is combined?

Family Size	***WOE***	***Calculation***
1	0.0000000	0.0000000
2&3	0.6931472	0.2079442
4	−0.4054651	0.0405464
5	−1.0986123	0.2197225
Information Value		**0.4682131**

Note that the information value stayed the same. If this variable is used in the model, the number of degrees-of-freedom can be reduced by combining the two categories, without loss of information. Also, the calculation for the combined category is just the sum of the calculations for the individual categories. This indicates the usefulness of WOE when binning.

Education level	***WOE***	***Calculation***
1	−1.0986123	0.4394449
2	0.0000000	0.0000000
3	1.6094379	0.6437752
Information Value		**1.0832201**

Gini index:

In order to know how to order the attributes from best to worst risk, one needs to know whether the variable is positively or negatively correlated with the outcome variable. If there is a positive correlation, the higher the values of the variable, the higher their levels of risk, and vice versa. The "slope" of the WOE values is another way of determining how to order the values. If the values of the WOE go from negative to positive, it indicates that smaller values of the variable indicate a higher proportion of defaults and therefore higher risk.

	Cumulative% of				
Other products	***Goods***	***Bads***	$\mathbf{cpN}_i+\mathbf{cpN}_{i-1}$	$\mathbf{cpP}_i-\mathbf{cpP}_{i-1}$	**Z**
1	60%	30%	60%	30%	18%
0	100%	100%	160%	70%	112%
	Gini $= \sum \mathbf{Z} - \mathbf{1}$		**30**%		130%

	Cumulative % of				
Family size	***Goods***	***Bads***	$\mathbf{cpN}_i+\mathbf{cpN}_{i-1}$	$\mathbf{cpP}_i-\mathbf{cpP}_{i-1}$	**Z**
1	10%	10%	10%	10%	1%
2	30%	20%	40%	10%	4%
3	70%	40%	100%	20%	20%
4	90%	70%	160%	30%	48%
5	100%	100%	190%	30%	57%
	Gini $= \sum \mathbf{Z} - \mathbf{1}$		**30**%		130%

The categories with the same WOE (family size of 2 and 3) are again combined, to investigate the effect on the Gini index:

	Cumulative % of				
Family size	***Goods***	***Bads***	$\mathbf{cpN}_i+\mathbf{cpN}_{i-1}$	$\mathbf{cpP}_i-\mathbf{cpP}_{i-1}$	**Z**
1	10%	10%	10%	10%	1%
2&3	70%	40%	80%	30%	24%
4	90%	70%	160%	30%	48%
5	100%	100%	190%	30%	57%
	Gini $= \sum \mathbf{Z} - \mathbf{1}$		**30**%		130%

Again, the Gini index is not influenced by the collapse of the two categories, so it is recommended that the two categories are combined. This lowers the degrees-of-freedom of the variable, without information loss.

	Cumulative % of				
Education level	***Goods***	***Bads***	$\mathbf{cpN}_i+\mathbf{cpN}_{i-1}$	$\mathbf{cpP}_i-\mathbf{cpP}_{i-1}$	**Z**
3	50%	10%	50%	10%	5%
2	80%	40%	130%	30%	39%
1	100%	100%	180%	60%	108%
	Gini $= \sum \mathbf{Z} - \mathbf{1}$		**52**%		152%

The techniques used for the bivariate analysis will determine the inclusion criteria for multivariate analysis. Any variable whose bivariate test has a p-value < 0.25 is a candidate for the multivariate model along with all variables of known business importance. There is not a "statistical" cut-off for the Gini index, but a Gini index of 10% or more, or for sparse data (low default rate), 5% or more can be used as cut-offs. In terms of the information value, characteristics with values of less than 0.1 are typically viewed as weak, while values over 0.3 are sought after.

One problem with a bivariate approach is that it ignores the possibility that a collection of variables, each of which is weakly associated with the outcome, can become an important predictor of outcome when taken together. If this is thought to be a possibility, one should choose a significance level large enough to allow the suspected variables to become candidates for inclusion in the multivariate model.

The issue of variable selection is made more complicated by different analytical philosophies as well as by different statistical methods. One can argue for the inclusion of all scientifically relevent variables into the multivariate model regardless of the results of the bivariate analysis. In general, the appropriateness of the decision to begin the multivariate model with all the possible variables depends on the overall sample size and the number in each outcome group relative to the total number of candidate variables. When the data is adequate to support such analysis, it may be useful to begin the multivariate modeling from this point. When the data is inadequate, however, this approach can produce a numerically unstable multivariate model.

Although bivariate analysis can offer much insight into the prediction ability of variables and is an important first step in scorecard development, it should be done with caution. Evaulating the association between a positive predictor and the outcome variable without taking into account the influence of other variables can sometimes be misleading. In short, bivariate analysis can lead one astray in certain cases if proper care is not taken. The best approach is always to further analyze any variable whose statistics run counter to expectations. On further analysis, one will often find that the relationship one expected between predictors and the outcome is borne if one accounts for the influence of other variables that are correlated with the predictor.

After all possible variables have been analyzed and those that appear to have no predictive ability been eliminated, one can reduce the number of candidate variables further by eliminating those with largely redundant information. It is important to do this, because including redundant variables in the multivariate analysis can:

- Destabilize the parameter estimates
- Increase the risk of over-fitting the model
- Confound interpretation of the coefficients
- Increase computation time

Variable redundancy occurs when two or more variables are predictive of the outcome variable are so correlated with each other that one adds no predictive ability beyond that contained in the other. One of the ways to deal with this issue is to use variable cluster analysis.

7.3 Variable cluster analysis

Dimension reduction is one of the most important data mining tasks to handle data sets with a very large number of variables. Some easy and common supervised dimension reduction tasks can be achieved through simple linear regression, that is, by using R^2 between dependent and independent variables, stepwise regression, and other variants of the regression method. Another popular method is an unsupervised technique that uses principal component analysis. This technique gives very successful dimension reduction results and remedies the multicollinearity problem. Principal component analysis is a popular dimension reduction technique. It provides a good remedy for the multicollinearity problem, but interpretation of the input space is not as good. To overcome the interpretation problem, cluster components are obtained through variable clustering.

There are two typical types of variable clustering techniques. One method is to apply common clustering techniques to any distance matrix of the variables. This method is usually applied to observation clustering. The other method is using variables structure from factor analysis or principal component analysis. The former is widely used and very intuituve. Its performance depends largely on the type of clustering algorithm that is used. The latter is a more expensive process than the former because it requires eigenvalue decomposition and a certain iterative optimization process such as factor rotation.

Both methods of clustering variables are widely used: one is clustering based on a distance matrix, and the other is using latent variables. Both methods are summarized below:

Method 1: Variable clustering that is based on a distance matrix

- Calculate any distance matrix of the variables (e.g. correlation matrix).
- Apply any (observational) clustering algorithm to the distance matrix.
- Obtain clusters that contain homogenous variables.
- (Optional) Calculate cluster components (or first principal components) from each cluster.

Method 2: Variable clustering that is based on latent variables

- Start with all variables to find the first two principal components.
- Perform an orthoblique rotation (quartimax rotation) on eigenvectors.
- Assign each variable to the rotated component with which it has the higher squared correlation.
- Variables are iteratively reassigned to clusters to try to maximize the variance accounted for by the cluster components.
- Stop the iterative assignment when some criteria are met.

One of the latter methods is used with PROC VARCLUS in SAS. The procedure can be used to tell if groups of variables among the total set of possible predictors are highly related. The clustering technique divides variables up into smaller clusters that are as correlated as possible among themselves and as uncorrelated as possible with variables in other clusters. The procedure generates variable cluster structures, identifies key variables within each cluster, and provides non-orthogonal principal components that are called cluster components. The cluster components give much better interpretation than regular principal components, because they consist of only the variables in each cluster.

The VARCLUS procedure attempts to divide a set of variables into non-overlapping clusters in such a way that each cluster can be interpreted as essentially undimensional. For each cluster, PROC VARCLUS computes a component that can be either the first principal component or the centroid component and tries to maximize the sum across all clusters of the variation accounted for by the cluster components.

The VARCLUS procedure is used as a variable reduction method. A large set of variables can often be replaced by a set of cluster components with little loss of information. A given number of cluster components does not generally explain as much variance as the same number of principal components on the full set of variables, but the cluster components are usually easier to interpret than the principal components, even if the latter are rotated.

PROC VARCLUS is a very powerful data analysis tool that can identify clusters, where each variable is assigned to a distinct cluster and provide an audit trail of how the clusters are derived. The resulting formulae could be used to create new variables, but rather than using this directly, one or two variables with the highest potential predictive power are chosen from each for credit scoring purposes.

The VARCLUS algorithm is both divisive and iterative. By default, PROC VARCLUS begins with all variables in a single cluster. The reassingment of variables to clusters occurs in two phases. The first is a nearest component sorting phase. In every iteration, the cluster components are computed and each variable is assigned to the component with which it has the highest squared correlation. The second phase involves a search algorithm in which each variable in turn is tested to see if assigning it to a different cluster increases the amount of variance explained. If a variable is reassigned during the search phase, the components of the two clusters involved are recomputed before the next variable is tested.

By default, PROC VARCLUS begins with all variables in a single cluster. It then repeats the following steps:

- A cluster is chosen for splitting. Depending on the options specified, the selected cluster has either the smallest percentage of variation explained by its cluster component (using the PERCENT= option) or the largest eigenvalue associated with the second principal component (using the MAXEIGEN= option).
- The chosen cluster is split into two clusters by finding the first two principal components, performing an orthoblique rotation (raw quartimax rotation on the eigenvector), and assigning each variable to the rotated component with which it has the higher squared correlation.
- Variables are iteratively reassigned to clusters to maximize the variance accounted for by the cluster components. The reassignment may be required to maintain a hierarchical structure.

The procedure stops splitting when either:

- the maximum number of clusters as specified is reached, or
- each cluster satisfies the stopping criteria specified by the percentage of variation explained and/or the second eigenvalue options.

By default, PROC VARCLUS stops when each cluster has only a single eigenvalue greater than one, thus satisfying the most popular criterion for determining the sufficiency of a single underlying factor dimension.

The iterative reassignment of variables to clusters proceeds in two phases. The first is a nearest component sorting (NCS) phase, similar in principle to the nearest centroid sorting algorithms. In each iteration, the cluster components are computed, and each variable is assigned to the component with which it has the highest squared correlation. The second phase involves a search algorithm in which each variable is tested to see if assigning it to a different cluster increases the amount of variance explained. If a cluster is reassigned during the search phase, the components of the two clusters involved are recomputed before the next variable is tested. The NCS phase is much faster than the search phase but is more likely to be trapped by a local optimum.

One can have the iterative reassignment phases restrict the reassignment of variables such that hierarchical clusters are produced. In this case, when a cluster is split, a variable in one of the two resulting clusters can be reassigned to the other cluster resulting from the split, but not to a cluster that is not part of the original cluster (the one that is split).

If principal components are used, the NCS phase is an alternating least-squares method and converges rapidly. The search phase is very time consuming for a large number of variables and is omitted by default. If the default initialization method is used, the search phase is rarely able to improve the results of the NCS phase. If random initialization is used, the NCS phase may be trapped by a local optimum from which the search phase can escape. If centroid components are used, the NCS phase is not an alternating least-square method and may not increase the amount of variance explained; therefore, it is limited by default, to one iteration.

Lee, Duling, Latour (2008) proposes two-stage variable clustering for large data sets. They propose to use a method that combines supervised and non-supervised methods to overcome the computational difficulties. They propose a method that combines variable clustering that is based on a distance matrix to get global clusters and variable clustering that is based on latent variables to get sub-clusters. Finally a single tree of global clusters and sub-clusters is created. The combined method is described in two stages below:

Stage 1: Variable clustering based on a distance matrix:

- Calculate the correlation matrix of the variables.
- Apply a hierarchical clustering algorithm to the correlation matrix.
- Using a predefined cluster number, assign cluster variables into homogenous groups. The cluster number is generally no more than the integer value of $(nvar/100) + 2$. These clusters are called global clusters.

Stage 2: Variable clustering based on latent variables:

- Run PROC VARCLUS with all variables within each global clusters as one would run a single-stage, variable clustering task.
- For each global cluster, calculate the global cluster components, which are the first principal component of the variables in its cluster.
- Create a global cluster structure using the global cluster components and the same method as the first step in stage 2.
- Form a single tree of variable clusters from the above steps in stage 2.

7.3.1 Interpreting VARCLUS procedure output

The scoring coefficients are coefficients applied to the standardized variables to compute component scores. The cluster structure contains the correlations between each variable and each cluster component. A cluster pattern is not displayed because it would be the same as the cluster structure, except that zeros would appear in the same places in which zeros appear in the scoring coefficients. The intercluster correlations are the correlations among cluster components.

PROC VARCLUS also displays a cluster summary and a cluster listing. The cluster summary gives the number of variables in each cluster and the variation explained by the cluster component. It includes contributions from only the variables in that cluster. The proportion variance explained is obtained by dividing the variance explained by the total variance of the variables in the cluster. If the cluster contains two or more variables and the CENTROID option is not used, the second largest eigenvalue of the cluster is also printed.

The cluster listing gives the variables in each cluster. Two squared correlations are calculated for each cluster. The output of PROC VARCLUS also displays the R^2 value of each variable with its own cluster and the R^2 value with its nearest cluster. The R^2 value for a variable with the nearest cluster should be low if the clusters are well separated. The last column displays the ratio of $(1 - R^2_{own})/(1 - R^2_{nearest})$ for each variable. Small values of this ratio indicate good clustering.

Bivariate analysis and variable clustering can create useful efficiencies in the scorecard building process by helping one pare an unwieldy number of variables down to a more manageable size.

However, knowing the variables one is working with and the business for which the scorecard is being developed should always take precedence over any "rules" of scorecard building. Any variable that experts in the business consider important in explaining the outcome variable should be retained for further analysis, no matter what the statistics and procedures says.

Chapter 8

Different modeling techniques for logistic regression

The previous chapters focused on estimating, testing and interpreting the coefficient in a logistic regression model. The examples discussed were characterized by having few independent variables, and there was perceived to be only one possible model. In the previous chapter, measures were discussed to reduce the number of independent variables, but not methods for fitting the model and obtaining a final model.

Two methods of modeling as shown by Hosmer and Lemeshow (2000) will be discussed in this chapter:

- Stepwise logistic regression
- Best subsets logistic regression

Both of these methods are utilized in credit scoring methodologies and it is often the choice of the developer which one is used.

8.1 Stepwise logistic regression

In stepwise logistic regression, variables are selected for inclusion or exclusion from the model in a sequential fashion based solely on statistical criteria. The stepwise approach is useful and intuitively appealing in that it builds models in a sequential fashion and it allows for the examination of a collection of models which might not otherwise have been examined.

The two main versions of the stepwise procedure are forward selection followed by a test for backward elimination or backward elimination followed by forward selection. Forward selection starts with no variables and selects variables that best explains the residual (the error term or variation that has not yet been explained.) Backward elimination starts with all the variables and removes variables that provide little value in explaining the response function. Stepwise methods are combinations that have the same starting point by consider inclusion and elimination of variables at each iteration. SAS has an option that allows the modeler to perform this type of analysis.

Any stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm that checks for the "importance" of variables and either includes or excludes them on the basis of a fixed decision rule. The "importance" of a variable is defined in terms of a measure of statistical significance of the coefficient for the variable. The statistic used depends on the assumptions of the model. In stepwise linear regression an F-test is used since the errors are assumed to be normally distributed. In logistic regression the errors are assumed to follow a binomial distribution, and the significance of the variable is assessed via the likelihood ratio chi-square test. At any step in the procedure the most important variable, in statistical terms, is the one that produces the greatest change in the log-likelihood relative to a model not containing the variable.

As discussed earlier, a polychotomous variable with k levels is appropriately modeled through its $k-1$ design/dummy variables. Since the magnitude of G depends on its degrees of freedom, any procedure based on the likelihood ratio test statistic, G, must account for the possible differences in degrees of freedom between variables. This is done by assessing the significance through the p-value for G.

The algorithm for forward selection followed by backward elimination in stepwise logistic regression is described below. Any variants of this algorithm are simple modifications of this procedure. The method is described by considering the statistical computations that a computer must perform at each step of the procedure.

8.1.1 Step 0

Suppose that a total of p possible independent variables are available, all of which are judged to be predictive in studying the outcome variable (using methods described in the previous chapter.) Step 0 starts with a fit of the "intercept only model" and an evaluation of its log-likelihood, L_0. This is followed by fitting each of the p possible univariate logistic regression models and comparing their respective log-likelihoods. Let the value of the log-likelihood for the model containing the variable x_j at step 0 be denoted by $L_j^{(0)}$. (The subscript j refers to the variable that has been added to the model and the superscript (0) refers to the step and this notation will be used to keep track of both the step number and the variable in the model.)

Let the value of the likelihood ratio test for the model containing x_j versus the intercept only model be denoted by $G_j^{(0)} = -2(L_0 - L_j^{(0)})$ and its p-value be denoted by $p_j^{(0)}$. This p-value is determined by the tail probability

$$P[\chi^2(\nu) > G_j^{(0)}] = p_j^{(0)}$$

where $\nu = 1$ if x_j is continuous and $\nu = k-1$ if x_j is polytochomous with k categories.

The most important variable is the one with the smallest p-value. If this variable is denoted by x_{e_1}, then $p_{e_1} = \min(p_j^{(0)})$, where min stands for selecting the minimum of the quantities enclosed in the brackets. The subscript e_1 is used to denote that the variable is a candidate for entry at step 1. For example, if x_2 had the smallest p-value, then $p_2^{(0)} = \min(p_j^{(0)})$, and $e_1 = 2$. Just because x_{e_1} is the most important variable, there is no guarantee that it is "statistically significant". For example, if $p_{e_1}^{(0)} = 0.83$, the most probable conclusion is that there is little point in continuing the analysis, because the "most important" variable is not related to the outcome. On the flip side, if $p_{e_1}^{(0)} = 0.003$, one would like to look at the logistic regression containing this variable and see if there are any other variables that are important, given that x_{e_1} is in the model.

A crucial aspect of using stepwise logistic regression is the choice of an "alpha" (α) level to judge the importance of the variables. Lee and Koval have done some research on the appropriate significance level for forward stepwise logistic regression and it will be discussed in the following chapter. Let p_E denote the significance level chosen, where E stands for entry. Whatever the choice for p_E, a variable is judged important enough to include in the model if the p-value for G is less than p_E. Thus, the program proceeds to step 1 if $p_{e_1}^{(0)} < p_E$, otherwise it stops.

8.1.2 Step 1

Step 1 commences with a fit of the logistic regression model containing x_{e_1}. Let $L_{e_1}^{(1)}$ denote the log-likelihood of this model. To determine whether any of the remaining $p-1$ variables are important once the variable x_{e_1} is in the model, the $p-1$ logistic regression models containing x_{e_1} and x_j are fitted, where $j = 1, 2, 3, ..., p$ and $j \neq e_1$. For the model containing x_{e_1} and x_j, let the log-likelihood be denoted by $L_{e_1 j}^{(1)}$, and let the likelihood ratio chi-square statistic of this model versus the model containing only x_{e_1}be denoted by $G_j^{(1)} = -2(L_{e_1}^{(1)} - L_{e_1 j}^{(1)})$. The p-value for this statistic is denoted by $p_j^{(1)}$. Let the variable with the smallest p-value at step 1 be $x_{e_2}$where $p_{e_2} = min(p_j^{(1)})$. If this value is less than p_E, proceed to step 2, otherwise stop.

8.1.3 Step 2

Step 2 begins with the fit of the model containing both x_{e_1} and x_{e_2}. It is possible that once x_{e_2} has been added to the model, x_{e_1} is no longer important. Thus, step 2 includes a check for backward elimination. In general this is accomplished by fitting models that delete one of the variables added in the previous steps and assessing the continued importance of the variable removed.

At step 2, let $L_{-e_j}^{(2)}$ denote the log-likelihood of the model with x_{e_j} removed. In similar fashion let the likelihood ratio test of this model versus the full model at step 2 be $G_{e_j}^{(2)} = -2(L_{-e_j}^{(2)} - L_{e_1 e_2}^{(2)})$ and $p_{-e_j}^{(2)}$ be its p-value. To ascertain whether a variable should be deleted from the model, the program selects that variable which, when removed, yields the maximum p-value. Denoting this variable as x_{r_2}, then $p_{r_2}^{(2)} = \max(p_{-e_j}^{(2)}, p_{-e_2}^{(2)})$. To decide whether x_{r_2} should be removed, the program compares $p_{r_2}^{(2)}$ to a second pre-chosen "alpha" (α) level, p_R, where "R" stands for remove. Whatever the value chosen for p_R, it must exceed the value of p_E to guard against the possibility of having the program enter and remove the same variable at successive steps. If one does not wish to exclude many variables once they have entered, then one might choose $p_R = 0.9$. A more stringent value would be used if a continued "significant" contribution were required. For example, if $p_E = 0.15$ is used, then one might choose $p_R = 0.2$. If the maximum p-value to remove, $p_{r_2}^{(2)}$, exceeds p_R then x_{r_2} is removed from the model. If $p_{r_2}^{(2)}$ is less than p_R, then x_{r_2} remains in the model. In either case, the program proceeds to the variable selection phase.

At the forward selection phase, each of the $p-2$ logistic regression models are fit containing x_{e_1}, x_{e_2} and x_j, $j = 1, 2, 3, ..., p$, $j \neq e_1, e_2$. The program evaluates the log-likelihood for each model, computes the likelihood ratio test versus the model containing only x_{e_1} and x_{e_2} and determines the corresponding p-value. Let x_{e_3} denote the variable with the minimum p-value, i.e., $p_{e_3}^{(2)} = min(p_j^{(2)})$. If this p-value is smaller than p_E, $p_{e_3}^{(2)} < p_E$, then the program proceeds to step 3, otherwise it stops.

8.1.4 Step 3

The procedure for step 3 is identical to that of step 2. The program fits the model including the variable selected during the previous step, performs a check for backward elimination followed by forward selection. The process continues in this manner until the last step, step S.

8.1.5 Step S

This step occurs when:

- All p-variables have entered the model, or
- All the variables in the model have p-values to remove that are less than p_R, and the variables not included in the model have p-values to enter that exceed p_E.

The model at this step contains those variables that are important relative to the criteria p_E and p_R. These may or may not be the variables reported in a final model. For instance, if the chosen values of p_E and p_R correspond to the modeler's belief for statistical significance, then the model at step S may well contain the significant variables. However, if the modeler used values for p_E and p_R which are less stringent, the business might choose to select the variables for a final model from a table that summarizes the results of the stepwise procedure.

There are two methods that may be used to select variables from a summary table; these are comparable to methods commonly used in a stepwise linear regression. The first method is based on the p-value of entry at each step, while the second is based on a likelihood ratio test of the model at the current step versus the model at the last step.

Let q denote the arbitrary step in the procedure. In the first method one compares $p_{e_q}^{(q-1)}$ to a prechosen significance level such as $\alpha = 0.15$. If the value of $p_{e_q}^{(q-1)}$ is less than α, then one moves to step q. When $p_{e_q}^{(q-1)}$ exceeds α, one stops. Consider the model at the previous step for further analysis. In this method the criterion for entry is based on a test of the significance of the coefficient for x_{e_q} conditional on $x_{e_1}, x_{e_2}, ..., x_{e_{q-1}}$ being in the model. The degrees-of-freedom for the test are 1 or $k-1$, depending on whether x_{e_q} is continuous or polychotomous with k categories.

In the second method, one compares the model at the current step, step q, not the model at the previous step, step q-1, but to the model at the last step, step S. One evaluates the p-value for the likelihood ratio test of these two models and proceed in this fashion until this p-value exceeds α. This tests that the coefficients for the variables added to the model from step q to step S are all equal to zero. At any given step it has more degrees-of-freedom than the test employed in the first method. For this reason the second method may possibly select a larger number of variables than the first method.

It is well known that the p-values calculated in stepwise selection procedures are not p-values in the traditional hypothesis testing context. Instead, they should be interpreted as indicators of relative importance among variables. It is recommended that one err in the direction of selecting a relatively rich model following stepwise selection. The variables so identified should then be subjected to more intensive analysis, using similar methods as in variable selection.

A common modification of the normal stepwise selection procedure is to begin with a model at step 0 which contains known important covariates. Selection is the performed from among other variables.

One disadvantage of the normal stepwise selection procedures is that the maximum likelihood estimates for the coefficients of all variables not in the model must be calculated at each step. For large data sets with large numbers of variables this can be quite time consuming. An alternative to a full maximum likelihood analysis is available in SAS which selects new variables based on the Score test for the variables not included in the model. Another alternative which is also less time consuming is based on a multivariate Wald test. Although these are different selection methods, it does seem likely that an important variable will be identified, regardless of the method used.

One must be cautious when considering a model with many variables as significant regressions may be obtained from "noise" variables, completely unrelated to the outcome variable. A thorough analysis that examines statistical and business significance is
essential following any stepwise method.

8.2 Best subsets logistic regression

An alternative to stepwise selection of variables for a model is best subset selection. In this approach a specified number of "best" models containing one, two, and three up to all variables is fitted.

Best subsets logistic regression may be performed in a straight forward manner using any program capable of best subsets linear regression. SAS have implemented a best subsets option in its logistic regression modules.

Applying best subsets linear regression software to perform best subsets logistic regression is most easily explained using vector and matrix notation. Let $\mathbf{X}$ denote the $n \times (p+1)$ matrix containing the values of all p independent variables for each subject with the first column containing $\mathbf{1}$ to represent the constant term.

The p variables may represent the total number of variables, or those selected after univariate analysis of the variables before model building. Let $\mathbf{V}$ denote a $n \times n$ diagonal matrix with general element $v_i = \widehat{\pi}_i(1-\widehat{\pi}_i)$, where $\widehat{\pi}_i$ is the estimated logistic probability computed using the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ and the data for the i^{th} case x_i.

The expression for $\mathbf{X}$ and $\mathbf{V}$ are as follows:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

and

$$\mathbf{V} = \begin{pmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & 0 & \ldots & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & 0 & \ldots & 0 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & \ldots & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{pmatrix}.$$

As noted before, the maximum likelihood estimate is determined iteratively. It can be shown that:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{z}$$

where $\mathbf{z} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{V}^{-1}\mathbf{r}$ and $\mathbf{r}$ is the vector of residuals, $\mathbf{r} = (\mathbf{y} - \widehat{\boldsymbol{\pi}})$.

This representation of $\widehat{\boldsymbol{\beta}}$ provides the basis for using linear regression software. Linear regression packages that allows weights, produce coefficient estimates identical to $\widehat{\boldsymbol{\beta}}$ when used with $\mathbf{z}_i$ as the dependent variable and case weights v_i equal to the diagonal elements of $\mathbf{V}$. To replicate the results of the maximum likelihood fit from a logistic regression package using a linear regression package, one can calculate for each case, the value of the dependent variable as follows:

$$\begin{aligned} z_i &= (1, \mathbf{x}_i')\widehat{\boldsymbol{\beta}} + \tfrac{(y_i - \widehat{\pi}_i)}{\widehat{\pi}_i(1-\widehat{\pi}_i)} \\ &= \widehat{\beta}_0 + \textstyle\sum_{j=1}^{p} \widehat{\beta}_j x_{ij} + \tfrac{(y_i - \widehat{\pi}_i)}{\widehat{\pi}_i(1-\widehat{\pi}_i)} \\ &= \ln(\tfrac{\widehat{\pi}_i}{1-\widehat{\pi}_i}) + \tfrac{(\mathbf{y}_i - \widehat{\pi}_i)}{\widehat{\pi}_i(1-\widehat{\pi}_i)} \end{aligned}$$

and a case weight

$$v_i = \widehat{\pi}_i(1-\widehat{\pi}_i).$$

Note that all one needs is access to the fitted values, $\widehat{\pi}_i$, to compute the values of z_i and v_i. Next, run a linear regression program using the values of z_i as the dependent variable, the values of $\mathbf{x}_i$ for the vector of independent variables and the values of v_i as the case weights.

Proceeding with the linear regression, the residuals from the fit are:

$$(z_i - \widehat{z}_i) = \frac{(y_i - \widehat{\pi}_i)}{\widehat{\pi}_i(1 - \widehat{\pi}_i)}$$

and the weighted residual sum-of-squares produced is:

$$\sum_{i=1}^{n} v_i(z_i - \widehat{z}_i)^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i(1 - \widehat{\pi}_i)}$$

which is χ^2, the Pearson chi-square statistic from a maximum likelihood logistic regression program. It follows that the mean residual sum-of-squares is $s^2 = \chi^2/(n-p-1)$. The estimates of the standard error of the estimated coefficients produced by the linear regression program are s times the square root of the diagonal elements of the matrix $(\mathbf{X'VX})^{-1}$. To obtain the correct values given by $SE(\widehat{\beta}_j) = \sqrt{\widehat{Var}(\widehat{\beta}_j)}$, one needs to divide the estimates of the standard error produced by the linear regression by s, the square root of the mean square error (standard error of the estimate).

The subsets of variables selected for "best" models depend on the criterion chosen for "best". In best subsets linear regression, three criteria have primarily been used to select variables. Two of these are based on the the concept of the proportion of the total variation explained by the model. These are R^2, the ratio of the regression sum-of-squares to the total sum-of-squares, and adjusted R^2 (or AR^2), the ratio of the regression mean squares to the total mean squares. Since the adjusted R^2 is based on mean squares rather than sums-of-squares, it provides a correction for the number of variables in the model. If R^2 is used, the best model is always the model containing all p variables, a result that is not very helpful. An extension for best subsets logistic regression is to base the R^2 measure on deviance, rather than the Pearson chi-square.

However, the use of R^2 is not recommended for best subsets logistic regression. The third measure is a measure of predictive squared error, developed by Colin Mallows, C_q. It is normally denoted as C_p, but here p refers to the total number of possible variables, so it is denoted as C_q where q refers here to some subset of variables. Mallows proposed this statistic as a way of facilitating comparisons among many alternative subset regressions in 1973.

For a subset q of p variables:

$$C_q = \frac{\chi^2 + \lambda^*}{\chi^2/(n-p-1)} + 2(q+1) - n$$

where $\chi^2 = \sum \left\{(y_i - \widehat{\pi}_i)^2 / \left[\widehat{\pi}_i(1 - \widehat{\pi}_i)\right]\right\}$, the Pearson chi-square statistic from the model with p variables and λ^* is the multivariate Wald test statistic for the hypothesis that the coefficients for the $p-q$ variables not in the model are equal to zero. Under the assumption that the model fit is the correct one, the approximate expected values of χ^2 and λ^* are $(n-p-1)$ and $p-q$ respectively. Substitution of these approximate expected values into the expression for C_q yields $C_q = q+1$. Hence, models with C_q near $q+1$ are candidates for the best model.

Use of best subset linear regression should help select, in the same way as its application in linear regression does, a core of q important covariates from the p possible covariates.

Some programs, like SAS Proc Logistic, provides best subsets selection of covariates based on the Score test for the variables in the model. For example, the best two variable model is the one with the largest Score test among all the two variable models. The output lists the covariates and Score test for a user specified number of best models of each size. The difficulty one faces when presented with this output is that the Score test increases with the number of variables in the model. An approximation for C_q can be obtained from Score test output in a survival time analysis. The first assumption is that the Pearson chi-square test statistic is equal to its mean, e.g.$\chi^2 \approx (n-p-1)$. The next assumption is that the Wald test statistic for the $p-q$ excluded covariates may be approximated by the difference between the values of the Score test for all p covariates and the Score test for q covariates, namely $\lambda_q^* \approx S_p - S_q$. This results in the following approximation:

$$\begin{aligned} C_q &= \frac{\chi^2+\lambda^*}{\chi^2/(n-p-1)} + 2(q+1) - n \\ &\approx \frac{(n-p-1)+(S_p-S_q)}{1} + 2(q+1) - n \\ &\approx S_p - S_q + 2q - p + 1 \end{aligned}$$

The value S_p is the Score test for the model containing all p covariates and is obtained from the computer output. The value of S_q is the Score test for the particular subset of q covariates and its value is also obtained from the output.

The advantage of best subsets logitic regression is that many more models can be quickly screened than was possible with other modeling appproaches to variable identification. There is, however, one potential disadvantage with the best subset approach: one must be able to fit the model containing all possible covariates. In analyses that include a large number of variables, this may not be possible. Numerical problems can occur when one over-fits a logistic regression model. If the model has many variables, one runs the risk that the data are too thin to be able to estimate all the parameters. If the full model proves to be too rich, then some selective weeding out of obviously unimportant variables with univariate tests may remedy this problem. Another approach is to perform the best subset analysis using several smaller "full" models.

As is the case with any statistical selection method, the business importance of all variables should be addressed before any model is accepted as the final model.

8.3 Testing the importance of variables in the model

Following the fit of the multivariate model, the importance of each variable included in the model should be verified. This should include an examination of the Wald statistic for each variable and a comparison of each estimated coefficient with the coefficient from the model containing only that variable. Variables that do not contribute to the model based on these criteria should be elminated and a new model should be fit. The new model should be compared to the old, larger model used in the likelihood ratio test. Also, the estimated coefficients for the remaining variables should be compared to those from the full model. In particular, one should be concerned about variables whose coefficients have changed markedly in magnitude. This indicates that one or more of the excluded variables was important in the sense of providing a needed adjustment of the effect of the variable that remained in the model. The other explanation could be that one or more of the excluded vairables have been highly correlated with the variables in the model and multicollinearity occurred.

Another check is to look at the overall Gini index of the model and the Gini index when the specific variable is excluded from the model. If there is not much of a difference whether the variable is included or not, it might indicate that the variable is not needed in the model. Keeping such variables in the model might result in an unstable model that is over-fitted to the data sample. In credit scoring, it is important that the model can be generalized, so a more parsimonious model is always preferred. This process of deleting, refitting and verifying continues until it appears that all of the important variables are included in the model and those excluded are clinically and/or statistically unimportant.

Any variable that was not selected for the initial multivariate model can now be added back into the model. This step can be helpful in identifying variables that, by themselves, are not significantly related to the outcome but make an important contribution in the presence of other variables.

8.4 Analysis of the variables in the model

Once a model has been obtained that is felt to contain all the essential variables, one should look more closely at the variables in the model. The question of appropriate categories should be addressed at the univariate stage. For continuous variables, one should check the assumption of linearity in the logit. It is common in credit scoring though, that even continuous variables are bucketed and used as categorical variables in the model. This eases implementation, which is a hugely important in credit scoring.

Chapter 9

Best significance level in forward stepwise logistic regression

The process of selecting a subset of variables from a large number of variables is called model-building. The main purpose of model-building in credit scoring is prediction. Forward stepwise logistic regression is widely used in credit scoring. This procedure, which was explained in detail in the previous chapter, involves selection and stopping criteria. The standard stopping criterion is the χ^2 based on a fixed α level. The usual conventional value for α has been 0.05. It is not known whether $\alpha = 0.05$ is the best value for the purpose of prediction in the logistic model. In this chapter research that was done by Lee and Koval(1997) will be given. They used Monte Carlo simulations to determine the best significance level for the stopping criterion $\chi^2(\alpha)$ in conjunction with forward stepwise logistic regression in terms of the estimated true error rate of prediction ($\widehat{ERR}$).

9.1 Performance criterion

Estimated true error rate of prediction ($\widehat{ERR}$) is used as the performance criterion and is defined as:

$$(\widehat{ERR}) = ARR + \widehat{\omega}$$

where ARR is the apparent error rate of the prediction and $\widehat{\omega}$ is an estimate for the bias of ARR. The apparent error rate is estimated by the resubstitution method and this tends to underestimate the true error rate because the data is used twice, both to fit the model and to evaluate its accuracy.

There are several nonparametric methods of estimating ω including cross-validation, jack-knifing and bootstrapping. However, estimates of ω must be computed at each step in the forward stepwise procedure. Thus, these nonparametric methods of estimating ω are not always feasible as it requires a great deal of computing time.

9.2 Selection criteria

Suppose $k-1$ variables have been previously selected and the k^{th} variable is considered for inclusion in the model. The components of $\boldsymbol{\beta}$ can be partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_{k-1}, \beta_k)$. The hypotheses of interest are $H_0 : \beta_k = 0$ and $H_1 : \beta_k \neq 0$. Let $\widehat{\boldsymbol{\beta}}^0$ denote the maximum likelihood estimation under the null hypothesis (restricted MLE) and $\widehat{\boldsymbol{\beta}}^1$ denote the maximum likelihood estimation under the alternative hypothesis (unrestricted MLE), that is $\widehat{\boldsymbol{\beta}}^0 = (\widehat{\boldsymbol{\beta}}_{k-1}, 0)$ and $\widehat{\boldsymbol{\beta}}^1 = (\widehat{\boldsymbol{\beta}}_{k-1}, \widehat{\beta}_k)$.

Three selection criteria were used in this study. The likelihood ratio statistic (LR) is:

$$\begin{aligned} LR &= 2\left[L(\widehat{\boldsymbol{\beta}}^1) - L(\widehat{\boldsymbol{\beta}}^0)\right] \\ &= 2\left[L(\widehat{\boldsymbol{\beta}}_{k-1}, \widehat{\beta}_k) - L(\widehat{\boldsymbol{\beta}}_{k-1}, 0)\right] \end{aligned}$$

where L is the log-likelihood function.

The Wald Statistic (WD) is defined as

$$WD = \widehat{\beta}_k \mathbf{C}^{-1}_{k\times k} \widehat{\beta}_k$$

where $\mathbf{C}$ is the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$.

The score statistic is

$$\begin{aligned} SC &= \mathbf{U}'(\widehat{\boldsymbol{\beta}}^0)\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}}^0)\mathbf{U}(\widehat{\boldsymbol{\beta}}^0) \\ &= \mathbf{U}'(\widehat{\boldsymbol{\beta}}_{k-1}, 0)\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}}_{k-1}, 0)\mathbf{U}(\widehat{\boldsymbol{\beta}}_{k-1}, 0) \end{aligned}$$

where $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta})$ is the efficient score vector, and $\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{U}(\boldsymbol{\beta})\right]$ is Fisher's information matrix.

These three criteria were used with the standard stopping criterion χ^2_α in the forward stepwise logistic regression. 19 possible values of α were considered, from 0.05 to 0.95 in steps of 0.05. The best α level is defined to be that for which the grand mean of $\widehat{ERR}$ over all sampling situations is a minimum.

9.3 Monte Carlo Experimental Design

The use of real data is of limited value in evaluating estimators, as they offer a limited range of sample size, number of predictor variables and distribution of dependent and predictor variables. The choice of the data set may influence the conclusions and limit the generalizability of the results. Simulation studies can be used to obtain results over a wide range of sampling situations.

There are four steps in the general design of a simulation experiment:

9.3.1 Step 1:

The process starts with the generation of predictor variables, $\mathbf{X}_i$, $i = 1, 2, ..., N$.

Generate $(\mathbf{X}_i|Y = 0)$, $i = 1, 2, ...n_0$ from population Π_0 and generate $(\mathbf{X}_i|Y = 1)$, $i = 1, 2, ...n_1$ from population Π_1, with $N = n_0 + n_1$.

9.3.2 Step 2:

This step computes $\widehat{\boldsymbol{\beta}}$, the maximum likelihood estimates of $\boldsymbol{\beta}$. $\widehat{\beta}_i$, $i = 1, 2, ..., p$ are obtained iteratively via reweighted least squares.

9.3.3 Step 3:

Computations of estimated probabilities $P(Y = 1|\mathbf{X}_i) = \widehat{\pi}_i(\mathbf{X})$. $\widehat{\pi}_i(\mathbf{X})$, $i = 1, 2, ..., N$ are computed using $\mathbf{X}_i$ from step 1 and $\boldsymbol{\beta}$ from step 2.

9.3.4 Step 4:

In this step, the predicted dependent variable is generated,$\widehat{Y}_i$. $\widehat{Y}_i$ is equal to 0 if $\widehat{\pi}_i(\mathbf{X}) \leq \frac{1}{2}$ and $\widehat{Y}_i$ is equal to 1 if $\widehat{\pi}_i(\mathbf{X}) > \frac{1}{2}$.

Repeat steps 1 through 4 twenty times.

Two multivariate distributions were considered for the predictor variables, the multivariate normal and the multivariate binary.

9.4 Multivariate normal case

Suppose that $\mathbf{X} \sim N_p(0, \Sigma)$ in population Π_0 and $\mathbf{X} \sim N_p(\mu, \Sigma)$ in population Π_1. The parameters μ and Σ are reparameterized in terms of four factors, P, V, Δ^2 and D.

The first factor P is the number of predictor variables. The second factor $V \epsilon (0, 1]$ determines the eigenvalues λ_i, $i = 1, 2, ..., P$ of $\boldsymbol{\Sigma}$ by the means of the expression

$$\lambda_i = aV^{i-1} + \delta, \text{ for } i = 1, 2, ..., P$$

where

$$a = \begin{cases} 0.9P(1 - V)(1 - V^P) \text{ if } 0 < V < 1 \\ 1 - \delta, \text{ if } V = 1 \end{cases}.$$

A value of $\delta = 0.1$ was chosen as a lower bound on the smallest eigenvalue λ_p to avoid the difficulties of nearly singular matrices. The eigenvalues reflect the degree of interdependence among the predictor variables. The variables are highly dependent near $V = 0$ and highly independent near $V = 1$. Since $\boldsymbol{\Sigma} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'$, where $\mathbf{E}$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues λ_i, then once the λ_i were specified, a random orthogonal matrix $\mathbf{E}$ was generated and used to create $\boldsymbol{\Sigma}$.

The third factor is the Mahalanobis distance between Π_0 and Π_1 defined by $\Delta^2 = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. It describes the separation of the two populations. The fourth factor D determines the elements μ_i of the vector $\boldsymbol{\mu}$. As D varies from 0 to 1, the rate of increase in Δ^2 decreases as the number of included variables increases from 1 to P. Let

$$\mu_i^* = (bD^{i-1})^{1/2} \text{ for } i = 1, 2, ..., P \text{ and } 0 < D \leq 1$$

where

$$b = \begin{cases} \Delta^2(1 - D)/(1 - D^p) \text{ if } 0 < D < 1 \\ \Delta^2/P, \text{ if } D = 1 \end{cases}.$$

Elements μ_i of $\boldsymbol{\mu}$ were then obtained from $\boldsymbol{\mu} = \mathbf{R}\boldsymbol{\mu}^*$ where $\boldsymbol{\Sigma} = \mathbf{R}\mathbf{R}'$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$.

For a P-variate observation $\mathbf{X}$ from $N_p \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, P independent $N(0,1)$ values $Z's$ were first generated. The vector $\mathbf{Z}$ was then transformed to the required vector $\mathbf{X}$ by $\mathbf{X} = \boldsymbol{\mu} + \mathbf{R}\mathbf{Z}$, with $\boldsymbol{\Sigma} = \mathbf{R}\mathbf{R}'$ as above.

The levels of the five factors P, V, Δ^2, D and N must be specified. In this study Lee and Koval (1997) used a second-order central surface design. This design allows for the fit of a model with linear and quadratic effects in all factors and first-order interactions between all factors. The former was of particular interest in this study because of the possibility of a linear trend of the best α in one or more of the factors. In addition this design permitted the examination of five levels of all five factors using only 48 sampling situations (combination of levels). The alternative of a 3^5 factorial design would only evaluate three levels of each factor, yet demand 243 sampling situations. This is very time-consuming, even on the fastest computer.

To describe the experimental design, each point in the factor space is regarded as a quintuplet of factor levels, written symbolically as (P, V, Δ^2, D, N). Each factor has five levels, which are taken to be equally spaced on a suitable scale and are coded as $(-2, -1, 0, 1, 2)$. These levels are termed "low star", "low factorial", "center", "high factorial", and "high star" respectively. The values of the 5 factors for this study is given in the table below:

	Level (Code)				
	Low star	Low factorial	Center	High factorial	High star
Factor	(-2)	(-1)	(0)	(1)	(2)
P	5	10	15	20	25
V	0.2	0.4	0.6	0.8	1
Δ^2	1.0	1.5	2.0	2.5	3.0
D	0.2	0.4	0.6	0.8	1.0
N	100	150	200	250	300

The design consisted of 48 sampling situations or points of three types:

- $2^5 = 32$ factorial points which were all possible combinations of the ± 1 levels for each factor
- 10 star points which have the -2 or $+2$ level, and
- 6 center points which have the 0 levels.

Note the following:

- The values of Δ^2 follows a narrow range, because small values of Δ^2 provide no discrimination between populations and large values lead to complete separation which means that the likelihood estimates are not unique.
- Logistic regression models require iterative solution, so the sample size must be a reasonable multiple of the number of parameters.
- Two equal-sized samples were drawn from each population, that is $n_0 = n_1 = N/2$.

9.5 Multivariate binary case

The multivariate binary variables were generated from the second-order Bahadur model. Let $\mathbf{X} = (X_1, X_2, ..., X_P)$, where X_j is a Bernoulli random variable with $p_j = P(X_j = 1)$ and $1 - p_j = P(X_j = 0)$, $j = 1, 2, ..., P$. Set $Z_j = (X_j - p_j)/\left[p_j(1-p_j)\right]^{1/2}$ and $\rho(jk) = E(Z_j Z_k)$. According to Lee and Koval (1997), the second-order Bahadur model for each population is then given by

$$\begin{aligned} f_i(\mathbf{x}) &= P(\mathbf{X} = \mathbf{x}|\Pi_i) \\ &= \prod_{j=1}^{P} p_{ij}^{x_{ij}}(1-p_{ij})^{1-x_{ij}} \left[1 + \textstyle\sum_{j<k} \rho_i(jk) z_{ij} z_{ik}\right], \; i = 0, 1 \end{aligned}$$

For simplicity, it was assumed that

$$\rho_i(jk) = \rho_i, \text{ for all } j \neq k, i = 0, 1$$

and

$$p_{ij} = p_i, \text{ for all } j, i = 0, 1.$$

When higher-order correlations are set to zero in the original Bahadur model to yield to his second-order model, severe limitations are placed on $\rho_i(jk)$ to ensure that the $f_i(\mathbf{x})$ is a proper density function. Letting $t = \sum_{j=1}^{P} x_j$, Bahadur has shown that the second-order Bahadur model yields a valid probability distribution if and only if

$$-\frac{2}{P(P-1)} Min\left(\frac{p}{1-p}, \frac{1-p}{p}\right) \leq \rho \leq \frac{2p(1-p)}{(P-1)p(1-p) + 1/4 - \gamma_0}$$

where

$$\gamma_0 = Min_t \left\{[t - (P-1)p - 1/2]^2\right\} \leq 1/4.$$

Let ρ_{upp_i} denote the upper bound of the above equation for population i. In any simulation in the multivariate binary after having defined all the other parameters, Lee and Koval set $\rho_i = \rho_{upp_i}$.

The simulations of the multivariate binary is made easier by the fact that the elements of the vector $\mathbf{x}$ can be generated recursively. Suppose that $x_1, x_2, ..., x_{k-1}$ have been generated. The conditional distribution of x_k can be expressed as

$$f(x_k|x_i, ..., x_{k-1}) = \frac{p_i^{x_k}(1-p_i)^{1-x_k}\left[1 + \sum_{j<k} \rho(jk) z_j z_k\right]}{\left[1 + \sum_{j<(k-1)} \rho(jk) z_j z_k\right]}.$$

If u is an observation from a uniform random distribution on $[0, 1]$, then x_k is defined according to the rule

$$x_k = \begin{cases} 0, & \text{if } u \leq f(x_k = 0|x_1, ..., x_{k-1}) \\ 1, & \text{otherwise} \end{cases}.$$

For the multivariate binary case, a full factorial design was defined with 3 levels for each of 4 factors. For the factors which the multivariate binary share with the multivariate normal, the levels are those central (rather than extreme) to the response surface design, that is, those denoted as low factorial, center and high factorial. Thus, Lee and Koval used $P = 10$, 15 and 20; $N = 150$, 200 and 250. p_0 and p_1 was also taken such that $p_0 < p_1$; $p_0 = 0.2$ with $p_1 = 0.3$, 0.4 and 0.5; $p_0 = 0.4$ with $p_1 = 0.5$, 0.6 and 0.7; and $p_0 = 0.6$ with $p_1 = 0.7$, 0.8 and 0.9. These 9 pairs of (p_0, p_1) give rise to 3 levels of p_0 (0.2, 0.4 and 0.6) and 3 levels of $(p_1 - p_0)$ (0.1, 0.2 and 0.3). Let the symbols B and M denote p_0 and $(p_1 - p_0)$, respectively.

9.6 Results of the simulations experiments

This section presents the results found by Lee and Koval (1997)of the sampling experiments in the multivariate normal and multivariate binary cases. It has two main purposes: (1), to recommend the best α level of significance for the $\chi^2_{(\alpha)}$ stopping criterion; and (2), to investigate the effects of the five factors P, V, Δ^2, D and N on the best α level in the multivariate normal case, and of the four factors P, B, M and N on the best α level in the multivariate binary case.

9.6.1 Multivariate Normal Case

For each of the three selection criterion, the mean (over all 48 sampling situations) of ARR and of Bias were plotted against the α level of significance. This gave a monotonically decreasing ARR and a monotonically increasing Bias function of the α level. In other words, ARR and Bias are monotonically decreasing and increasing, respectively, functions of the number of predictor variables in the model.

The mean (over all 48 sampling situations) of $\widehat{ERR} = (ARR + Bias)$ decreases from a value of 0.167 at $\alpha = 0.05$ to a minimum of 0.164 at $\alpha = 0.20$ increasing again to a maximum of 0.178 at $\alpha = 0.95$. The minimum value of $\widehat{ERR}$ occurs at $\alpha = 0.20$, for all three selection criteria. The best α levels are between 0.05 and 0.4; and are the same for all three selection criteria.

Analysis of variance was done on the response surface design. The lack-of-fit test was not significant ($p = 0.308$) and it was therefore concluded that the quadratic surface fits the data well. Only the factor P was statistically significant ($p < 0.01$) while the factor D is marginally not significant ($p = 0.09$). A plot of the best α level against P showed a linear function of the form "best $\alpha = P/100$".

9.6.2 Multivariate Binary Case

As for the multivariate normal case, the mean of ARR is an increasing function, and the mean of Bias is a decreasing function of the α of the stopping criterion. Moreover, the mean of $\widehat{ERR} = (ARR + Bias)$ decreases from a value of 0.260 at $\alpha = 0.05$ to a minimum of 0.257 at $\alpha = 0.15$ and increasing again to a maximum of 0.266 at $\alpha = 0.95$. The minimum value of $\widehat{ERR}$ occurs at $\alpha = 0.15$ for all three selection criteria. The best α level for 81 sampling situations were between 0.05 and 0.40 as in the multivariate normal case.

An analysis of variance was employed to assess the effects of the four factors P, B, M and N on the best α levels. It was assumed that the effects of third-order and fourth-order interactions could be ignored.

The factor M and P were highly significant ($p < 0.001$) and the factor B was moderately significant ($p = 0.020$), whereas the interaction PB was marginally non-significant ($p = 0.076$).

9.7 Conclusion

Lee and Koval (1997) studied the determinants of the best a level for the $\chi^2_{(\alpha)}$ stopping criterion for the purpose of prediction. The best choice of α varies between 0.05 and 0.4 in both multivariate normal and multivariate binary cases.

In the multivariate normal case, the choice of α depends upon the factor P. The result for the factor P suggests that α should increase with the number of predictor variables; in particular, α should be set at $P/100$, at least for P in the range 5 to 25. This increase of α with P is in agreement with the idea that when there are many predictor variables, a particular variable has less chance of selection; in this situation a larger α level should be used to give each variable a chance of being selected comparable to that when there are a smaller number of variables.

In the multivariate binary case, the choice of α depends upon the factors P, $M(= p_1 - p_0)$ and $B(= p_0)$. For the factor P, the best α level increases with the number of predictor variables in the data according to the formula best $\alpha = P/100$; this agrees with the multivariate normal case. For the factor M, the difference between the means of the binary variables in population 1 and population 0, the non-linearity of the change of best α with M makes it difficult to formulate a rule. Similarly for the factor B, the mean of the binary variables for the first population, although the best α changes with B, it does so in a non-linear way.

These results imply that uniform specification of the best α level for the standard $\chi^2_{(\alpha)}$ stopping criterion cannot be made. However, if a recommendation had to be made, Lee and Koval would recommend that $0.15 \leq \alpha \leq 0.20$ be used, with a further refinement that, if $5 \leq P \leq 25$, then $\alpha = P/100$.

Chapter 10

Assessing the fit and predictive power of the model

This chapter is started with the assumption that one is at least preliminarily satisfied with the efforts at the model building stage. Now the question is how effectively the model one has describes the outcome variable. This is referred to as its goodness-of-fit, or in credit scoring, as the predictive power of the model.

If one intends to assess the goodness-of-fit of the model, then there should be some specific ideas about what it means to say that a model fits.

Suppose the observed sample values of the outcome variable in vector form is denoted as $\mathbf{y}$ where $\mathbf{y}'= (y_1, y_2, y_3, ..., y_n)$. Denote the values predicted by the model, or the fitted values as $\widehat{\mathbf{y}}$ where $\widehat{\mathbf{y}}' = (\widehat{y}_1, \widehat{y}_2, \widehat{y}_3, ..., \widehat{y}_n)$.The conclusion is that the model fits if summary measures of the distance between $\mathbf{y}$ and $\widehat{\mathbf{y}}$ are small and the contribution of each pair $(y_i, \widehat{y}_i)$, $i = 1, 2, ..., n$ to these summary measures is unsystematic and is small relative to the error structure of the model. A complete assessment of the fitted model involves both the calculation of summary measures of the distance between $\mathbf{y}$ and $\widehat{\mathbf{y}}$ and a thorough examination of the individual components of these measures.

When the model building stage has been completed, a series of logical steps may be used to assess the fit of the model. This includes the computation and evaluation of overall measures of fit, examination of the individual components of the summary statistics and examination of other measures of the difference or distance between the components of $\mathbf{y}$ and $\widehat{\mathbf{y}}$.

10.1 Summary measures of Goodness-of-fit

This section is largely based on the work done by Hosmer and Lemeshow (2000). Summary measures are routinely provided as output with any fitted model and give an overall indication of the fit of the model. Summary statistics, by nature, may not provide information about the individual model components. A small value for one of these statistics does not rule out the possibility of some substantial and thus interesting deviation from fit for a few subjects. On the other hand, a large value for one of these statistics is a clear indication of a substantial problem with the model.

Before discussing specific goodness-of-fit statistics, one must first consider the effect the fitted model has on the degrees-of-freedom available for the assessment of model performance. The term "covariate pattern" is used to describe a single set of values for the covariates in the model. For example, in a data set containing values of age, spend, number of credit cards and other products with the bank, the combination of these factors may result in as many different covariate patterns as there are observations. On the other hand, if the model contains only number of credit cards and other products with the bank, both coded at two levels each, there are only four possible covariate patterns. During model development, it is not necessary to be concerned about the number of covariate patterns. The degrees-of-freedom for tests are based on the difference in the number of parameters in competing models, not on the number of covariate patterns. However, the number of covariate patterns may be an issue when the fit of a model is assessed.

Goodness-of-fit is assessed over the constellation of fitted values determined by the covariates in the model, not the total collection of covariates. Suppose that our fitted model contains p independent variables, $\mathbf{x}' = (x_1, x_2, x_3, ..., x_p)$ and let J denote the number of distinct values of $\mathbf{x}$ observed. If some observations have the same value of $\mathbf{x}$, then $J < n$. Denote the number of observations with $\mathbf{x} = \mathbf{x}_j$ by m_j, $j = 1, 2, ..., J$.

It follows that $\sum m_j = n$. Let y_j denote the number of events, $y = 1$, among the m_j subjects with $\mathbf{x} = \mathbf{x}_j$. It follows that $\sum y_j = n_1$, the total number of observations with $y = 1$.The distribution of the goodness-of-fit statistics is obtained by letting n become large. If the number of covariate patterns also increases with n then each value of m_j tends to be small. Distributional results obtained under the condition that only n becomes large are said to be n-asymptotics. If $J < n$ is fixed and let n become large then each value of m_j also tends to become large. Distributional results based on each m_j becoming large are said to be based on m-asymptotics.

Initially assume that $J \approx n$. This presents the greatest challenge in developing distributions of goodness-of-fit statistics.

10.1.1 Pearson Chi-square and Deviance

In linear regression, summary measures of fit as well as diagnostics for case-wise effect on the fit, are functions of a residual defined as the difference between the observed and fitted value $(y - \widehat{y})$. In logistic regression there are several possible ways to measure the difference between the observed and fitted values. To emphasize the fact that the fitted values in logistic regression are calculated for each covariate pattern and depend on the estimated probability for that covariate pattern, denote the fitted value for the j^{th} covariate pattern as y_j where

$$\widehat{y}_j = m_j \widehat{\pi}_j = m_j \frac{e^{\widehat{g}(\mathbf{x}_j)}}{1 + e^{\widehat{g}(\mathbf{x}_j)}}$$

where $\widehat{g}(\mathbf{x}_j)$ is the estimated logit

Consider two measures of of the difference between the observed and fitted values: the Pearson residual and the deviance residual. For a particular covariate pattern the Pearson residual is defined as follows:

$$r(y_j, \widehat{\pi}_j) = \frac{(y_j - m_j \widehat{\pi}_j)}{\sqrt{m_i \widehat{\pi}_j (1 - \widehat{\pi}_j)}}.$$

The summary statistic based on these residuals is the Pearson chi-square statistic

$$\chi^2 = \sum_{j=1}^{J} r(y_j, \widehat{\pi}_j)^2.$$

The deviance residual is defined as

$$d(y_j, \widehat{\pi}_j) = \pm \left\{ 2 \left[y_j \ln \left(\frac{y_j}{m_j \widehat{\pi}_j} \right) + (m_j - y_j) \ln \left(\frac{(m_j - y_j)}{m_j (1 - \widehat{\pi}_j)} \right) \right] \right\}^{\frac{1}{2}}$$

where the $\pm$ sign is the same sign as $(y_j - m_j \widehat{\pi}_j)$. For covariate patterns with $y_j = 0$ the deviance residual is

$$d(y_j, \widehat{\pi}_j) = -\sqrt{2 m_j |\ln(1 - \widehat{\pi}_j)|}$$

and the deviance residual when $y_j = m_j$ is

$$d(y_j, \widehat{\pi}_j) = -\sqrt{2 m_j |\ln(\widehat{\pi}_j)|}.$$

The summary statistic based on the deviance residuals is the deviance:

$$D = \sum_{j=1}^{J} d(y_j, \widehat{\pi}_j)^2.$$

The distribution of the statistics χ^2 and D under the assumption that the fitted model is correct in all aspects is supposed to be chi-square with degrees-of-freedom equal to $J - (p + 1)$. For the deviance this statement follows from the fact that is the likelihood ratio test statistic of a saturated model with J parameters versus the fitted model with $p + 1$ parameters. Similar theory provides the null distribution of χ^2. The problem is that when $J \approx n$, the distribution is obtained under n-asymptotics, and hence the number of parameters is increasing at the same rate as the sample size. Thus, p-values calculated for these two statistics when $J \approx n$, using the $\chi^2(J - p - 1)$ distribution, are incorrect.

One way to avoid these difficulties with the distribution of χ^2 and D when $J \approx n$ is to group the data in such a way that m-asymptotics can be used. To understand the rationale behind the various grouping strategies that have been proposed, it is helpful to think of χ^2 as the Pearson and D as the log-likelihood chi-square statistics that result from a $2 \times J$ table. The rows of the table correspond to the two values of the outcome variable, $y = 1, 0$. The J columns correspond to the J possible covariate patterns. The estimate of the expected value under the hypothesis that the logistic model in question is the correct model for the cell corresponding to the $y = 1$ row and the j^{th} column is $m_j \widehat{\pi}_j$. It follows that the estimate of the expected value for the cell corresponding to the $y = 0$ row and the j^{th} column is $m_j(1 - \widehat{\pi}_j)$. The statistics χ^2 and D are calculated in the usual manner from this table.

Thinking of the statistics as arising from the $2 \times J$ table gives some intuitive insight as to why one cannot expect them to follow the $\chi^2(J - p - 1)$ distribution. When chi-square tests are computed from a contingency table the p-values are correct under the null hypothesis when the estimated expected values are "large" in each cell. This condition holds under m-asymptotics.

In the $2 \times J$ table described above ($J \approx n$) the expected values are always quite small since the number of columns increases as n increases. To avoid this problem the columns may be collapsed into a fixed number of groups, g, and then observed and expected frequencies can be calculated. By fixing the number of columns, the estimated expected frequencies become large as n becomes large and thus m-asymptotics hold.

In credit scoring, continuous variables are mostly also bucketed, so a fixed number of covariate patterns will exist. Therefore m-asymptotics will hold.

10.1.2 The Hosmer-Lemeshow Tests

Hosmer and Lemeshow (2000) proposed grouping based on the values of the estimated probabilities. Suppose that $J = n$. In this case the n columns are thought of as corresponding to the n values of the estimated probabilities, with the first column corresponding to the smallest value, and the n^{th} column to the largest value. Two grouping strategies were proposed as follows:

- Collapse the table based on percentiles of the estimated probabilities, and
- Collapse the tables based on fixed values of the estimated probability

With the first method, use of $g = 10$ groups result in the first group containing the $n_1 = n/10$ observations having the smallest estimated probabilities and the last group containing the $n_{10} = n/10$ observations having the largest estimated probabilities. With the second method, use of $g = 10$ groups result in cut-off points defined at the values $k/10$, $k = 1, 2, \ldots, 9$, and the groups contain all observations with estimated probabilities between adjacent cut-off points. For example, the first group contains all observations whose estimated probability is less than 0.1, while the tenth group contains those observations whose estimated probabilities is greater than 0.9. For the $y = 1$ row, estimates of the expected values are obtained by summing the estimated probabilities over all observations in a group. For the $y = 0$ row, the estimated expected value is obtained by summing, over all observations in the group, one minus the estimated probability.

For either grouping strategy, the Hosmer-Lemeshow goodness-of-fit statistic, $\widehat{C}$, is obtained by calculating the Pearson chi-square statistic from the $g \times 2$ table of observed and estimated expected frequencies. A formula defining the calculation of $\widehat{C}$ is as follows:

$$\widehat{C} = \sum_{g=1}^{k} \frac{(o_k - n_k \overline{\pi}_k)^2}{n_k \overline{\pi}_k (1 - \overline{\pi}_k)}$$

where n_k is the total number of observations in the k^{th} group, c_k denotes the number of covariate patterns in the k^{th} decile,

$$o_k = \sum_{j=1}^{c_k} y_j$$

is the number of responses among the c_k covariate patterns, and

$$\overline{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \widehat{\pi}_j}{n_k}$$

is the average estimated probability.

Hosmer and Lemeshow (2000) demonstrated that, when $J = n$ and the fitted logistic regression model is the correct model, the distribution of the statistic $\widehat{C}$ is well approximated by the chi-square distribution with $g - 2$ degrees of freedom, $\chi^2(g - 2)$. It is also likely that $\chi^2(g - 2)$ approximates the distribution when $J \approx n$.

An alternative to the denominator shown in the above equation for $\widehat{C}$, is obtained if one considers o_k to be the sum of independent non-identically distributed random variables. This suggests that one should standardize the squared difference between the observed and expected frequency, where the estimated expected frequency is given by

$$\sum_{j=1}^{c_k} m_j \widehat{\pi}_j (1 - \widehat{\pi}_j),$$

then

$$\sum_{j=1}^{c_k} m_j \widehat{\pi}_j (1 - \widehat{\pi}_j) = n_k \overline{\pi}_k (1 - \overline{\pi}_k) - \sum_{j=1}^{c_k} m_j (\widehat{\pi}_j - \overline{\pi}_k)^2.$$

Additional research has shown that the grouping method based on percentiles of the estimated probabilities is preferable to the one based on fixed cut-off points in the sense of better adherence to the $\chi^2(g-2)$.distribution, especially when many of the estimated probabilities are small. Thus, unless stated otherwise, $\widehat{C}$ is based on the percentile-type of grouping, usually with $g = 10$ groups. Most logistic regression software packages provide the capability to obtain $\widehat{C}$ and its p-value, usually based on 10 groups. In addition many packages provide the option to obtain the 10×2 table listing the observed and estimated expected frequencies in each decile.

The distribution of $\widehat{C}$ depends on m-asymptotics, thus the appropriateness of the p-value depends on the validity of the assumption that the estimated expected frequencies are large. In general, all expected frequencies must be greater than 5.

When the number of covariate patterns are less than n, the possibility exists that one or more of the empirical deciles will occur at a pattern with $m_j > 1$. If this happens, the value of $\widehat{C}$ will depend, to some extent, on how these ties are assigned to deciles. The use of different methods to handle ties by different software packages is not likely to be an issue unless the number of covariate patterns is so small that assigning all tied values to one decile results in a huge imbalance in decile size, or worse, considerably fewer than 10 groups. In addition, when too few groups are used to calculate $\widehat{C}$, there is the risk that one will not have the sensitivity needed to distinguish observed from expected frequencies. When $\widehat{C}$ is calculated from fewer than 6 groups, it will almost always indicate that the model fits.

The advantage of a summary goodness-of-fit statistic like $\widehat{C}$ is that it provides a single, easily interpretable value that can be used to assess the fit. The great disadvantage is that in the process of grouping one may miss an important deviation from fit due to a small number of data points. Therefore, before finally accepting that a model fits, it is advised to perform an analysis of the individual residuals and relevant diagnostic statistics.

Tables listing the observed and estimated expected frequencies in each decile contain valuable descriptive information for assessing the adequacy of the fitted model over the deciles. Comparison of the observed to expected frequencies within each cell may indicate regions where the model does not perform statistically.

A complete assessment of fit is a multi-faceted investigation involving summary tests and measures as well as diagnostic statistics. This is especially important to keep in mind when using overall goodness-of-fit tests. The desired outcome for most investigators is the decision not the reject the null hypothesis that the model fits. With this decision one is subject to the possibility of the Type II error and hence the power of the test becomes an issue. It has been shown that none of the overall goodness-of-fit tests is especially powerful for small to moderate sample sizes $n < 400$.

The large sample normal approximation to the distribution of the Pearson chi-square statistic developed by Osius and Rojek(1992) may be easily computed in any statistical package that has the option to save the fitted values from the logistic regression model and do a weighted linear regression. The essential steps in the procedure when one has J covariate patterns are:

- Retain the fitted values from the model, denoted as $\widehat{\pi}_j$, $j = 1, 2, 3...J$.
- Create the variable $v_j = m_j\widehat{\pi}_j(1 - \widehat{\pi}_j)$, $j = 1, 2, 3, ...J$.
- Create the variable $c_j = \frac{(1-2\widehat{\pi}_j)}{v_j}$, $j = 1, 2, 3, ..., J$.
- Compute the Pearson chi-square statistic, namely:

$$\chi^2 = \sum_{j=1}^{J} \frac{(y_j - m_j\widehat{\pi}_j)^2}{v_j}.$$

- Perform a weighted linear regression of c, defined above, on $\mathbf{x}$, the model covariates, using weights v, also defined above. Note that the sample size for this regression is J, the number of covariate patterns. Let RSS denote the residual sum-of-squares from this regression. Some statistical packages scale the weights to sum to 1. In this case, the reported residual sum-of-squares must be multiplied by the mean of the weights to obtain the correct RSS.
- Compute the correction factor, denoted A as follows:

$$A = 2(J - \sum_{j=1}^{J} \frac{1}{m_j}).$$

- Compute the standardized statistic

$$z = \frac{[\chi^2 - (J - p - 1)]}{\sqrt{A + RSS}}.$$

- Compute a two-tailed p-value using the standard normal distribution.

To carry out the above analysis, it is necessary to form an aggregated data set. The essential steps in any package are:

- Define as aggregation the main effects in the model. This defines the covariate patterns.
- Calculate the sum of the outcome variable and the number of terms in the sum over the aggregation variables. This produces y_j and m_j for each covariate pattern.
- Output a new data set containing the values of the aggregation variables, covariate patterns and two calculated variables y_j and m_j.

A two degrees-of-freedom test was proposed that determines whether two parameters in a generalized logistic model are equal to zero. The two additional parameters allow the tail of the logistic regression model (i.e. the small and large probabilities) to be either heavier/longer or lighter/shorter that the standard logistic regression model. This test is not a goodness-of-fit test since it does not compare observed and fitted values. However it does provide a test of the basic logistic regression model assumption and it is useful in adjunction to the Hosmer-Lemeshow and Osius-Rojek goodness-of-fit (1992) tests.

This test can be easily obtained from the following procedure:

- Retain the fitted values from the model, denoted as $\widehat{\pi}_j$, $j = 1, 2, 3, ..., J$.
- Compute the estimated logit

$$\widehat{g}_j = \ln(\frac{\widehat{\pi}_j}{1-\widehat{\pi}_j}) = \mathbf{x}_j`\widehat{\boldsymbol{\beta}} \ ,j = 1, 2, 3, ..., J.$$

- Compute two new covariates:

$$z_{1j} = 0.5 \times \widehat{g}_j^2 \times I(\widehat{\pi}_j \geq 0.5)$$

and

$$z_{2j} = 0.5 \times \widehat{g}_j^2 \times I(\widehat{\pi}_j < 0.5)$$

$j = 1, 2, 3, ..., J$, where $I(condition) = 1$ if the condition is true and 0 if the condition is false. Note that in a setting when all the fitted values are either less than or greater than 0.5 only one variable is created.

- Perform the partial likelihood ratio test for the addition z_{1j} and/or z_{2j} to the model.
- Calculate a p-value from the partial likelihood ratio test using two degrees of freedom. The p-value is then examined against a chosen significance level for the null hypothesis.

10.1.3 Classification tables

An intuitively appealing way to summarize the results of a fitted logistic regression model is via a classification table. This table is the result of cross-classifying the outcome variable, y, with a dichotomous variable whose values are derived from the estimated logistic probabilities.

To obtain the derived dichotomous variable one must define a cut-off point, c, and compare each estimated probability to c. If the estimated probability exceeds c then let the derived variable be equal to 1; otherwise it is equal to zero. A classification table is then drawn up:

	Actual		
Predicted	$y = 1$	$y = 0$	Total
$y = 1$	a	b	$a + b$
$y = 0$	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

The correctly classified cases are called the true positives and true negatives. If they not correspond, they are labeled the false positives (type I error: predicted $y = 1$, actual $y = 0$) and false negatives (type II error: predicted $y = 0$, actual $y = 1$).

If the model predicts the true positives and negatives accurately, this is thought to provide evidence that the model fits. Unfortunately, this may or may not be the case.

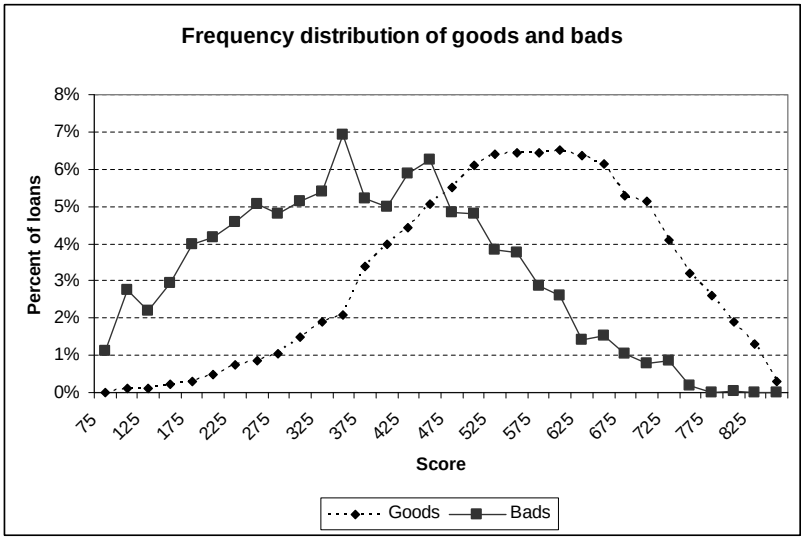
The overall rate of correct classification is calculated as $(a + d)/(a + b + c + d) \times 100\%$. Sensitivity is calculated as $a/(a + c) \times 100\%$ (percentage of $y = 1$ observed correctly predicted) and specificity is calculated as $d/(b + d) \times 100\%$ (percentage of $y = 0$ observed correctly predicted).

The problem with using the misclassification rate as a criterion is that it reduces a probabilistic model where the outcome is measured on a continuum, to a dichotomous model where predicted outcome is binary. For practical purposes there is little difference between the values of $\widehat{\pi} = 0.48$ and $\widehat{\pi} = 0.52$, yet use of a cut-off point of 0.5 would establish these two observations as markedly different.The classification table is most appropriate when classification is a stated goal of the analysis, otherwise it should only supplement more rigorous methods of assessments of fit.

Classification tables are rarely used as a measure in credit scoring. The purpose of a credit scoring model is ranking, rather than classification.

10.2 Separation statistics

What one expects of a scorecard/ credit scoring model is that it will assign different scores to loans that have an undesirable outcome than to those that have a desirable outcome. In essence, bad loans (according to the chosen default definition) should have lower scores than good loans.



Goods and bads distributions

The above graph shows the score frequency distribution for a group of bad loans (on the left) and group of good loans. Clearly, the distribution for the bad loans is centered over a lower scoring range that the good loans. The further apart these two distributions are, the more successful the scorecard has been in distinguishing bad loans from good. In fact, if a scorecard did a perfect job, there would be no overlap at all in the distributions- they would be side by side. Conversely, a scorecard doing a poor job would not be able to distinguish between the two groups of loans at all. At the extreme, the two distributions would lie on top of one another.

When people talk about how well a scorecard "separates", this is exactly what they are talking about – separating the distributions of good and bad loans.

10.2.1 Divergence statistic

The divergence statistic is very straightforward. It is simply the square of the difference of the mean of the goods and the mean of the bads, divided by the average variance of the score distributions.

The formula for the divergence statistic is:

$$D^2 = (\mu_G - \mu_B)^2/\sigma^2$$

where μ_G is the mean score of the goods, μ_B is the mean score of the bads, and $\sigma^2 = (\sigma_G^2 + \sigma_B^2)/2$ with σ_G^2 the score variance of the goods and σ_B^2 the score variance of the bads

The square of the difference in the means is divided by the average variance. For distributions that are very spread out, there must be a large difference in the means before the distributions are said to be different.

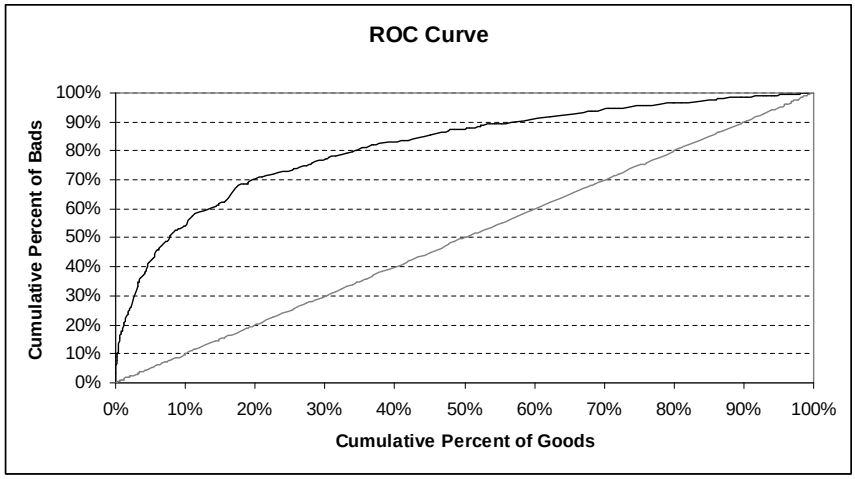
The divergence statistic is closely related to the information value statistic described earlier, where it was discussed as a statistic that could be used to evaluate the predictiveness of individual scorecard characteristic. To evaluate separation ability, either the information value or the divergence statistic may be calculated for any individual predictive variable or for the score itself. For a continuous predictive variable, divergence and information value are equal if the scores of the good loans and the bad loans are normally distributed and have equal variances.

One benefit of the divergence is its simplicity and ease of interpretation. It does not, however, capture all the important information about the shape of the distributions. It is important to recognize that the statistics used to evaluate scorecards are simply summary measures of some aspect of the difference between the good and bad distributions. They do not, and cannot, tell everything on how well a scorecard is doing its job. If one relies on summary statistics alone, there may be cases where it will be misleading. It is therefore good practice to always generate plots of the distributions as well when reporting any scoring statistics.

10.2.2 Area under the ROC curve

Sensitivity and specificity as defined earlier in the section on classification tables rely on a single cut-off point to classify a test result as positive. A more complete description of classification accuracy is given by the area under the ROC (Receiver Operating Characteristic) curve. This curve, originating from signal detection theory, shows how the receiver operates the existence of signal in the presence of noise. It plots the probability of detecting true signal (sensitivity) and false signal (1-specificity) for an entire range of cut-off points. Sensitivity is thus the ability to mark true positives and specificity is the ability to identify true negatives.

A plot of sensitivity versus 1-specificity over all possible cut-off points is shown below. The curve generated by these points is called the ROC curve and the area under the curve provides a measure of discrimination which is the likelihood that an observation who has $y = 1$ will have a higher $P(y = 1)$ than an observation who has $y = 0$.



ROC curve

The ROC curve is also known as the trade-off curve because it shows the trade-off between goods and bads - the percentage of total bads that must be accepted in order to accept a given percentage of total goods. The 45 degree line on the graph shows a ROC curve for a score with no ranking ability. The point below which 20% of the bads are found is the same score below which 20% of goods are found.

The area under the ROC curve, which ranges from zero to one, provides a measure of the model's ability to discriminate between those observations that experience the outcome of interest versus those who do not. In credit scoring, this is called a measure of the model's predictive power. This area under the ROC curve is also referred to as the C statistic.

In this particular graph, at the score which 10% of the goods are found, about 55% of the bads are found. One would like this curve to rise as quickly as possible because it means that more bad loans are assigned low scores relative to good loans. The faster it rises, the greater the area under it and the larger the C statistic. In practice the C statistic usually ranges from 0.3 to 0.9 for credit scorecards. Computer algorithms are used to calculate this statistic, and SAS generates this by default every time a logistic regression is run.

It is possible that a poorly fitting model (i.e. poorly calibrated as assessed by the goodness-of-fit measures discussed earlier) may still have good discrimination. For example, if one added 0.25 to every probability in a good fitting logistic regression model with a good ability to discriminate, the new model would be poorly calibrated whereas the discrimination would not be affected at all. Thus, model performance should be assessed by considering both calibration and discrimination.

Another more intuitive way to understand the meaning of the area under the ROC curve is as follows: recall that n_1 denote the number of observations with $y = 1$ and n_0 denote the number of observations with $y = 0$. Pairs are then created: each observation with $y = 1$ is paired with each observation with $y = 0$. Of these $n_1 \times n_0$ pairs, determine the proportion of the time that the observation with $y = 1$ had the higher of the two probabilities. This proportion is equal to the area under the ROC curve. Note that when the probability is the same for both observations, $\frac{1}{2}$ is added to the count.

10.2.3 KS statistic

Another measure of discrimination used in credit scoring, is the Kolmogorov-Smirnov statistic. It is mainly used in measuring the predictive power of rating systems, but in other environments, the Gini or area under the ROC curve seem to be more prevalent.

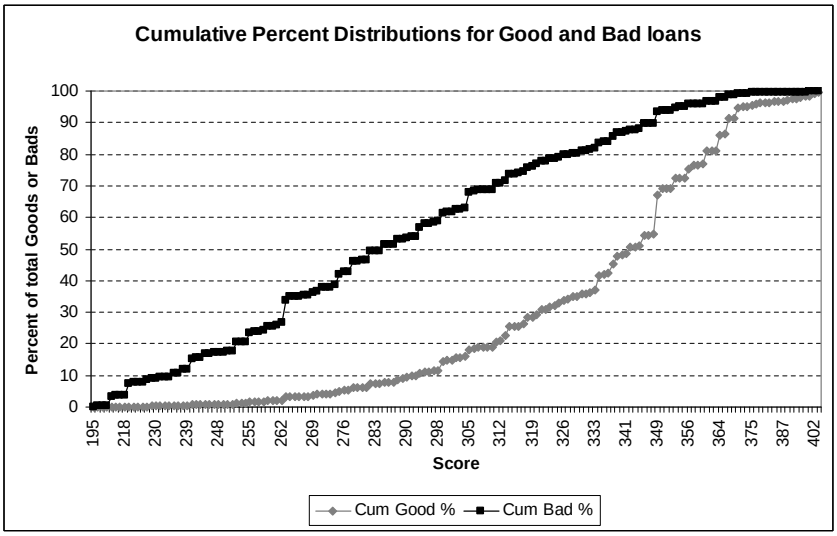
Traditionally the KS statistic is used to compare an unknown, observed distribution to a known, theoretical distribution. The maximum distance between the cumulative distributions are calculated and measured against a critical value. If the maximum distance is less than the critical value, there is good chance that the distributions are the same.

In credit scoring, it is mostly used as a data-visualization tool to illustrate the model's effectiveness. The cumulative distributions of events and non-events are plotted against the score (output from the model, transformed, can also be seen as the different covariate patterns, ranked in a way). The maximum distance is then calculated between the two distributions. It is calculated as:

$$D_{KS} = \max[abs(cp_1 - cp_0)]$$

where cp_1 is the cumulative percentage for events and cp_0 is the cumulative percentage of non-events, *abs* indicates taking the absolute value and the maximum is taken of all the differences over all possible scores.

KS is also referred to as the fish-eye graph. It may make the calculation of separation and the KS even more clear if one looks at the graph below, which plots the cumulative percentage of goods and bads.



KS graph

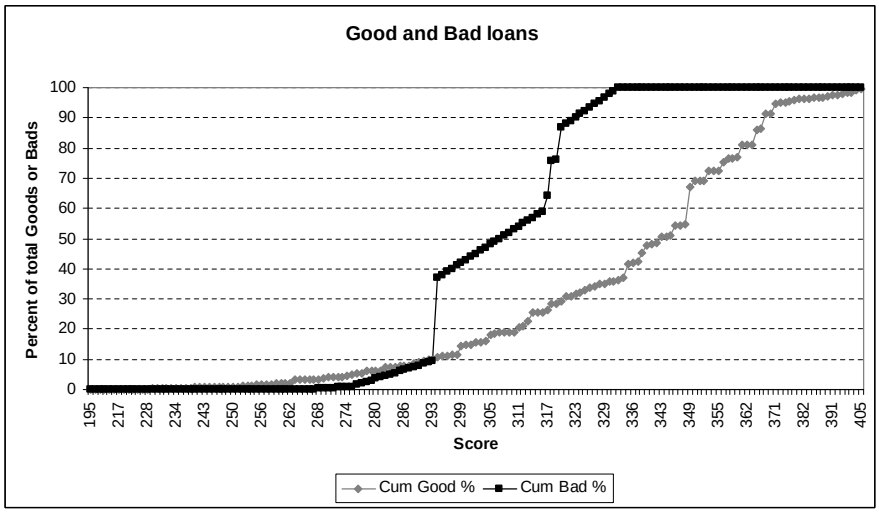
The bad distribution is the one on top. It rises more quickly than the good distribution because more of the bads are found among the lowest-scoring loans. The vertical distance between the two is the separation at each point. The maximum difference occurs at a score of 305, where 68% of the bads lie below that score, but only 18% of the goods. The KS is $68 - 18 = 50$.

There are no hard and fast rules on what the expected value of the KS should be or how big it should be before we can be confident we have a good scorecard. The KS value expected to be achieved will vary depending on the product the scorecard is developed for and on the data available. A relatively low KS does not necessarily mean someone has done a poor job of building the scorecard – it may be the best that can be obtained in that particular situation

While KS's can theoretically range from 0 to 100, in practice the range is generally from about 20 to about 70. If the KS is lower than 20, it would be reasonable to question whether the scorecard is worth using. Above 70, it is probably too good to be true and one should suspect problems with the way it is being calculated or with the scorecard itself.

The KS is based on a single point on the good and bad distributions – the point where the cumulative distributions are the most different. There may thus be cases where it would be a major mistake to blindly rely on the KS to tell whether it is a good scorecard or not, without taking a careful look at the distributions of goods and bads to see how well the scorecard is ranking them.

The following graph duplicates the above graph for the goods, but the bads have been rearranged.



Undesirable KS graph

At 63.7, the KS is higher than it was before, but in this case the KS is misleading, because clearly the scorecard is not ranking the loans well in the bottom of the score distribution. That is why, just as with the divergence statistic, it is necessary to look at the distributions to get the whole story.

One final point should be made about the KS statistic. After a new scorecard is implemented and a cut-off score set, the lender may observe a decline in the KS when the score is validated compared to its value in the model-building sample. This happens when the lender has adhered to the new cut-off and booked fewer low-scoring loans than in the past. The KS is thus calculated on a narrower distribution of scored loans. That is, the newly-booked population is truncated compared to the lender's booked population before the new scorecard was implemented. The KS is sensitive to population truncation and can fall from one sample to the next even though the ranking ability of the score may not have declined.

10.3 Logistic regression diagnostics

The goodness-of-fit measures described above provide a single number that summarizes the agreement between observed and fitted values. The advantage of these statistics is that a single number is used to summarize considerable information. It is therefore also a disadvantage of these measures. Thus, before concluding that the model "fits", it is crucial that other measures be examined to see if the fit is supported over the entire set of covariate patterns. This is accomplished through a series of specialized measures falling under the general heading of regression diagnostics.

The derivation of logistic regression diagnostics will be described briefly (Hosmer and Lemeshow (2000)). In this development it was assumed that the fitted model contained p covariates and that they form J covariate patterns. The key quantities for logistic regression diagnostics, as in linear regression, are the concepts of the "residual sum-of-squares". In linear regression a key assumption (as described in an earlier chapter) is that the error variance does not depend on the conditional mean, $E(Y_j|\mathbf{x}_j)$.

However, in logistic regression one has binomial errors and, as a result, the error variance is a function of the conditional mean:

$$var(Y_j|\mathbf{x}_j) = m_j E(Y_j|\mathbf{x}_j) \times [1 - E(Y_j|\mathbf{x}_j)]$$

thus $var(Y_j|\mathbf{x}_j) = m_j\pi_j \times [1 - \pi_j]$.

Begin with the residuals defined in the previous section (Pearson and deviance) which have been "divided' by estimates of their standard errors.

The Pearson residual is defined as follows and denoted it by r_j:

$$r(y_j, \widehat{\pi}_j) = \frac{(y_j - m_j\widehat{\pi}_j)}{\sqrt{m_j\widehat{\pi}_j(1 - \widehat{\pi}_j)}}.$$

The deviance residual, denoted as d_j is defined as:

$$d(y_j, \widehat{\pi}_j) = \pm \left\{ 2 \left[y_j \ln(\frac{y_j}{m_j\widehat{\pi}_j}) + (m_j - y_j) \ln(\frac{(m_j - y_j)}{m_j(1 - \widehat{\pi}_j)}) \right] \right\}^{\frac{1}{2}}.$$

Thus r_j and d_j are the Pearson and deviance residuals for the covariate pattern $\mathbf{x}_j$. Since each residual has been divided by an approximate estimate of its standard error, it is expected that if the logistic regression model is correct, these quantities will have a mean approximately equal to zero and a variance approximately equal to 1.

In addition to the residuals for each covariate pattern, other quantities central to the formation and interpretation of linear regression diagnostics are the "hat" matrix and the leverage values derived from it. In linear regression the hat matrix is the matrix that provides the fitted values as the projection of the outcome variable into the covariate space.

Let $\mathbf{X}$ denote the $J \times (p+1)$ matrix containing the values for all J covariate patterns formed from the observed values of the p covariates, with the first column being one to reflect the presence of an intercept in the model. The matrix $\mathbf{X}$ is also known as the design matrix. In linear regression, the hat matrix is $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$; for example $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.
The linear regression residuals $(\mathbf{y} - \widehat{\mathbf{y}})$ expressed in terms of the hat matrix are $(\mathbf{I} - \mathbf{H})\mathbf{y}$ where $\mathbf{I}$ is the $J \times J$ identity matrix. Using weighted least squares linear regression as a model, a linear approximation to the fitted values can be derived, which yields a hat matrix for logistic regression. This matrix is:

$$\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{1/2}$$

where $\mathbf{V}$ is a $J \times J$ diagonal matrix with general element

$$v_j = m_j\widehat{\pi}(\mathbf{x}_j)[1 - \widehat{\pi}(\mathbf{x}_j)].$$

In linear regression the diagonal elements of the hat matrix are called the leverage values and are proportional to the distance from $\mathbf{x}_j$ to the mean of the data. This concept of distance to the mean is important in linear regression, as points that are far from the mean may have considerable influence on the values of the estimated parameters. The extension of the concept of leverage to logistic regression requires additional explanation.

Let the quantity h_j denote the j^{th} diagonal element of the matrix $\mathbf{H}$ as defined above. It can be shown that

$$h_j = m_j\widehat{\pi}(\mathbf{x}_j)[1-\widehat{\pi}(\mathbf{x}_j)]\mathbf{x}_j'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_j' = v_j \times b_j$$

where

$$b_j = \mathbf{x}_j'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_j'$$

and $\mathbf{x}_j' = (1, x_{1j}, x_{2j}, ..., x_{pj})$ is the vector of covariate values defining the j^{th} covariate pattern.

The sum of the diagonal elements of $\mathbf{H}$ is, as is the case in linear regression, $\sum h_j = (p+1)$, the number of parameters in the model. In linear regression the dimension of the hat matrix is usually $n \times n$ and thus ignores any common covariate patterns in the data. With this formulation, any diagonal element in the hat matrix has an upper bound of $1/k$ where k is the number of observations with the same covariate pattern. If the hat matrix for logistic regression is formulated as an $n \times n$ matrix then each diagonal element is bounded from above by $1/m_j$, where m_j is the total number of observations with the same covariate pattern. When the hat matrix is based upon data grouped by covariate patterns, the upper bound for any diagonal element is 1.

It is important to know whether the statistical package being used calculates the diagnostic statistics by covariate pattern. SAS's logistic procedure computes diagnostic statistics based on the data structure in the model statement. If one assumes that there are n covariate patterns (and the outcome is either 0 or 1) then diagnostic statistics are based on individual observations. However, if data have been previously collapsed or grouped into covariate patterns and binomial trails input (y_j/m_j) is used, then diagnostic statistics are by covariate pattern. It is recommended that diagnostic statistics are computed taking covariate patterns into account. This is especially important when the number of covariate patterns, J, is much smaller than n, or if some values of m_j are larger than 5.

When the number of covariate patterns is much smaller than n there is the risk that one may fail to identify influential and/or poorly fit covariate patterns. Consider a covariate pattern with m_j observations, $y_j = 0$ and estimated logistic probability $\widehat{\pi}_j$. The Pearson residual defined above, computed individually for each observation with this covariate pattern, is

$$\begin{aligned} r_j &= \frac{(0-\widehat{\pi}_j)}{\sqrt{\widehat{\pi}_j(1-\widehat{\pi}_j)}} \\ &= -\sqrt{\frac{\widehat{\pi}_j}{(1-\widehat{\pi}_j)}} \end{aligned}$$

while the Pearson residual based on all observations with this covariate pattern is

$$\begin{aligned} r_j &= \frac{(0-m_j\widehat{\pi}_j)}{\sqrt{m_j\widehat{\pi}_j(1-\widehat{\pi}_j)}} \\ &= -\sqrt{m_j}\sqrt{\frac{\widehat{\pi}_j}{(1-\widehat{\pi}_j)}} \end{aligned}$$

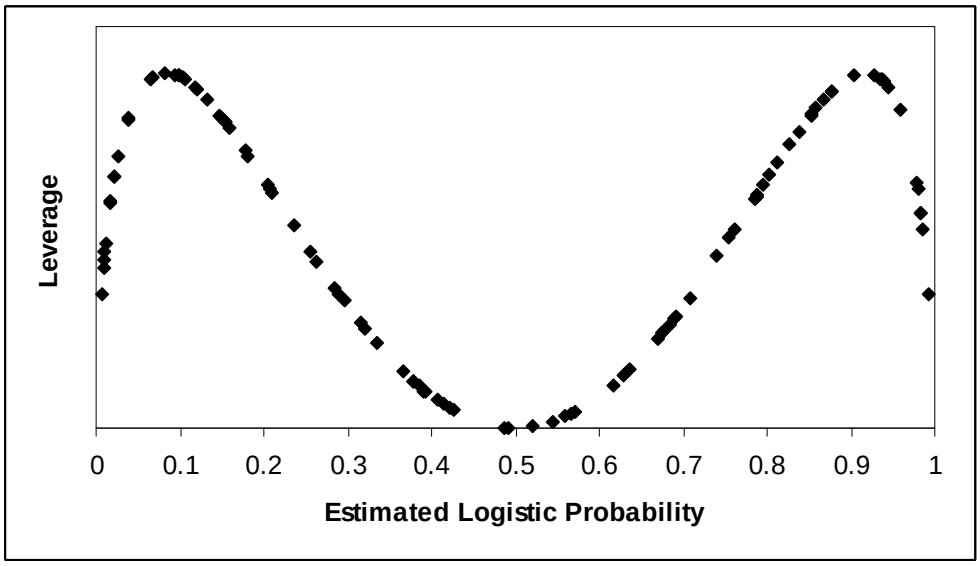
which increases negatively as m_j increases.

If $m_j = 1$ and $\widehat{\pi}_j = 0.5$ then $r_j = -1$ which is not a large residual. On the other hand, if there were $m_j = 16$ observations with this covariate pattern, then $r_j = -4$, which is quite large. Thus, if one performs the analysis using the covariate patterns, the Pearson residual would be -4 for each of the 16 observations in the covariate pattern. If the analysis is performed with a sample size of n, ignoring the covariate pattern, then the Pearson residual would be -1 for all 16 observations. Thus the diagnostic statistics are different even though the same fitted model was used.

A major point that must be kept in mind when interpreting the magnitude of the leverage is the effect that v_j has on h_j in the equation

$$h_j = m_j\widehat{\pi}(\mathbf{x}_j)[1-\widehat{\pi}(\mathbf{x}_j)]\mathbf{x}_j'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_j' = v_j \times b_j.$$

It can be argued that the fit determines the estimated coefficients and, since the estimated coefficients determine the estimated probabilities, points with large values of h_j are extreme in the covariate space and thus lie far from the mean. On the other hand, the term v_j in the expression of h_j cannot be ignored.. The following example demonstrates that, up to a point, both approaches are correct:

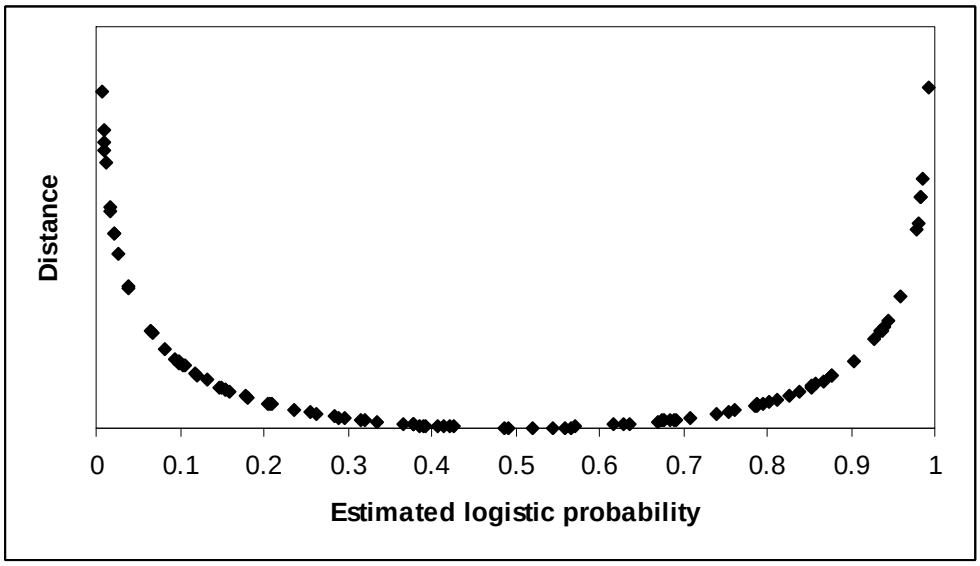


Leverage graph

This above graph presents a plot of the leverage values versus the estimated probabilities for a sample of 100 observations from a logistic model with $g(x) = 0.8x$ and $x \sim N(0, 9)$.

One can see that the leverage values increases as the estimated probability gets further from 0.5 (x gets further from its mean, nominally zero) until the estimated probabilities become less than 0.1 or greater than 0.9. At that point the leverage decreases and rapidly approaches zero. This example shows that the most extreme points in the covariate space may have the smallest leverage. This is exactly the opposite of the situation in linear regression, where the leverage is a monotonic increasing function of the distance of the covariate pattern from the mean. The practical consequence of this is that to interpret a particular value of the leverage in logistic regression correctly, one needs to know whether the estimated probability is small (< 0.1) or large (> 0.9). If the estimated probability lies between 0.1 and 0.9, then the leverage gives a value that may be thought of as distance. When the estimated probability lies outside the interval 0.1 to 0.9, then the value of the leverage may not measure distance in the sense that further from the mean implies a larger value.

A quantity that does not increase with the distance from the mean is $b_j = \mathbf{x}_j'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_j'$. Thus, if one is only interested in the distance then one should focus on b_j. A plot of the b_j versus the estimated probability for the same example is shown below:



Distance

In the above graph it is clear that b_j provides a measure of distance in the covariate space and, as a result, is more like the leverage values in linear regression. However, since the most useful diagnostic statistics for logistic regression are functions of the full leverage, h_j, the distance portion, b_j, is not discussed further.

If the following linear regression-like approximation for the residual for the j^{th} covariate pattern is used, $[y_j - m_j\widehat{\pi}(\mathbf{x}_j)] \approx (1-h_j)y_j$, then the variance of the residual is $m_j\widehat{\pi}(\mathbf{x}_j)\left[1-\widehat{\pi}(\mathbf{x}_j)\right](1-h_j)$ which suggests that the Pearson residuals do not have a variance equal to 1 unless they are further standardized. The standardized Pearson residual for the covariate pattern $\mathbf{x}_j$ is

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}}.$$

Another useful diagnostic statistic is one that examines the effect that deleting all observations with a particular covariate pattern has on the value of the estimated coefficients and the overall summary measures of fit, χ^2 and D. The change in the value of the estimated coefficients is obtained as the standardized difference between $\widehat{\beta}$ and $\widehat{\beta}_{(-j)}$, where these represent the maximum likelihood estimates computed using all J covariate patterns and excluding the m_j observations with pattern $\mathbf{x}_j$ respectively, and standardizing via the estimated covariance matrix of $\widehat{\beta}$. To a linear approximation, this quantity for logistic regression is:

$$\begin{aligned}\triangle\widehat{\beta}_j &= (\widehat{\beta}-\widehat{\beta}_{(-j)})'(\mathbf{X}'\mathbf{V}\mathbf{X})(\widehat{\beta}-\widehat{\beta}_{(-j)}) \\ &= \frac{r_j^2 h_j}{(1-h_j)^2} \\ &= \frac{r_{sj}^2 h_j}{(1-h_j)}\end{aligned}$$

Using similar linear approximations it can be shown that the decrease in the value of the Pearson chi-square statistic due to the deletion of the observations with the covariate pattern $\mathbf{x}_j$ is

$$\begin{aligned}\triangle\chi_j^2 &= \frac{r_j^2}{(1-h_j)} \\ &= r_{sj}^2\end{aligned}$$

A similar quantity may be obtained for the change in the deviance,

$$\triangle D_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)}.$$

If r_j^2 is replaced by d_j^2, it yields the approximation

$$\triangle D_j = \frac{d_j^2}{(1-h_j)}.$$

These diagnostic statistics are conceptually quite appealing, as they allow one to identify those covariate patterns that are poorly fit (large values of $\triangle\chi_j^2$ and/or $\triangle D_j$), and those that have a great deal of influence on the values of the estimated parameters (large values of $\triangle\widehat{\beta}_j$). After identifying these influential patterns (observations), one can begin to address the role they play in the analysis.

So, what is expected to be learnt from the application of diagnostics? First, consider the measure of fit, $\triangle\chi_j^2$. This measure is smallest when y_j and $m_j\widehat{\pi}(\mathbf{x}_j)$ are close. This is most likely to happen when $y_j = 0$ and $\widehat{\pi}(\mathbf{x}_j) < 0.1$ or $y_j = m_j$ and $\widehat{\pi}(\mathbf{x}_j) > 0.9$. Similarly $\triangle\chi_j^2$ is largest when y_j is furthest from $m_j\widehat{\pi}(\mathbf{x}_j)$. This is most likely to occur when $y_j = 0$ and $\widehat{\pi}(\mathbf{x}_j) > 0.9$ or with $y_j = m_j$ and $\widehat{\pi}(\mathbf{x}_j) < 0.1$. These same covariate patterns are not likely to have a large $\triangle\widehat{\beta}_j$ since, when $\widehat{\pi}(\mathbf{x}_j) < 0.1$ or $\widehat{\pi}(\mathbf{x}_j) > 0.9$, $\triangle\widehat{\beta}_j \approx \triangle\chi_j^2 h_j$ and h_j is approaching zero. The influence diagnostic, $\triangle\widehat{\beta}_j$ is large when both $\triangle\chi_j^2$ and h_j are at least moderate. This is most likely to occur when $0.1 < \widehat{\pi}(\mathbf{x}_j) < 0.3$, or $0.7 < \widehat{\pi}(\mathbf{x}_j) < 0.9$. From the above graph, where leverage is plotted against estimated probability ($\widehat{\pi}$), it is known that these are the intervals where the leverage, h_j, is largest. In the region where $0.3 < \widehat{\pi}(\mathbf{x}_j) < 0.7$ the chances are not as great that either $\triangle\chi_j^2$ or h_j is large. The following table summarizes these observations:

	Diagnostic statistic		
$\widehat{\pi}$	$\triangle\chi^2$	$\triangle\widehat{\beta}$	h
< 0.1	Large or small	Small	Small
$0.1 - 0.3$	Moderate	Large	Large
$0.3 - 0.7$	Moderate to small	Moderate	Moderate to small
$0.7 - 0.9$	Moderate	Large	Large
> 0.9	Large or small	Small	Small

Note that this table reports what might be expected, not what may actually happen in any particular example. It should therefore only be used as a guide to further understanding and interpretation of the diagnostic statistics.

In linear regression essentially two approaches are used to interpret the value of the diagnostics often in conjunction with each other. The first is graphical. The second employs the distribution theory of the linear regression model to develop the distribution of the diagnostics under the assumption that the fitted model is correct. In the graphical approach, large values of diagnostics either appear as spikes or reside in the extreme corners of plots. A value of the diagnostic statistic for a point appearing to lie away from the balance of the points is judged to be extreme if it exceeds some percentile of the relevant distribution.

This may sound a little too hypothesis-testing orientated but, under the assumption of linear regression with normal errors, there is a known statistical distribution whose percentiles provide some guidance as to what constitutes a large value. Presumably, if the model is correct and fits, then no values should be exceptionally large and the plots should appear as expected under the distribution of the diagnostic.

In logistic regression one has to rely primarily on visual assessment, as the distribution of the diagnostics under the hypothesis that the model fits is known only in certain limited settings. For instance, consider the Pearson residual, r_j. It is often stated that the distribution of this quantity is approximately $N(0,1)$, when the model is correct. This statement is only true when m_j is sufficiently large to justify that the normal distribution provides an adequate approximation to the binomial distribution, a condition obtained under m-asymptotics. For example, if $m_j = 1$ then r_j has only two possible values and can hardly be expected to be normally distributed. All of the diagnostics are evaluated by covariate pattern; hence any approximations to their distributions based on the normal distribution, under binomial errors, depends on the number of observations with the pattern. When a fitted model contains some continuous covariates then the number of covariate patterns, J, is of the same order as n, and m-asymptotic results cannot be relied upon. In practice, an assessment of "large" is, of necessity, a judgment call based on experience and the particular set of data analyzed. Using the $N(0,1)$, or equivalently, the $\chi^2(1)$ distribution for squared quantities may provide some guidance as to what large is. However, these percentiles should be used with extreme caution. There is no substitute for experience in the effective use of diagnostic statistics.

Seven diagnostic statistics have been defined, which may be divided into three categories:

- The basic building blocks, which are of interest in themselves, but are also used to form other diagnostics, (r_j, d_j, h_j).
- Derived measures of the effect of each covariate pattern on the fit of the model, $(r_{sj}, \triangle\chi^2_j, \triangle D_j)$.
- A derived measure of the effect of each covariate pattern on the value of the estimated parameters, $(\triangle\widehat{\beta}_j)$.

Most logistic regression software packages provide the capability to obtain at least one of the measures within each group.

A number of different types of plots have been suggested for use, each directed at a particular aspect of fit. Some are formed from the seven diagnostics while others require additional computation, for example they are based on grouping and smoothing. It is impractical to consider all possible suggested plots, so only a few of the more easily obtained ones that are meaningful in logistic regression analysis are considered. These can be seen as the core of an analysis of diagnostics. These consist of the following:

- Plot $\triangle\chi^2_j$ versus $\widehat{\pi}_j$.
- Plot $\triangle D_j$ versus $\widehat{\pi}_j$.
- Plot $\triangle\widehat{\beta}_j$ versus $\widehat{\pi}_j$.

Other plots that are sometimes useful include:

- Plot $\triangle\chi^2_j$ versus h_j.
- Plot $\triangle D_j$ versus h_j.
- Plot $\triangle\widehat{\beta}_j$ versus h_j.

These allow direct assessment of the contribution of leverage to the value of the diagnostic statistic.

Note that if the summary statistics indicated that the model fits, one does not expect an analysis of diagnostics to show large numbers of covariate patterns being fit poorly. One might uncover a few covariate patterns which do not fit, or which have a considerable influence on the estimated parameters.

Note that it is preferable to plot the diagnostics $\triangle\chi^2$ and $\triangle D$ versus the estimated logistic probabilities, instead of plots of r_j and d_j versus $\widehat{\pi}_j$. The reasons for this choice are as follows:

- When $J \approx n$, most positive residuals correspond to a covariate pattern where $y_j = m_j$ (e.g.=1) and negative residuals to those with $y_j = 0$. Hence the sign of the residual is not useful.
- Large residuals, regardless of sign, correspond to poorly fit points. Squaring these residuals further emphasizes the lack of fit and it removes the issue of sign.
- The shape of the plot allows one to determine which patterns have $y_j = 0$ and which have $y_j = m_j$.

Covariate patterns that are poorly fit will generally be represented by points falling in the top left or right corners of the plots. Look for points that fall some distance from the balance of the data plotted. Assessment of the distance is partly based on numeric value and partly based on visual impression. The range of $\triangle\chi^2$ is much larger than $\triangle D$. This is a property of Pearson versus deviance residuals. Whenever possible, it is preferable to use plots of both $\triangle\chi^2$ and $\triangle D$ versus $\widehat{\pi}$.

One problem with the influence diagnostic $\triangle\widehat{\beta}$ is that it is a summary measure of change over all coefficients in the model simultaneously. For this reason it is important to examine the changes in the individual coefficients due to specific covariate patterns identified as influential.

Suppose there is a model where the summary statistics indicate that there is substantial deviation from fit. In this situation, there is evidence that for more than a few covariate patterns, y_j differs from $m_j\widehat{\pi}_j$. One or more of three things has likely happened:

- The logistic regression model does not provide a good approximation to the correct relationship between the conditional mean, $E(Y|\mathbf{x}_j)$ and $\mathbf{x}_j$.
- An important covariate in this model was not measured and/or included.
- At least one of the covariates in the model has not been entered in the correct scale.

The logistic regression model is remarkably flexible. Unless one is dealing with a set of data where most of the probabilities are very small or very large, or where the fit is extremely poor in an identifiable systematic manner, it is unlikely that any alternative model will provide a better fit. If one suspects, based on clinical or other reasons (such as graphical representations or statistical tests) that the logistic model is the wrong one, then careful thought should be given to the choice of the alternative model. Particular attention should be given to the issues of interpretation. Are the coefficients clinically interpretable? The approach that tries all other possible models and selects the "best fitting" one is not recommended, as no thought is given to the clinical implications of the selected model. In some situations, inadequacy of a fitted logistic regression model can be corrected by returning to model building and rechecking variable selection and scale identification. Model fitting is an iterative procedure. One rarely obtains a final model on the first pass through the data. In credit scoring, the fit of the model is not as important as its ranking ability (ranking customers with worst to best risk). Even if the logistic model is not theoretically correct, if the results make business sense and are easily interpretable and usable, it is unlikely that an alternative modeling technique will be used.

When performing an analysis, one hopes that the study was designed carefully so that the data on all major covariates were collected. However, it is possible that the risk factors associated with the outcome variable are not well known and in this case a key variable may not be present in the observed data. The potential biases and pitfalls of this oversight are enormous and little can be done in this case, except to go back and collect the data, but this is often impractical. This problem is common in a credit scoring environment. Typically, the users of the models are not the developers. The user tries to score the client as quickly as possible and might not input all possible covariates that can be used in future modeling. The data is therefore skewed and the modeler is restricted in terms of the available covariates. Luckily the credit bureaus in South Africa have substantially increased their databases and data is normally available from them retrospectively.

10.4 Assessment of fit via external validation

In some situations it may be possible to exclude a sub-sample of the observations, develop a model based on the remaining observations, and then test the model on the originally excluded observations. This is typically done in credit scoring, if enough events/defaults are available. The data is split into a build and a holdout sample. The model is developed on the build sample and evaluated on the holdout sample.

In other situations it may be possible to obtain a new sample of data to assess the goodness-of-fit of a previously developed model. This is also common in credit scoring, where the model is evaluated on an out-of-time sample, usually from a more recent time period than the development data.

This type of assessment is often called model validation, and is it especially important when the fitted model is used to predict outcome for future observations. The reason for considering this type of assessment of model performance, is that the fitted model almost always performs in an optimistic manner on the development data set. In credit scoring, this is referred to testing whether the model can be generalized.

The use of validation data amounts to an assessment of goodness-of-fit where the fitted model is considered to be theoretically known, and no estimation is performed. Some of the diagnostics discussed earlier like ($\triangle\chi^2, \triangle D, \triangle\widehat{\beta}$) mimic this idea by computing, for each covariate pattern, a quantity based on the exclusion of the particular covariate pattern. With a new data set a more thorough assessment is possible.

The methods for assessment of fit in the validation sample parallel those described earlier for the development sample. The major difference is that the values of the coefficients in the model are regarded as fixed constants rather than estimated values.

Suppose that the validation sample consists of n_v observations $(y_i, \mathbf{x}_i)$, $i = 1, 2, ..., n_v$, which may be grouped into J_v covariate patterns. Let y_j denote the number of positive responses among the m_j observations with covariate pattern $\mathbf{x} = \mathbf{x}_j$ for $j = 1, 2, 3, ..., J_v$. The logistic probability for the j^{th} covariate pattern is π_j, the value of the previously estimated logistic model using the covariate pattern, $\mathbf{x}_j$ from the validation sample. These quantities become the basis for the computation of the summary measures of fit, χ^2, D and C, from the validation sample.

The computation of the Pearson chi-square follows directly from the equation:

$$\chi^2 = \sum_{j=1}^{J_v} r(y_j, \widehat{\pi}_j)^2$$

with the obvious substitution of quantities from the validation sample.

In this case, χ^2 is computed as the sum of J_v terms. If each $m_j\pi_j$ is large enough to use the normal approximation to the binomial distribution, then χ^2 is distributed as $\chi^2(J_v)$ under the hypothesis that the model is correct. If the observed numbers of observations are small within each covariate pattern, with most $m_j = 1$, m-asymptotics cannot be applied.In this case the Osius-Rojek goodness-of-fit test is used to obtain a statistic that follows the standard normal distribution under the hypothesis that the model is correct and J_v is sufficiently large. The procedure is similar to the one presented earlier. Specifically one computes the standardized statistic

$$z = \frac{[\chi^2 - J_v]}{\sigma_v}$$

where

$$\sigma_v = 2J_v + \sum_{j=1}^{J_v} \frac{1}{m_j\pi_j(1-\pi_j)} - 6\sum_{j=1}^{J_v} \frac{1}{m_j}.$$

The test uses a two-tailed p-value based on z.

The same line of reasoning discussed earlier to develop the Hosmer-Lemeshow test may be used to obtain an equivalent statistic for the validation sample. Assume that 10 groups composed of the deciles of risk is used. Any other grouping strategy can be used with obvious modifications to the calculation. Let n_k denote the approximately $n_v/10$ observations in the k^{th} decile of risk. Let $o_k = \sum y_j$ be the number of positive responses among the covariate patterns falling in the k^{th}decile of risk. The estimate of the expected value of o_k under the assumption that the model is correct is $e_k = \sum m_j\pi_j$ where the sum is over the covariate patterns in the decile of risk. The Hosmer-Lemeshow statistic is obtained as the Pearson chi-square statistic computed from the observed and expected frequencies

$$C_v = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{n_k\overline{\pi}_k(1-\overline{\pi}_k)}$$

where

$$\overline{\pi}_k = \sum m_j\pi_j/n_k.$$

The subscript, v, has been added to C to emphasize that the statistic has been calculated from a validation sample. Under the hypothesis that the model is correct, and the assumption that each e_k is sufficiently large for each term in C_v to be distributed as $\chi^2(1)$, it follows that C_v is distributed as $\chi^2(10)$. In general, if g groups are used then the distribution is $\chi^2(g)$. In addition to calculating a p-value to assess overall fit, it is recommended that each term in C_v be examined to assess the fit within each decile of risk.

In credit scoring, validation of a model is not so focused on the fit of the model, but rather the stability of the population and the model's predictive power and accuracy. To track the stability of the population, the distributions of the different covariates are compared to its distribution at development. To determine whether the model retained its predictive power, the Gini coefficient and/or ROC curve can be used to compare to the development data. The Kolmogorov-Smirnov statistic can also be used here.

Chapter 11

Multinomial and Ordinal Regression

11.1 The Multinomial Logistic Regression model

In the previous chapters, the focus was on the use of the logistic regression model when the outcome variable is dichotomous or binary. The model can easily be extended and modified to handle the case where the outcome variable is nominal with more than two levels as shown by Hosmer and Lemeshow (2000). For example, consider the choice of insurance plans on a credit card from among three plans offered to the customers of a bank. The outcome variable has three levels indicating which plan, A, B, or C, is chosen. Possible covariates might include gender, age, income, family size and others. The goal is to model the odds of plan choice as a function of the covariates and to express the results in terms of odds ratios for choice of different plans. A model with these options is often referred to as a discrete choice model, a multinomial, polychotomous or a polytomous logistic regression model. In this study, the term multinomial will be used.

An outcome variable with any number of levels can be used to illustrate the extension of the model and the methods. However, the details are most easily illustrated with three categories. Further generalization to more than three categories is more a problem of notation than of concepts. For the remainder of this study, only the situation where the outcome variable has three categories will be considered.

When one considers a regression model for a discrete outcome variable with more than two responses, one must pay attention to the measurement scale. The logistic regression model for a case in which the outcome is nominal scale will be discussed first. Later in this chapter logistic regression models for ordinal scale outcomes will be discussed.

For this study, the assumption will be that the categories of the outcome variable, Y, are coded as 0, 1 or 2. Some software packages might not allow a zero code and might require the codes to begin with 1, so it is worth checking before jumping into modeling.

The logistic regression model used for a binary outcome variable is parameterized in terms of the logit of $Y = 1$ versus $Y = 0$. In the three outcome category model two logit functions are needed. One has to decide which outcome categories to compare. The obvious extension is to use $Y = 0$ as the referent or baseline outcome and to form logits comparing $Y = 1$ and $Y = 2$ to it. It will be shown later that the logit function for $Y = 2$ versus $Y = 1$ is the difference between these two logits.

To develop the model, assume p covariates and a constant term denoted by the vector, $\mathbf{x}$, of length $p+1$ where $x_0 = 1$, is available. The two logit functions is denoted as

$$\begin{aligned} g_1(\mathbf{x}) &= \ln\left[\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] \\ &= \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + ... + \beta_{1p}x_p \\ &= \mathbf{x}'\boldsymbol{\beta}_1 \end{aligned}$$

and

$$\begin{aligned} g_2(\mathbf{x}) &= \ln\left[\frac{P(Y=2|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] \\ &= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + ... + \beta_{2p}x_p \\ &= \mathbf{x}'\boldsymbol{\beta}_2 \end{aligned}$$

It follows that the conditional probabilities of each outcome category given the covariate vector are

$$P(Y=0|\mathbf{x}) = \frac{1}{1+e^{g_1(\mathbf{x})}+e^{g_2(\mathbf{x})}} \ ,$$

$$P(Y=1|\mathbf{x}) = \frac{e^{g_1(\mathbf{x})}}{1+e^{g_1(\mathbf{x})}+e^{g_2(\mathbf{x})}}$$

and

$$P(Y=2|\mathbf{x}) = \frac{e^{g_2(\mathbf{x})}}{1+e^{g_1(\mathbf{x})}+e^{g_2(\mathbf{x})}}.$$

Following the convention for the binary model, let $\pi_j = P(Y=j|\mathbf{x})$ for $j=0,1,2$. Each probability is a function of the vector of $2(p+1)$ parameters $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$.

A general expression for the conditional probability in the three category model is

$$P(Y=j|\mathbf{x}) = \frac{e^{g_j(\mathbf{x})}}{\sum_{k=0}^{2} e^{g_k(\mathbf{x})}}$$

where the vector $\boldsymbol{\beta}_0 = \mathbf{0}$ and $g_0(\mathbf{x})$ =0.

To construct the likelihood function three binary variables coded 0 or 1 need to be created, to indicate the group membership of an observation. Note that these variables are introduced only to clarify the likelihood function and are not used in the actual multinomial logistic regression analysis.

The variables are coded as follows:

- If $Y=0$ then $Y_0=1$, $Y_1=0$ and $Y_2=0$.
- If $Y=1$ then $Y_0=0$, $Y_1=1$ and $Y_2=0$.
- If $Y=2$ then $Y_0=0$, $Y_1=0$ and $Y_2=1$.

No matter what value Y takes on, the sum of these variables is $\sum_{j=0}^{2} Y_j = 1$. Using this notation it follows that the conditional likelihood function for a sample of n independent observations is

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\pi_0(\mathbf{x}_i)^{y_{0i}} \pi_1(\mathbf{x}_i)^{y_{1i}} \pi_2(\mathbf{x}_i)^{y_{2i}}\right].$$

Taking the log and using the fact that $\sum y_{ij} = 1$ for each i, the log-likelihood function is:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_{1i} g_1(\mathbf{x}_i) + y_{2i} g_2(\mathbf{x}_i) - \ln\left[1 + e^{g_1(\mathbf{x}_i)} + e^{g_2(\mathbf{x}_i)}\right].$$

The likelihood equations are found by taking the first partial derivatives of $L(\boldsymbol{\beta})$ with respect to each of the $2(p+1)$ unknown parameters. To simplify the notation a bit, let $\pi_{ji} = \pi_j(\mathbf{x}_i)$. The general form of these equations is:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ki}(y_{ji} - \pi_{ji})$$

for $j = 1, 2$ and $k = 0, 1, 2, ..., p$, with $x_{01} = 1$ for each subject.

The maximum likelihood estimator, $\widehat{\boldsymbol{\beta}}$, is obtained by setting these equations equal to zero and solving for $\boldsymbol{\beta}$. The solution requires the same type of iterative computation that is used to obtain the estimate in the binary outcome case.

The matrix of second partial derivatives is required to obtain the information matrix and the estimator of the covariance matrix of the maximum likelihood estimator. The general form of the elements in the matrix of second partial derivatives is as follows:

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jk} \partial_{jk'}} = -\sum_{i=1}^{n} x_{k'i} x_{ki} \pi_{ji}(1 - \pi_{ji})$$

and

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jk} \partial_{j'k'}} = \sum_{i=1}^{n} x_{k'i} x_{ki} \pi_{ji} \pi_{j'i'}$$

for j and $j' = 1, 2$ and k and $k' = 0, 1, 2, ..., p$.

The observed information matrix, $\mathbf{I}(\widehat{\boldsymbol{\beta}})$, is the $2(p+1) \times 2(p+1)$ matrix whose elements are the negatives of the values in the two above equations evaluated at $\widehat{\boldsymbol{\beta}}$. The estimator of the covariance matrix of the maximum likelihood estimator is the inverse of the observed information matrix,

$$\widehat{\mathbf{Var}}(\widehat{\boldsymbol{\beta}}) = \mathbf{I}(\widehat{\boldsymbol{\beta}})^{-1}.$$

A more concise representation for the estimator of the information matrix may be obtained by using a form similar to the binary outcome case. Let the matrix $\mathbf{X}$ be the $n \times (p+1)$ matrix containing the values of the covariates for each subject, let the matrix $\mathbf{V}_j$ be the $n \times n$ diagonal matrix with general element $\widehat{\pi}_{ji}(1 - \widehat{\pi}_{ji})$ for $j = 1, 2$ and $i = 1, 2, 3, ..., n$, and let $\mathbf{V}_3$ be the $n \times n$ diagonal matrix with general element $\widehat{\pi}_{1i}\widehat{\pi}_{2i}$. The estimator of the information matrix may be expressed as

$$\widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})_{11} & \widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})_{12} \\ \widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})_{21} & \widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})_{22} \end{bmatrix}$$

where

$$\widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})_{11} = (\mathbf{X}'\mathbf{V}_1\mathbf{X}),$$

$$\widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})_{22} = (\mathbf{X}'\mathbf{V}_2\mathbf{X})$$

and

$$\widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})_{12} = \widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})_{21} = -(\mathbf{X}'\mathbf{V}_3\mathbf{X}).$$

11.1.1 Interpreting and Assessing the Significance of the Estimated Coefficients

To simplify the discussion of the estimation and interpretation of odds ratios in the multinomial outcome setting, the generalization of the notation used in the binary outcome is needed to include the outcomes being compared as well as the values of the covariate. Assume that the outcome labeled with $Y = 0$ is the reference outcome. The subscript on the odds ratio indicates which outcome is being compared to the reference outcome. The odds ratio of outcome $Y = j$ versus the outcome $Y = 0$ for covariate values of $x = a$ versus $x = b$ is

$$OR_j(a, b) = \frac{P(Y = j|x = a)/P(Y = 0|x = a)}{P(Y = j|x = b)/P(Y = 0|x = b)}.$$

In the special case when the covariate is binary, coded 0 or 1, the notation can be simplified further and let $OR_j = OR_j(1, 0)$.

To explain the basic concepts, consider a model containing a single dichotomous covariate coded 0 or 1. In the binary outcome model, the estimated slope coefficient is identical to the log-odds ratio obtained from the 2×2 table cross-classifying the outcome and the covariate. As noted earlier in this chapter, when the outcome has three levels, there are two logit functions. These functions are defined in such a way that the two estimated coefficients, one from each logit function, are equal to the log-odds ratios from the pair of 2×2 tables obtained by cross-classifying the $y = j$ and $y = 0$ outcomes by the covariate with $y = 0$ as the reference outcome value.

Refer to the earlier example of the insurance plan and suppose that gender is available as the independent variable/ covariate. The data are in the below table:

	Gender			
Insurance product	Male(0)	Female(1)	Total	$\widehat{OR}$
A(0)	220	14	234	1.0
B(1)	85	19	104	3.51
C(2)	63	11	74	2.74
Total	368	44	412	

If insurance product=0 (A) is used as the reference outcome, the two odds ratios calculated from the data in the above table are

$$\widehat{OR}_1 = \frac{19 \times 220}{85 \times 14} = 3.51$$

and

$$\widehat{OR}_2 = \frac{11 \times 220}{63 \times 14} = 2.74.$$

Note in SAS there are two ways to fit the three-category logistic regression model using Proc Logistic, dependent on the data. If the data is in a frequency table, as above the resulting SAS code is:

```
proc logistic data=combine;
  freq f;
  class gender/order=data param=ref ref=first;
  model IP (ref='0')=gender/link=glogit
     aggregate scale=none covb;
run;
```

where f refers to the cell frequencies in the above table.

If the data is in its raw form, i.e. using the example where the data set has 412 observations, the needed SAS code is:

```
proc logistic data=multi;
  class gender/order=data param=ref ref=first;
  model IP (ref='0')=gender/ link=glogit
        scale=none covb ;
run;
```

Note that, irrespective of which way the data set is constructed, if the same reference category is used, the results from fitting the model will be the same.

The results from fitting a three-category logistic regression in SAS, gives the following output:

```
                Analysis of Maximum Likelihood Estimates

                                          Standard          Wald
Parameter       IP     DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept       1       1     -0.9510      0.1277       55.4474        <.0001
Intercept       2       1     -1.2505      0.1429       76.5842        <.0001
gender     1    1       1      1.2564      0.3747       11.2448        0.0008
gender     1    2       1      1.0093      0.4275        5.5744        0.0182


                          Odds Ratio Estimates

                                     Point          95% Wald
          Effect                IP   Estimate      Confidence Limits

          gender 1 vs 0         1       3.513       1.685       7.320
          gender 1 vs 0         2       2.744       1.187       6.342
```

The results are summarized in the following table:

Logit	**Variable**	**Coefficient**	**Std error**	$\widehat{OR}$	95% **CI**
1	Gender	1.256	0.3747	3.51	1.685, 7.321
	Constant	−0.951	0.1277		
2	Gender	1.009	0.4275	2.74	1.187, 6.342
	Constant	−1.250	0.1429		

The $\widehat{OR}$ in the above table is obtained by taking the exponent of the estimated slope coefficients. Note that they are identical to the values calculated and tabled above in the previous table. As is the case in the binary outcome setting with a dichotomous covariate, the estimated standard error of the coefficient is the square root of the sum of the inverse of the cell frequencies. For example, the estimated standard error of the coefficient gender in the first logit is:

$$SE(\widehat{\beta}_{11}) = \left[\frac{1}{19} + \frac{1}{220} + \frac{1}{85} + \frac{1}{14}\right]^{0.5} = 0.3747$$

which is identical to the above table.

The endpoints of the confidence interval are obtained in exactly the same manner as for the binary outcome case. First the confidence interval for the coefficient is obtained, then the exponents of the endpoints are taken to obtain the confidence interval for the odds ratio. For example, the 95% CI for the odds ratio of IP=1 versus IP=0 shown in the above tables is calculated as follows:

$$\exp(1.256 \pm 1.96 \times 0.3747) = (1.685, 7.321).$$

The endpoints for the confidence interval for IP=2 versus IP=0 in the above table are obtained in a similar matter.

Each estimated odds ratio and its corresponding confidence interval is interpreted as if it came from a binary outcome setting. In some cases is may further support the analysis to compare the magnitude of the two estimated odds ratios. This can be done with or without the support tests of equality.

The interpretation of the effect of gender on chosen insurance product is as follows:

- The odds among customers who are female choosing product B are 3.5 times greater than the odds among customers who are male. In other words, customers who are female are 3.5 times more likely to choose product B than customers who are male. The confidence interval indicates that the odds could be as little as 1.7 times or as much as 7.3 times larger with 95 percent confidence.
- The odds among customers who are female choosing product C are 2.7 times greater than the odds among customers who are male. Put in another way, customers who are female are 2.7 times more likely to choose product C than customers who are male. The odds could be as little as 1.2 times or as much as 6.3 times larger with 95 percent confidence.

Thus, one can see that gender is a significant factor in choice of insurance product.

Note that the test of the equality of two odds ratios, $OR_1 = OR_2$, is equivalent to a test that the log-odds for IP=2 versus IP=1 is equal to zero. The simplest way to obtain the point and interval estimate is from the difference between the two estimated slope coefficients in the logistic regression model. For example, using the above example:

$$\begin{aligned}\widehat{\beta}_{21} - \widehat{\beta}_{11} &= 1.009 - 1.256 \\ &= -0.247 \\ &= \ln(\tfrac{11\times 85}{19\times 63})\end{aligned}$$

The estimator of the variance of the difference between the two coefficients, $\widehat{\beta}_{21} - \widehat{\beta}_{11}$, is

$$\widehat{Var}(\widehat{\beta}_{21} - \widehat{\beta}_{11}) = \widehat{Var}(\widehat{\beta}_{21}) + \widehat{Var}(\widehat{\beta}_{11}) - 2\widehat{Cov}(\widehat{\beta}_{21}, \widehat{\beta}_{11}).$$

The values for the estimates of the variances and covariances are obtained from the estimated covariance matrix, which is provided by most statistical software packages, like SAS:

Estimated Covariance Matrix

Parameter	Intercept_ 1	Intercept_ 2	gender1_1	gender1_2
Intercept_1	0.01631	0.004545	-0.01631	-0.00455
Intercept_2	0.004545	0.020418	-0.00455	-0.02042
gender1_1	-0.01631	-0.00455	0.14037	0.075974
gender1_2	-0.00455	-0.02042	0.075974	0.182756

The estimates can also be calculated using the general formula:

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}})^{-1}$$

where the variance matrix is the inverse of the information matrix and it is calculated as explained earlier.

The form of this matrix is a little different from the covariance matrix in the binary setting. There are two matrices containing the estimates of the variances and covariances of the estimated coefficients in each logit and a third containing the estimated covariances of the estimated coefficients from the different logits.

The matrix for the above model is summarized from the SAS output and is shown in this next table, where Logit 1 is the logit function for IP=1 versus IP=0 and Logit 2 is the logit function for IP=2 versus IP=0.

		Logit1		**Logit2**	
		Gender	Constant	Gender	Constant
Logit1	Gender	0.1404			
	Constant	−0.0163	0.0163		
Logit2	Gender	0.0760		0.1828	
	Constant	−0.0045	0.0045	−0.0204	0.0204

Using the results from the above table the estimate of the variance of the difference in the two estimated coefficients is obtained as

$$\widehat{Var}(\widehat{\beta}_{21} - \widehat{\beta}_{11}) = 0.1404 + 0.1828 - 2 \times 0.0760 = 0.1712.$$

The endpoints of a 95 percent confidence interval for this difference are

$$-0.247 \pm 1.96 \times \sqrt{0.1712} = (-1.058, 0.564).$$

Since the confidence interval includes zero one cannot conclude that the log odds for IP=1 is different from the log odds for IP=2. Equivalently, one can express these results in terms of odds ratios by taking the exponent of the point and interval estimates. This yields an odds ratio for IP=2 versus IP=1 as $\widehat{OR} = 0.781$ and a confidence interval of $(0.347, 1.758)$. Note that this confidence interval includes 1, so one cannot conclude that the odds ratio for IP=1 is different from the odds ratio for IP=2. The interpretation of the odds ratio is that the odds of product B is 22 percent lower than the odds of product C for customers that are female, i.e. $\widehat{OR}_2 \approx 0.78 \times \widehat{OR}_1$.

In practice, if there was no difference in the separate odds ratios over all covariates, the modeler might consider pooling outcome categories 1 and 2 into a binary (product A versus not product A) outcome. In a model with many covariates the extra computations for these additional comparisons could become a burden. Luckily software packages like SAS, has the option to provide the modeler with these comparisons.

An indication of the importance of the variable may be obtained from the two Wald statistics, but as is the case with any multi-degree of freedom variable, one should rather use the likelihood ratio test for the significance of the coefficients. For example, to test for the significance of the coefficients for gender, the log-likelihood function from the model containing gender to the log-likelihood for the model containing only the two constant terms is compared, one for each logit. Under the null hypothesis that the coefficients are zero, minus twice the change in log-likelihood follows a chi-square distribution with 2 degrees of freedom.

The output from SAS:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
-2 Log L	805.198	792.340

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12.8581	2	0.0016

The value of the likelihood test statistic is 12.8581, which yields a p-value of 0.0016. Thus, from a statistical point of view, the variable gender is significantly associated with the chosen insurance product.

In general, the likelihood ratio test for the significance of the coefficients for a variable has degrees of freedom equal to the number of outcome categories minus one times the degrees of freedom for the variable in each logit. For example, if one has a four category outcome variable and a covariate that is modeled as continuous then the degrees of freedom is $(4-1) \times 1 = 3$. If one has a categorical covariate coded at five levels then the covariate has 4 design variables within each logit and the degrees of freedom for the test are $(4-1) \times (5-1) = 12$. This is easy to keep track of if one can remember that one is modeling one logit for comparing the reference outcome category to each other outcome category.

For a polytomous covariate the number of odds ratios is expanded to include comparisons of each level of the covariate to a reference level for each to a reference level for each possible logit function. To illustrate this, consider the variable "Underwriter" modeled via two design variables using the value of 1 (Company X) as the reference category.

The cross classification of the insurance product (IP) by Underwriter (UW) is given in the below table:

	UW			
IP	X	Y	Z	Total
A(0)	13	77	144	234
B(1)	1	12	91	104
C(2)	4	16	54	74
Total	18	105	289	412

Using the value of IP=0 as the reference outcome category and UW=1 as the reference covariate value, the four odds ratios are as follows:

$$\widehat{OR}_1(2,1) = \frac{12 \times 13}{77 \times 1} = 2.03,$$

$$\widehat{OR}_1(3,1) = \frac{91 \times 13}{144 \times 1} = 8.22,$$

$$\widehat{OR}_2(2,1) = \frac{16 \times 13}{77 \times 4} = 0.68,$$

and

$$\widehat{OR}_2(3,1) = \frac{54 \times 13}{144 \times 4} = 1.22.$$

Using the following SAS code, the logistic regression model was fitted. Note that the data was read in as a frequency table (like the table above).

```
data combine;
  input IP UW f;
  datalines;
  0 1 13
  0 2 77
  0 3 144
  1 1 1
  1 2 12
  1 3 91
  2 1 4
  2 2 16
  2 3 54
  ;
run;

proc logistic data=combine;
  freq f;
  class UW/order=data param=ref ref=first;
  model IP (ref='0')=UW/link=glogit
     aggregate scale=none covb;
run;
```

The *freq f* statement indicates that the data is in the form of a frequency/contingency table and the data is aggregated (the option *aggregate* on the model statement).

Fitting the logistic regression model gives the following results:

Analysis of Maximum Likelihood Estimates

Parameter	IP	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-2.5649	1.0377	6.1091	0.0134
Intercept	2	1	-1.1787	0.5718	4.2494	0.0393
UW	2	1	0.7060	1.0831	0.4248	0.5145
UW	2	2	-0.3926	0.6344	0.3830	0.5360
UW	3	1	2.1059	1.0463	4.0509	0.0441
UW	3	2	0.1978	0.5936	0.1111	0.7389

The results are summarized in the following table:

Logit	**Variable**	**Coefficient**	**Std error**	$\widehat{OR}$	**95% CI**
1	UW_2	0.706	1.0831	2.03	0.242, 16.928
	UW_3	2.106	1.0463	8.22	1.057, 63.864
	Constant	−2.565	1.0377		
2	UW_2	−0.393	0.6344	0.68	0.195, 2.341
	UW_3	0.198	0.5936	1.22	0.381, 3.901
	Constant	−1.179	0.5718		

The odds ratios are obtained by taking the exponent of the estimated logistic regression coefficients for instance, $e^{0.706} = 2.03$.

These are equal to the odds ratios formed from the 2 by 2 tables obtained from the main 3 by 3 contingency table. The odds ratios for logit 1 are obtained from the 2 by 3 table containing the rows corresponding to the IP=0 and IP=1 and the 3 columns. The odds ratios for logit 2 are obtained from the 2 by 3 table containing the rows corresponding to IP=0 and IP=2 and the 3 columns.

To asses the significance of the variable UW, the likelihood ratio test is used, with the following output from SAS given:

```
                    Model Fit Statistics

                                          Intercept
                               Intercept        and
                Criterion          Only Covariates
                -2 Log L        805.198    778.401


          Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        26.7969        4         <.0001
```

The test statistic for the likelihood ratio test has a value $G = 26.7969$ which, with 4 degrees of freedom, yields a p-value of less than 0.001. Thus one can conclude that the choice of insurance product is significantly associated with the underwriter issuing the policy. Examining the estimated odds ratios and their confidence intervals one can see that the association is strongest when comparing insurance product B to insurance product A, and comparing underwriter X to underwriter Z. The interpretation is that the odds of choosing insurance product A when the underwriter is company Z, is 8.22 times larger than the odds of choosing insurance product A when the underwriter is company X.

The confidence interval estimates for the logit of IP=1 versus IP=0 are quite wide. These estimates are a function of the cell counts and for this particular case there is a cell with only one subject in (see above table). It follows from the fact that the estimated standard errors are equal to the square root of the sum of the inverse of the cell counts. For example, the estimated standard error of the coefficient for the log odds of UW=3 vs. UW=1 in the first logit is:

$$SE(\widehat{\beta}_{12}) = \left[\frac{1}{91} + \frac{1}{13} + \frac{1}{144} + \frac{1}{1}\right]^{0.5} = 1.0463.$$

Continuous covariates that are modeled as linear in the logit have a single estimated coefficient in each logit function. This coefficient, when the exponent is taken, gives the estimated odds for a change of one unit in the variable. Knowing what a single unit is, and estimation of odds ratios for clinically meaningful change apply directly to each logit function in the multinomial logistic regression model.

11.1.2 Model building strategies for multinomial logistic regression

In principle, the strategies and methods for multivariate modeling with a multinomial outcome variable is identical to those for the binary outcome variable discussed earlier.

Since a full multinomial logistic model can be approximated by separate binary logistic models, it opens up the possibility of performing variable selection using the stepwise or best subsets approaches discussed. Thus, in absence of software capable of fitting a multinomial logistic regression model, one could use the results of individual logistic regressions, realizing of course that the resulting estimates are approximations to the maximum likelihood estimates.

One problem that one was not faced with in the binary outcome case but which can be an issue in a multinomial logistic regression model, occurs when a covariate is significant for some but not all logit functions. If one models using the principle that one would like to minimize the number of parameters, then one should force the coefficients to be zero in some logit functions and estimate their values for the other logit functions. As in all modeling situations, clinical considerations should play an important role in variable selection.

11.1.3 Assessing the fit and diagnostics for the multinomial logistic regression model

As with any fitted model, before it is used to make inferences, its overall fit should be assessed and the contribution of each subject to the fit examined. In multinomial logistic regression, the multiple outcome categories make this a more difficult problem than was the case with a model for a binary outcome variable. When one models a binary outcome variable there is a single fitted value, the estimated logistic probability of the outcome being present, $P(Y = 1|\mathbf{x})$. When the outcome variable has three categories there is two estimated logistic probabilities, the estimated probabilities of categories 1 and 2, $P(Y = 1|\mathbf{x})$ and $P(Y = 2|\mathbf{x})$. There are extensions of the tests for goodness-of-fit and logistic regression diagnostics to the multinomial logistic regression model, however these methods are not that easy to calculate and not readily available with most of the available software. It is therefore recommended to assess the fit and calculating the logistic regression diagnostics using the individual logistic regressions approach.

For an outcome variable with three categories, it is suggested to assess the fit of the two logistic regression models and then integrating the results, usually descriptively, to make a statement about the fit of the multinomial logistic regression model. The procedure for assessing the fit of each individual logistic regression model is described in an earlier chapter. Integration of the results requires thoughtful consideration of the effects of influential and poorly fit covariate patterns on each logit function. In particular, covariate patterns that are influential for only one logit should be examined closely with due consideration to business issues before they are excluded from analyses. While this process requires more computation than for a single logistic regression model for a binary outcome variable, there is nothing new conceptually.

Summary goodness-of-fit statistics are calculated using the observed covariate patterns generated by the variables in the model. These summary statistics include the Hosmer-Lemeshow statistic and the Pearson χ^2.

The leverage, h, and diagnostic statistics $\Delta\widehat{\boldsymbol{\beta}}, \Delta X^2$ and ΔD defined earlier are calculated for each covariate pattern for each of the individually fit logistic regression models. Plots of the diagnostic statistics versus the estimated probabilities are then used to identify patterns with large values for one or more of the statistics. Exclusions of observations in covariate patterns should be based on clinical plausibility. Also, try not to discard data on a large number of observations representing a fairly common response pattern without first trying to improve the model.

Note that individual logistic regressions are performed and the diagnostic statistics are examined for each logit. Computations are done on separate data sets, thus if pattern numbers are assigned to covariate pattern (to facilitate discussion of the values of the diagnostic statistics), the pattern numbers for the different logits will not refer to the same covariate patterns.

The real challenge when fitting a multinomial logistic regression model is the fact that there are multiple odds ratios for each model covariate. This certainly complicates the discussion. On the other hand, using a multinomial outcome can provide more complete description of the process being studied. If one combines two of the categories for a three category outcome variable, one might miss differences in odds ratios. Thus, from a statistical point of view, one should not pool the outcome categories, unless the estimated coefficients in the logits are not significantly different from each other.

In summary, fitting and interpreting the results from a multinomial logistic regression model follows the same basic paradigm as was the case for the binary model. The difference is that the user should be aware of the possibility that informative comparative statements may be required for the multiple odds ratios for each covariate.

11.2 Ordinal logistic regression models

There are occasions when the scale of a multiple category outcome is not nominal but ordinal. Common examples of ordinal outcomes include variables such as extent of disease (none, some, severe), job performance (very poor, poor, average, above average, outstanding) and opinions in a survey on some issue (strongly disagree, disagree, agree, strongly agree). In such a setting one could use the multinomial logistic model described in the previous section. This analysis, however, would not take into account the ordinal nature of the outcome and hence the estimated odds ratios may not address the questions asked of the analysis.

What complicates this model is that there are more than one logistic regression to choose from. Three of the most commonly used models will be discussed: the adjacent-category, the continuation-ratio and the proportional odds model.

Assume that the ordinal outcome variable, Y, can take on $K+1$ values coded $0, 1, 2, ..., K$. Denote a general expression for the probability that the outcome is equal to k conditional on the vector $\mathbf{x}$ of p covariates as $P[Y = k|\mathbf{x}] = \phi_k(\mathbf{x})$. If one assumes that the model is the multinomial logistic model in the previous section, then $\phi_k(\mathbf{x}) = \pi_k(\mathbf{x})$. In the context of ordinal logistic regression models the multinomial model is frequently called the baseline logit model. This term arisis from the fact that the model is usually parameterized so that the coefficients are log-odds comparing category $Y = k$ to a "baseline" category, $Y = 0$. The fully parameterized baseline logistic regression model has $K \times (p+1)$ coefficients. Under this model the logits are:

$$g_k(\mathbf{x}) = \ln\left[\frac{\pi_k(\mathbf{x})}{\pi_0(\mathbf{x})}\right] = \beta_{k0} + \mathbf{x}'\boldsymbol{\beta}_k$$

for $k = 1, 2, ...K$.

When moving to an ordinal model one has to decide what outcomes to compare and what the most reasonable model is for the logit. For example, suppose that one wishes to compare each response to the next larger response. This model is called the adjacent-category logistic model. If it is assumed that the log odds does not depend on the response and the log odds is linear in the coefficients then the adjacent category logits are as follows:

$$a_k(\mathbf{x}) = \ln\left[\frac{\phi_k(\mathbf{x})}{\phi_{k-1}(\mathbf{x})}\right] = \alpha_k + \mathbf{x}'\boldsymbol{\beta}$$

for $k = 1, 2, ..., K$.

The adjacent-category logits are a constrained version of the baseline logits. To see this, express the baseline logits in terms of the adjacent-category logits as follows:

$$\begin{aligned}
\ln\left[\tfrac{\phi_k(\mathbf{x})}{\phi_0(\mathbf{x})}\right] &= \ln\left[\tfrac{\phi_1(\mathbf{x})}{\phi_0(\mathbf{x})}\right] + \ln\left[\tfrac{\phi_2(\mathbf{x})}{\phi_1(\mathbf{x})}\right] + ... + \ln\left[\tfrac{\phi_k(\mathbf{x})}{\phi_{k-1}(\mathbf{x})}\right] \\
&= a_1(\mathbf{x}) + a_2(\mathbf{x}) + ... + a_k(\mathbf{x}) \\
&= (\alpha_1 + \mathbf{x}'\boldsymbol{\beta}) + (\alpha_2 + \mathbf{x}'\boldsymbol{\beta}) + ... + (\alpha_k + \mathbf{x}'\boldsymbol{\beta}) \\
&= (\alpha_1 + \alpha_2 + ... + \alpha_k) + k\mathbf{x}'\boldsymbol{\beta}
\end{aligned}$$

Thus the model in the above equation is a version of the baseline model with intercept $\beta_0 = (\alpha_1 + \alpha_2 + ... + \alpha_k)$ and slope coefficients $\boldsymbol{\beta}_k = k\boldsymbol{\beta}$. An easy way to fit the adjacent-category model is thus via a constrained baseline logistic model.

Suppose instead of comparing each response to the next larger response, one compares each response to all lower responses, that is $Y = k$ versus $Y < k$ for $k = 1, 2, ..., K$. This model is called the continuation-ratio logistic model. The logit for this model is defined as follows:

$$\begin{aligned}
r_k(\mathbf{x}) &= \ln\left[\tfrac{P(Y=k|\mathbf{x})}{P(Y<k|\mathbf{x})}\right] \\
&= \ln\left[\frac{\phi_k(\mathbf{x})}{\phi_0(\mathbf{x})+\phi_1(\mathbf{x})+...+\phi_{k-1}(\mathbf{x})}\right] \\
&= \theta_k + \mathbf{x}'\boldsymbol{\beta}_k
\end{aligned}$$

for $k = 1, 2, ..., K$.

Under the parameterization in the above equation, the continuation-ratio logits have different constant terms and slopes for each logit. The advantage of this unconstrained parameterization is that the model can be fit via K ordinary binary logistic regression models.

The model given above can also be constrained to have a common vector of slope coefficients and different intercepts namely:

$$r_k(\mathbf{x}) = \theta_k + \mathbf{x}'\boldsymbol{\beta}$$

Note that it is also possible to define the continuation ratio in terms of $Y = k$ versus $Y > k$ for $k = 1, 2, ..., K$. Unfortunately the results from the two parameterizations are not equivalent. The parameterization $Y = k$ versus $Y < k$ is usually preferred since, if $K = 1$ the model simplifies to the usual logistic regression model where the odds ratio compare response $Y = 1$ to $Y = 0$.

The third ordinal logistic regression model considered is the proportional odds model. With this model one compares the probability of an equal or smaller response, $Y \leq k$, to the probability of a larger response $Y > k$

$$\begin{aligned} c_k(\mathbf{x}) &= \ln\left[\frac{P(Y \leq k|\mathbf{x})}{P(Y > k|\mathbf{x})}\right] \\ &= \ln\left[\frac{\phi_o(\mathbf{x})+\phi_1(\mathbf{x})+...+\phi_k(\mathbf{x})}{\phi_{k+1}(\mathbf{x})+\phi_{k+2}(\mathbf{x})+...+\phi_K(\mathbf{x})}\right] \\ &= \tau_k - \mathbf{x}`\boldsymbol{\beta} \end{aligned}$$

for $k = 1, 2, ..., K$.

Note that in the case when $K = 1$ the model as defined here simplifies to the complement of the usual logistic regression model in that it yields odds ratios of $Y = 0$ versus $Y = 1$.

The method used to fit each of the models, except the unconstrained continuation-ratio model, is based on an adaptation of the multinomial likelihood and its log as shown here for $K = 2$:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_{1i}g_1(\mathbf{x}_i) + y_{2i}g_2(\mathbf{x}_i) - \ln(1 + e^{g_1(\mathbf{x}_i)} + e^{g_2(\mathbf{x}_i)})$$

The basic procedure involves the following steps:

1. The expressions defining the model specific logits are used to create an equation defining $\phi_k(\mathbf{x})$ as a function of the unknown parameters.
2. The values of a $K + 1$ dimensional multinomial outcome, $\mathbf{z}` = (z_0, z_1, ..., z_K)$, are created from the ordinal outcome as $z_k = 1$ if $y = k$ and $z_k = 0$ otherwise. It follows that only one value of z is equal to one.

The general form of the likelihood for a sample of n independent observations, $(y_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$, is

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\phi_0(\mathbf{x}_i)^{z_{0i}}\phi_1(\mathbf{x}_i)^{z_{1i}} \times ... \times \phi_K(\mathbf{x}_i)^{z_{Ki}}]$$

where $\boldsymbol{\beta}$ is used to denote both the p slope coefficients and the K model-specific intercept coefficients. It follows that the log-likelihood function is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} z_{0i} \ln[\phi_0(\mathbf{x}_i)] + z_{1i} \ln[\phi_1(\mathbf{x}_i)] + ... + z_{Ki} \ln[\phi_K(\mathbf{x}_i)].$$

The MLE's of the parameters are obtained by differentiating the above equation with respect to each of the unknown parameters, setting each of the $K + p$ equations equal to zero and solving for $\widehat{\boldsymbol{\beta}}$. The estimator of the covariance matrix of the estimated coefficients is obtained by evaluating the inverse of the negative of the matrix of the second partial derivatives at $\widehat{\boldsymbol{\beta}}$.

An ordinal outcome can arise in a number of different ways. For example, an ordinal outcome can be created by categorizing an observed continuous outcome variable. Alternatively, categories may be observed that can be hypothesized to have come from categorizing a hypothetical and unobserved continuous outcome. This is often a useful way to envision outcome scales in categories ranging from strongly disagree to strongly agree. Another possibility is that the outcome is a composite of a number of other scored variables. Common examples are health status or extent of disease, which arise from many individual clinical indicators.

In general, the continuation-ratio model might be preferred over the baseline and adjacent-category model when the conditioning used in defining and fitting the model makes clinical sense. A common example is one where the number of attempts to pass a test or attain some binary outcome is modeled. The first logit models the log odds of passing the test the first time it is taken. The second logit models the log odds of passing the test on the second attempt, givent that it was not passed on the first attempt. This process continues until one is modeling the Kth attempt.

The choice of what model to ultimately used in any problem should consider which odds ratios are most informative for the problem as well as an assessment of model adequacy.

The baseline model will now be illustrated with an example to aid in the interpretation of the odds ratios that result from it.

Example 17 *As an example, form a four category outcome from household disposable monthly income using the following cut-off points:* $R4,000$, $R8,000$ *and* $R12,000$. *This example is not typical of many ordinal outcomes that use loosely defined "low", "medium" or "high" categorizations of some measurable quantity. Instead, here this variable was explicitly derived from a measured continuous variable.*

First, one needs to give some thought to the assignment of codes to the outcome variable, as this has implications on the definition of the odds ratio calculated by the various ordinal models. The obvious choice is to use the naturally increasing sequence of codes: 0 *if HMI (household monthly income)*$\leq 4,000$, 1 *if* $4,000 < HMI \leq 8,000$, 2 *if* $8,000 < HMI \leq 12,000$ *and* 3 *if* $HMI > 12,000$. *This coding is appropriate if one wants low or lower income as the reference outcome.*

However, say for instance that a decreasing sequence of codes makes more sense and one wants the highest income to be the reference outcome. One will then code the variable as follows: 3 *if HMI (household monthly income)* $\leq 4,000$, 2 *if* $4,000 < HMI \leq 8,000$, 1 *if* $8,000 < HMI \leq 12,000$ *and* 0 *if* $HMI > 12,000$. *This coding is used for HMI as the outcome variable for the example. The actual coding, for the most part, does not make a difference, as long as one is able to figure out how to correct the signs of the coefficients obtained by software packages.*

The data for the example:

Household	*Residential status (RS)*		
monthly income	*Own*(0)	*Rent*(1)	*TOTAL*
$0: HMI > 12,000$	35	11	46
$1: 8,000 < HMI \leq 12,000$	29	17	46
$2: 4,000 < HMI \leq 8,000$	22	16	38
$3: HMI \leq 4,000$	29	30	59
TOTAL	115	74	189

As a starting point, consider the cross-classification of HMI versus residential status. The odds ratios for the multinomial or baseline logit model defined by:

$$g_k(\mathbf{x}) = \ln\left[\frac{\pi_k(\mathbf{x})}{\pi_0(\mathbf{x})}\right] = \beta_{k0} + \mathbf{x}'\boldsymbol{\beta}_k$$

is given by

$$\widehat{OR}(1,0) = \frac{17 \times 35}{29 \times 11} = 1.87,$$

$$\widehat{OR}(2,0) = \frac{16 \times 35}{22 \times 11} = 2.31$$

and

$$\widehat{OR}(3,0) = \frac{30 \times 35}{29 \times 11} = 3.29.$$

The increase in the odds ratio demonstrates an increase in odds of a progressively lower income among people that rent their home/residence.

The adjacent-category model postulates that the log odds of each successively higher comparison of the baseline log odds is a constant multiple of the log odds of $Y = 1$ versus $Y = 0$. Under the adjacent-category model, the relationship required is $\ln[OR(k,0)] = k \times \ln[OR(1,0)]$. The adjacent-category model was fit via the constrained baseline model in SAS Proc IML, using a procedure proposed by Matthews and Crowther (1995).

The frequencies are read into Proc IML as a vector and the matrix C is defined to compare each category with the first, and the 3 logits are defined. The parameter estimates are determined under the unconstrained model.

```
proc iml worksize=50;
options linesize=120;
reset nolog;

f={11,35,17,29,16,22,30,29};

 C={-1  0  1  0  0  0  0  0,
    0 -1  0  1  0  0  0  0,
   -1  0  0  0  1  0  0  0,
    0 -1  0  0  0  1  0  0,
   -1  0  0  0  0  0  1  0,
    0 -1  0  0  0  0  0  1};

logitf=C*log(f);
A={1   1,
   1   0};
A=block(A,A,A);
lambda=inv(A`*A)*A`*logitf;

print lambda;
```

The unconstrained parameters are given by the vector lambda.

```
lambda

-0.188052
0.6233703
-0.464306
0.8389991
-0.188052
1.1913543
```

Two constraints are placed on the parameters, to obtain the relationship $\ln[OR(k,0)] = k \times \ln[OR(1,0)]$. *This is done by using the matrix AH. The second slope parameter is constrained to be 2 times the first slope parameter, and the third slope parameter is constrained to be 3 times the first slope parameter. The constrained parameters are estimated iteratively:*

```
AH=A*inv(A`*A);
AH=(AH[,4]-2#AH[,2])||(AH[,6]-3#AH[,2]);


gf=AH`*logitf;
do i=1 to 5;
f=f-C`*AH*ginv(AH`*C*diag(1/f)*C`*AH)*AH`*C*log(f);
ft=f`;
print i ft[format=8.4];
end;
lambdah=inv(A`*A)*A`*C*log(f);

print lambdah[format=8.4];
```

The constrained parameters are given by the vector lamdah:

```
lambdah

-0.1100
 0.3696
-0.4414
 0.7392
-0.1750
 1.1087
```

In summary, the estimated coefficients are given in the table below:

Logit	*Variable*	*Coefficient*
1	*RS*	0.370
	constant	−0.110
2	*RS*	0.739
	constant	−0.441
3	*RS*	1.109
	constant	−0.175

The equations for the adjacent-category logits are obtained by using the algebraic relationship between the constrained baseline and adjacent-category models $(\ln\left[\frac{\phi_k(\mathbf{x})}{\phi_0(\mathbf{x})}\right] = (\alpha_1+\alpha_2+...+\alpha_k)+k\mathbf{x}'\boldsymbol{\beta})$. *It follows that the first estimated adjacent-category logit is identical to the first estimated baseline logit:*

$$\widehat{\alpha}_1(RS) = -0.110 + 0.370 \times RS.$$

The estimated coefficient for RS in the second adjacent-category logit is the sames as in the first. The estimated coefficient for logit 2 in the above table is twice the value in logit 1 and reflects the constraint placed on the fitted baseline logit model. It follows that the estimate of the constant term for the second adjacent-category logit is equal to the difference between the two estimated constant terms in the above table:

$$\widehat{\alpha}_2 = \widehat{\beta}_{20} - \widehat{\beta}_{10} = -0.441 - (-0.110) = -0.331.$$

Hence, the equation for the second adjacent-category logit is

$$\widehat{\alpha}_2(RS) = -0.331 + 0.370 \times RS.$$

The equation for the third adjacent-category logit is obtained in a similar manner. In particular, the estimated coefficient for RS shown in the third logit in the above table is three times the estimated coefficient for the first logit. It follows that the estimate of the constant term is $\widehat{\alpha}_3 = \widehat{\beta}_{30} - \widehat{\beta}_{20} = -0.175 - -(0.441) = 0.266$. *Hence the third estimated adjacent-category logit is*

$$\widehat{\alpha}_3(RS) = 0.266 + 0.370 \times RS.$$

Under the adjacent-category model the estimate of the odds ratio for residential status over the income groups is

$$\widehat{OR}(k, k-1) = \exp(0.370) = 1.45$$

for $k = 1, 2, 3$. *The interpretation of this estimate is that the odds of income in the next lower income category among those who rent their home are* 1.45 *times the odds for those who own their home.*

11.2.1 Model building strategies for Ordinal Logistic Regression Models

The steps in model building for an ordinal logistic model are the same as for the binary logistic model. Unfortunately, however, a full array of modeling tools is not available in all software packages.

For ordinal models, a sensible approach to model building involves the following steps:

- Perform the usual purposeful or stepwise selection of main effects.
- Check for the scale of continuous covariates using design variables in the ordinal model. In addition, one could check for nonlinearity using fractional polynomial analyses with K seperate binary regressions of $y = k$ versus $y = 0$. Any nonlinear transformation found should, of course, make business/clinical sense, be reasonably similar across the separate logistic regressions and make a significant improvement over treating the covariate as linear in the ordinal model.
- Check to make sure all omitted covariates are neither significant nor confounders of main effects in the model.
- Check the need to include interactions using the usual selection methods.
- Check any model assumptions of constant coefficients by comparing the constrained model to its unconstrained version. This can be done via a likelihood ratio comparison of the fitted model versus the baseline model.

Diagnostic statistics and goodness-of-fit tests have not been extended for use with ordinal models. Thus one has to use the separate binary regressions approach. The big disadvantage of this approach is that one is really not checking the actual fitted model, only an approximation to it. However this method may help identify influential and poorly fitted subjects. In general this approach is a bit ad-hoc and all results should be checked by deleting identified subjects and refitting the ordinal model. Inferential statements based on estimated odds ratios and their confidence intervals should be worded in such a way that it is clear which ordinal model has been used.

Chapter 12

Practical example

Data was obtained from a leading South African bank, to demonstrate the concepts presented in this study. Note that, for confidentiality purposes, none of the variable names can be disclosed. The variables are typically delinquency indicators, payment behaviour, e.g. timing and sizes of payments. The variables were measured over a period of six months.

The data set consists of 47 independent variables and 1 dependent variable. There are $83,685$ observations in the data set. Note that the dependent variable was coded as 1 to indicate default (according to a default definition chosen by the bank) and 0 to indicate non-default. There are $1,145$ defaults (events) and $82,540$ non-defaults in the data set and no indeterminants were used in this definition.

The data set has been cleaned to eliminate the effect of outliers (a ceiling and a floor was applied to the data). Missing values in variables were coded as "#" for text variables and 99998 for numerical variables. This allows the modeler to decide whether the missing values in the data are informative or not.

After choosing the default definition, obtaining the data sample and cleaning the data, the next step in the process is to do bivariate analysis of the independent variables. The purpose of this is to obtain the most predictive variables that will be used in the model building process.

The following tests will be performed for each of the variables:

- Likelihood Ratio test
- Weight of evidence (WOE)
- Information Value (IV)
- Gini index

Note that the WOE and IV requires all the variables to be categorized. This will be manually done for continues variables. Twenty buckets (if possible) will be created for each variable, with approximately 5% of the sample in each category if possible, except for the special category for missing values, which will be kept in its own bucket to be able to decide whether it is informative or not.

12.1 Initial bivariate analysis

The results of the initial bivariate analysis between each variable and the outcome variable (more detail available in Appendix 1 of this chapter):

Variable	IV	Likelihood ratio	DF	p-value	Gini index
var2	0.152775	171.3066	19	<0.0001	21.4
var3	0.112276	121.4153	19	<0.0001	17.9
var4	0.305701	362.527	16	<0.0001	28.7
var5	0.27684	396.8523	4	<0.0001	23.3
var6	1.66826	2286.7761	13	<0.0001	57
var7	0.380608	476.7248	16	<0.0001	29.4
var8	0.032222	37.9197	7	<0.0001	7.7
var9	0.046889	55.4253	7	<0.0001	8.1
var10	0.063771	68.4198	6	<0.0001	11.4
var11	1.474597	1925.1076	13	<0.0001	55.1
var12	0.774483	1058.5159	16	<0.0001	38.3
var13	0.314774	393.27	16	<0.0001	27.4
var14	1.670507	2232.554	19	<0.0001	58.8
var15	1.138026	1610.269	20	<0.0001	44.8
var16	0.445321	592.5461	20	<0.0001	31.1
var17	0.219975	209.5439	19	<0.0001	21.3
var18	0.173245	175.0854	20	<0.0001	20.5
var19	0.19469	206.7808	20	<0.0001	23.6
var20	1.456144	2022.4336	14	<0.0001	52.2
var21	0.674781	915.4379	14	<0.0001	36.9
var22	0.034826	38.5158	8	<0.0001	8.6
var23	0.044289	48.9257	7	<0.0001	9.9
var24	0.051586	53.725	7	<0.0001	10
var25	1.263112	1710.5847	14	<0.0001	49.7
var26	0.949072	1300.1428	15	<0.0001	42.5
var27	0.649905	876.2582	16	<0.0001	35.5
var28	1.466602	1999.9089	19	<0.0001	54.2
var29	1.164503	1649.7996	20	<0.0001	46.1
var30	0.786893	1095.3414	20	<0.0001	39.5
var31	0.109244	113.6209	19	<0.0001	16.6
var32	0.14109	146.8998	20	<0.0001	18.9
var33	0.121368	125.0006	20	<0.0001	16.9
var36	0.005603	6.4139	2	0.0405	3.5
var37	0.204883	244.7821	3	<0.0001	21.8
var38	0.417436	545.3355	19	<0.0001	30.9
var39	0.193634	223.6264	20	<0.0001	24.2
var40	1.127038	1686.6871	5	<0.0001	32.9
var41	1.400599	2046.1559	5	<0.0001	41.9
var42	0.204645	241.2851	9	<0.0001	22.8
var43	0.294919	425.734	5	<0.0001	13.7
var44	0.332201	473.1598	5	<0.0001	16.2
var45	0.255963	279.5701	7	<0.0001	27.2
var46	0.803913	979.8675	5	<0.0001	44.1
var47	0.183814	201.4935	4	<0.0001	21.9
var48	0.786852	996.4839	3	<0.0001	39.4
var49	0.112438	132.9193	9	<0.0001	15.4
var50	0.090889	101.2763	9	<0.0001	16.6

Using the p-value of the likelihood ratio test, it can be seen that all the variables are significant using a 5% level of significance. The information value and Gini index should therefore rather be used to determine whether the variable should be kept to be considered for the model.

Recall from an earlier chapter, that buckets with the same WOE can be grouped together without information loss, as measured by the IV. It is also common practice in credit scoring to have a minimum of 5% of the sample distribution in a bucket (if possible). The buckets were combined using the WOE, default rate per bucket and the distribution (graphs in Appendix 2). Note that business hypotheses were also considered when the variables were bucketed again, as this is probably the most important factor to consider when variables are bucketed for a credit scoring model.

Note that, once the variables are bucketed to combine groups with similar WOE's and default rates, at least 5% of the sample in the bucket and makes business sense, a lot of the variables might lose predictive power. This will reduce the list of possible factors.

12.2 Secondary bivariate analysis

The results of the secondary categorization of the variables: (more details can be found in Appendix 2)

Variable	IV	Likelihood ratio	DF	p-value	Gini index
var4	0.240313142	287.3887	1	<.0001	22
var5	0.265610393	325.2883	1	<.0001	22.6
var6	1.568360138	2203.8668	3	<.0001	51.9
var7	0.237819024	278.0614	3	<.0001	26.4
var8	0.005697519	6.0743	1	0.0137	1.5
var9	0.011234185	11.6483	1	0.0006	2.1
var10	0.03090221	29.664	1	<.0001	3.1
var11	0.606281981	666.0219	1	<.0001	37.8
var12	0.314619163	375.237	2	<.0001	26.3
var13	0.166420533	199.7647	2	<.0001	19.4
var14	0.950569858	1128.2624	2	<.0001	47.7
var15	0.421124132	487.6195	2	<.0001	33.8
var16	0.25036743	291.0015	2	<.0001	26.5
var17	0.005395444	6.0418	1	0.014	3.5
var18	0.004633053	5.191	1	0.0227	3.2
var19	0.035196782	39.9684	2	<.0001	9.9
var20	0.723927083	825.5545	3	<.0001	43.9
var21	0.428299116	507.614	3	<.0001	33.9
var22	0.016497216	16.7275	1	<.0001	2.4
var23	0.019400381	19.4411	1	<.0001	2.6
var24	0.031001916	29.7534	1	<.0001	3.1
var25	0.56310624	636.2295	3	<.0001	38.4
var26	0.410149448	484.3302	3	<.0001	32.1
var27	0.232052083	263.0845	2	<.0001	24.9
var28	0.912158452	1106.7424	3	<.0001	47.7
var29	0.543302527	639.2373	3	<.0001	38.4
var30	0.451269613	540.333	3	<.0001	34.9
var31	0.011712311	13.2866	2	0.0013	3.7
var32	0.017544887	19.9812	2	<.0001	7.1
var33	0.034144706	38.6559	2	<.0001	9.9
var36	0.005223775	5.9435	1	0.0148	3.4
var37	0.200695023	238.6738	2	<.0001	21.4
var38	0.217676643	278.6626	2	<.0001	16.8
var40	1.011491911	1485.4551	1	<.0001	32.6
var41	1.227686419	1737.6879	1	<.0001	41.1
var42	0.144612036	181.5336	1	<.0001	13.9
var43	0.264186287	370.6593	1	<.0001	13.6
var44	0.295453208	407.5791	1	<.0001	16
var45	0.207702624	223.4275	1	<.0001	21.5
var46	0.752581537	910.9186	1	<.0001	40.6
var47	0.179410875	195.4425	1	<.0001	20.4
var48	0.786748063	995.9671	1	<.0001	39.3

As can be seen from the above table, a few variables were excluded from the analysis (variables 2, 3, 39, 49 and 50), as the statistics and business hypotheses could not be matched. The bucketing has also caused the variables to lose some predictive power and information, but all the variables were bucketed to make business sense and to reduce the chance of over fitting the model to the data.

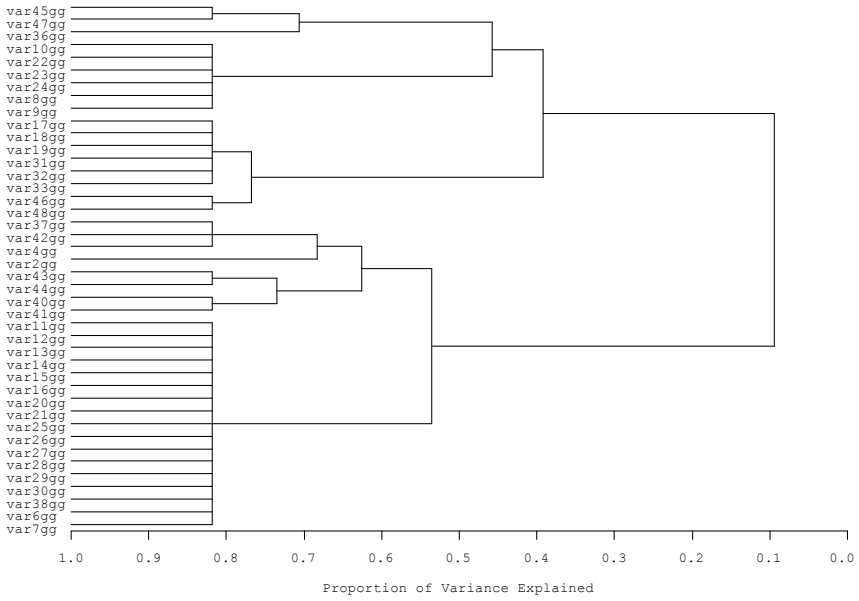
12.3 Variable clustering

Before any other exclusions were made, variable clustering was performed. Ten clusters were created using the SAS VARCLUS procedure. For more detail on the clustering, please refer to the Appendix 3 at the end of this chapter.

Cluster	Total var explained by cluster	Prop of var explained by cluster
1	3.944637	0.0939
2	16.460871	0.3919
3	19.230147	0.4579
4	22.494613	0.5356
5	26.28778	0.6259
6	28.662154	0.6824
7	29.652482	0.706
8	30.869113	0.735
9	32.243267	0.7677
10	34.370544	0.8183

Cluster	Min prop explained by cluster	Min R^2 for a var
1	0.0939	0.0031
2	0.3037	0.004
3	0.3783	0.0126
4	0.4519	0.0354
5	0.2379	0.1551
6	0.5054	0.1743
7	0.5295	0.1743
8	0.5641	0.1743
9	0.5659	0.1743
10	0.6706	0.5493

Cluster	Max $1 - R^2$ ratio for a var
1	1.1847
2	1.1044
3	1.1509
4	1.1052
5	1.1052
6	1.1052
7	1.1052
8	1.1052
9	1.1052
10	0.8089



Cluster results

12.3.1 Cluster 1

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var11	0.5576	0.3485	0.6791
var12	0.809	0.2378	0.2506
var13	0.7633	0.3632	0.3717
var14	0.5747	0.4742	0.8089
var15	0.8607	0.3182	0.2043
var16	0.8057	0.2783	0.2693
var20	0.8623	0.3409	0.209
var21	0.8746	0.367	0.198
var25	0.8536	0.3535	0.2264
var26	0.8935	0.3684	0.1686
var27	0.785	0.215	0.2738
var28	0.7992	0.381	0.3244
var29	0.9176	0.3462	0.1261
var30	0.8693	0.3761	0.2096
var38	0.5493	0.1683	0.5418
var6	0.6784	0.3695	0.5101
var7	0.8136	0.3366	0.281

12.3.2 Cluster 2

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var10	0.7043	0.3016	0.4234
var22	0.8887	0.2021	0.1394
var23	0.9216	0.2165	0.1001
var24	0.8346	0.2658	0.2252
var8	0.8365	0.1851	0.2006
var9	0.9046	0.2123	0.1211

12.3.3 Cluster 3

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var17	0.7138	0.3698	0.4542
var18	0.7906	0.3634	0.329
var19	0.7731	0.1925	0.281
var31	0.8857	0.3039	0.1643
var32	0.8903	0.2941	0.1555
var33	0.8602	0.2496	0.1862

12.3.4 Cluster 4

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var45	0.9025	0.1187	0.1106
var47	0.9025	0.1674	0.1171

12.3.5 Cluster 5

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var43	0.9261	0.0547	0.0782
var44	0.9261	0.0547	0.0782

12.3.6 Cluster 6

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var37	0.6939	0.1888	0.3774
var42	0.654	0.156	0.4099
var4	0.6642	0.2004	0.42

12.3.7 Cluster 7

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var2	1	0.0575	0

12.3.8 Cluster 8

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var36	1	0.0017	0

12.3.9 Cluster 9

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var43	0.8892	0.1126	0.1248
var44	0.8892	0.1383	0.1285

12.3.10 Cluster 10

	R^2 with		
Variable	Own	Next closest	$1 - R^2$ ratio
var46	0.8742	0.2676	0.1717
var48	0.8892	0.2393	0.1653

The results of the cluster analysis will be kept in consideration when the final model is chosen.

12.4 Stepwise logistic regression

12.4.1 Significance levels

The significance level for entry was set at 0.05 and the significance level for removal was set at 0.1. The significance levels are low due to the large sample, to avoid over-fitting the model to the data.

12.4.2 Design/dummy variables

Design/dummy variables were created for all the variables considered for modelling. They were assigned as follows, dependent on the number of buckets created for the variable:

		Design/dummy variable		
Number of buckets	Value	1	2	3
2	1	1		
	2	0		
3	1	1	0	
	2	0	1	
	3	0	0	
4	1	1	0	0
	2	0	1	0
	3	0	0	1
	4	0	0	0

12.4.3 Stepwise summary

A summary of the stepwise procedure is presented here. Full output of the stepwise procedure (as well as the SAS code used) is available in Appendix 4 of this chapter.

Step	Variable	DF	Score χ^2	p-value
1	var6	3	6597.4182	<.0001
2	var4	1	227.9883	<.0001
3	var40	1	176.3495	<.0001
4	var42	1	48.5738	<.0001
5	var44	1	40.4273	<.0001
6	var28	3	41.5973	<.0001
7	var10	1	32.913	<.0001
8	var41	1	30.5176	<.0001
9	var27	2	31.216	<.0001
10	var38	2	24.4035	<.0001
11	var37	2	14.4028	0.0007
12	var29	3	16.1425	0.0011
13	var12	2	14.2426	0.0008
14	var26	3	12.0342	0.0073
15	var23	1	7.2224	0.0072
16	var7	3	11.4575	0.0095

Sixteen variables entered the model and no variables were removed during the procedure. The Gini index for this model is 67.9.

12.4.4 Model coefficients

The coefficients for the model, as given by the stepwise procedure:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7707	0.2898	91.4271	<.0001
var10gg 1	1	1.0565	0.2651	15.8778	<.0001
var12gg 1	1	0.3575	0.1634	4.7883	0.0287
var12gg 2	1	-0.1107	0.1396	0.6287	0.4278
var23gg 1	1	0.6477	0.2451	6.9852	0.0082
var26gg 1	1	0.1343	0.2955	0.2065	0.6495
var26gg 2	1	-0.4591	0.2850	2.5951	0.1072
var26gg 3	1	-0.4578	0.2042	5.0280	0.0249
var27gg 1	1	0.7376	0.2245	10.7939	0.0010
var27gg 2	1	0.5635	0.1616	12.1539	0.0005
var28gg 1	1	-1.1418	0.2397	22.6895	<.0001
var28gg 2	1	-0.4188	0.1380	9.2117	0.0024
var28gg 3	1	-0.3962	0.1142	12.0338	0.0005
var29gg 1	1	-0.8770	0.2634	11.0903	0.0009
var29gg 2	1	-0.5576	0.2348	5.6388	0.0176
var29gg 3	1	-0.00238	0.1690	0.0002	0.9888
var37gg 1	1	0.2150	0.0898	5.7329	0.0167
var37gg 2	1	0.4952	0.1336	13.7342	0.0002
var38gg 1	1	-0.2154	0.1059	4.1411	0.0419
var38gg 2	1	-0.4063	0.0906	20.0942	<.0001
var40gg 1	1	-0.2683	0.1238	4.6964	0.0302
var41gg 1	1	-0.6426	0.1287	24.9357	<.0001
var42gg 1	1	0.5058	0.0950	28.3642	<.0001
var44gg 1	1	-0.4965	0.0942	27.7596	<.0001
var4gg 1	1	0.7111	0.0761	87.3460	<.0001
var6gg 1	1	-2.4217	0.2298	111.0196	<.0001
var6gg 2	1	-2.5763	0.1816	201.2410	<.0001
var6gg 3	1	-1.9644	0.0974	406.3438	<.0001
var7gg 1	1	-0.0818	0.1928	0.1800	0.6713
var7gg 2	1	0.3502	0.1617	4.6867	0.0304
var7gg 3	1	0.1199	0.1036	1.3388	0.2472

As can be seen from the above coefficients, not all coefficients were significant. A more parsimonious model can therefore be fit when collapsing some of the less significant buckets within variables.

12.4.5 Model refinement

Note that all the changes to the stepwise model were not made at once, but the model was tested after every one or two changes.

First, group 1 and group 2 of variable var12 was combined, as well as group 1 and group 2 of variable var26.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.5930	0.2847	82.9632	<.0001
var10gg	1	1	1.0211	0.2688	14.4296	0.0001
var12gg	1	1	0.0403	0.1306	0.0951	0.7578
var23gg	1	1	0.4722	0.2427	3.7853	0.0517
var26gg	1	1	-0.1441	0.2646	0.2964	0.5862
var26gg	2	1	-0.4438	0.2012	4.8655	0.0274
var27gg	1	1	0.8010	0.2160	13.7498	0.0002
var27gg	2	1	0.5507	0.1600	11.8444	0.0006
var28gg	1	1	-1.1433	0.2388	22.9229	<.0001
var28gg	2	1	-0.5067	0.1367	13.7277	0.0002
var28gg	3	1	-0.3934	0.1142	11.8646	0.0006
var29gg	1	1	-0.7323	0.2539	8.3192	0.0039
var29gg	2	1	-0.7255	0.2298	9.9642	0.0016
var29gg	3	1	-0.0626	0.1701	0.1355	0.7128
var37gg	1	1	0.2263	0.0893	6.4223	0.0113
var37gg	2	1	0.4921	0.1333	13.6210	0.0002
var38gg	1	1	-0.1810	0.1046	2.9924	0.0837
var38gg	2	1	-0.4267	0.0907	22.1436	<.0001
var40gg	1	1	-0.2550	0.1238	4.2436	0.0394
var41gg	1	1	-0.6388	0.1288	24.5787	<.0001
var42gg	1	1	0.5412	0.0945	32.8185	<.0001
var44gg	1	1	-0.4867	0.0942	26.7116	<.0001
var4gg	1	1	0.7277	0.0758	92.2220	<.0001
var6gg	1	1	-2.3660	0.2301	105.7340	<.0001
var6gg	2	1	-2.4197	0.1738	193.8959	<.0001
var6gg	3	1	-1.9720	0.0974	410.0792	<.0001
var7gg	1	1	0.1734	0.1810	0.9179	0.3380
var7gg	2	1	0.3440	0.1593	4.6658	0.0308
var7gg	3	1	0.1115	0.1039	1.1515	0.2832

The Gini index decreased slightly from 67.9 to 67.5, but the degrees-of-freedom of the model also decreased from 30 to 28.

Next, variable var12 was removed from the model as it had only 1 coefficient that was not significant according to the Wald test. Also note that var12 forms part of the first cluster in the cluster analysis, and that cluster is well represented by other variables in the current model.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.5917	0.2846	82.9459	<.0001
var10gg	1	1	1.0227	0.2688	14.4761	0.0001
var23gg	1	1	0.4678	0.2422	3.7299	0.0534
var26gg	1	1	-0.1182	0.2507	0.2222	0.6374
var26gg	2	1	-0.4203	0.1861	5.0992	0.0239
var27gg	1	1	0.7931	0.2144	13.6857	0.0002
var27gg	2	1	0.5475	0.1597	11.7454	0.0006
var28gg	1	1	-1.1418	0.2387	22.8849	<.0001
var28gg	2	1	-0.5041	0.1365	13.6401	0.0002
var28gg	3	1	-0.3951	0.1141	11.9978	0.0005
var29gg	1	1	-0.7416	0.2519	8.6649	0.0032
var29gg	2	1	-0.7342	0.2279	10.3759	0.0013
var29gg	3	1	-0.0663	0.1695	0.1529	0.6958
var37gg	1	1	0.2250	0.0892	6.3629	0.0117
var37gg	2	1	0.4905	0.1333	13.5488	0.0002
var38gg	1	1	-0.1786	0.1044	2.9292	0.0870
var38gg	2	1	-0.4251	0.0905	22.0571	<.0001
var40gg	1	1	-0.2576	0.1235	4.3500	0.0370
var41gg	1	1	-0.6359	0.1285	24.4796	<.0001
var42gg	1	1	0.5430	0.0943	33.1843	<.0001
var44gg	1	1	-0.4877	0.0941	26.8647	<.0001
var4gg	1	1	0.7281	0.0758	92.3624	<.0001
var6gg	1	1	-2.3647	0.2300	105.7146	<.0001
var6gg	2	1	-2.4171	0.1736	193.9588	<.0001
var6gg	3	1	-1.9684	0.0967	414.7011	<.0001
var7gg	1	1	0.2013	0.1567	1.6511	0.1988
var7gg	2	1	0.3697	0.1357	7.4190	0.0065
var7gg	3	1	0.1225	0.0974	1.5831	0.2083

The Gini index of the above model is 67.7, which increased from the previous step as well as lowered the degrees-of-freedom of the model to 27.

The next step was to collapse groups 3 and 4 for both variables, var29 and var7.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.5598	0.2832	81.7099	<.0001
var10gg	1	1	1.0406	0.2685	15.0194	0.0001
var23gg	1	1	0.4600	0.2419	3.6156	0.0572
var26gg	1	1	-0.1170	0.2351	0.2479	0.6186
var26gg	2	1	-0.4189	0.1494	7.8595	0.0051
var27gg	1	1	0.7864	0.2138	13.5227	0.0002
var27gg	2	1	0.5521	0.1600	11.9087	0.0006
var28gg	1	1	-1.1396	0.2379	22.9447	<.0001
var28gg	2	1	-0.5028	0.1323	14.4506	0.0001
var28gg	3	1	-0.3917	0.1074	13.3018	0.0003
var29gg	1	1	-0.7027	0.2150	10.6847	0.0011
var29gg	2	1	-0.6880	0.1759	15.2896	<.0001
var37gg	1	1	0.2272	0.0892	6.4834	0.0109
var37gg	2	1	0.4956	0.1331	13.8636	0.0002
var38gg	1	1	-0.1742	0.1042	2.7951	0.0946
var38gg	2	1	-0.4195	0.0904	21.5469	<.0001
var40gg	1	1	-0.2625	0.1234	4.5253	0.0334
var41gg	1	1	-0.6338	0.1285	24.3207	<.0001
var42gg	1	1	0.5460	0.0942	33.5704	<.0001
var44gg	1	1	-0.4919	0.0940	27.3592	<.0001
var4gg	1	1	0.7317	0.0757	93.4082	<.0001
var6gg	1	1	-2.3665	0.2300	105.8521	<.0001
var6gg	2	1	-2.4153	0.1736	193.6802	<.0001
var6gg	3	1	-1.9708	0.0965	416.9984	<.0001
var7gg	1	1	0.1255	0.1447	0.7518	0.3859
var7gg	2	1	0.2844	0.1178	5.8292	0.0158

The Gini index of the above model is 67.7, which stayed the same from the previous version of the model, but lowered the degrees-of-freedom in the model to 25. This means that two less parameters were fitted to obtain the same predicitve power.

It is clear from the above regression coefficients, that the first and second groups of the following variables are not significantly different and can therefore be combined: var26, var38 and var7.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.5516	0.2829	81.3293	<.0001
var10gg	1	1	1.0435	0.2689	15.0590	0.0001
var23gg	1	1	0.4442	0.2412	3.3917	0.0655
var26gg	1	1	-0.3890	0.1449	7.2118	0.0072
var27gg	1	1	0.9551	0.1839	26.9686	<.0001
var27gg	2	1	0.5423	0.1574	11.8724	0.0006
var28gg	1	1	-1.1095	0.2366	21.9959	<.0001
var28gg	2	1	-0.4643	0.1304	12.6797	0.0004
var28gg	3	1	-0.3920	0.1073	13.3372	0.0003
var29gg	1	1	-0.6953	0.1870	13.8280	0.0002
var29gg	2	1	-0.5811	0.1626	12.7670	0.0004
var37gg	1	1	0.2233	0.0892	6.2739	0.0123
var37gg	2	1	0.4755	0.1328	12.8263	0.0003
var38gg	1	1	-0.1626	0.1039	2.4501	0.1175
var38gg	2	1	-0.4225	0.0903	21.8909	<.0001
var40gg	1	1	-0.2570	0.1234	4.3394	0.0372
var41gg	1	1	-0.6325	0.1285	24.2293	<.0001
var42gg	1	1	0.5532	0.0942	34.5183	<.0001
var44gg	1	1	-0.4987	0.0939	28.2011	<.0001
var4gg	1	1	0.7286	0.0757	92.5676	<.0001
var6gg	1	1	-2.3927	0.2284	109.7300	<.0001
var6gg	2	1	-2.3971	0.1721	193.9137	<.0001
var6gg	3	1	-1.9683	0.0964	416.5067	<.0001
var7gg	1	1	0.2270	0.1105	4.2223	0.0399

The Gini index of this model decreased slightly to 67.4, but less parameters were fit, so the degrees-of-freedom decreased to 23.

From the above regression coefficients and their respective p-values, it is clear that the variable, var23, is no longer significant. Var23 forms part of cluster 2, which is already represented in the model by var10.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3645	0.2598	82.8465	<.0001
var10gg	1	1	1.3067	0.2322	31.6808	<.0001
var26gg	1	1	-0.3954	0.1450	7.4357	0.0064
var27gg	1	1	0.9356	0.1835	25.9926	<.0001
var27gg	2	1	0.5436	0.1575	11.9142	0.0006
var28gg	1	1	-1.1117	0.2362	22.1548	<.0001
var28gg	2	1	-0.4553	0.1304	12.1836	0.0005
var28gg	3	1	-0.3851	0.1073	12.8924	0.0003
var29gg	1	1	-0.6936	0.1869	13.7652	0.0002
var29gg	2	1	-0.5731	0.1627	12.4022	0.0004
var37gg	1	1	0.2129	0.0891	5.7093	0.0169
var37gg	2	1	0.4683	0.1328	12.4365	0.0004
var38gg	1	1	-0.1716	0.1038	2.7316	0.0984
var38gg	2	1	-0.4229	0.0903	21.9356	<.0001
var40gg	1	1	-0.2638	0.1234	4.5724	0.0325
var41gg	1	1	-0.6304	0.1284	24.0880	<.0001

The degrees-of-freedom for the above model has decreased to 22 and the Gini index is now 67.1.

The coefficient of the first group of var38 in the above table is not significant, it was therefore regrouped to combine group 1 and 2.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3473	0.2599	81.5642	<.0001
var10gg	1	1	1.2600	0.2316	29.5992	<.0001
var26gg	1	1	-0.4129	0.1451	8.0932	0.0044
var27gg	1	1	1.0419	0.1802	33.4361	<.0001
var27gg	2	1	0.5659	0.1576	12.8878	0.0003
var28gg	1	1	-1.0777	0.2360	20.8474	<.0001
var28gg	2	1	-0.4737	0.1303	13.2173	0.0003
var28gg	3	1	-0.4016	0.1069	14.1233	0.0002
var29gg	1	1	-0.6516	0.1865	12.2121	0.0005
var29gg	2	1	-0.5855	0.1628	12.9348	0.0003
var37gg	1	1	0.2378	0.0887	7.1865	0.0073
var37gg	2	1	0.4873	0.1327	13.4823	0.0002
var38gg	1	1	-0.3418	0.0851	16.1472	<.0001
var40gg	1	1	-0.2715	0.1233	4.8491	0.0277
var41gg	1	1	-0.6329	0.1286	24.2056	<.0001
var42gg	1	1	0.5701	0.0937	37.0057	<.0001
var44gg	1	1	-0.4766	0.0934	26.0208	<.0001
var4gg	1	1	0.7443	0.0749	98.7501	<.0001
var6gg	1	1	-2.3811	0.2285	108.5624	<.0001
var6gg	2	1	-2.3447	0.1713	187.4182	<.0001
var6gg	3	1	-1.9685	0.0965	415.7093	<.0001
var7gg	1	1	0.2156	0.1099	3.8499	0.0497

The Gini index of this model is now 66.6 and the degrees-of-freedom is now 21.

The coefficient for var7 is barely significant, so the model was tested when var7 is removed.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3549	0.2600	82.0526	<.0001
var10gg 1	1	1.2751	0.2316	30.3095	<.0001
var26gg 1	1	-0.4070	0.1448	7.9049	0.0049
var27gg 1	1	1.1434	0.1723	44.0489	<.0001
var27gg 2	1	0.6093	0.1556	15.3319	<.0001
var28gg 1	1	-1.0906	0.2362	21.3127	<.0001
var28gg 2	1	-0.4822	0.1303	13.6910	0.0002
var28gg 3	1	-0.4136	0.1067	15.0200	0.0001
var29gg 1	1	-0.5700	0.1827	9.7327	0.0018
var29gg 2	1	-0.5000	0.1580	10.0111	0.0016
var37gg 1	1	0.2357	0.0888	7.0510	0.0079
var37gg 2	1	0.5004	0.1326	14.2448	0.0002
var38gg 1	1	-0.3344	0.0850	15.4812	<.0001
var40gg 1	1	-0.2785	0.1233	5.1049	0.0239
var41gg 1	1	-0.6276	0.1288	23.7556	<.0001
var42gg 1	1	0.5707	0.0938	37.0206	<.0001
var44gg 1	1	-0.4646	0.0933	24.7886	<.0001
var4gg 1	1	0.7459	0.0749	99.2429	<.0001
var6gg 1	1	-2.3647	0.2288	106.7731	<.0001
var6gg 2	1	-2.3342	0.1715	185.2251	<.0001
var6gg 3	1	-1.9671	0.0967	413.9830	<.0001

The Gini index increased to 66.7 and the degrees-of-freedom also decreased. Hence, the model is more predictive with less variables. As predictive power (as measured by the Gini index) is a more important consideration in credit scoring than the significance of individual variables, var 7 can be removed and it will improve the predictive power of the model.

It is clear from the above coefficients that they are all significant.

Although all the coefficients in the above output is significant, each variable will be tested to make sure it adds to the predictive power of the model. Note that the aim of the model is not to explain the underlying data, but to make predictions about future observations.

Each variable was removed in turn from the model to assess the effect of that particular variable on the model's predictive power.

Variable	Gini index	Difference
All variables	66.7	
Remove var10	65.7	-1
Remove var26	66.8	0.1
Remove var27	66.9	0.2
Remove var28	66	-0.7
Remove var29	66.7	0
Remove var37	67	0.3
Remove var38	66.7	0
Remove var40	66.8	0.1
Remove var41	66.8	0.1
Remove var42	66.7	0
Remove var44	66.6	-0.1
Remove var4	66.3	-0.4
Remove var6	65.3	-1.4

It is clear from the above table that some variables does not add to the predictive power of the model, because when removed, the Gini index stays the same. Some variables have a negative impact on the predictive power of the model, as the Gini index increases as these variables are removed one at a time. The variable that has the biggest influence on the power of the model is var37 and will be removed. Note that var37 is part of cluster3, which is already represented by var42 and var4 in the model.

Analysis of Maximum Likelihood Estimates

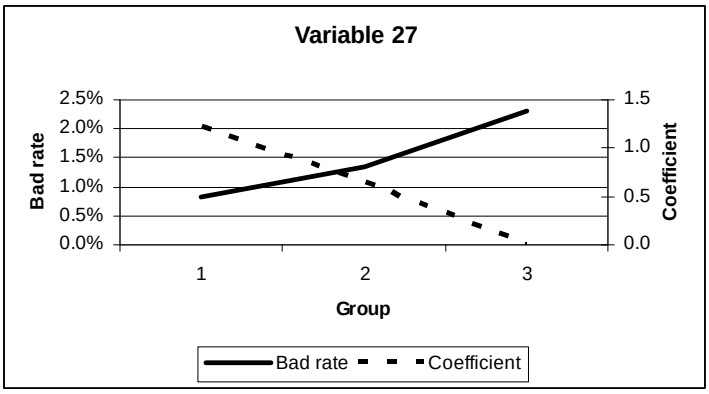
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2675	0.2589	76.7118	<.0001
var10gg 1	1	1.2463	0.2315	28.9867	<.0001
var26gg 1	1	-0.4053	0.1455	7.7573	0.0053
var27gg 1	1	1.2240	0.1714	50.9708	<.0001
var27gg 2	1	0.6466	0.1558	17.2195	<.0001
var28gg 1	1	-1.1316	0.2361	22.9778	<.0001
var28gg 2	1	-0.5155	0.1304	15.6204	<.0001
var28gg 3	1	-0.4156	0.1068	15.1481	<.0001
var29gg 1	1	-0.5331	0.1824	8.5429	0.0035
var29gg 2	1	-0.5022	0.1581	10.0943	0.0015
var38gg 1	1	-0.3216	0.0849	14.3681	0.0002
var40gg 1	1	-0.3153	0.1227	6.6000	0.0102
var41gg 1	1	-0.6698	0.1281	27.3546	<.0001
var42gg 1	1	0.6310	0.0870	52.5475	<.0001
var44gg 1	1	-0.4120	0.0921	20.0132	<.0001
var4gg 1	1	0.8033	0.0720	124.4348	<.0001
var6gg 1	1	-2.3775	0.2294	107.4166	<.0001
var6gg 2	1	-2.3751	0.1715	191.6880	<.0001
var6gg 3	1	-1.9822	0.0970	417.3622	<.0001

All the remaining coefficients in the model are significant. Again, each variable was removed from the model to ascertain its effect on the predictive power of the model.

Variable	Gini index	Difference
All variables	67	
Remove var10	65.5	-1.5
Remove var26	66.6	-0.4
Remove var27	66.1	-0.9
Remove var28	65.9	-1.1
Remove var29	66.5	-0.5
Remove var38	66.7	-0.3
Remove var40	66.7	-0.3
Remove var41	66.5	-0.5
Remove var42	66.2	-0.8
Remove var44	66.7	-0.3
Remove var4	64.6	-2.4
Remove var6	64.6	-2.4

The Gini index decreases when any of the remaining variables in the model is removed, hence all the remaining variables add to the predictive power of the model. The next step is to analyze the relationship between the default rate sloping and the coefficients of the variables. As the default (**model def(event='1'**) in the SAS code) was modeled, one expects the coefficients to have the same direction as the bad rate sloping of the variables (univariately). If this is not the case, it might require further analysis of the relationships between the variables in the model. If removing the variable from the model does not drop the predictive power of the model too much, it might be argued that the variable rather be removed than doing complex, time-consuming analysis of different relationships between the variables.

For the variable var27, the coefficients were counter to what was expected, as illustrated by the graph below:



Var27

The direction of the bad rate and the coefficients are not the same, which indicates that the variable is not predicting as expected if this was the only variable in the model. This might indicate that there is an interaction between this variable and one or more of the other variables in the current model, which would require in-depth analysis. Var27 forms part of the first cluster, which is already presented in the model by five other variables. This variable will be removed from the model, to test whether exclusion of this variable will have a big impact on the predictiveness of the model.If the variable does not have a big impact on the predictive power of the model, it can be removed. If it does have a big impact on the predictiveness of the model, in-depth analysis of the relationships between this variable and those also in the model should be performed to explain why the variable is sloping the way it is. All the other variables had the correct relationship between the default rate sloping and the coefficients. The graphs are presented in Appendix 5.

Var27 was removed from the model:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2188	0.2589	73.4670	<.0001
var10gg	1	1	1.2101	0.2316	27.3099	<.0001
var26gg	1	1	0.1958	0.1034	3.5846	0.0583
var28gg	1	1	-1.1610	0.2371	23.9837	<.0001
var28gg	2	1	-0.5569	0.1301	18.3100	<.0001
var28gg	3	1	-0.5193	0.1077	23.2576	<.0001
var29gg	1	1	0.0639	0.1577	0.1641	0.6854
var29gg	2	1	0.0816	0.1280	0.4066	0.5237
var38gg	1	1	-0.2720	0.0841	10.4720	0.0012
var40gg	1	1	-0.3280	0.1223	7.1946	0.0073
var41gg	1	1	-0.6924	0.1279	29.2973	<.0001
var42gg	1	1	0.6375	0.0870	53.7164	<.0001
var44gg	1	1	-0.4054	0.0920	19.4079	<.0001
var4gg	1	1	0.8237	0.0719	131.1085	<.0001
var6gg	1	1	-2.3649	0.2307	105.1125	<.0001
var6gg	2	1	-2.3380	0.1710	186.9001	<.0001
var6gg	3	1	-1.9880	0.0971	419.0391	<.0001

The Gini index has decreased from 67 to 66.1, which does not make a complex analysis worthwhile in order to keep the variable.. From the above regression coefficients it is clear that variable var29 has lost its predictive power as both the coefficients for this variable have large p-values. The variable will be removed from the model.

Remove var29:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2225	0.2588	73.7746	<.0001
var10gg	1	1	1.2125	0.2315	27.4317	<.0001
var26gg	1	1	0.2196	0.0954	5.3046	0.0213
var28gg	1	1	-1.1361	0.2104	29.1577	<.0001
var28gg	2	1	-0.5478	0.1266	18.7387	<.0001
var28gg	3	1	-0.5273	0.1068	24.3892	<.0001
var38gg	1	1	-0.2697	0.0839	10.3264	0.0013
var40gg	1	1	-0.3281	0.1223	7.2039	0.0073
var41gg	1	1	-0.6909	0.1279	29.1846	<.0001
var42gg	1	1	0.6373	0.0870	53.6979	<.0001
var44gg	1	1	-0.4031	0.0919	19.2315	<.0001
var4gg	1	1	0.8246	0.0719	131.5345	<.0001
var6gg	1	1	-2.3536	0.2304	104.3531	<.0001
var6gg	2	1	-2.3005	0.1618	202.2352	<.0001
var6gg	3	1	-1.9906	0.0971	420.2571	<.0001

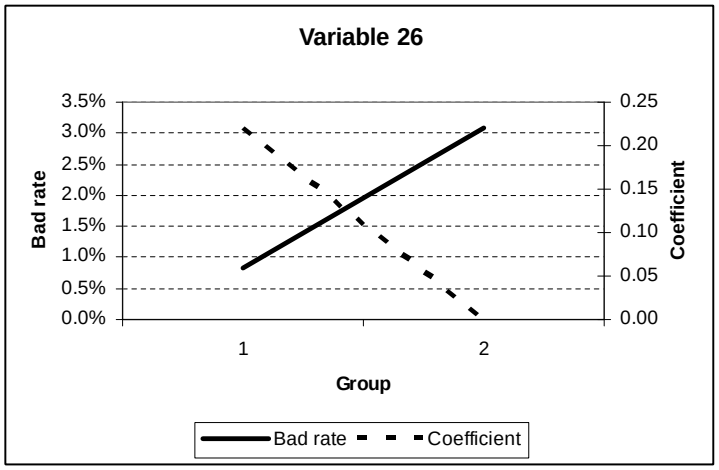
The Gini index remained at 66.1, even though the variable was removed, which means that the variable did not add to the predictive power of the model. All the coefficients in the model is again significant.

Each variable's effect on the Gini index is again tested.

Variable	Gini index	Difference
All variables	66.1	
Remove var10	64.6	-1.5
Remove var26	65.9	-0.2
Remove var28	65.2	-0.9
Remove var38	65.5	-0.6
Remove var40	66	-0.1
Remove var41	65.6	-0.5
Remove var42	64.9	-1.2
Remove var44	65.8	-0.3
Remove var4	63.7	-2.4
Remove var6	64	-2.1

The Gini index decreases when any of the remaining variables in the model is removed, hence all the remaining variables add to the predictive power of the model. The relationship between the default rate sloping and the coefficients of the variables is analyzed again. Recall that one expects the coefficients to have the same direction as the bad rate sloping of the variables (univariately).

Again, one variable had coefficients that sloped in the different direction than the default rate, variable 26. This again might indicate interaction between this variable and one or more of the other variables in the model. If the variable has a big impact on the overall power of the model, the relationships between this model and all other variables in the model mus t be analyzed in order to explain why it is sloping in a different way as it would univariately. All the other graphs for the other variables is available in Appendix 6.



Var26

Variable 26 was removed from the model, as it does not add too much to the model in terms of predictive power and therefore does not warrant a complex analysis of variable relationships.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2184	0.2588	73.4956	<.0001
var10gg	1	1	1.2221	0.2316	27.8536	<.0001
var28gg	1	1	-0.9741	0.1993	23.8870	<.0001
var28gg	2	1	-0.3809	0.1051	13.1394	0.0003
var28gg	3	1	-0.4677	0.1036	20.3960	<.0001
var38gg	1	1	-0.2400	0.0828	8.4073	0.0037
var40gg	1	1	-0.3224	0.1222	6.9657	0.0083
var41gg	1	1	-0.6991	0.1280	29.8384	<.0001
var42gg	1	1	0.6341	0.0871	53.0630	<.0001
var44gg	1	1	-0.4075	0.0919	19.6653	<.0001
var4gg	1	1	0.8296	0.0719	133.1734	<.0001
var6gg	1	1	-2.3344	0.2309	102.1918	<.0001
var6gg	2	1	-2.2835	0.1619	198.9078	<.0001
var6gg	3	1	-1.9735	0.0972	412.5563	<.0001

The Gini index decreased from 66.1 to 65.9.

Again, the impact of each variable on the Gini index was determined.

Variable	Gini index	Difference
All variables	65.9	
Remove var10	64.5	-1.4
Remove var28	65.2	-0.7
Remove var38	65.9	0
Remove var40	66.3	0.4
Remove var41	65.8	-0.1
Remove var42	64.6	-1.3
Remove var44	66.2	0.3
Remove var4	64.2	-1.7
Remove var6	63.8	-2.1

Removing some of the variables increases the predictive power of the model. As the aim of the model is to predict, rather than to explain the underlying data, this should be taken into consideration. When Var40 is removed, the Gini index increases from 65.9 to 66.3. Var40 is part of cluster 9, which is also represented by Var41 in the model.

Variable 40 is removed from the model:

Analysis of Maximum Likelihood Estimates

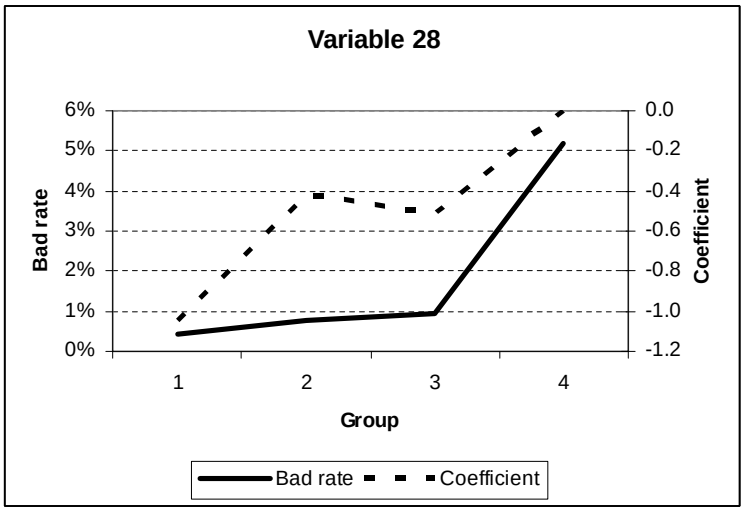
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2975	0.2571	79.8251	<.0001
var10gg	1	1	1.2228	0.2314	27.9150	<.0001
var28gg	1	1	-1.0486	0.1971	28.2978	<.0001
var28gg	2	1	-0.4206	0.1034	16.5585	<.0001
var28gg	3	1	-0.5129	0.1015	25.5528	<.0001
var38gg	1	1	-0.2504	0.0824	9.2488	0.0024
var41gg	1	1	-0.9142	0.0955	91.7280	<.0001
var42gg	1	1	0.6587	0.0863	58.2651	<.0001
var44gg	1	1	-0.4197	0.0914	21.0720	<.0001
var4gg	1	1	0.8356	0.0717	135.7759	<.0001
var6gg	1	1	-2.2808	0.2301	98.2784	<.0001
var6gg	2	1	-2.2510	0.1613	194.7284	<.0001
var6gg	3	1	-1.9593	0.0965	412.3503	<.0001

All the coefficients are significant in the model and the Gini index increased from 65.9 to 66.3.

Again, the effect of each variable on the overal predictive power of the model was tested using the Gini index.

Variable	Gini index	Difference
All variables	66.3	
Remove var10	64.6	-1.7
Remove var28	65.2	-1.1
Remove var38	65.8	-0.5
Remove var41	65.6	-0.7
Remove var42	65.2	-1.1
Remove var44	66.1	-0.2
Remove var4	64.1	-2.2
Remove var6	63.8	-2.5

Removing all the remaining variables one by one decreases the Gini index. The relationship between the default rate sloping and coefficients were again checked.



Var28

The coefficient sloping for var28 does not correspond for group 2 and group 3 to the bad rate sloping. These two groups will be combined as analysis of the reason why this occurred will be more complex and time-consuming than the added predictive power of the extra category in the model can justify. The graphs for all the other variables can be found in Appendix 7 of this chapter.

Combining group 2 and 3 for Variable 28:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3018	0.2571	80.1487	<.0001
var10gg	1	1	1.2218	0.2315	27.8603	<.0001
var28gg	1	1	-1.0633	0.1964	29.3067	<.0001
var28gg	2	1	-0.4690	0.0852	30.2856	<.0001
var38gg	1	1	-0.2416	0.0815	8.7876	0.0030
var41gg	1	1	-0.9120	0.0954	91.4740	<.0001
var42gg	1	1	0.6616	0.0862	58.8482	<.0001
var44gg	1	1	-0.4207	0.0914	21.1885	<.0001
var4gg	1	1	0.8331	0.0716	135.3132	<.0001
var6gg	1	1	-2.2703	0.2299	97.5456	<.0001
var6gg	2	1	-2.2113	0.1540	206.0595	<.0001
var6gg	3	1	-1.9599	0.0964	413.0086	<.0001

The Gini index decreased from 66.3 to 65.9.

Again, the impact of each variable on the Gini index of the model was assessed:

Variable	Gini index	Difference
All variables	65.9	
Remove var10	64.9	-1
Remove var28	65.2	-0.7
Remove var38	65.9	0
Remove var41	65.5	-0.4
Remove var42	65.2	-0.7
Remove var44	66	0.1
Remove var4	64.1	-1.8
Remove var6	63.8	-2.1

Removing var44 will result in an increase in the Gini index of the model.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.5803	0.2503	106.2537	<.0001
var10gg	1	1	1.2627	0.2315	29.7472	<.0001
var28gg	1	1	-1.0941	0.1961	31.1227	<.0001
var28gg	2	1	-0.4760	0.0850	31.3450	<.0001
var38gg	1	1	-0.3203	0.0791	16.4035	<.0001
var41gg	1	1	-0.9533	0.0957	99.2501	<.0001
var42gg	1	1	0.6539	0.0864	57.3116	<.0001
var4gg	1	1	0.8388	0.0717	136.8091	<.0001
var6gg	1	1	-2.2826	0.2296	98.8405	<.0001
var6gg	2	1	-2.2412	0.1541	211.5099	<.0001
var6gg	3	1	-2.0090	0.0963	435.4554	<.0001

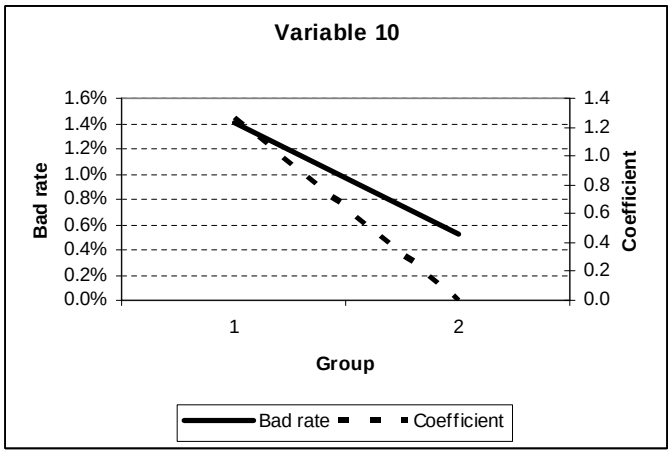
The Gini index for the above model is 66, which is an improvement on the previous Gini index of 65.9, although a variable was removed from the model.

The effect of each variable on the predictive power of the model was determined again, using the Gini index.

Variable	Gini index	Difference
All variables	66	
Remove var10(IBR)[1]	64.6	-1.4
Remove var28(AIR)	65	-1
Remove var38(BLR)	65.5	-0.5
Remove var41(DS1)	65.2	-0.8
Remove var42(PBR)	64.8	-1.2
Remove var4(PPM)	63.5	-2.5
Remove var6(DS2)	63.2	-2.8

From the above table, it is clear that each variable adds to the predictive power of the model. What remains now is to assess the relationship between the default rate sloping and coefficients of each variable.

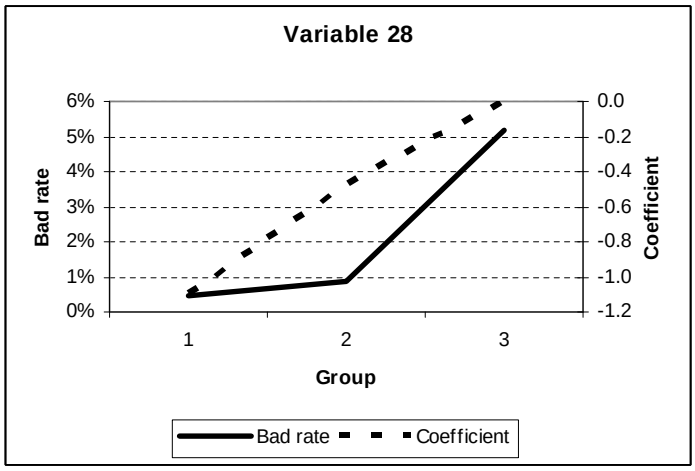
Variable 10 (IBR):



Var10

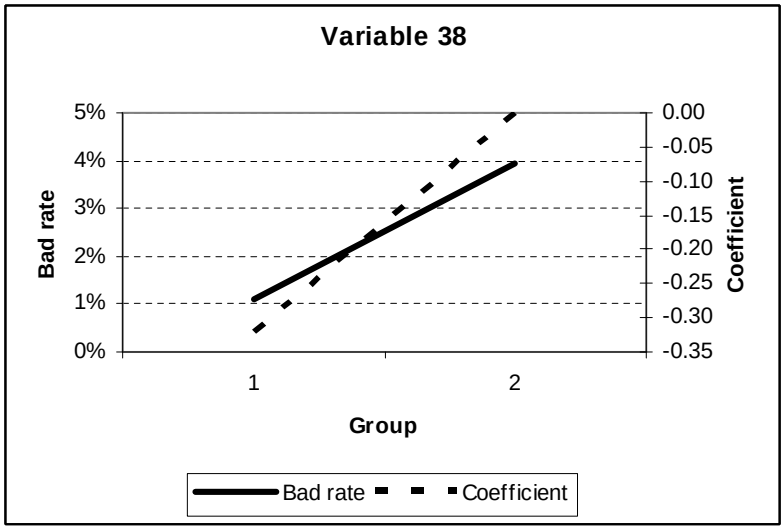
[1] Descriptions of the variable names can be found in the following section.

Variable 28(AIR):



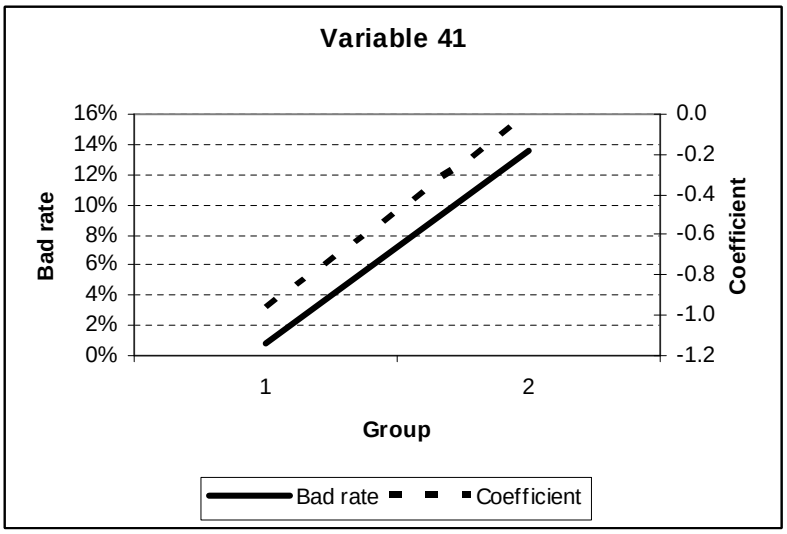
Var28

Variable 38 (BLR):



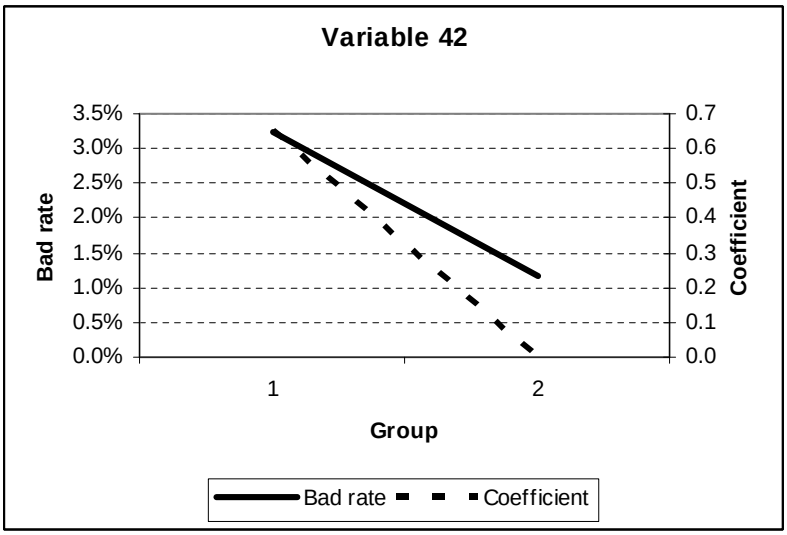
Var38

Variable 41:



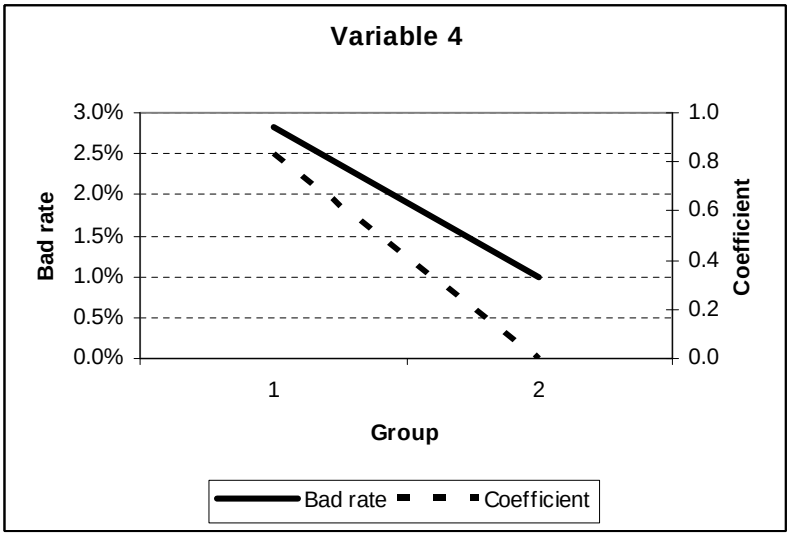
Var41

Variable 42 (PBR):



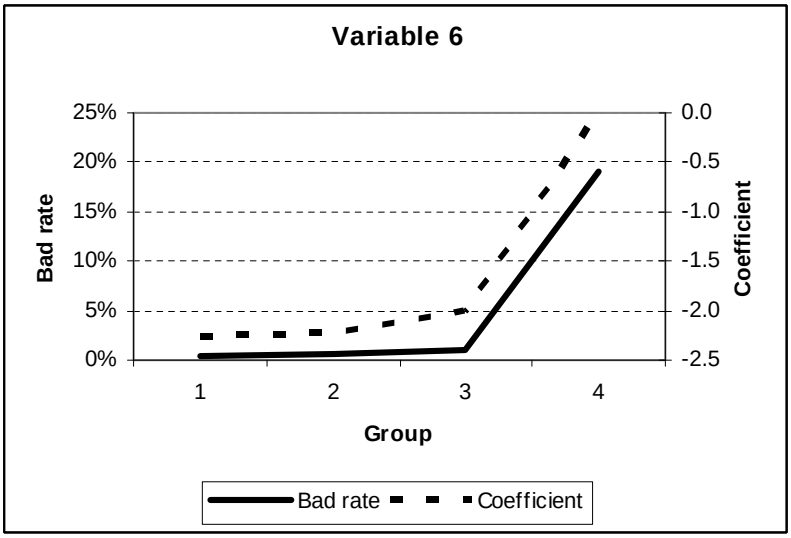
Var42

Variable 4 (PPM):



Var4

Variable 6 (DS2):



Var6

As all the variables add to the model in terms of predictive power (measured by the Gini index) and all the variables have the correct relationship between the coefficients and the default rate sloping, one can conclude that the final model has been reached that will be put forward to the business for consideration.

12.5 Final model

The final model contains the following variables from the clusters as indicated:

Variable	Cluster	Description	Sign of coefficient
var10	2	Interest to balance ratio (IBR)	Positive
var28	1	Arrears to instalment ratio(AIR)	Negative
var38	1	Balance to Limit Ratio (BLR)	Positive
var41	9	Delinquency status (DS1)	Positive
var42	3	Payment to Balance Ratio (PBR)	Negative
var4	3	Percentage Payment made (PPM)	Positive
var6	1	Delinquency status (DS2)	Negative

The variable with the highest positive coefficient in the model is var10, interest to balance ratio. This variable indicates that a client with a high value for interest to balance, is a higher risk than a client with a low value. This is also an indicator of how long a person has been a client and how far he/she has progressed with the loan. A high value for this ratio indicates that a high portion of the payment is still for interest, rather than the principal of the loan. This means that the client probably is not far down the loan term, with lots of potential to default.

The highest negative coefficients in the model is for var6, deliquency status2. The lower the delinquency status, as measured in days past due (or number of payments missed), the less risky the client is, and therefore a lower chance of default. As the client goes to more days past due, he/she gets riskier, and that is also reflected in the coefficient assigned.

The Hosmer-Lemeshow Goodness-of-Fit test

Partition for the Hosmer and Lemeshow Test

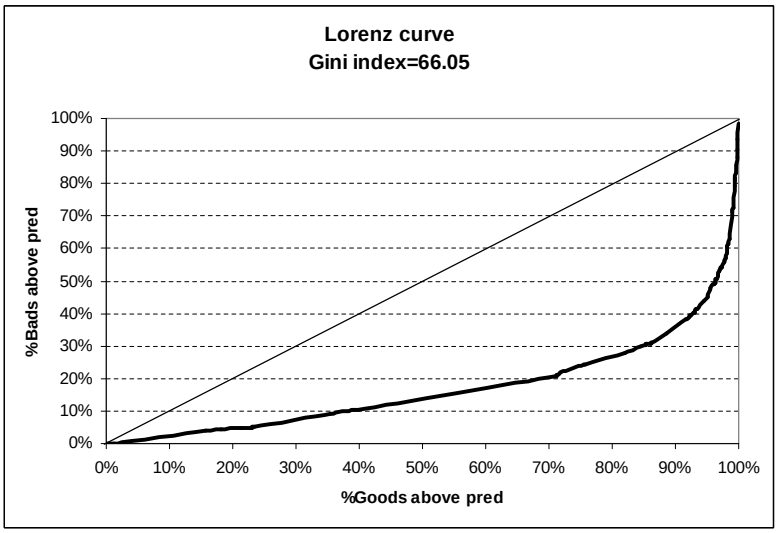
		def = 1		def = 0	
Group	Total	Observed	Expected	Observed	Expected
1	13064	44	32.42	13020	13031.58
2	6322	14	20.01	6308	6301.99
3	10012	49	49.31	9963	9962.69
4	3484	10	20.61	3474	3463.39
5	26058	121	161.68	25937	25896.32
6	9129	83	85.06	9046	9043.94
7	8337	123	108.33	8214	8228.67
8	7279	701	667.58	6578	6611.42

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
25.6567	6	0.0003

From the above SAS output one can conclude that the model fits the data reasonably well.

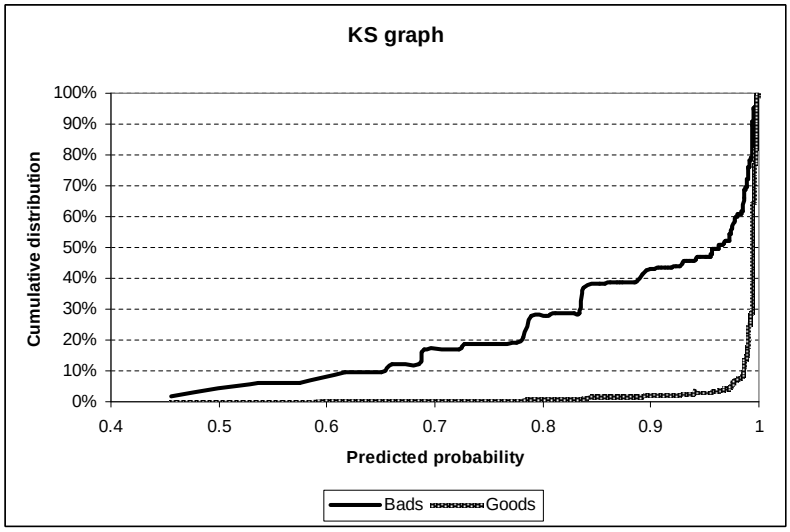
Lorenz curve and Gini index



Lorenz curve

The above graph shows that the model has significant predictive power.

KS graph



KS graph

The KS statistic is the maximum difference between the distributions of the goods and the bads. The KS for the above model is 55. The shape of the curves also indicate that the model predicts well over all ranges of the predicted probability.

12.6 Next steps

The next step in the model development would be to present the model to business experts. The business experts can then provide additional information regarding the expected impact of variables and their sloping. The business can also add any variables that they deem important and the modeler will test these. The experts can also decide to remove any of the variables.

Another important step is to assess the model on either a holdout sample, or a completely independent sample from a different time period. This will determine whether the model can be generalized and whether the model was over-fit on the development data. A drop in the Gini index of about 5% or more is normally alarming.

The last remaining step, once a final model is chosen, is to calibrate the model to give the correct prediction. Recall that a model can be developed on any default definition, but for capital purposes, the Basel II default definition must be used. The model is then calibrated to reflect this default definition.

12.7 Closing thoughts

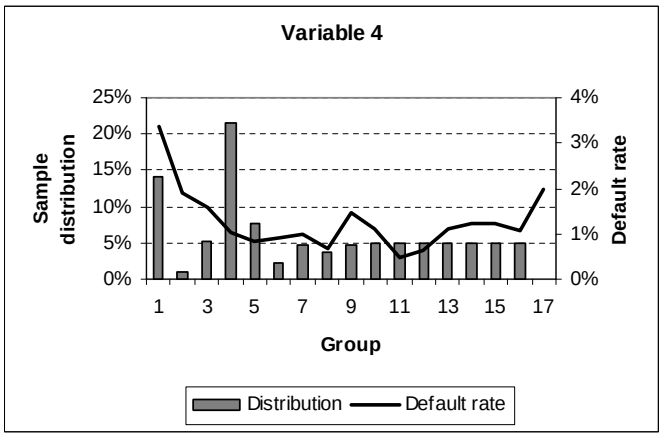
From the above example, it should be clear that credit scoring model development is more than just statistical methods. Careful thought should be applied to ensure that the statistics makes sense. It is important that the modeler understands the business for which the model is developed, to be sure that the trends can be explained. The development process involves checking several criteria at each step in the model, and not to just accept the model from the stepwise process.

Although credit scoring will never be able to predict with absolute certainty the performance of an individual borrower/loan, its benefits and usefulness have been proven time and time again. Therefore credit scoring has become essential for any bank in the world.

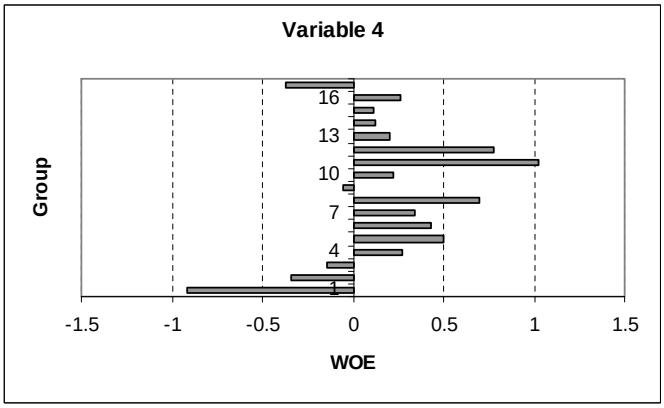
12.8 Appendix 1: Initial bivariate analysis

In order to avoid repitition, only the results of the bivariate analysis between the variable and the outcome variable is shown for the variables that were selected in the stepwise procedure.

Variable 4:



Var4 Distribution and bad rate

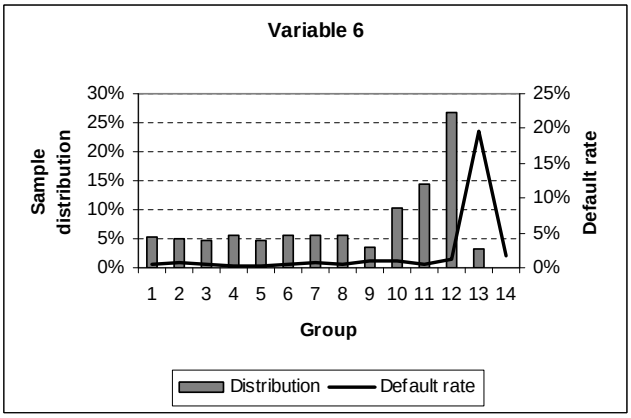


Var4 WOE

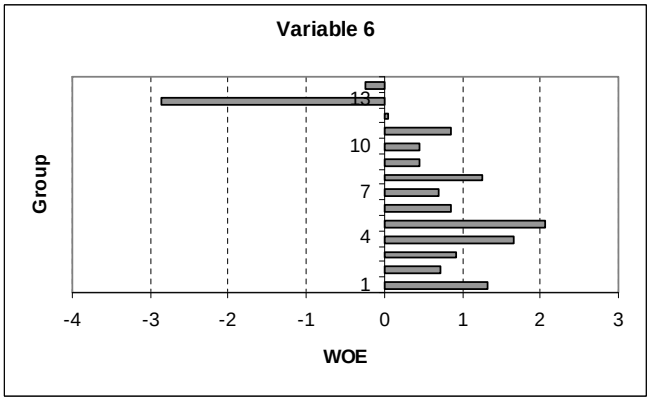
The following is considered for the secondary bucketing:

- The business hypothesis is that the higher the group number, the lower the risk.
- Group 1 to 3 will be combined.
- Bucket 17 represents missing observations and will be grouped with bucket 1 to 3 as it has a similar WOE and default rate.
- Group 4 to 16 will be grouped together.
- The variable does support the business hypothesis, but not very clearly, so only 2 buckets will be created.

Variable 6:



Var6 Distribution and bad rate

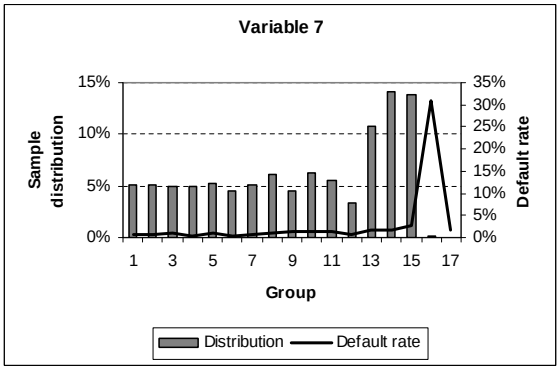


Var6 WOE

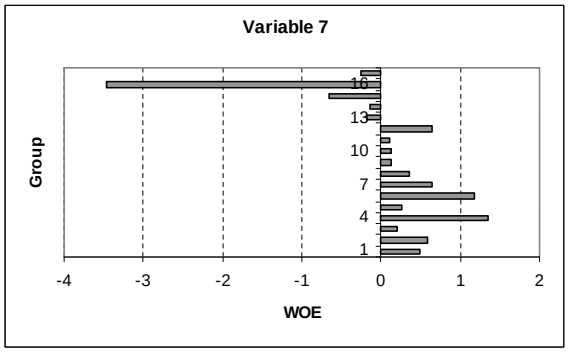
The following is considered for the secondary bucketing:

- The business hypothesis is that the higher the group number, the higher the default rate should be.
- Group 1 to 5 will be combined, based on the WOE.
- Group 6 to 8 will be grouped together.
- Group 10-12 will be added together.
- Group 13 and 14 are the only ones with negative WOE and will be combined.
- No missing observations were observed for this variable.

Variable 7:



Var7 Distribution and bad rate

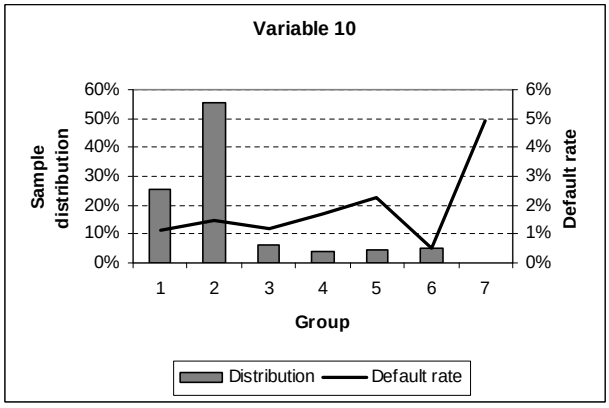


Var7 WOE

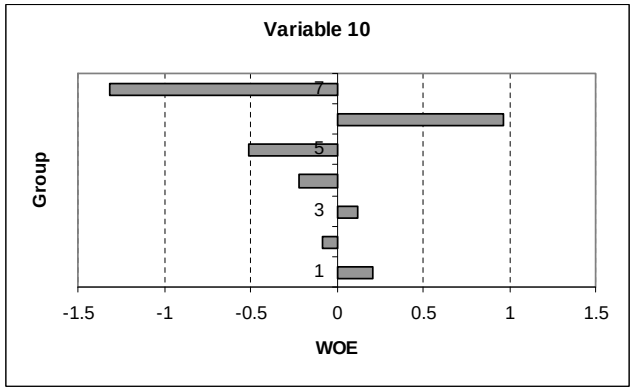
The following is considered for the secondary bucketing:

- The business hypothesis is that the higher the group number, the higher the default rate should be.
- Group 1 to 6 will be combined, based on the WOE.
- Group 7 to12 will be added together.
- Group 13 and 14 are the only will be combined.
- Bucket 17 represents missing observations and will be grouped with bucket 13 to 14 as it has a similar WOE and default rate.
- Group 15 and group 16 is combined.
- There will be some loss of predictive power and information, but the variable will be a lot more stable and less chance of over-fitting on the data.

Variable 10:



Var10 Distribution and bad rate

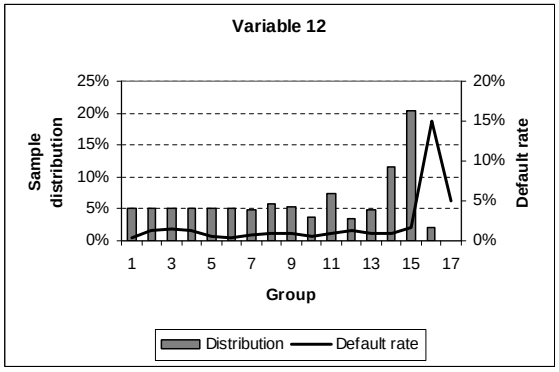


Var10 WOE

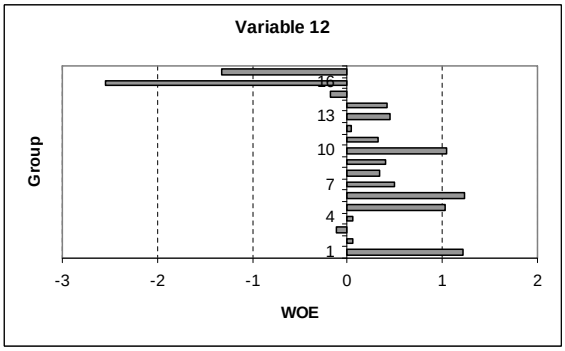
The following is considered for the secondary bucketing:

- The business hypothesis is that the higher the group number, the lower the default rate should be.
- Group 1 to 5 will be combined
- Group 6 will be kept separately.
- Bucket 7 represents missing observations and will be grouped with bucket 1 to 5 as it makes the most sense and does not have enough of the sample distribution in the bucket to be kept separately.
- There will be some loss of predictive power and information, but the variable will be a lot more stable and less chance of over-fitting on the data and it supports the business hypothesis.

Variable 12:



Var12 Distribution and bad rate

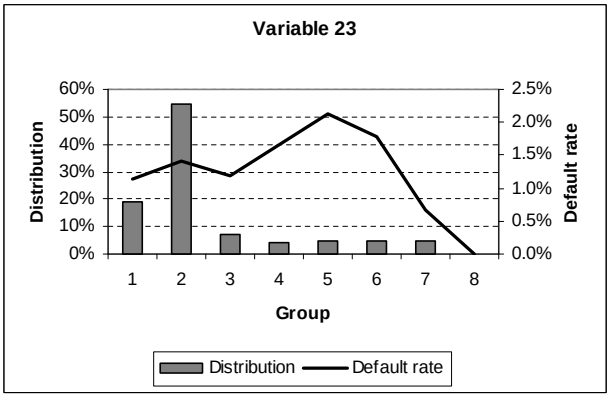


Var12 WOE

The following is considered for the secondary bucketing:

- The business hypothesis is that the higher the group number, the higher the default rate should be.
- There is no business reason for groups 2 to 4 to be different from group 1 and 5, so groups 1 to 10 will be grouped together.
- Group 11 to 14 will be added together.
- Group 15 to 17 will be combined.
- Group 17 contains all the missing values and it does not seem to be informative, especially because of the very low number of observations in the bucket.

Variable 23:



Var23 Distribution and bad rate

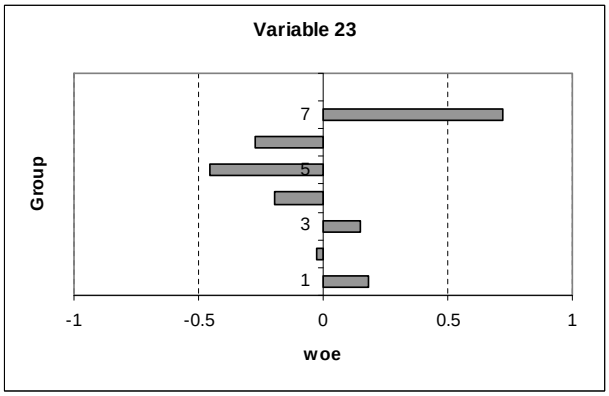
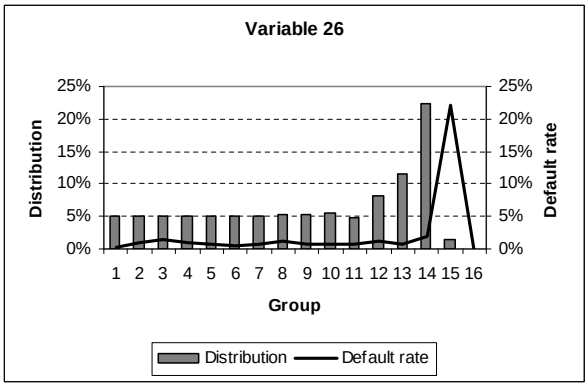


Figure 12.1: Var23 WOE

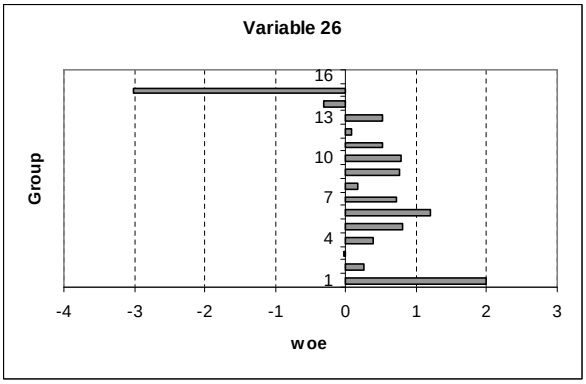
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the lower the default rate should be.
- Note that group 8 indicates missing values.
- The data does show the trend, but not clearly, so only 2 buckets will be created for this variable.
- Group 1 to 6 will be added together.
- Group 7 and 8 will be combined.

Variable 26:



Var26 Distribution and bad rate

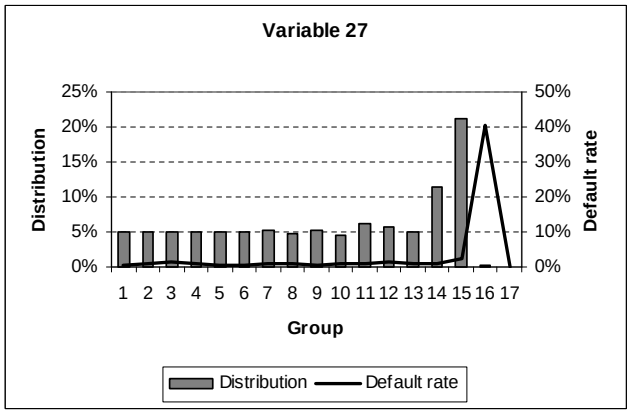


Var26 WOE

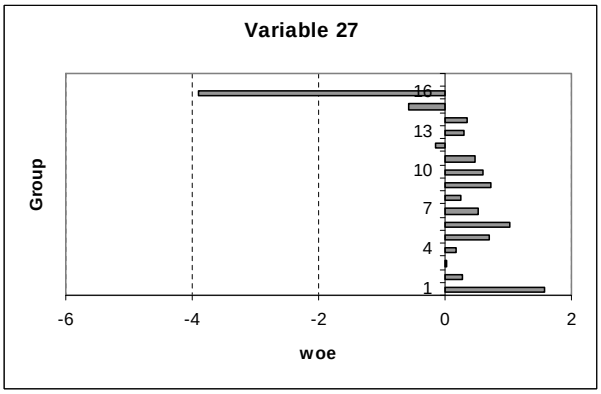
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the higher the default rate should be.
- Note that group 16 indicates missing values.
- Group 1 to 6 will be added together.
- Group 7 to 11 will be combined.
- Group 12 and 13 will form a new group.
- Group 14 to 16 will make up the final group.

Variable 27:



Var27 Distribution and bad rate

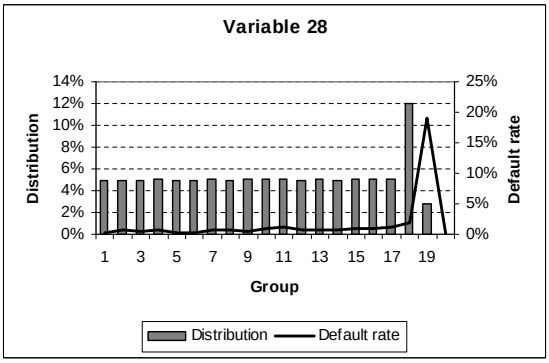


Var27 WOE

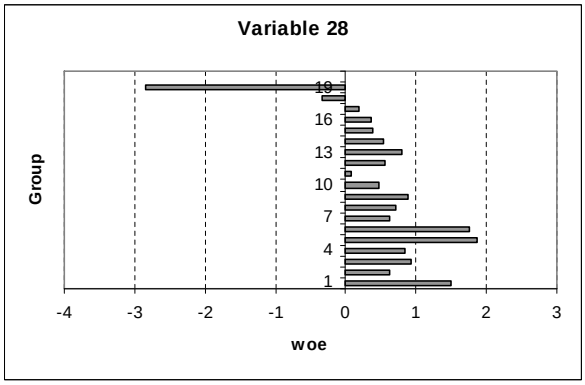
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the higher the default rate should be.
- Note that group 17 indicates missing values.
- Group 1 to 11 will be added together.
- Group 12 to 14 will form a new group.
- Group 15 to 17 will make up the final group.

Variable 28:



Var28 Distribution and bad rate

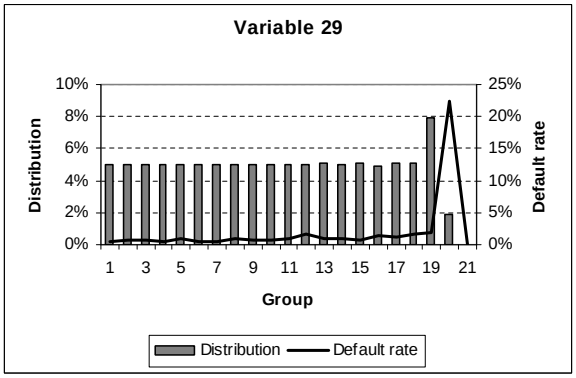


Var28 WOE

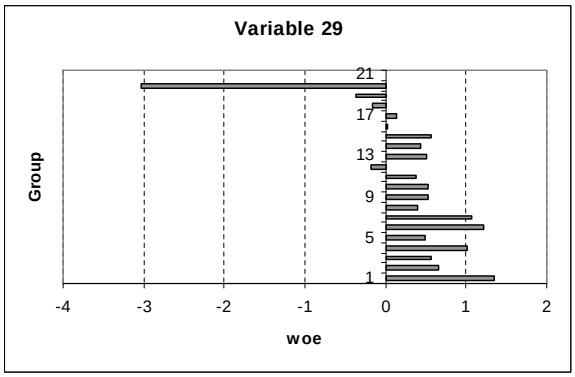
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the higher the default rate should be.
- Note that group 20 indicates missing values.
- Group 1 to 6 will be added together.
- Group 7 to 13 will be combined.
- Group 14 to 17 will form a new group.
- Group 18 to 20 will make up the final group.

Variable 29:



Var29 Distribution and bad rate

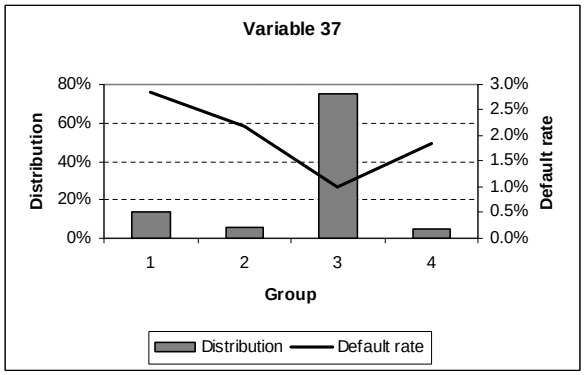


Var29 WOE

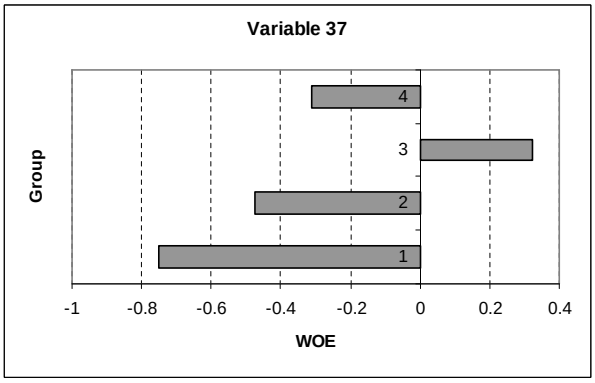
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the higher the default rate should be.
- Note that group 21 indicates missing values.
- Group 1 to 7 will be added together.
- Group 8 to 11 will be combined.
- Group 12 to 16 will form a new group.
- Group 17 to 21 will make up the final group.

Variable 37:



Var37 Distribution and bad rate

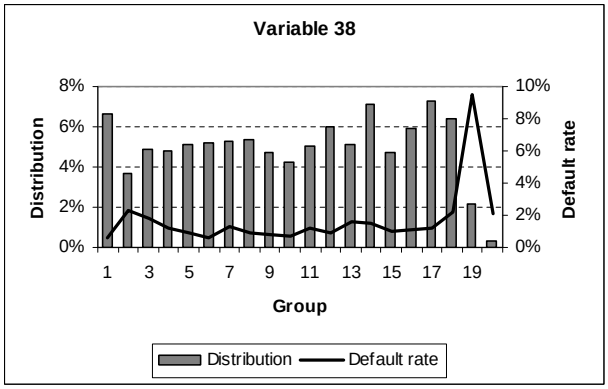


Var37 WOE

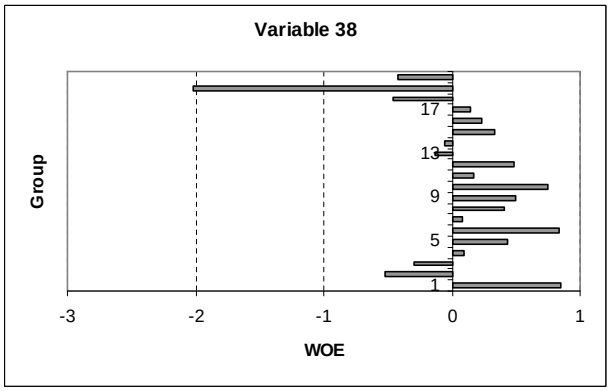
The following is considered for the secondary bucketing:

- Note that group 1 indicates missing values. As the default rate is the highest for this group, and there is a substantial percentage of the sample in that group, this is a clear example of an informative missing value.
- Group 1 and 2 will be combined.
- Group 4 will become group 2 and group 3 will remain group 3.

Variable 38:



Var38 Distribution and bad rate

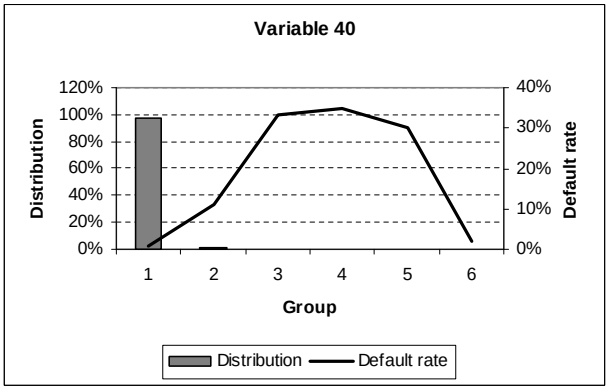


Var38 WOE

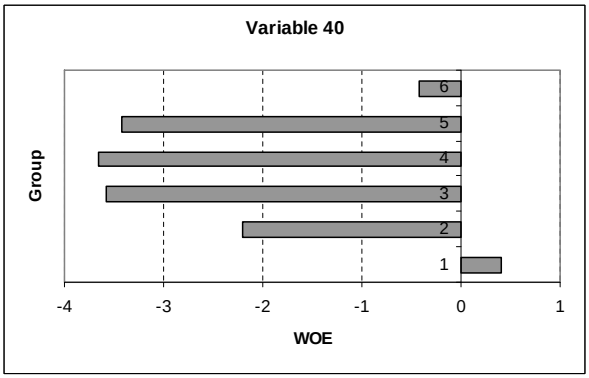
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the higher the default rate should be.
- Note that group 20 indicates missing values.
- Group 1 to 10 will be added together.
- Group 11 to 17 will be combined.
- Group 18 to 20 will form a new group.

Variable 40:



Var40 Distribution and bad rate

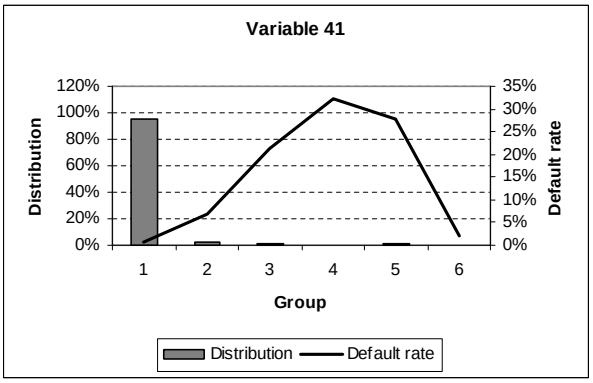


Var40 WOE

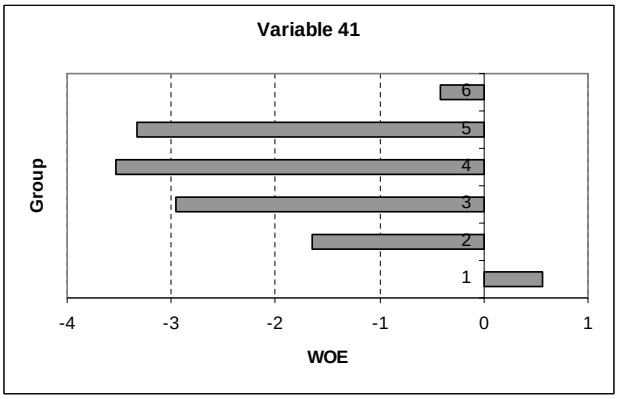
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the higher the default rate should be.
- Note that group 6 indicates missing values.
- The distribution of the data is a concern as about 97% of the sample lies in the first group.
- Nevertheless, group 2 to 6 will be combined and group 1 will be its own group.

Variable 41:



Var41 Distribution and bad rate

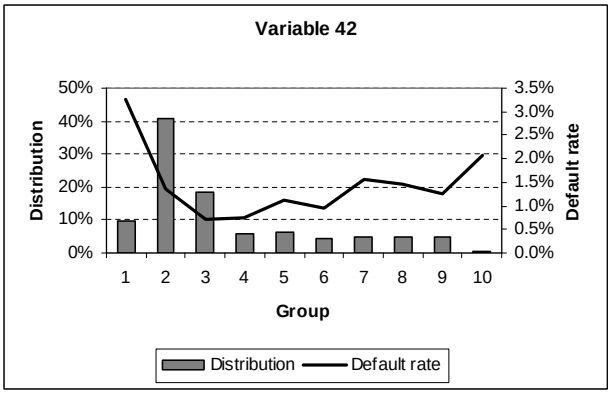


Var41 WOE

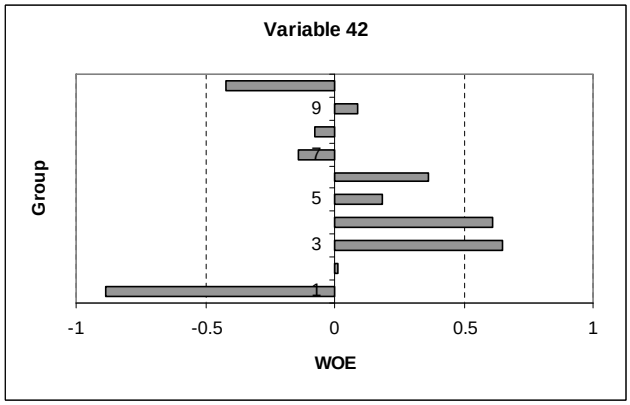
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the higher the default rate should be.
- Note that group 6 indicates missing values.
- The distribution of the data is a concern as about 95% of the sample lies in the first group.
- Nevertheless, group 2 to 6 will be combined and group 1 will be its own group.

Variable 42:



Var42 Distribution and bad rate

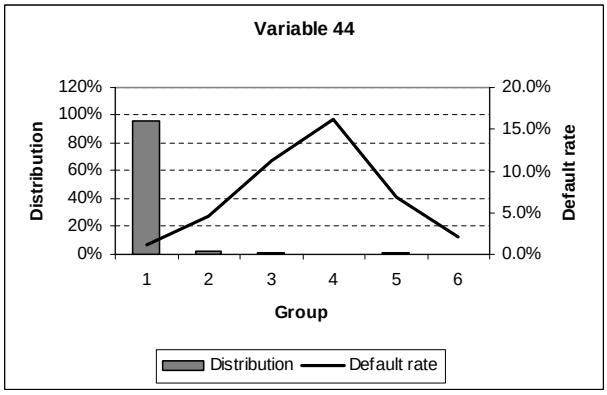


Var42 WOE

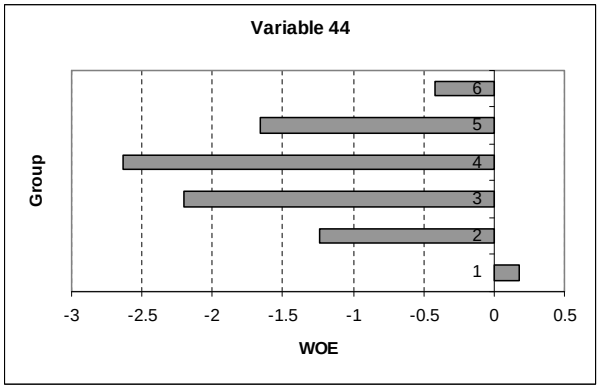
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the lower the default rate should be.
- Note that group 10 indicates missing values.
- Group 7 to 9 does not support the business hypothesis, so only 2 groups will be created for this variable.
- Group 1 and 10 will be combined and the rest of the groups will be combined.

Variable 44:



Var44 Distribution and bad rate



Var44 WOE

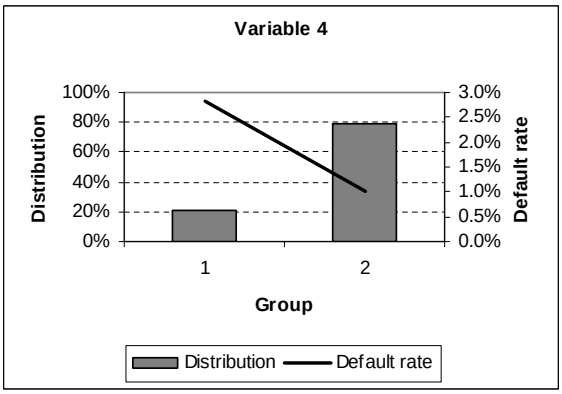
The following is considered for the secondary bucketing:

- The variable's business hypothesis is that the higher the group number, the higher the default rate should be.
- Note that group 6 indicates missing values.
- The distribution of the data is a concern as about 96% of the sample lies in the first group.
- Nevertheless, group 2 to 6 will be combined and group 1 will be its own group.

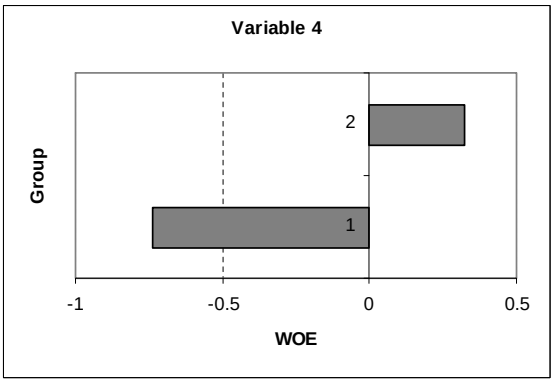
12.9 Appendix 2: Secondary bucketing of variables

Again, to avoid repitition, only the results of the bivariate analysis between the variable and the outcome variable is shown for the variables that were selected in the stepwise procedure.

Variable 4:

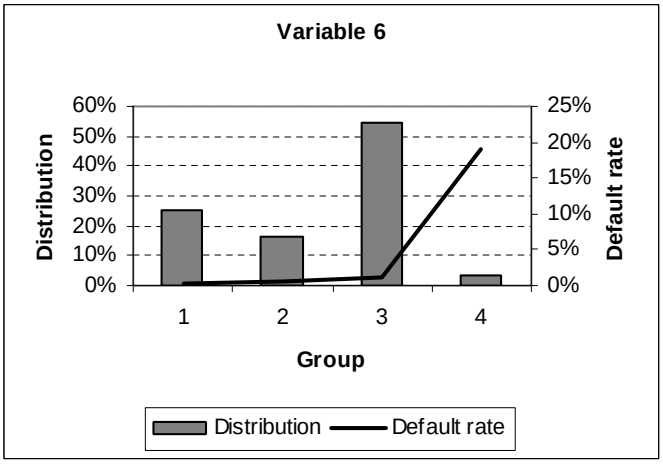


Var4 Distribution and bad rate

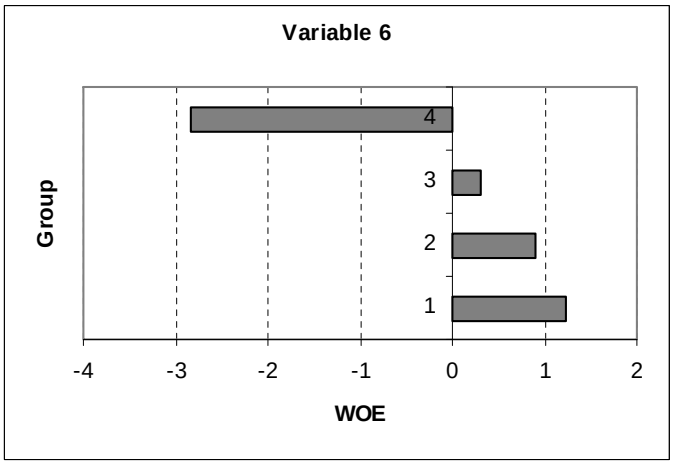


Var4 WOE

Variable 6:

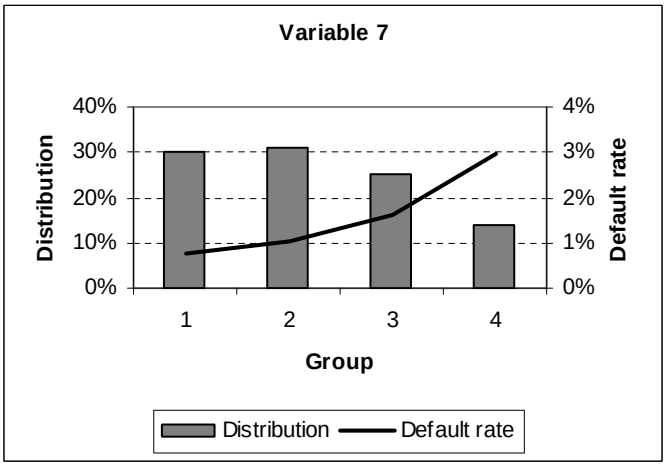


Var6 Distribution and bad rate

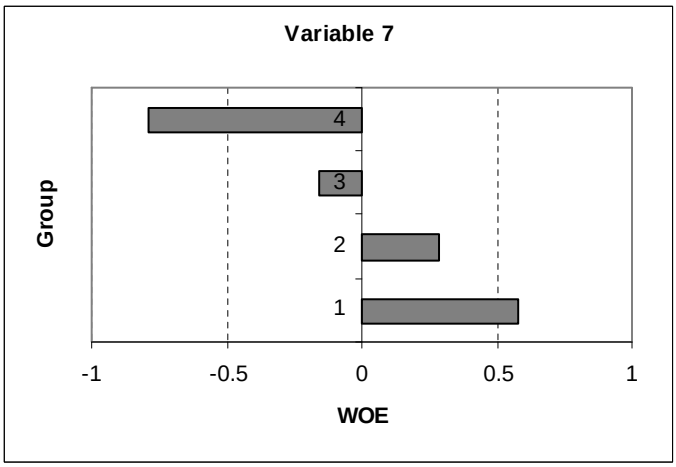


Var6 WOE

Variable 7:

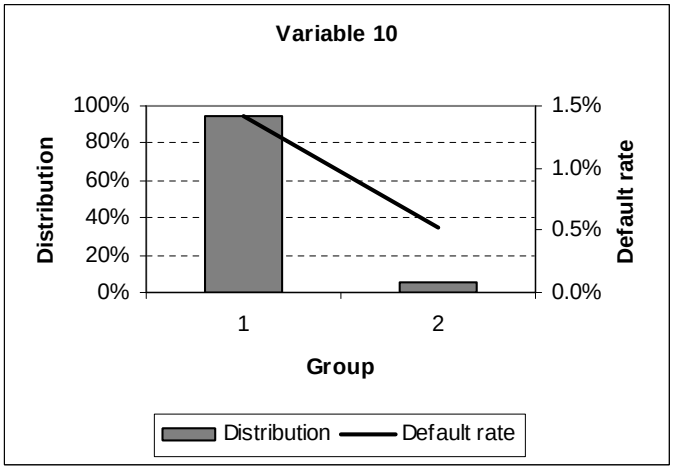


Var7 Distribution and bad rate

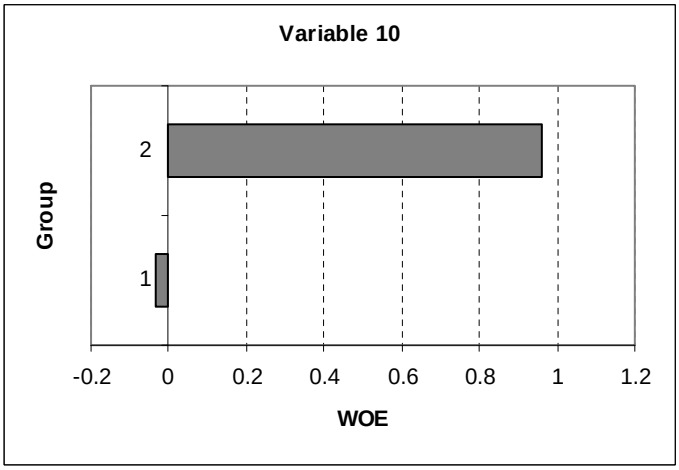


Var7 WOE

Variable 10:

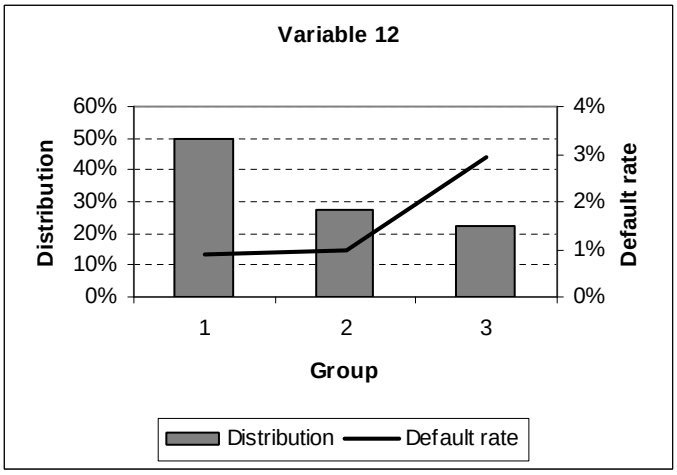


Var10 Distribution and bad rate

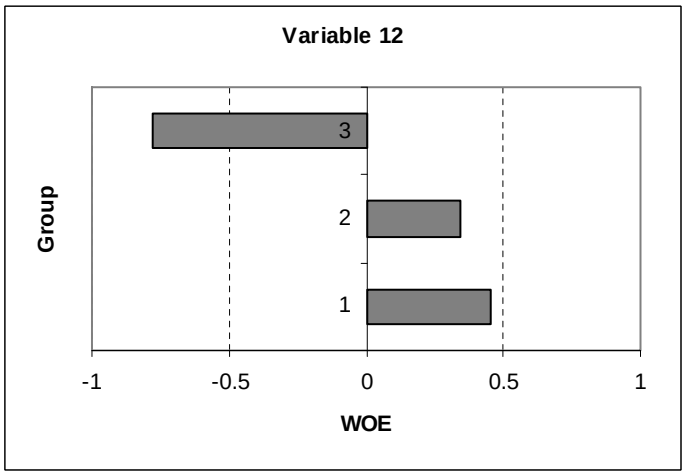


Var10 WOE

Variable 12:

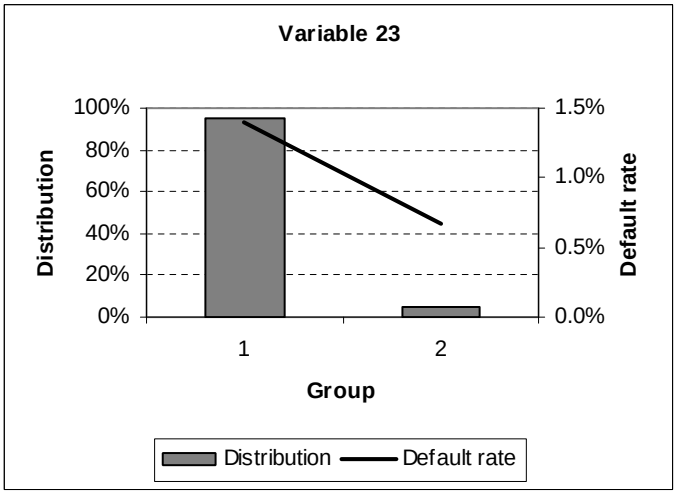


Var12 Distribution and bad rate

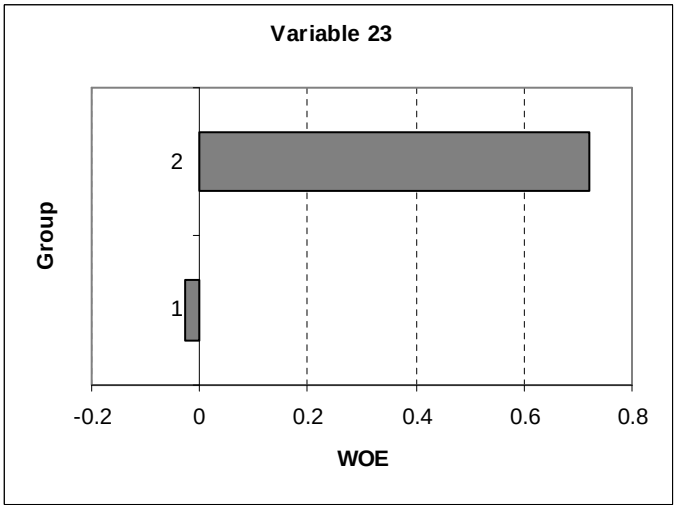


Var12 WOE

Variable 23:

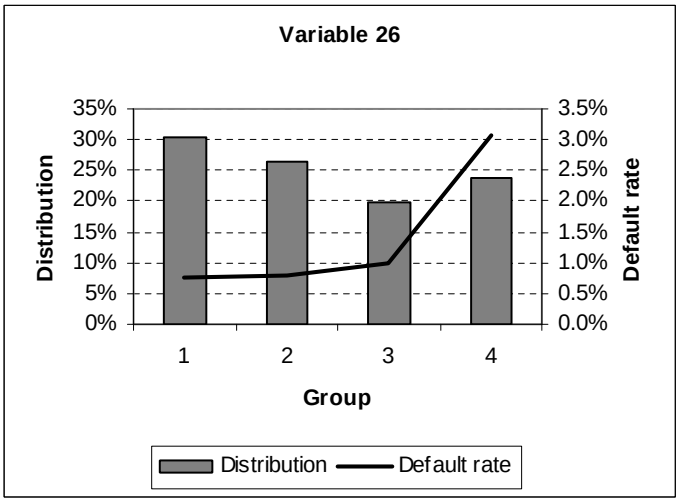


Var23 Distribution and bad rate

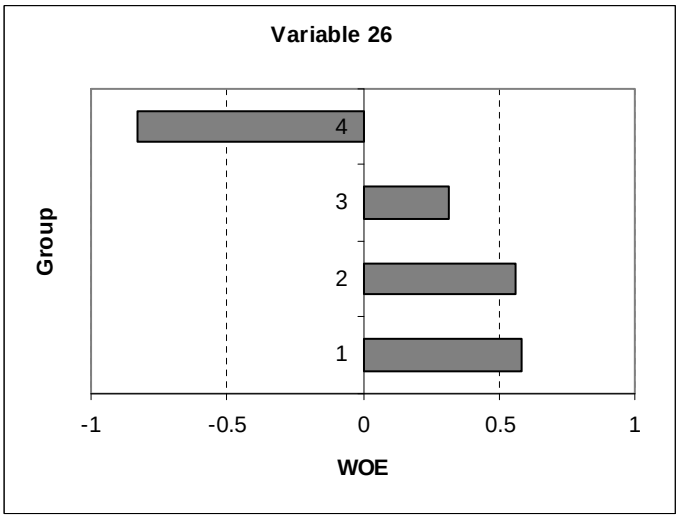


Var23 WOE

Variable 26:

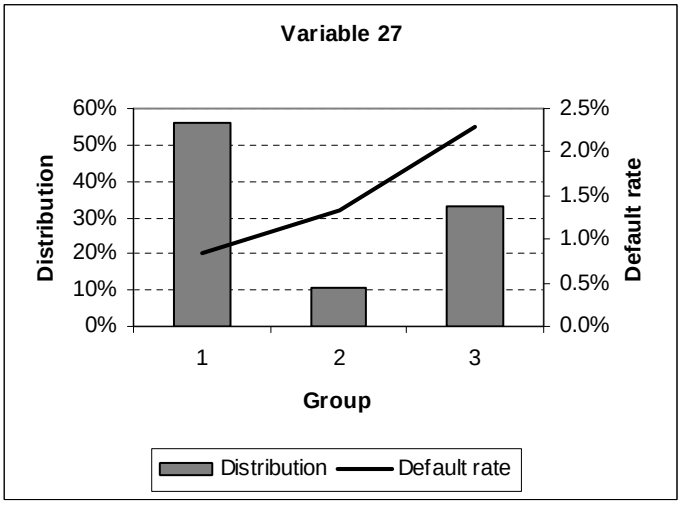


Var26 Distribution and bad rate

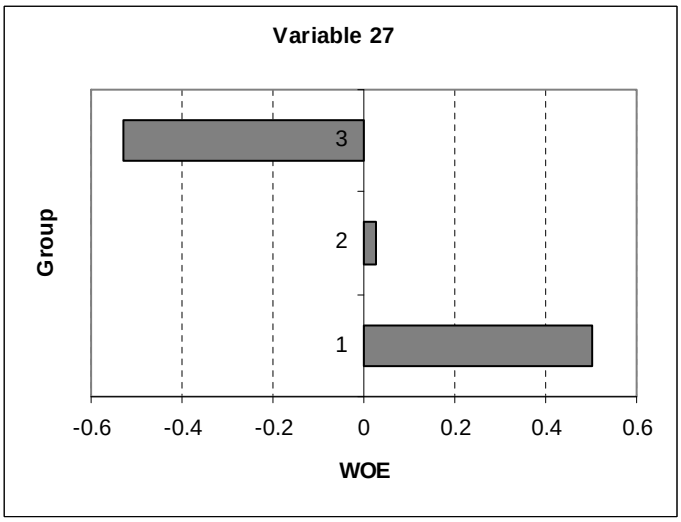


Var26 WOE

Variable 27:

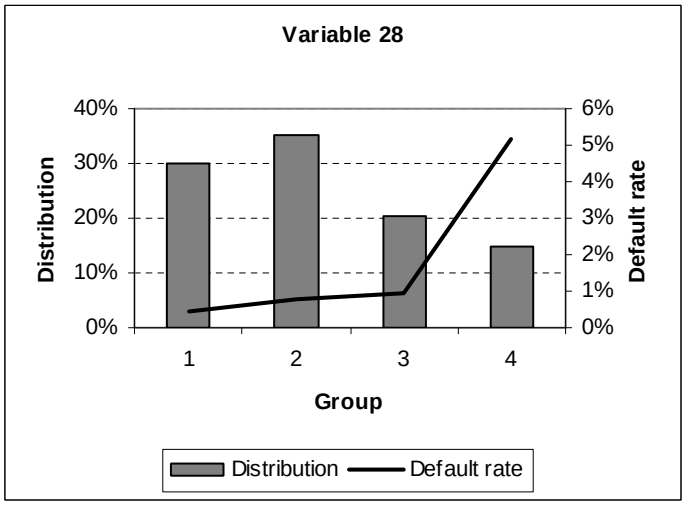


Var27 Distribution and bad rate

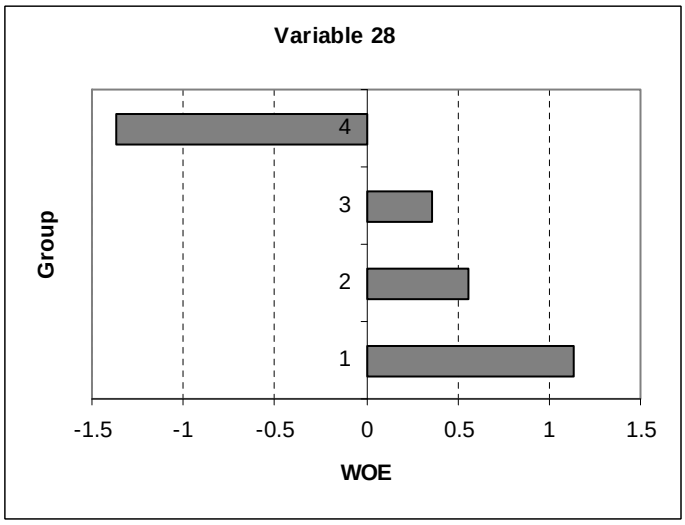


Var27 WOE

Variable 28:

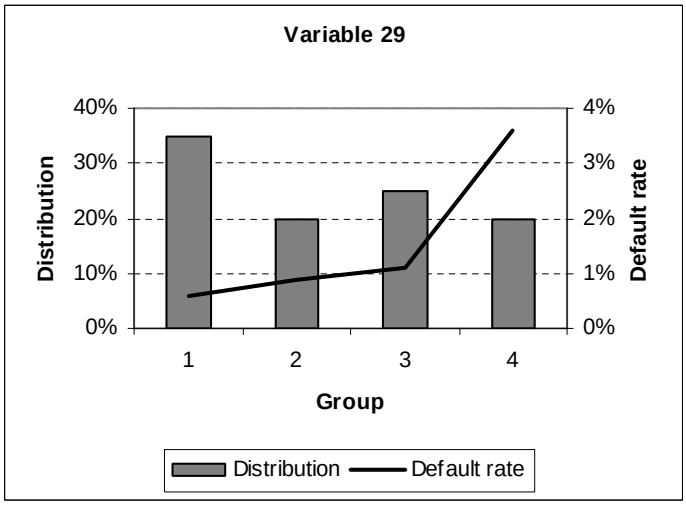


Var28 Distribution and bad rate

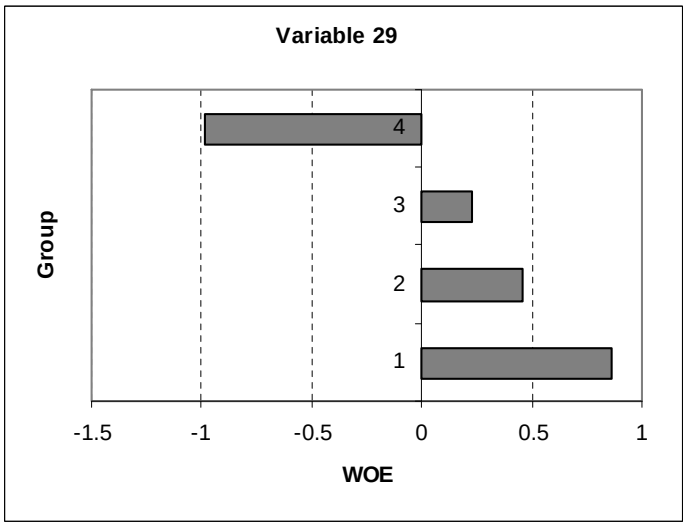


Var28 WOE

Variable 29:

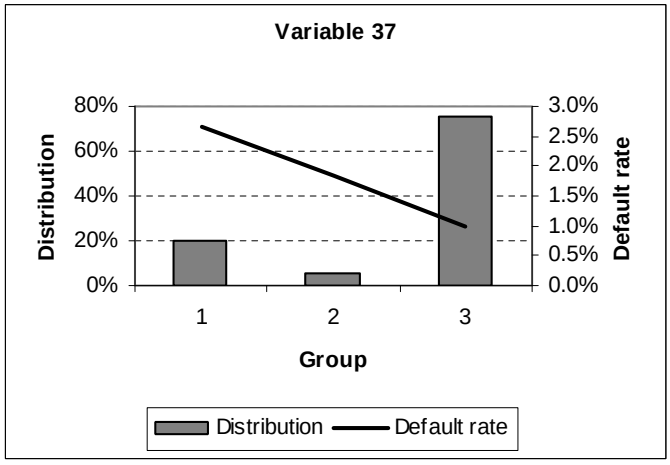


Var29 Distribution and bad rate

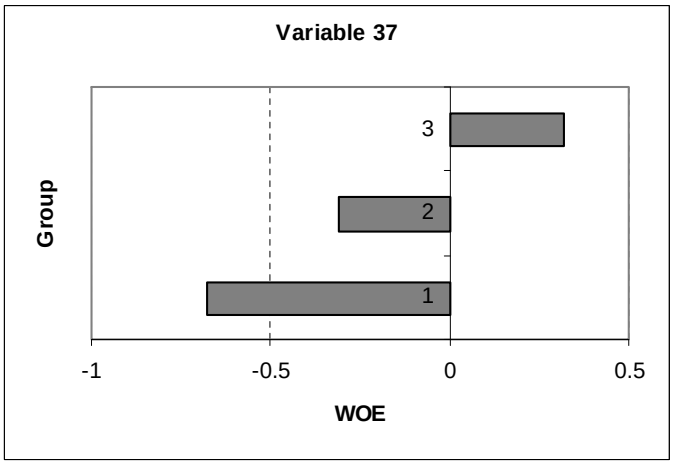


Var29 WOE

Variable 37:

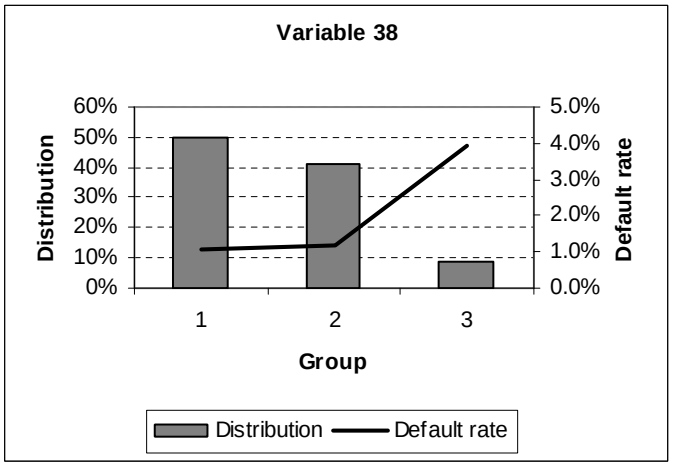


Var37 Distribution and bad rate

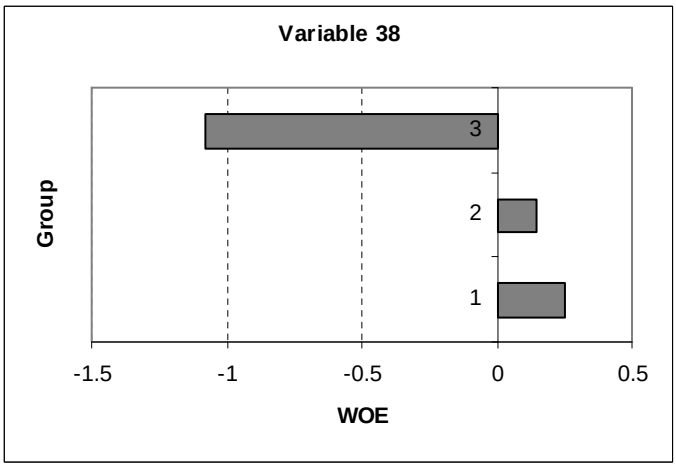


Var37 WOE

Variable 38:



Var38 Distribution and bad rate



Var38 WOE

Variable 40:

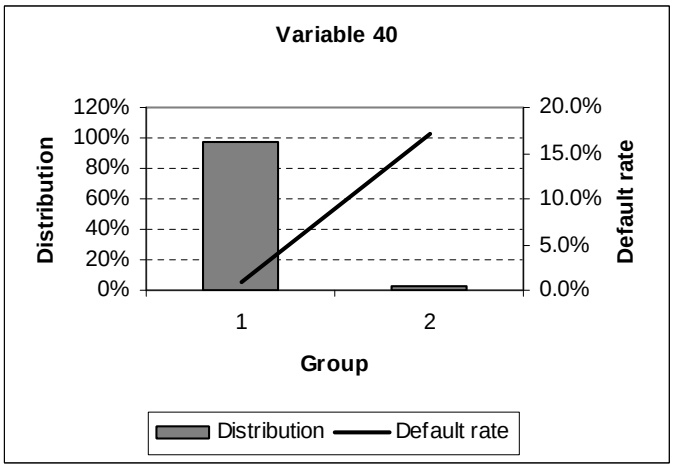
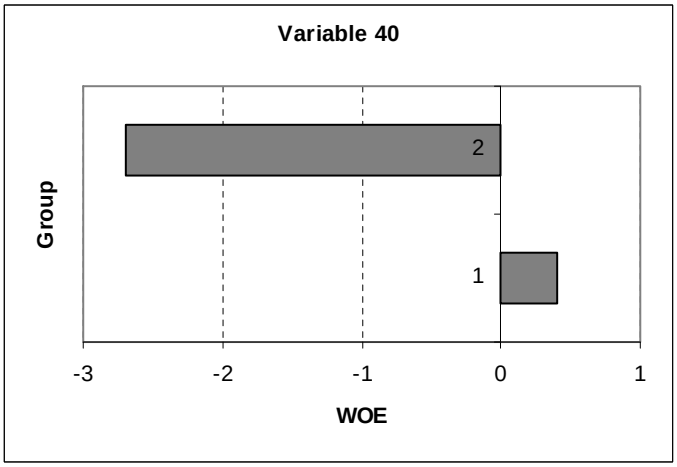
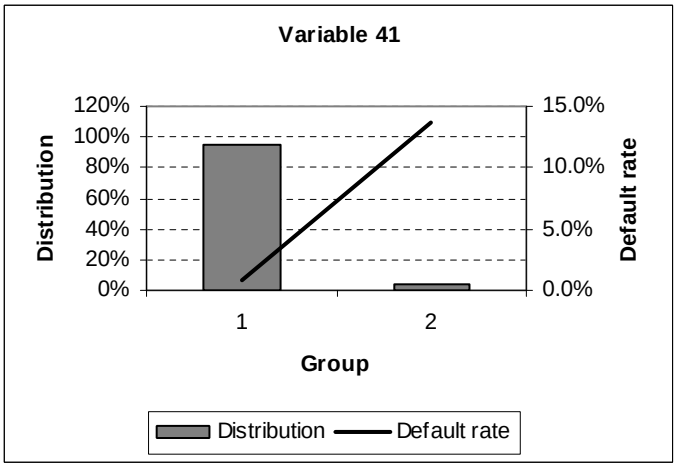


Figure 12.2: Var40 Distribution and bad rate

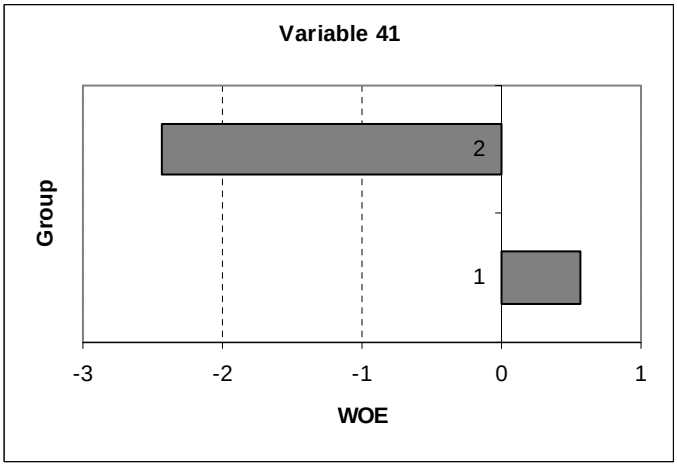


Var40 WOE

Variable 41:

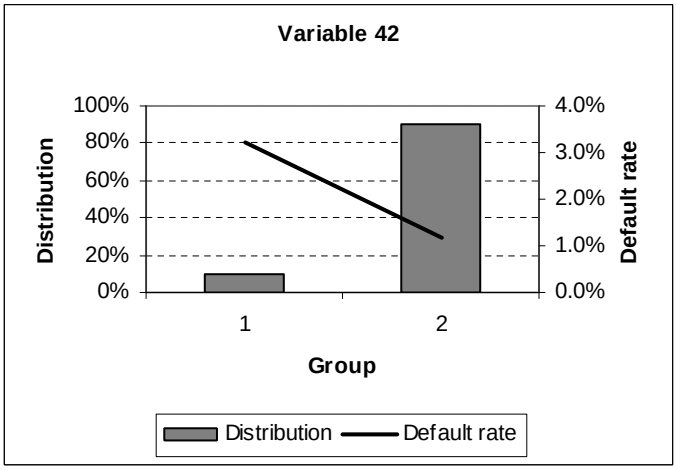


Var41 Distribution and bad rate

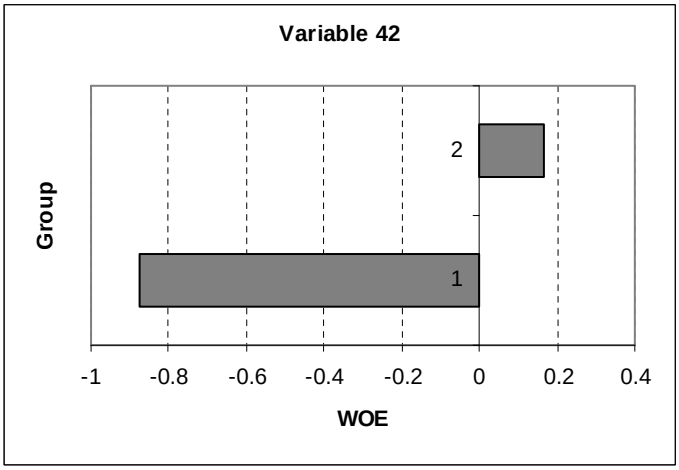


Var41 WOE

Variable 42:

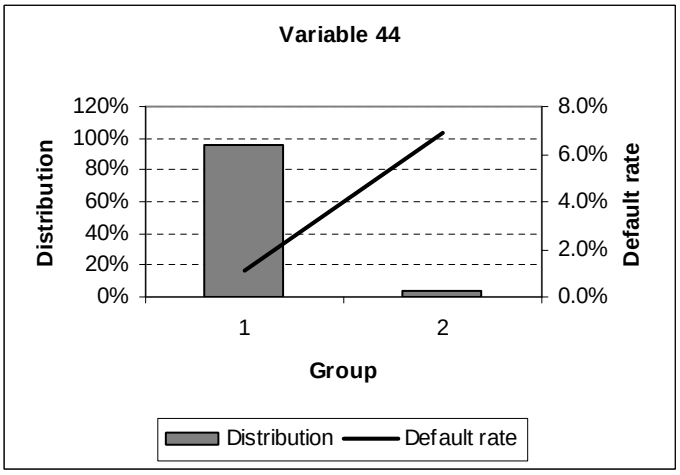


Var42 Distribution and bad rate

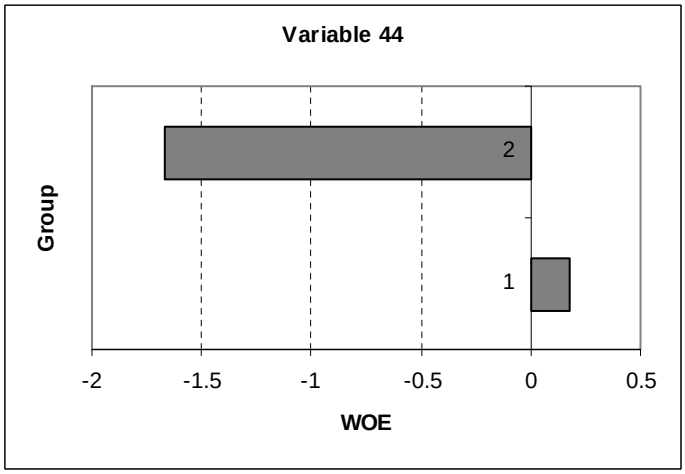


Var42 WOE

Variable 44:



Var44 Distribution and bad rate



Var44 WOE

12.10 Appendix 3: Cluster analysis SAS code and output

The following SAS code was used to do the cluster analysis and to draw the cluster diagram:

```
proc varclus data=prac.bva5 outtree=tree centroid
                            maxclusters=10;
      var var10gg var11gg var12gg var13gg var14gg
          var15gg var16gg var17gg var18gg var19gg
          var20gg var21gg var22gg var23gg var24gg
          var25gg var26gg var27gg var28gg var29gg
          var2gg  var30gg var31gg var32gg var33gg
          var36gg var37gg var38gg var40gg var41gg
          var42gg var43gg var44gg var45gg var46gg
          var47gg var48gg var4gg  var6gg  var7gg
          var8gg var9gg;
run;

axis1 label=none;
proc tree data=tree horizontal vaxis=axis1;
   height _propor_;
run;
```

Step 1:

Cluster Summary for 1 Cluster

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	42	42	3.944637	0.0939

Cluster 1 will be split because it has the smallest proportion of variation explained, 0.09392, which is less than the PROPORTION=1 value.

Step 2:

Cluster Summary for 2 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	25	25	11.29849	0.4519
2	17	17	5.16238	0.3037

Total variation explained = 16.46087 Proportion = 0.3919

Cluster 2 will be split because it has the smallest proportion of variation explained, 0.303669, which is less than the PROPORTION=1 value.

Step 3:

Cluster Summary for 3 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	25	25	11.29849	0.4519
2	9	9	3.404679	0.3783
3	8	8	4.526977	0.5659

Total variation explained = 19.23015 Proportion = 0.4579

Cluster 2 will be split because it has the smallest proportion of variation explained, 0.378298, which is less than the PROPORTION=1 value.

Step 4:

Cluster Summary for 4 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	25	25	11.29849	0.4519
2	6	6	5.080724	0.8468
3	8	8	4.526977	0.5659
4	3	3	1.588421	0.5295

Total variation explained = 22.49461 Proportion = 0.5356

Cluster 1 will be split because it has the smallest proportion of variation explained, 0.45194, which is less than the PROPORTION=1 value.

Step 5:

Cluster Summary for 5 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	17	17	13.18809	0.7758
2	6	6	5.080724	0.8468
3	8	8	4.526977	0.5659
4	3	3	1.588421	0.5295
5	8	8	1.903571	0.2379

Total variation explained = 26.28778 Proportion = 0.6259

Cluster 5 will be split because it has the smallest proportion of variation explained, 0.237946, which is less than the PROPORTION=1 value.

Step 6:

Cluster Summary for 6 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	17	17	13.18809	0.7758
2	6	6	5.080724	0.8468
3	8	8	4.526977	0.5659
4	3	3	1.588421	0.5295
5	4	4	2.256519	0.5641
6	4	4	2.021426	0.5054

Total variation explained = 28.66215 Proportion = 0.6824

Cluster 6 will be split because it has the smallest proportion of variation explained, 0.505356, which is less than the PROPORTION=1 value.

Step 7:

Cluster Summary for 7 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	17	17	13.18809	0.7758
2	6	6	5.080724	0.8468
3	8	8	4.526977	0.5659
4	3	3	1.588421	0.5295
5	4	4	2.256519	0.5641
6	3	3	2.011754	0.6706
7	1	1	1	1.0000

Total variation explained = 29.65248 Proportion = 0.7060

Cluster 4 will be split because it has the smallest proportion of variation explained, 0.529474, which is less than the PROPORTION=1 value.

Step 8:

Cluster Summary for 8 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	17	17	13.18809	0.7758
2	6	6	5.080724	0.8468
3	8	8	4.526977	0.5659
4	2	2	1.805052	0.9025
5	4	4	2.256519	0.5641
6	3	3	2.011754	0.6706
7	1	1	1	1.0000
8	1	1	1	1.0000

Total variation explained = 30.86911 Proportion = 0.7350

Cluster 5 will be split because it has the smallest proportion of variation explained, 0.56413, which is less than the PROPORTION=1 value.

Step 9:

Cluster Summary for 9 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	17	17	13.18809	0.7758
2	6	6	5.080724	0.8468
3	8	8	4.526977	0.5659
4	2	2	1.805052	0.9025
5	2	2	1.852209	0.9261
6	3	3	2.011754	0.6706
7	1	1	1	1.0000
8	1	1	1	1.0000
9	2	2	1.778463	0.8892

Total variation explained = 32.24327 Proportion = 0.7677

Cluster 3 will be split because it has the smallest proportion of variation explained, 0.565872, which is less than the PROPORTION=1 value.

Step 10:

Cluster Summary for 10 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	17	17	13.18809	0.7758
2	6	6	5.080724	0.8468
3	6	6	4.905775	0.8176
4	2	2	1.805052	0.9025
5	2	2	1.852209	0.9261
6	3	3	2.011754	0.6706
7	1	1	1	1.0000
8	1	1	1	1.0000
9	2	2	1.778463	0.8892
10	2	2	1.748479	0.8742

Total variation explained = 34.37054 Proportion = 0.8183

12.11 Appendix 4: Stepwise logistic regression SAS code and output

The following SAS code was used for the stepwise logistic regression:

```
proc logistic data=prac.bva5 outest=betas covout;
  class var10gg(ref='2') var11gg(ref='2') var12gg(ref='3')
        var13gg(ref='3') var14gg(ref='3') var15gg(ref='3')
        var16gg(ref='3') var17gg(ref='2') var18gg(ref='2')
        var19gg(ref='3') var20gg(ref='4') var21gg(ref='4')
        var22gg(ref='2') var23gg(ref='2') var24gg(ref='2')
        var25gg(ref='4') var26gg(ref='4') var27gg(ref='3')
        var28gg(ref='4') var29gg(ref='4') var30gg(ref='4')
        var31gg(ref='3') var32gg(ref='3') var33gg(ref='3')
        var36gg(ref='2') var37gg(ref='3') var38gg(ref='3')
        var40gg(ref='2') var41gg(ref='2') var42gg(ref='2')
        var43gg(ref='2') var44gg(ref='2') var45gg(ref='2')
        var46gg(ref='2') var47gg(ref='2') var48gg(ref='2')
        var4gg(ref='2')  var6gg(ref='4')  var7gg(ref='4')
        var8gg(ref='2') var9gg(ref='2') /param=ref;
  model def(event='1')=var10gg var11gg var12gg var13gg
                       var14gg var15gg var16gg var17gg
                       var18gg var19gg var20gg var21gg
                       var22gg var23gg var24gg var25gg
                       var26gg var27gg var28gg var29gg
                       var30gg var31gg var32gg var33gg
                       var36gg var37gg var38gg var40gg
                       var41gg var42gg var43gg var44gg
                       var45gg var46gg var47gg var48gg
                       var4gg var6gg var7gg var8gg var9gg
                    / selection=stepwise
                      slentry=0.05
                      slstay=0.1
                      details
                      lackfit;
     output out=pred p=phat lower=lcl upper=ucl
            predprob=(individual crossvalidate);
  run;
```

Due to the vast amount of output generated by SAS, only selected output will be shown here:

Step 0: Intercept entered

```
                    The LOGISTIC Procedure

                     -2 Log L = 12102.152
```

Step 1: Var6 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9906.285
SC	12113.487	9943.624
-2 Log L	12102.152	9898.285

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2203.8668	3	<.0001
Score	6597.4182	3	<.0001
Wald	2876.1125	3	<.0001

- The Gini index for the model is 51.9.
- No variables were eligible for removal at this stage.

Step 2: Var4 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9703.954
SC	12113.487	9750.628
-2 Log L	12102.152	9693.954

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2408.1977	4	<.0001
Score	6837.0466	4	<.0001
Wald	2997.8123	4	<.0001

- The Gini index for the model is 60.1.
- No variables were eligible for removal at this stage.

Step 3: Var40 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9536.936
SC	12113.487	9592.945
-2 Log L	12102.152	9524.936

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2577.2163	5	<.0001
Score	7811.9143	5	<.0001
Wald	3200.0686	5	<.0001

- The Gini index for the model is 60.8.
- No variables were eligible for removal at this stage.

Step 4: Var42 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9493.516
SC	12113.487	9558.860
-2 Log L	12102.152	9479.516

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2622.6359	6	<.0001
Score	7835.0139	6	<.0001
Wald	3189.5847	6	<.0001

- The Gini index for the model is 61.4.
- No variables were eligible for removal at this stage.

Step 5: Var44 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9458.254
SC	12113.487	9532.933
-2 Log L	12102.152	9442.254

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2659.8979	7	<.0001
Score	7875.1140	7	<.0001
Wald	3206.9139	7	<.0001

- The Gini index for the model is 62.
- No variables were eligible for removal at this stage.

Step 6: Var28 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9421.751
SC	12113.487	9524.434
-2 Log L	12102.152	9399.751

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2702.4014	10	<.0001
Score	7914.5486	10	<.0001
Wald	3224.9579	10	<.0001

- The Gini index for the model is 63.9.
- No variables were eligible for removal at this stage.

Step 7: Var10 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9384.449
SC	12113.487	9496.467
-2 Log L	12102.152	9360.449

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2741.7027	11	<.0001
Score	7995.7836	11	<.0001
Wald	3244.5974	11	<.0001

- The Gini index for the model is 65.4.
- No variables were eligible for removal at this stage.

Step 8: Var41 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9358.523
SC	12113.487	9479.875
-2 Log L	12102.152	9332.523

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2769.6292	12	<.0001
Score	8003.5998	12	<.0001
Wald	3260.5738	12	<.0001

- The Gini index for the model is 65.9.
- No variables were eligible for removal at this stage.

Step 9: Var27 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9332.846
SC	12113.487	9472.869
-2 Log L	12102.152	9302.846

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2799.3058	14	<.0001
Score	8023.1493	14	<.0001
Wald	3261.7680	14	<.0001

- The Gini index for the model is 66.
- No variables were eligible for removal at this stage.

Step 10: Var38 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9312.669
SC	12113.487	9471.361
-2 Log L	12102.152	9278.669

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2823.4832	16	<.0001
Score	8052.0655	16	<.0001
Wald	3270.0117	16	<.0001

- The Gini index for the model is 65.8.
- No variables were eligible for removal at this stage.

Step 11: Var37 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9302.965
SC	12113.487	9480.326
-2 Log L	12102.152	9264.965

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2837.1872	18	<.0001
Score	8057.6232	18	<.0001
Wald	3270.2604	18	<.0001

- The Gini index for the model is 66.3.
- No variables were eligible for removal at this stage.

Step 12: Var29 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9292.990
SC	12113.487	9498.356
-2 Log L	12102.152	9248.990

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2853.1622	21	<.0001
Score	8083.8395	21	<.0001
Wald	3278.8370	21	<.0001

- The Gini index for the model is 67.3.
- No variables were eligible for removal at this stage.

Step 13: Var12 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9282.900
SC	12113.487	9506.935
-2 Log L	12102.152	9234.900

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2867.2524	23	<.0001
Score	8092.1424	23	<.0001
Wald	3273.5799	23	<.0001

- The Gini index for the model is 67.6.
- No variables were eligible for removal at this stage.

Step 14: Var26 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9276.816
SC	12113.487	9528.856
-2 Log L	12102.152	9222.816

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2879.3360	26	<.0001
Score	8098.2022	26	<.0001
Wald	3267.7694	26	<.0001

- The Gini index for the model is 67.7.
- No variables were eligible for removal at this stage.

Step 15: Var23 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9271.012
SC	12113.487	9532.386
-2 Log L	12102.152	9215.012

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2887.1404	27	<.0001
Score	8103.7566	27	<.0001
Wald	3268.2559	27	<.0001

- The Gini index for the model is 68.
- No variables were eligible for removal at this stage.

Step 16: Var7 entered

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12104.152	9265.756
SC	12113.487	9555.135
-2 Log L	12102.152	9203.756

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2898.3961	30	<.0001
Score	8126.6833	30	<.0001
Wald	3275.1239	30	<.0001

- The Gini index for the model is 67.9.
- No variables were eligible for removal at this stage.
- No variables were eligible for entry any further.

Hosmer-Lemeshow Goodness-of-fit test

Partition for the Hosmer and Lemeshow Test

		def = 1		def = 0	
Group	Total	Observed	Expected	Observed	Expected
1	6054	10	11.20	6044	6042.80
2	8369	35	21.44	8334	8347.56
3	8455	26	28.63	8429	8426.37
4	8440	31	35.44	8409	8404.56
5	7818	25	39.99	7793	7778.01
6	8658	42	48.58	8616	8609.42
7	7845	43	52.24	7802	7792.76
8	8399	55	69.04	8344	8329.96
9	8368	108	89.60	8260	8278.40
10	11279	770	748.85	10509	10530.15

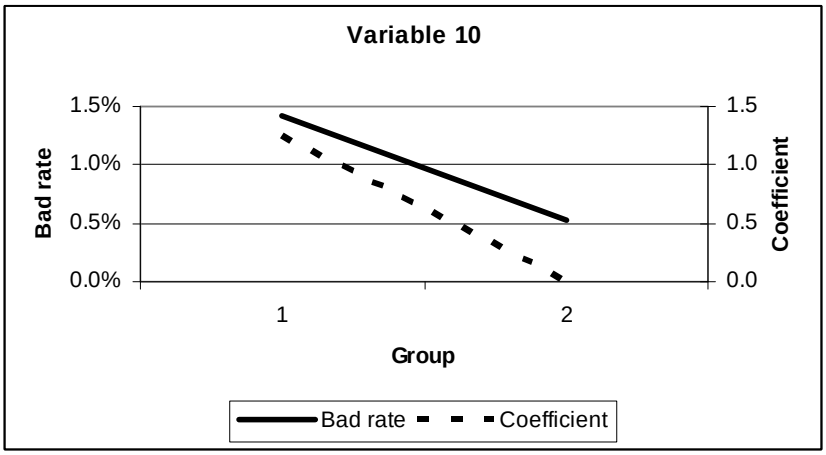
The LOGISTIC Procedure

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
25.0492	8	0.0015

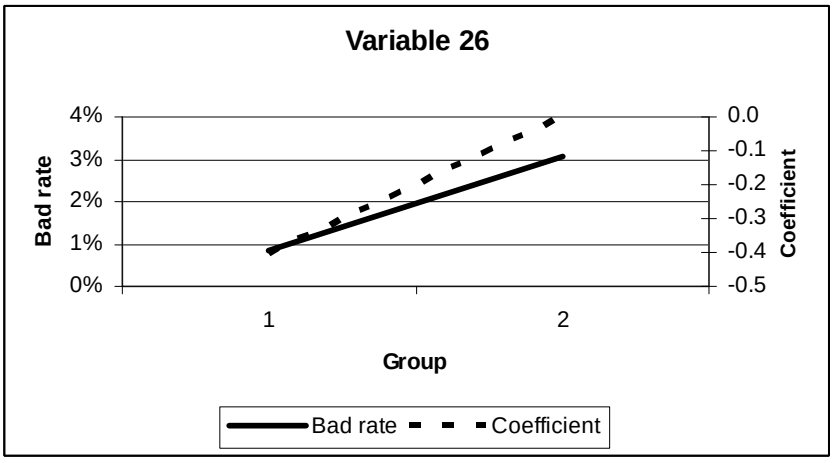
12.12 Appendix 5: Default rate sloping and model coefficients

Variable 10:



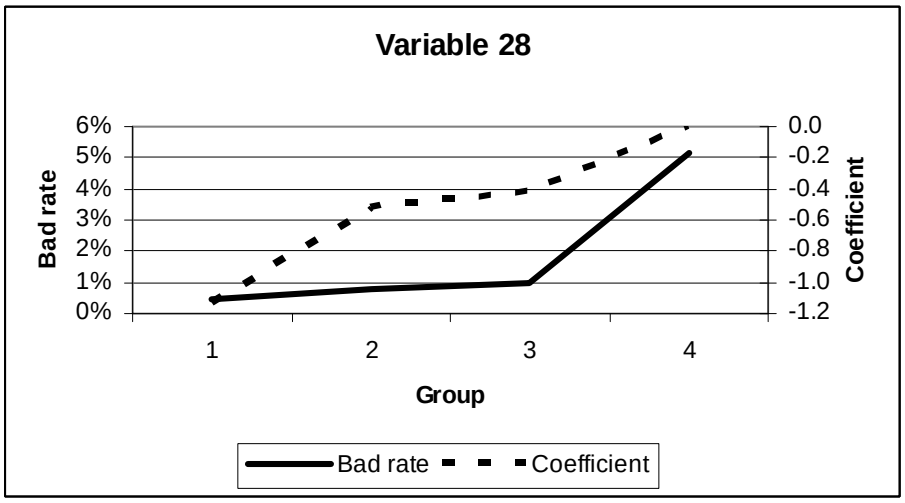
Var10

Variable 26:



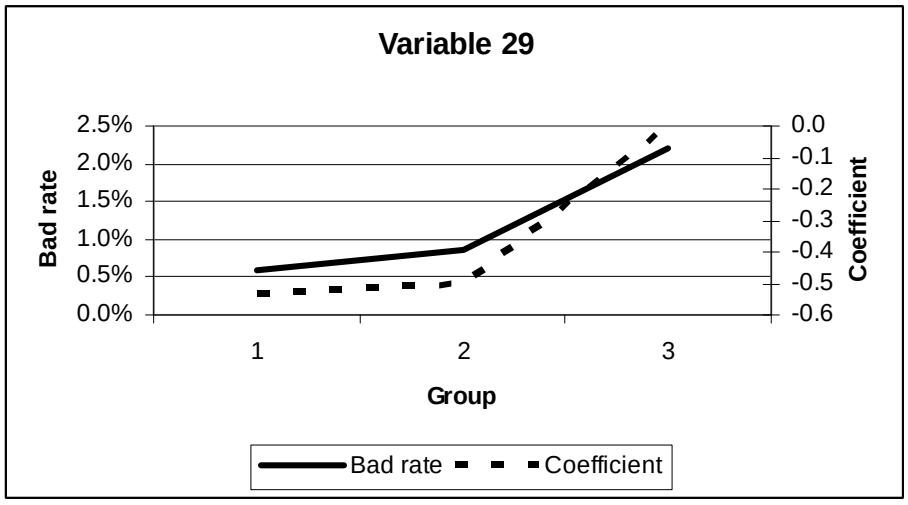
Var26

Variable 28:



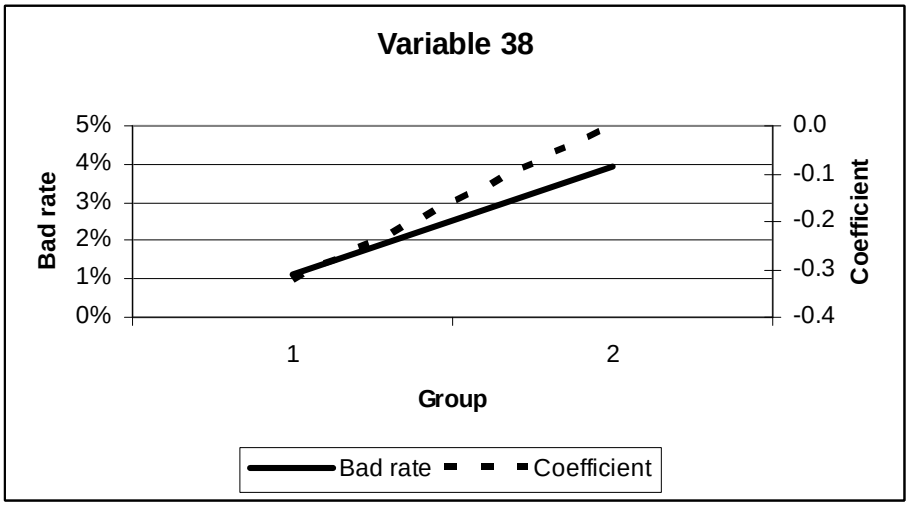
Var28

Variable 29:



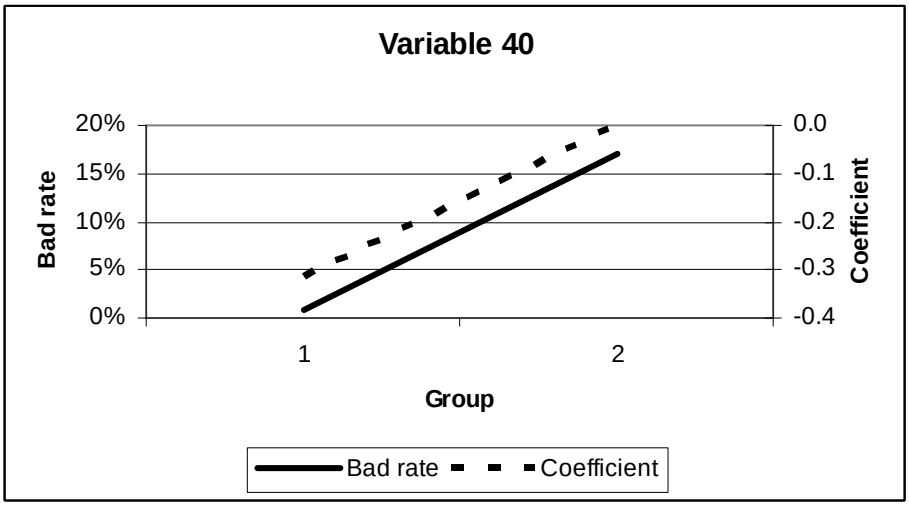
Var29

Variable 38:



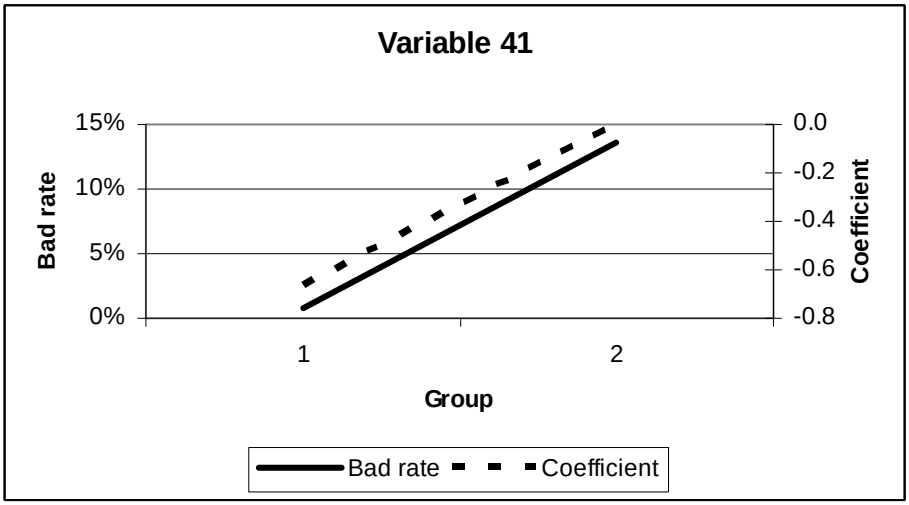
Var38

Variable 40:



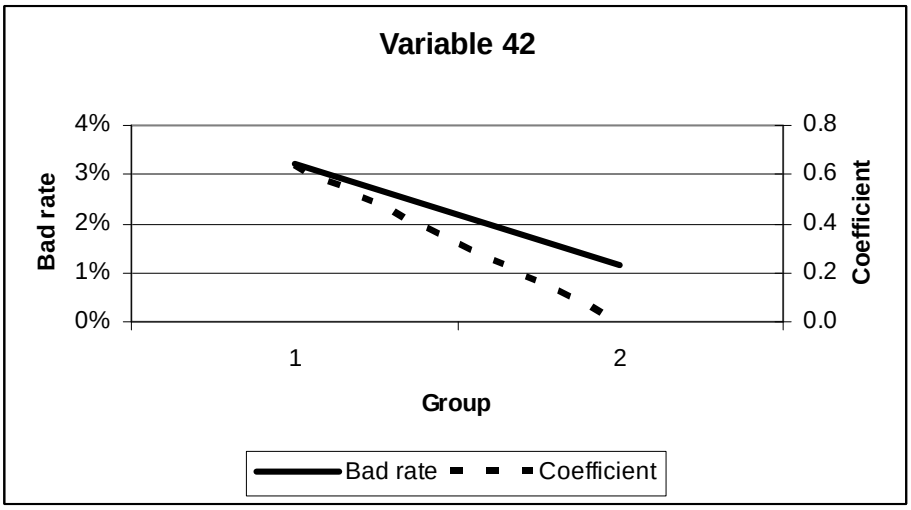
Var40

Variable 41:



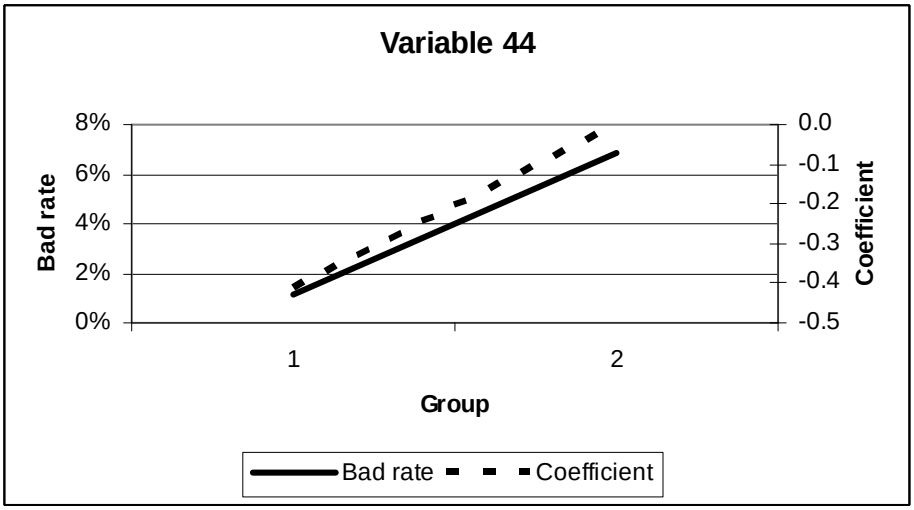
Var41

Variable 42:



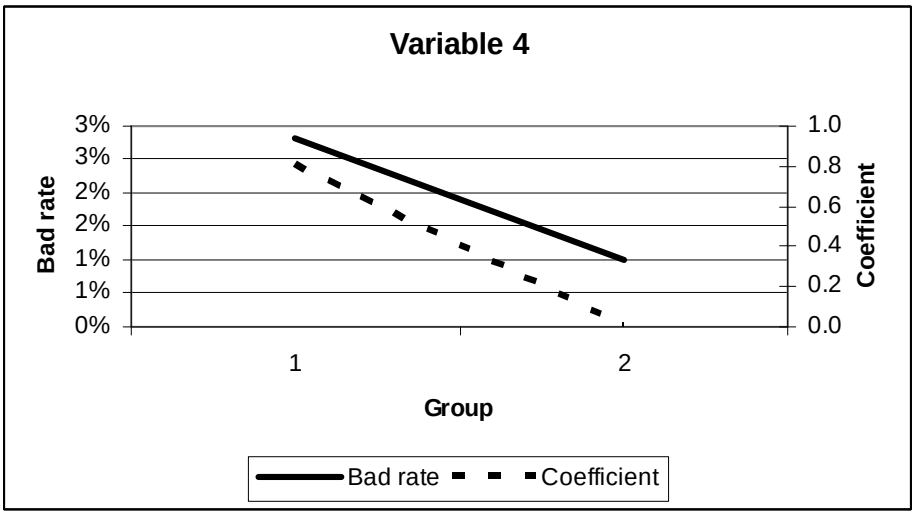
Var42

Variable 44:



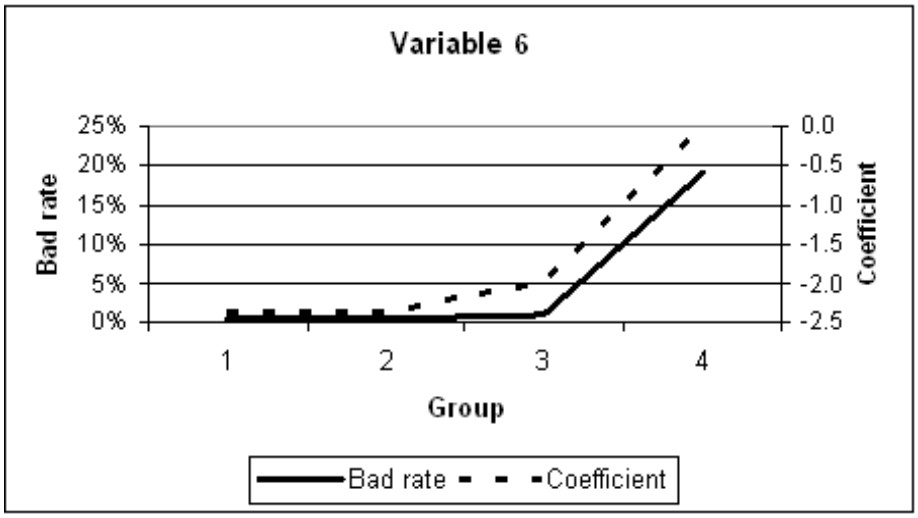
Var44

Variable 4:



Var4

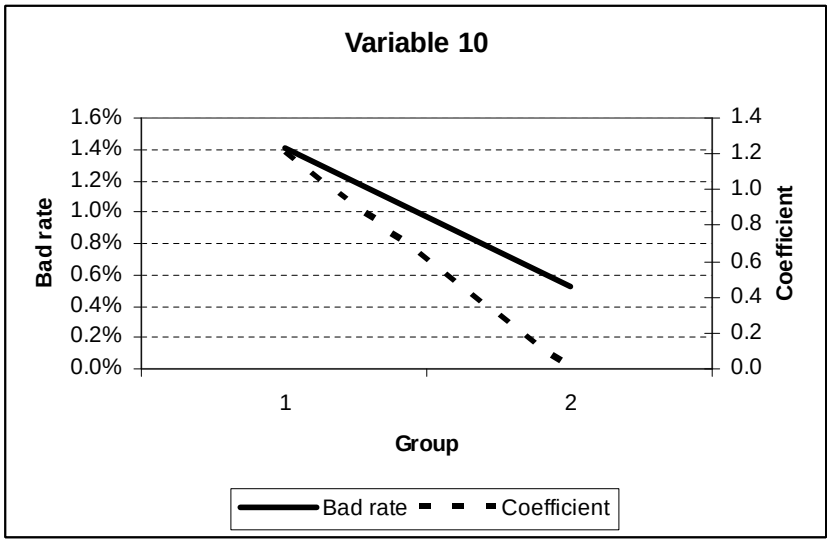
Variable 6:



Var6

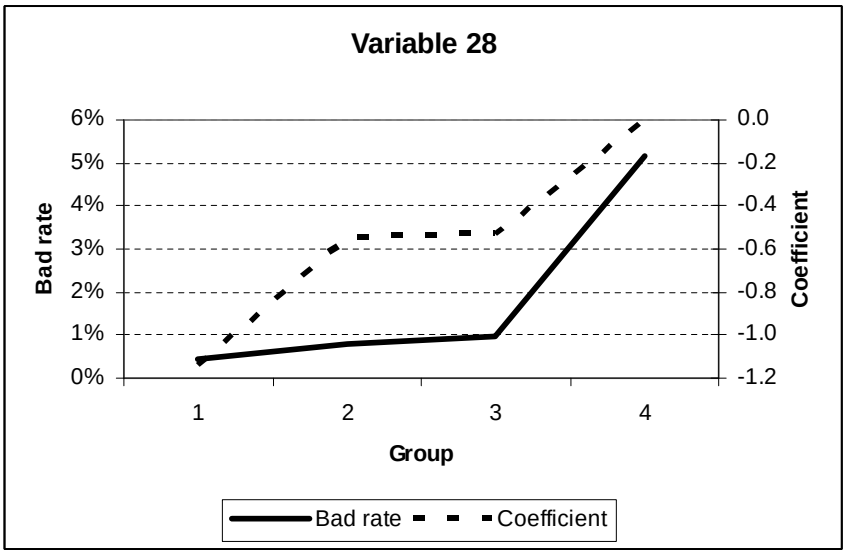
12.13 Appendix 6: Default rate sloping and model coefficients version 2

Variable 10:



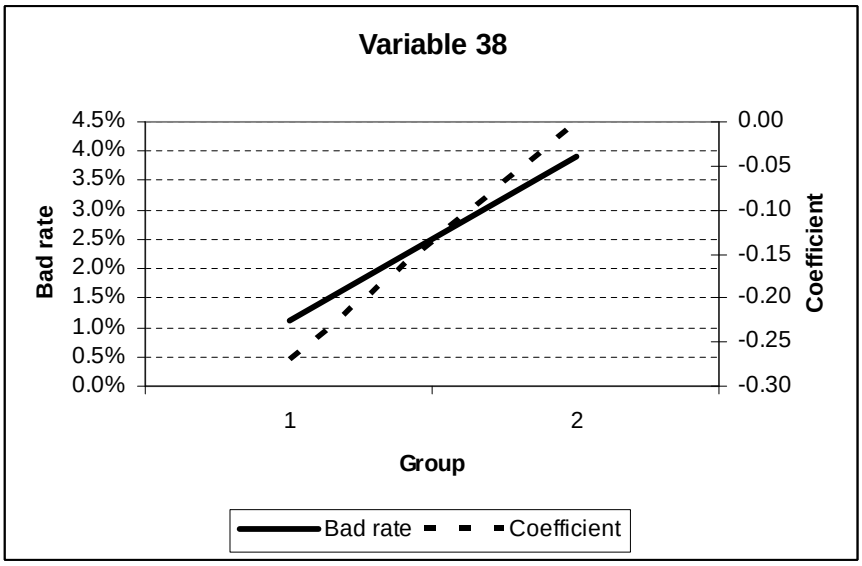
Var10

Variable 28:



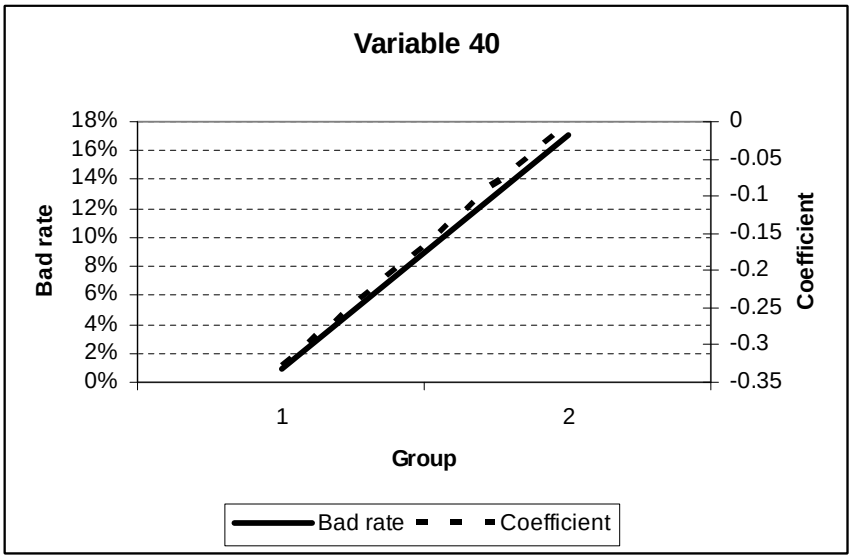
Var28

Variable 38:



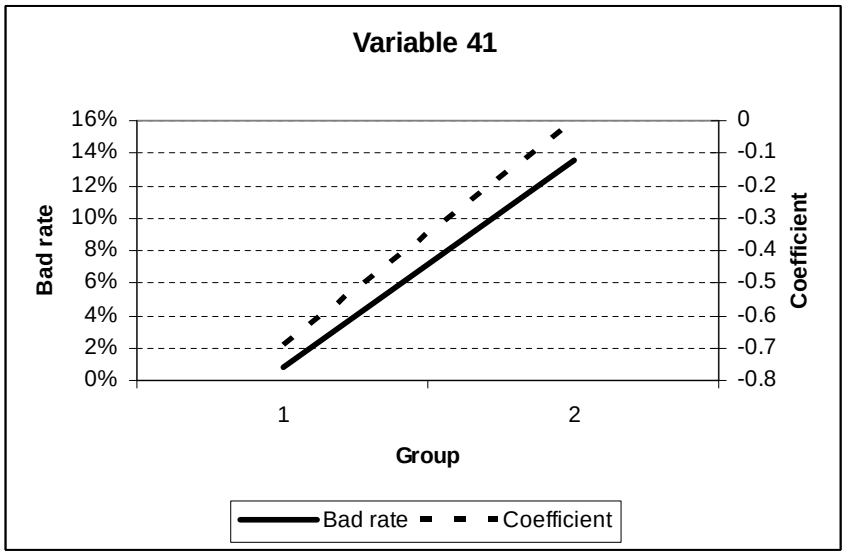
Var38

Variable 40:



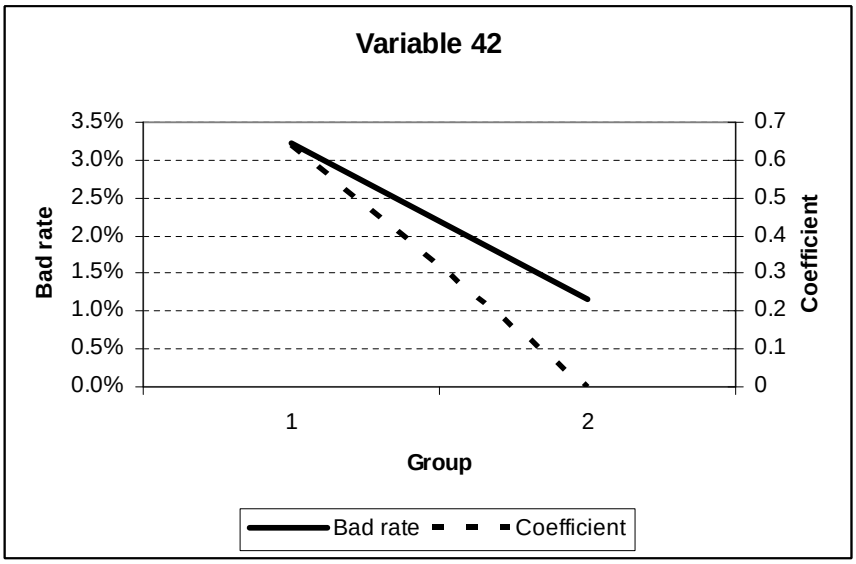
Var40

Variable 41:



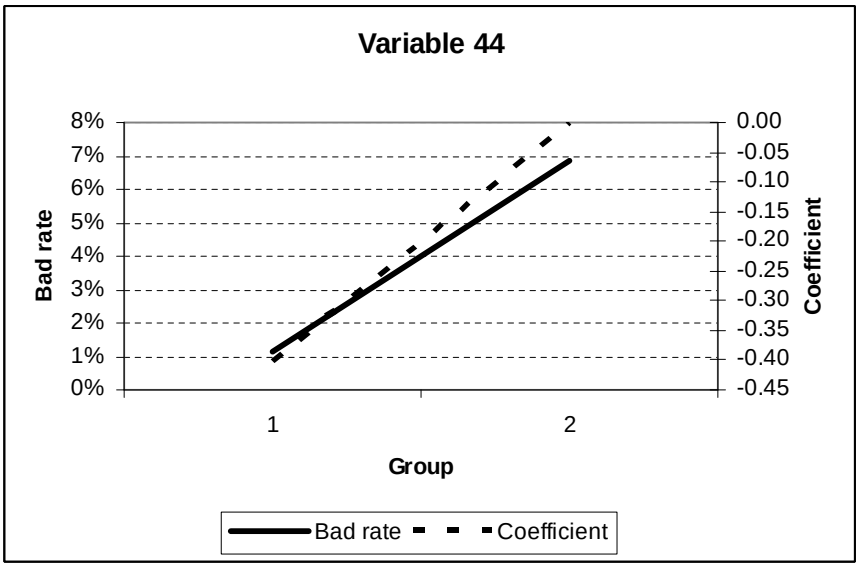
Var41

Variable 42:



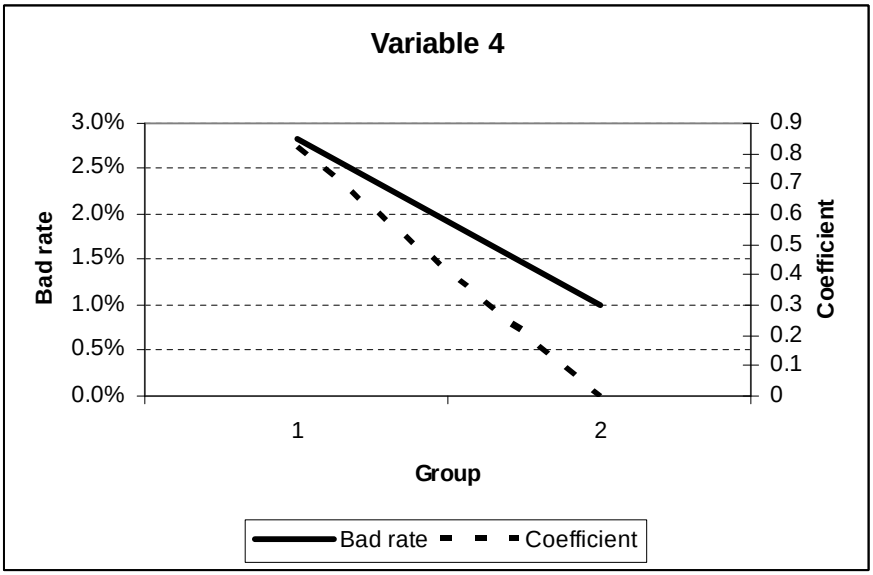
Var42

Variable 44:



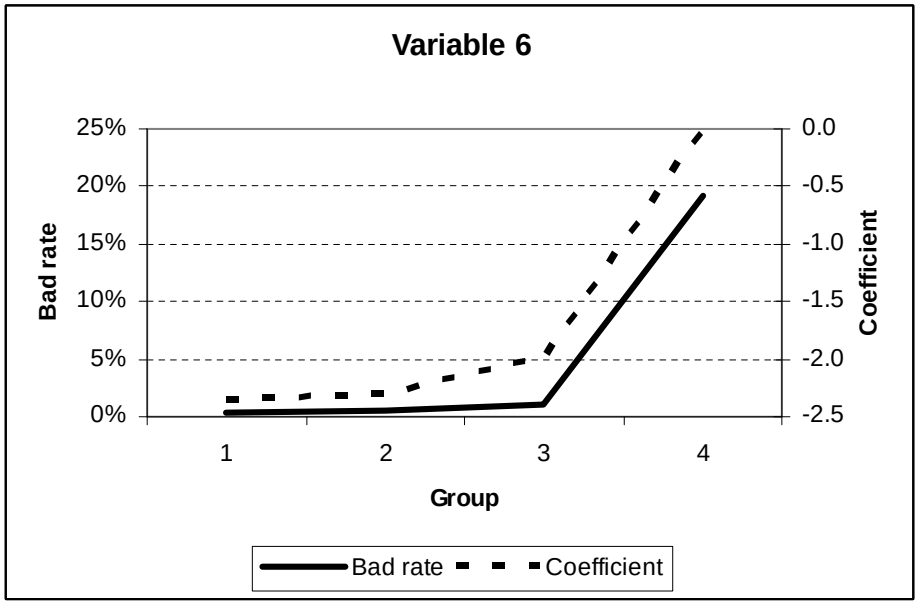
Var44

Variable 4:



Var4

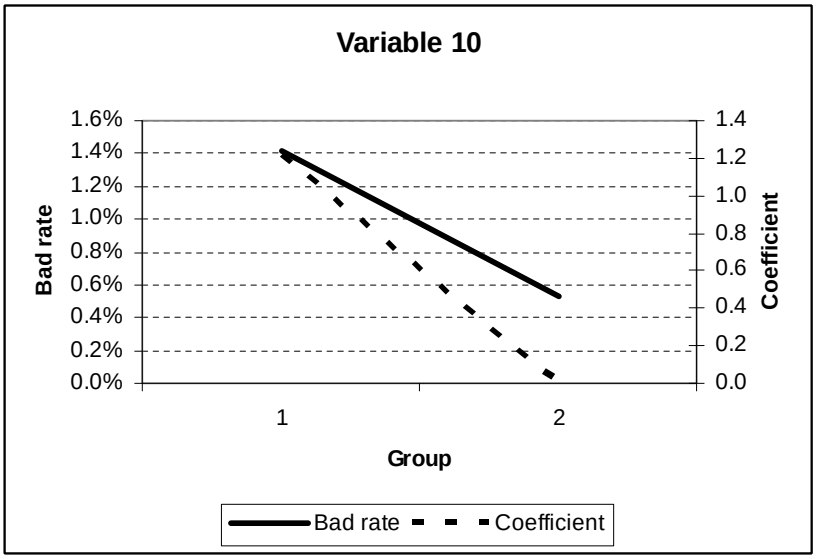
Variable 6:



Var6

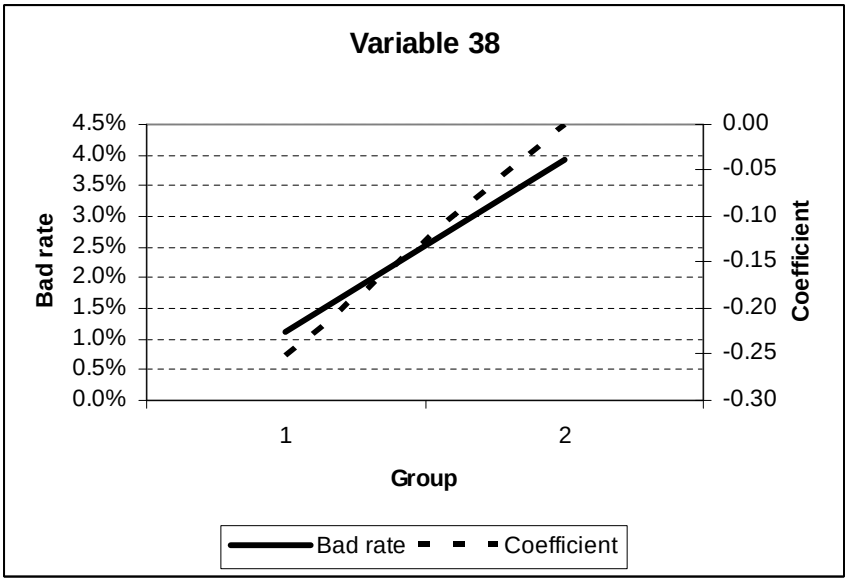
12.14 Appendix 7: Default rate sloping and model coefficients version 3

Variable 10:



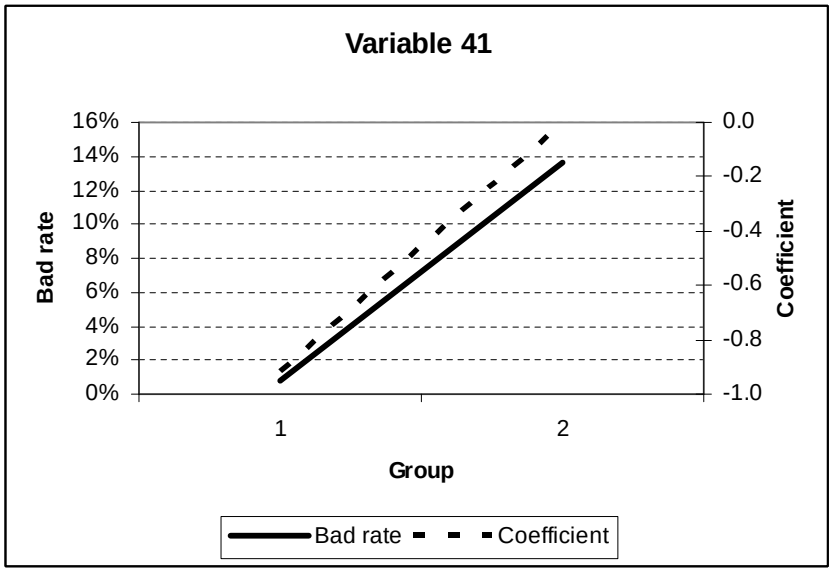
Var10

Variable 38:



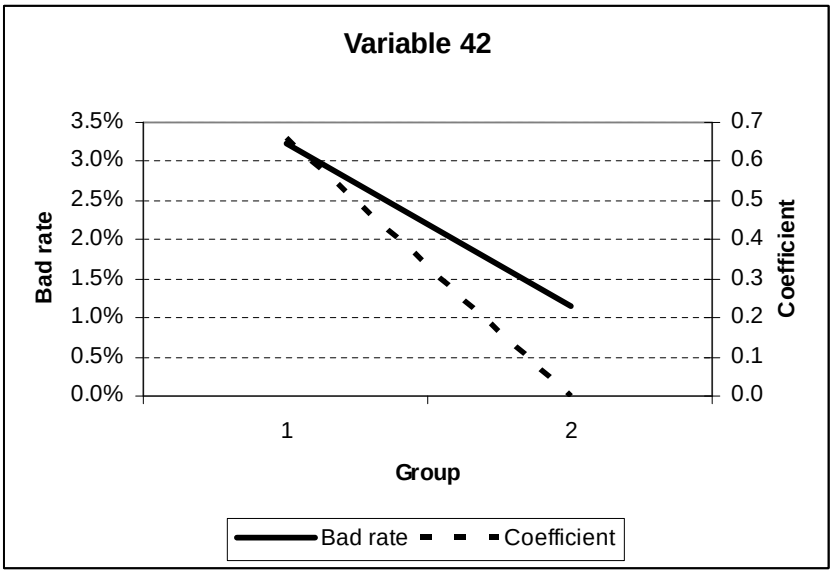
Var38

Variable 41:



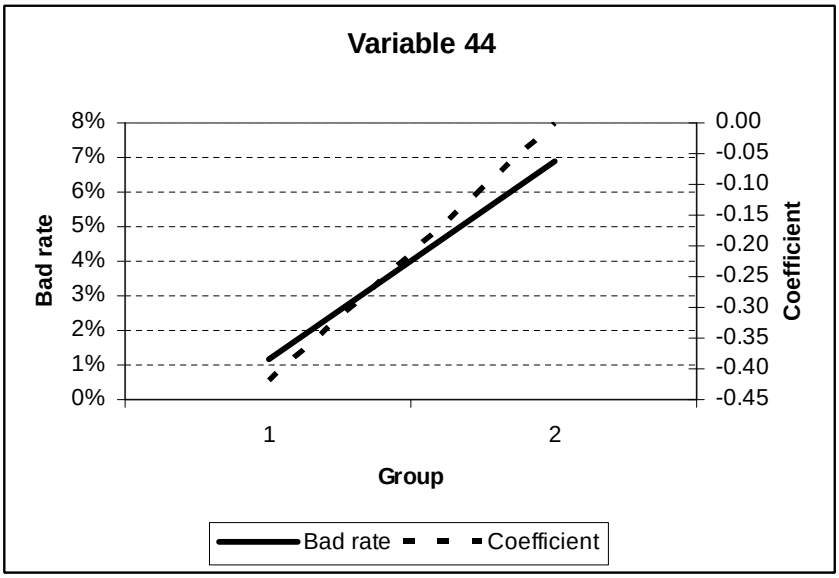
Var41

Variable 42:



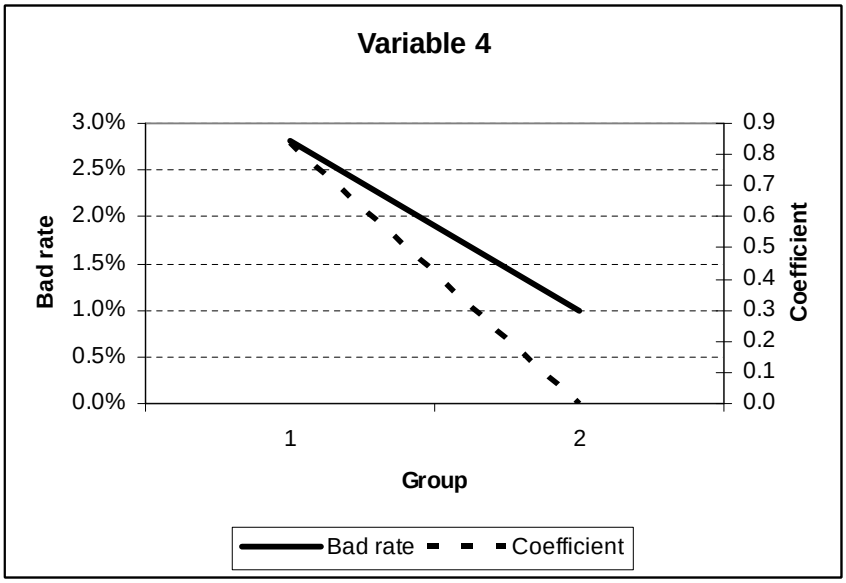
Var42

Variable 44:



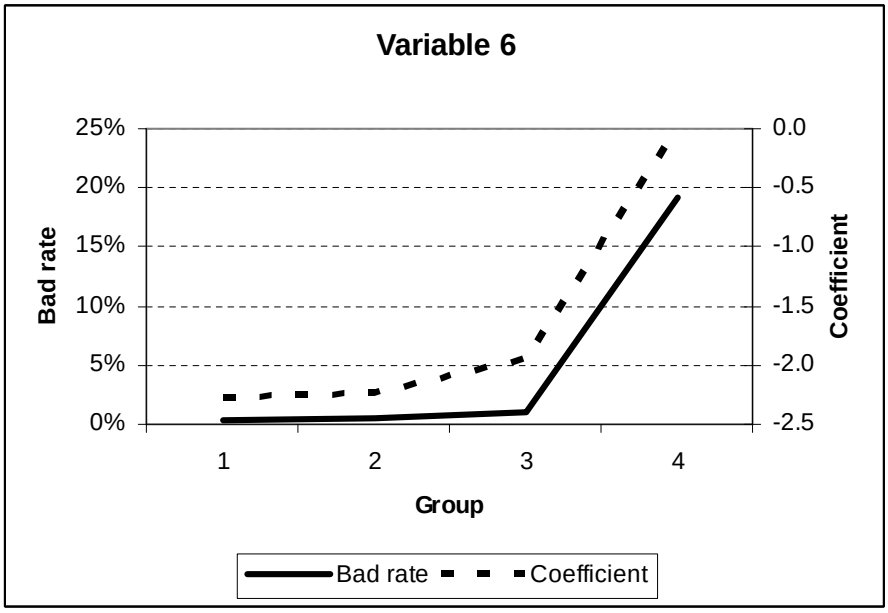
Var44

Variable 4:



Var4

Variable 6:



Var6

Chapter 13

Glossary

Bias: "Bias" as used in the field of statistics refers to directional error in an estimator. Statistical bias is error you cannot correct by repeating the experiment many times and averaging together the results.

Bonferroni Correction: The Bonferroni correction is a multiple comparison correction used when several dependent or independent statistical tests are performed simultaneously since, while a given α may be appropriate for each individual comparison, it is not for the set of all comparisons. In order to avoid a lot of spurious positives the alpha value needs to be lowered to account for the number of comparisons being performed. The Bonferroni correction sets the alpha value for the entire set of n comparisons equal to α by taking the alpha value for each comparison equal to α/n. Explicitly, given n tests T_i for hypotheses $H_i(1 \leq i \leq n\ )$ under the assumption H_0 that all hypotheses H_i are false, and if the individual test critical values are $\leq \alpha/n$, then the experiment-wide critical value is $\leq \alpha$. In equation form, if $P(T_i\ passes|H_0) \leq \alpha/n$ for $1 \leq i \leq n$, then $P(some\ T_i\ passes) \leq \alpha$ which follows from the Bonferroni inequalities.

Cholesky decomposition: If $\mathbf{X}$ is a positive definite matrix with row and column dimensions n, then $\mathbf{X}$ can be factored into an upper triangular matrix $\mathbf{R}$ (also of dimension n) such that: $\mathbf{X} = \mathbf{R}'\mathbf{R}$ where $\mathbf{R}'$ refers to the transpose of $\mathbf{R}$.

Coefficient of determination (R^2): Measure of the proportion of variation in the dependent variable about its mean that is explained by the independent variables. It varies between 0 and 1. If the regression model is properly applied and estimated, one can assume that the higher the value of R^2, the greater the explanatory power of the regression equation and the better the predictions.

Confidence limits: A statistical term for a pair of numbers that predict the range of values within which a particular parameter lies for a given level of confidence (probability).

Correlation coefficient(r): Indicates the strength of association between the dependent and independent variables. The sign (+ or -) indicates the direction of the relationship, The value range from -1 to 1, with 1 indicating a perfect positive relationship, 0 indicating no relationship and -1 indicating a perfect negative/reverse relationship.

Credit bureau: A Credit bureau is an organisation that keeps a record of a person's credit information. A credit record shows how that person manages his/her debts and is used by credit providers and moneylenders to decide if the person can afford to borrow money or pay back a new loan.

Credit policy: Guidelines that spell out how to decide which customers are sold on open account, the exact payment terms, the limits set on outstanding balances and how to deal with delinquent accounts.

Credit risk: The risk that the borrower may be unable or unwilling to honor his obligations under the terms of the contract for credit.

Customer attrition: Also known as customer churn, customer turnover, or customer defection is a business term used to describe loss of clients or customers.

Debt collection agency: A company hired by lenders to recover funds that are past due or accounts that are in default. A collection agency is hired after a company has made multiple attempts to collect what is owed to it. A collection agency usually earns a percentage of the funds or assets it recovers for the lender. Some collection agencies will actually purchase the debts from the lender, in which case the collection agency receives the full amount of whatever owed funds it manages to collect.

Delinquency: Failure to make a payment when due. Delinquency occurs when all or part of the borrower's monthly installment of principal and interest is unpaid after the due date.

Dependent variable: Variable being predicted or explained by the independent variables.

Euclidean distance: The straight line distance between two points. In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is $\sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)}$.

Factor analysis: Factor analysis is a correlational technique to determine meaningful clusters of shared variance. Factor analysis begins with a large number of variables and then tries to reduce the interrelationships amongst the variables to a few number of clusters or factors. Factor analysis finds relationships or natural connections where variables are maximally correlated with one another and minimally correlated with other variables and then groups the variables accordingly.

Homoscedacity and heteroscedacity: When the variance of the error terms (e) appears constant over a range of predictor data, the data are said to be homoscedastic. The assumption of equal variance of the population error, ε (estimated from e), is critical to the proper application of linear regression. When the error terms have increasing or modulating variance, the data are said to be heteroscedastic. The term means "differing variance" and comes from the Greek "hetero" ('different') and "skedasis" ('dispersion').

Independent variables: Variables selected as predictors and potential explanatory variables of the dependent variable.

Intercept: Value on the y-axis where the line defined by the regression equation. $y = \beta_0 + \beta_1 x$ crosses the axis. It is described by the constant term β_0 in the equation. In addition to its role in prediction, the intercept may or may not have a managerial interpretation. If the complete absence of the predictor variable (independent variable) has meaning, then the intercept represents that amount. For example, when estimating sales from the past advertising expenditures, the intercept represents the level of sales expected if advertising is eliminated. But, in many instances, the intercept only has predictive value because there might be no situation where all the predictor variables are absent.

Liquidity: The ability of an asset to be converted into cash quickly and without any price discount.

Mahalanobis distance: The Mahalanobis distance from a group of values with mean $\mathbf{m} = (\mu_1, \mu_2, ..., \mu_N)'$ and covariance matrix $\mathbf{\Sigma}$ for a multivariate vector $\mathbf{x} = (x_1, x_2, ..., x_N)'$ is defined as $D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \mathbf{m})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{m})}$. Mahalanobis distance (or "generalized squared interpoint distance" for its squared value) can also be defined as dissimilarity measure between two random vectors $\mathbf{x}$ and $\mathbf{y}$ of the same distribution with the covariance matrix $\mathbf{\Sigma}$: $d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}$. If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance.

Maximum likelihood: The method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data.

Monte Carlo simulation: An analytical technique for solving a problem by performing a large number of trail runs, called simulations, and inferring a solution from the collective results of the trial runs. Method for calculating the probability distribution of possible outcomes.

Multicollinearity: Condition that exists when independent variables are highly correlated with each other. In the presence of multicollinearity, the estimated Regression Coefficients may be unreliable. The presence of multicollinearity can be tested by investigating the correlation (r) between the independent variables.

Multivariate analysis: all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation.

Null plot: Plot of residuals vs. predicted values that exhibit a random pattern. A null plot is indicative of no identifiable violations of the assumptions underlying regression analysis.

Obligor: An individual or company that owes debt to another individual or company (the creditor), as a result of borrowing or issuing bonds, also called debtor.

Odds ratio: The odds ratio is a way of comparing whether the probability of a certain event is the same for two groups. An odds ratio of 1 implies that the event is equally likely in both groups. An odds ratio greater than one implies that the event is more likely in the first group. An odds ratio less than one implies that the event is less likely in the first group

Over-fitting: In statistics, over-fitting is fitting a statistical model that has too many parameters. An absurd and false model may fit perfectly if the model has enough complexity by comparison to the amount of data available. The model contains very specific random features of the data, that have no causal relation to the independent variable.

Partial regression plots: Graphical representation of the relationship between the dependent variable and a single independent variable. The scatter plot of points depicts the partial correlation between the two variables, with the effects of the other independent variables held constant.

Principal component analysis: Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

Regression coefficients: Numerical value of any parameter estimate directly associated with the independent variables. In multiple predictor models, the regression coefficients are partial because each takes into account not only the relationship between that variable and the independent variable, but also between the respective variables. The regression coefficients are not limited in range, as it is based on both the degree of association and the scale units of the predictor variable.

Residual (e or ε): Error in prediction the sample data is called the residual. Predictions will seldom be perfect. It is assumed that random error will occur, and the assumption is extended that this error is an estimate of the true random error in the population (ε), not just the error in prediction for the sample (e). Another assumption is made that the error in the population that is being estimated is distributed with a mean of zero and a constant variance.

Saturated model: A model that contains as many parameters as there are data points. This model contains all main effects and all possible interactions between factors. For categorical data, this model contains the same number of parameters as cells and results in a perfect fit for a data set. The (residual) deviance is a measure of the extent to which a particular model differs from the saturated model.

Standard error of the estimate: Measure of the variation in the predicted values that can be used to develop confidence intervals around any predicted value. This is similar to the standard deviation of a variable around the mean.

Statistically significant: Statistical significance is a mathematical tool used to determine whether the outcome of an experiment is the result of a relationship between specific factors or due to chance.

Sum of squared errors (SSE): The sum of the squared prediction errors (residuals) across all observations. It is used to denote the variance in the dependent variable that is not yet accounted for by the regression model.

Sum of squares regression: The sum of the squared differences between the mean and the predicted values of the dependent variable for all observations. This represents the amount of improvement when using independent variables in prediction vs. only using the mean.

Total sum of Squares (SST): Total amount of variation that exists to be explained by the independent variables. This "baseline" is calculated by summing the squared differences between the mean and the actual values for the dependent variable across all observations.

Chapter 14

References

Anderson, R. (2007). *The Credit Scoring Toolkit : Theory and Practice for Retail Credit Risk Management and Decision Automation*

Barman, R.B. *Estimation of Default Probability for Basel II on Credit Risk* (2005)

Basel Committee on Banking Supervision International Convergence of Capital Measurements and Capital Standards A revised Framework June 2004 (2004)

Bhatia, M. (2006).*Credit Risk Management and Basel II, An Implementation Guide*

Bowerman, B.L. and O'Connell ,R.T. (1990). *Linear Statistical Models: an applied Approach* Second Edition

Carling, K; Jacobson, T; Linde, J and Roszbach, K. (2002). *Capital Charges under Basel II: Corporate Credit Risk Modeling and the Macro Economy* Sveriges Riksbank Working Paper Series No. 142

Constitution of the Republic of South Africa No. 108 of 1996

Hair, J.F. ;Anderson, R.E. ; Tatham, R.L. and Black, W.C. (1995) *Multivariate Data Analysis with Readings* Fourth Edition

Hand, D.J. and Henley, W.E. (1997). *Statistical Classification Methods in Consumer Credit Scoring: a Review* J.R. Statist. Soc. 160, Part3, pp523-541

Hand, D.J.(2001) *Modeling Consumer Credit Risk*, IMA Journal of Management Mathematics,12,137-255

Hoare, R. (2004). *Using CHAID for classicification problems*

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression* Second Edition

Kendall, M.G. and Stuart, A. (1958). *The Advanced Theory of Statistics* Volume 1, Distribution Theory

Lee, K.I. and Koval, J.J. (1997). *Determination of the best significance level in forward stepwise logistic regression.* Communications in Statistics- Simulation and Computation, 26:2, 559-575

Lee, T. ;Duling, D. ;Lui, S. and Latour, D. (2008). *Two-stage variable clustering for large data sets* SAS Global Forum 2008

Lewis, R.J. (2000). *An Introduction to Classification and Regression Tree (CART) Analysis*

Matthews, G.B. and Crowther, N.A.S. (1995) *A maximum likelihood estimation procedure when modelling in terms of constraints.* S.A. Statist. J.,29, 29-51

Mays, E. (2003) *Credit Scoring for Risk Managers: The handbook for lenders*

Mohr, P. (1998). *Economic Indicators*

National Credit Act (2005) Government Gazette, 15 March 2006, Act No 34, 2005

Ong, M.K. (2002). *Credit Ratings, Methodologies, Rationale and Default Risk*

Osius G, Rojek D (1992). *Normal Goodness-of-Fit Tests for Multinomial Models With Large Degrees of Freedom.* JASA, 87:1145-1152, 1992.

Steyn, H.S.; Crowther, N.A.S.; Joubert, H.M. ; Barnardt, M. and Raath, E.L. (1990). *Statistiese modelle vir Inkomstestrukture* Verslag nr.1 in die reeks oor inkomstestrukture, Verslag WS-47

Vojtek, M. and Kočenda, E. (2005). *Credit Scoring Methods*

The following websites were also used:

http://ats.ucla.edu/stat/sas/output/sas_ologit_output

http://ats.ucla.edu/stat/sas/seminars/sas_logistic/logistic1

http://dictionary.babylon.com/Saturated%20model

http://en.wikipedia.org/wiki/Customer_attrition

http://faculty.chass.ncsu.edu/garson/PA765/cluster

http://financial-dictionary.thefreedictionary.com/Monte+Carlo+simulation

http://knowledgerush.com/kr/encyclopedia/Overfitting/

http://lesswrong.com/lw/ha/statistical_bias/

http://luna.cas.usf.edu/~mbrannic/files/regression/Logistic

http://online.sfsu.edu/efc/classes/biol710/logistic/logisticreg

http://wilderdom.com/research/factoranalysis.html

http://wordnetweb.princeton.edu/perl/webwn?s=delinquency

http://www.answers.com/topic/collection-agency

http://www.answers.com/topic/multicollinearity

http://www.blacksash.org.za/index.php?option=com_content&view=
article&id=967&Itemid=182

http://www.childrens-mercy.org/stats/definitions/or.htm

http://www.entrepreneur.com/encyclopedia/term/82124.html

http://www.everythingbio.com/glos/definition.php?word=confidence+limits

http://www.fon.hum.uva.nl/praat/manual/Principal_component_analysis.html

http://www.investorwords.com

http://www.investorwords.com/3375/obligor.html

http://www.itl.nist.gov/div897/sqg/dads/HTML/euclidndstnc.html

http://www.itl.nist.gov/div898/software/dataplot/refman2/ch4/cholesky.pdf

http://www.statsoft.com/textbook/stchaid

http://www.statsoft.com/textbook/stcluan