

A Practical Approach to Validating a PD Model

Lydian Medema Ruud H. Koning Robert Lensink

June 6, 2007

Corresponding author: Lydian Medema. *Adress:* University of Groningen, P.O. Box 800, 9700 AV, Groningen, The Netherlands. *Phone:* +31503633811, *Telefax:* +31503633850. *E-mail addresses:* l.medema@rug.nl, r.h.koning@rug.nl, b.w.lensink@rug.nl.

Abstract

The capital adequacy framework Basel II aims to promote the adoption of stronger risk management practices by the banking industry. The implementation makes the validation of credit risk models more important. Lenders therefore need a validation methodology to convince their supervisors that their credit scoring models are performing well. In this paper we take up the challenge to propose and implement a simple validation methodology that can be used by banks to validate their credit risk modelling exercise. We will contextualise the proposed methodology by applying it to a default model of mortgage loans of a commercial bank in the Netherlands.

1 Introduction

Since June 1999 the Basel Committee on Banking Supervision has published several proposals for revising the existing Basel I capital adequacy framework. The revised framework, known as Basel II (Basel Committee on Banking Supervision (2006)), is based on three pillars: minimum capital requirements, supervisory review, and market discipline. It aims to promote the adoption of stronger risk management practices by the banking industry. One of the main differences between the Basel I and Basel II frameworks is that banks' possibilities to use internal risk assessments as inputs to capital requirements are considerably enlarged. Duffie and Singleton (2003) categorize the risk faced by banks into: market risk, credit risk, liquidity risk, operational risk and systemic risk. In this paper we focus on credit risk. Within the framework of Basel II, banks can opt for different approaches to assess their credit risk. More specifically, banks may choose between a standardized approach where fixed risk weights are used and no differentiation is made on the basis of actual risk, and the internal ratings based approach (IRB), for which risk weights are based on the actual risk of transactions and banks can use own estimates of probability of default (PD).

The implementation of Basel II raises many technical questions regarding the development and calibration of credit risk models. It also makes the validation of credit risk models much more important, e.g. since the framework requires strong efforts by banks to assess their capital adequacy and by supervisors to review such assessments. Bank regulators will pay more and more attention to testing model validation processes in order to examine the accuracy of banks' credit scoring models. Lenders therefore need a solid and generally accepted validation methodology to convince their supervisors that their credit scoring models are performing well. This especially holds for banks that opt for the IRB approach of capital adequacy. A citation from the Basel Retail Guidance clarifies the utmost importance of validation, "A bank must establish policies for all aspects of validation. A bank must comprehensively validate risk segmentation and quantification at least annually, document the results, and reports its findings to senior management" (Internal Ratings-Based Systems for Retail Credit Risk for Regulatory Credit; 69 Federal Register, pp. 62, 748 ff, October 27, 2004).

Nowadays banks pay a lot of attention to the validation process, but still a generally accepted validation methodology does not exist. Validation requires e.g. quantifiable expectations about the impact of changing economic conditions. However, these dynamic effects are often not taken into account in the model constructing process. Moreover, the model construction is in many instances hampered by missing observations and because banks have not historically documented all important indicators of creditworthiness comprehensively. Facing these and other practical problems, the question then arises as to how validation should take place. Supervisors, like the Dutch Central Bank (DNB), give some guidance on how to validate credit risk models (De Nederlandsche Bank N.V. (2005)). However this guidance only gives an introduction to model validation.

In this paper we take up the challenge to propose and implement a simple validation methodology that can be used by banks to validate their credit risk modelling exercise. The methodology we propose is supposed to be general enough to be useful for a diversity of banks, and aims to be especially helpful for those banks that are planning to implement the IRB capital adequacy method. Our validation methodology is more profound than the guidance of DNB (De Nederlandsche Bank N.V. (2005)).

Validation is obviously not only a statistical exercise. Managerial judgement and a qualitative analysis of the model are also highly important. However, the initial validation will primarily be technical and model based. Moreover, statistical validation is needed to obtain scientific rigor and a common yardstick for the validation exercise. For these reasons, this article will focus on a quantitative validation technique and propose a statistical validation methodology. In addition, this article will contextualise the proposed methodology by applying it to a default model of mortgage loans of the Friesland Bank, a commercial bank in the Netherlands.

The remainder of the paper is organised as follows. Section 2 pro-

vides some background information on the Basel II accord and discusses several models that can be used for modelling credit risk. In section 3 our proposed validation methodology will be set out. We base our methodology on Harrell (2001) who validates a logit model with an application in the medical science. We will explain several statistical techniques that are available to validate models, and apply these techniques to validate the default model of mortgage loans of Friesland Bank in section 4. Section 5 surveys the article and provides some areas for further research.

2 Credit Risk

2.1 The Basel Capital Accord

The Basel Committee on Banking Supervision (Basel Committee) introduced the Capital Accord of 1988, also referred to as Basel I. Basel I aims to provide methods by which financial institutions can determine their minimum capital requirements. In the accord a capital measurement system is introduced according to which banks have to divide their activa into four classes: OECD governments, loans to OECD banks, mortgages and all other loans.

A risk weight has to be assigned to the total exposure in each class. Basel I sets the weights to the four classes equal to 0%, 20%, 50% and 100% respectively. The product of the total exposure and risk weight in each class is called the risk-weighted activa. Basel I sets a minimum ratio of capital to the risk-weighted activa of 8%.

In 1999 the Basel Committee proposed a new accord to replace the existing Basel I accord. This new accord, known as Basel II, is intended to improve the way capital requirements reflect the underlying risks. There are three approaches distinguished in Basel II: the Standardized Approach, the Foundation IRB approach and the Advanced IRB approach.

The Standardised Approach uses the same concepts contained in Basel I (see Basel Committee on Banking Supervision (2001b)). According to the Standardised Approach banks have to divide their credit exposures into classes based on observable characteristics of the exposures (for example whether it is a corporate loan or a mortgage loan). For all classes a fixed risk weight is determined by the supervisor. The minimum ratio of capital to the total weighted exposure is 8%.

Under IRB approaches, four inputs are needed for credit risk determination and capital calculations: the probability of default, an estimate of the loss given default, the exposure at default and the remaining maturity of the loan (see Basel Committee on Banking Supervision (2001a)). IRB approaches permits a bank to use internal ratings as primary inputs to capital calculations. This will lead to more diverse risk weights and a greater risk sensitivity. The banks are not allowed to determine all the elements needed to calculate their own capital requirements. The Basel Committee specified formulas which has to be used in combination with information provided by the banks

to determine the risk weights.

In the Foundation IRB Approach a bank determines the probability of default for a particular borrower and the supervisor supplies the other inputs, like the loss given default, the exposure at default and the maturity. The Advanced IRB Approach permits banks to estimate all four inputs needed for credit risk determination and capital calculations: the probability of default, the loss given default, the exposure at default and the maturity.

For a bank to be permitted to use an IRB approach, they must meet a set of minimum requirements. An overview of the most important requirements of Basel II is given in appendix A. Since in this paper we focus on PD models, we only list requirements which are applicable to PD models. One of the requirements is that banks have to estimate the probability of default for each loan. Typically, the portfolio on loans can consist of several classes of loans: loans to retail, mortgages, loans to small business and loans to large business. Banks are allowed to estimate separate PD models for each class of loans (section 395 of Basel II). According to the Basel Accord a default takes place when the borrower is past due more than 90 days on any credit obligation.

2.2 Notation

Before we describe models which can be used to model default, we introduce some notation to be used throughout the paper.

i , index of clients, $i = 1, \dots, n_t$.

n_t , number of observations in period t ,

t , time index, $t = 1, \dots, T$,

$N = \sum_{t=1}^T n_t$, total number of observations.

Note that n_t is not constant over time since not all clients are measured in each time period. Some contracts start in a period later than period 1 and some contracts mature before period T . So in practice a data set will be an unbalanced panel data set.

X_{it} , $(k+1)$ -vector of explanatory variables of client i at time t ,

$X_{it} = (1, x_{it,1}, \dots, x_{it,k})$.

X_{it} includes an intercept, the explanatory variables may be time varying (for example age of the client) or client specific (for example sex of the client).

$Y_{it} = \begin{cases} 1, & \text{if client } i \text{ defaults between time } t \text{ and } t+1, \\ 0, & \text{if not,} \end{cases}$

$\Pr(Y_{it} = y)$, probability that Y_{it} equals y , $y = 0, 1$,

$p_{it} = \Pr(Y_{it} = 1)$, probability that Y_{it} equals 1,

β , $(k+1)$ -parameter vector,

g , index of borrower grade g , $g = 1, \dots, G$,

n_{1g} , number of loans in borrower grade g that defaulted,

n_{0g} , number of loans in borrower grade g that did not default,
 $n_g = n_{1g} + n_{0g}$, number of loans in borrower grade g ,
 P_g , default probability in borrower grade g .

2.3 Default Models

Two main types of statistical models for modelling defaults are duration models and classification models. In duration models, one focusses on the time to default. Usually, this is done through modelling the hazard function: what is the probability of default in a short time interval starting at t , given that default has not occurred until t . The advantage of a duration model is that it provides instantaneous information. At each point in time, the time to default can be determined through the duration model. However, in the practice of defaults the data sets are often too limited to estimate a duration model. To estimate a duration model observations on the time of default are necessary. The data set we have at hand is censored in the sense that of the total number of observations only a small part defaulted on their contract. This censoring complicates the estimation of the model (Kalbfleisch and Prentice (1980)). A second problem is the problem of omitted variables or unobserved heterogeneity. Omitted variables can occur in two ways, conditional on the response variable default, the omitted variables can be either dependent or independent on the observed explanatory variables. Both cases omitted variables will cause problems in duration models (Cameron and Trivedi (2005)). Omitted variables will cause unobserved heterogeneity and with duration models this results in a serious specification error (Kalbfleisch and Prentice (1980)). Another disadvantage is that a duration model does not provide the probability of default in the next period directly. The estimation of default probabilities is a requirement for banks who use an IRB approach.

The other main approach in modelling the probability of default is through classification models (an excellent overview is given in Hastie, Tibshirani, and Friedman (2001)). The most popular models in this category are discriminant analysis and probability models (Duffie and Singleton (2003)). Discriminant analysis assumes that the overall population of borrowers consists of two subpopulations, a group of defaulters and a group of nondefaulters. Each borrower is assumed to be a draw from one of these populations and the bank wants to determine which. Based on the borrower characteristics the bank determines to which population the borrower belongs. Discriminant analysis assumes that the independent variables are each normally distributed and the joint distribution of the variables is assumed to be multivariate normal. In practice this assumption of normality is often violated. Another disadvantage of discriminant analysis is that it results in the subpopulation each borrower belongs to. As said before, Basel II explicitly requires banks to determine the probability of default when an IRB approach is used for capital calculations. There is no direct and obvious method to determine the default probabilities based on discriminant analysis.

Models that result directly in probabilities are probability models. In a probability model the probability of default is modelled as a function of the characteristics of the borrower. Let the true model be

$$\Pr(Y_{it} = 1) = G(X_{it}; \beta), \quad (1)$$

where β are unknown parameters to be estimated. Examples are the logit model

$$G(X_{it}; \beta) = \Lambda(\beta' X_{it}) = \frac{1}{1 + \exp(-\beta' X_{it})},$$

and the probit model

$$G(X_{it}; \beta) = \Phi(\beta' X_{it}),$$

with $\Phi(\cdot)$ the standard normal distribution function. In practice the logit model is often assumed. The assumption of a logit model is not restrictive. Equation (1) can be rewritten as

$$\Pr(Y_{it} = 1) = G(X_{it}; \beta) = \Lambda(\Lambda^{-1}(G(X_{it}; \beta))),$$

because $\Lambda(\cdot)$ is an invertible function. Therefore, the linear term in the logit model can be interpreted as a first-order Taylor expansion of $\Lambda^{-1}(G(X_{it}; \beta))$. Whether or not this approximation is precise enough, can be examined by adding non-linear terms and interactions to the index of the logit model. Note that the approximation is exact if the true model is a logit model. Of course, this argument can be applied to other choices of $G(\cdot)$ as well. In any case, the assumption of a logit model is not restrictive, as long as one allows for enough flexibility in the systematic part of the model. One of the advantages of the logit is that the parameters can be easily estimated using the maximum likelihood method. Of course, also with logit models the problem of omitted variables might occur. However, in contrast with duration models, omitted variables will not cause biased estimates if these variables are independent of the observed explanatory variables (proven by Lee (1982)).

As said before, banks are allowed to estimate PD models for each loan class separately. Moreover, banks may estimate hybrid models for a specific class. A hybrid model is a combination of two (or more) models, this type of modelling is also known as mixed models. One possibility that is applied in practice is the combination of a statistical model (for example a logit model) and a so-called expert model. An expert model is a model which is based on knowledge of an expert as opposed to a statistical model which is based on historical data. An expert can have information on the loans which is not available in the data set. Based on this information a minimum PD can be set for a particular group of loans. So banks do not have to rely on the results of statistical models completely. In fact, the outcome of a model may be overruled based on expert judgements. However, the bank must have clear guidelines on how and to what extent overruling can be used and whose responsible for it (section 417 and 428 of Basel II).

The models described above all result in a continuous outcome of the probability of default. Or, stated differently, one specific probability of default for each loan. In practice banks divide the loans into

borrower grades or risk buckets. At minimum banks must have seven borrower grades for non-defaulters and one grade for defaulters (section 404 of Basel II). As said in section 2.1 banks are allowed to estimate separate models for each class of loans (loans to retail, mortgages, loans to small business and loans to large business). Based on these separate PD models borrower grades for each class of loans have to be determined. There are two possible ways to determine the borrower grades. The first is to consider each class separately and determine borrower grades such that the fit in each class is best. However, it is very likely this will result in different borrower grades for each loan class and hence comparison of borrower grades for loans of different classes will be impossible. Figure 1 shows an example of different borrower grades for two loan classes. In the figure two classes of loans are considered, A and B . For both classes separate PD models are estimated and independently of each other the loans are divided into 8 risk buckets. Now compare a loan of class A with a loan of class B which both fall into risk bucket 6. In this example the PD of the loan in class A is higher than the PD in class B . This example illustrates that comparison of PDs is not possible when the risk buckets are determined for each class separately. The second way to determine risk buckets is to require in advance that the risk buckets are the same for each class of loans. In this case two loans in a specific risk bucket will have the same PD.

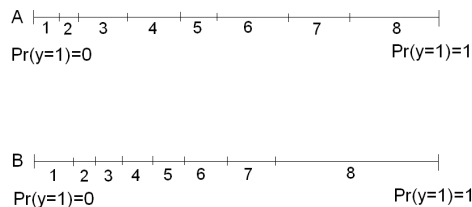


Figure 1: Example of borrower grades

3 Model Validation

3.1 General Ideas

The IRB approaches of Basel II requires banks to model the risk associated with their portfolios. Banks have all kinds of information

available on their portfolios, for example in computer dataware houses, but also in the form of documents. It is required to use all relevant information to determine the risk of the portfolio (section 411 Basel II). All relevant information available in different sources within the bank is merged into a data set. Often this data set is not suitable for statistical analysis. The next step is to use this data set to form a final data set which can be used for the calculations. This data set will be the basis for the statistical model. Finally, based on this statistical model, banks determine the risk associated with the portfolio. Once a credit risk model is implemented in the risk management of the bank this process can be repeated on a regular basis (for example once per year). The process described above is schematically summarized in figure 2.

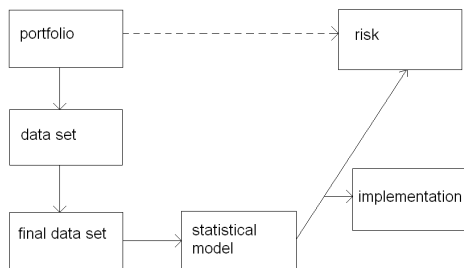


Figure 2: Determine Risk of Portfolio

Basel II requires the validation of this process (section 500): “Banks must have a robust system in place to validate the accuracy and consistency of rating systems, processes, and estimation of all relevant risk components.” The requirements a PD model must meet are set out in Basel II. Appendix A lists the most important requirements for the IRB approach. Validating a PD model means to verify to what extent the model meets the minimum requirements of Basel II. In order to do this, we distinguish three forms of validation: theoretical validity, data validity and statistical validity. This classification is also made by Gass and Thompson (1980). The methodology we develop in this section focusses mainly on probability models. The reasons for focussing on probability models are set out in section 2.3. We specifically focus on logit models, since in our application we have the task to validate a PD model of Friesland Bank, which uses a logit specification to estimate default probabilities. This section will first explain the three forms of validation briefly and next statistical validation will be discussed more extensively.

3.1.1 Theoretical validation

Theoretical validation requires the review of the theories and assumptions underlying the proposed model. This corresponds with section 402 of Basel II where a detailed outline of the theory and assumptions underlying the model is required.

Theories associated with PD models can be thought of as economic theories about the important risk drivers of default occurrence. If an important risk driver is missing the bank has to be conservative with the final estimates (requirement 411 of Basel II).

Section 2.3 discussed the types of models which can be used to model default. Underlying each model there are several assumptions. Reviewing these assumptions is part of theoretical validation. The use of the logit model to model PD assumes the observations to be independent. However, the data available for model estimation often contains observations at several points in time. Consequently, most mortgages are included in the data set more than once. So clearly the assumption of independence is violated.

3.1.2 Data validation

Data validation is about the data underlying the model. The data must be validated (section 417 of Basel II) and banks must show the data used are representative for the underlying population. To validate the data used to develop the model we distinguish three parts of data validation: representative data, appropriateness of the variables and completeness of the data set. In the following these three parts of data validation will be discussed.

Representative Data

In general, banks have two options regarding the data used to estimate the model. The first option is to use internal data and the second is to use external data. Basel II allows the use of external data (sections 448 and 463). The use of external data requires banks to demonstrate the data are representative for the underlying population of the bank. When the bank uses internal data on the complete portfolio the data are clearly representative. In practice, data sets on a complete portfolio can be too large to estimate a model, in this case a subset can be used instead. A subset has to be taken before doing any analysis and it can be obtained by taking a random draw of the complete data. The sampling procedure has to be reviewed to determine whether the sample is representative of the underlying population.

Appropriateness of the Variables

At a minimum borrower characteristics, transaction risk characteristics and delinquency of exposure has to be considered as explanatory variables in a PD model (section 402 of Basel II). Examples of borrower characteristics are age, income, marital status, occupation etcet. Transaction risk characteristics are for example mortgage type, loan to value, payment history etcet. Several problem arise with the variables.

The values of a variable can change over time. For instance, the variable income is very likely to change over time. The data set typically contains the income of the borrower at the moment the contract is made. However, it is reasonable to state that in determining the PD future income is important instead of the income at the moment the contract is made.

Completeness of the Data Set

Basel II requires the length of the underlying historical observation period to be at least five years (Basel II section 463). In practice it might be that banks have information on less than five periods. This means the data set is incomplete. Of course, this problem of incomplete data will be solved over time as more information becomes available. When the underlying observation period is less than five years banks are allowed to use external data to estimate the model. Where external data is used the bank must add a margin of conservatism (Basel II sections 451 and 462).

Incomplete data also occur in another way. Often information is missing for some variables for a number of observations. Missing data means less available information and consequently the results have to be interpreted conservatively (section 411 of Basel II). Conservatism may imply that the PD outcome of the model is considered as a lower bound. The final estimate of the PD can be set somewhat higher than this lower bound. From a statistical point of view missing data are a problem since all standard statistical methods require complete data sets. Several statistical methods exist to handle missing data. The method to be used for analyzing missing data depends on the missing-data mechanism. The missing-data mechanism is about the reasons why data are missing. To formally define the mechanisms we introduce some notation (we use the definition as given by Little and Rubin (2002)).

Let (Y_i, X_i) be the complete data vector for client i , $i = 1, \dots, n$. To keep notation simple, we omit the time index. Suppose the response variable Y_i is observed for each i and let Y be the n -vector of responses. Suppose some of the explanatory variables are missing for some clients and partition the vector X_i as $X_i = (X_{i,obs}, X_{i,mis})$, where $X_{i,obs}$ is the observed part and $X_{i,mis}$ is the missing part of X_i . Note that this partitioning can differ across clients. The missing-data indicator variable M_{ij} can take on two values, $M_{ij} = 1$ if x_{ij} is observed and $M_{ij} = 0$ if x_{ij} is missing, $i = 1, \dots, n$ and $j = 1, \dots, k$, $M_{i0} = 1$ for all i . Let M be the $(n \times (k+1))$ missing-data indicator matrix. The first column of M corresponds to the missingness in first column of X which represents the intercept of the model, so clearly $M_{i0} = 1$. The missing-data mechanism can be viewed as the conditional distribution of M given the data (Y, X) . Denote the conditional distribution by $f(M|Y, X, \phi)$, where ϕ is the unknown vector of parameters. If the missingness does not depend on the data, missing or observed, the missing-data mechanism is called missing completely at random (MCAR). So the missing-data mechanism is MCAR if $f(M|Y, X, \phi) = f(M|\phi)$. If the missing-data mechanism depends only on the observed data and not on the miss-

ing data the mechanism is called missing at random (MAR). Formally, the data are MAR if $f(M|Y, X, \phi) = f(M|Y, X_{obs}, \phi)$. The most general mechanism is not missing at random (NMAR), in this case the mechanism depends on the missing data.

Now we describe several methods to handle missing data and we also indicate which mechanism allows the use of certain methods. A first method is to exclude all cases that have missing values for any of the variables. This method is called the complete case analysis or listwise deletion. Besides this method available case analysis or pairwise deletion can be used. All the available cases are used when the method of pairwise deletion is applied to missing data. These two simple methods can be applied if relatively a small number of values are missing or if the missingness MCAR. However the MCAR assumption is often violated and in practice the mechanism is MAR or NMAR.

In general, the missing data mechanisms MAR or NMAR will lead to biased estimates in a complete case analysis. However, in the particular case of MAR where the missingness is not dependent on the response variable, a complete case analysis will not lead to bias (Little and Rubin (2002)). NMAR is the most general missing data mechanism. In order to get unbiased estimates in this case, a model for the missing data mechanism must be specified (Ibrahim, Chen, Lipsitz, and Herrin (2005)). It is possible to determine whether or not the missingness is MCAR. The basic idea of this test can be explained in the simplest case where only one of the variables is subject to missingness. The observations are split into cases with that variable missing and cases with that variable observed. The values of the other variables in the two groups can be compared by means of a two sample t -test. This procedure results in $k - 1$ t -tests for each variable subject to missing values. The difficulty with this procedure is to determine the significance level. Although the significance level of each individual test is known, the significance level of all tests simultaneously is not known. Instead of performing this t -test for each variable which is subject to missingness, Little (1988) proposed a single test statistic for testing MCAR. Unfortunately, there exists no such test for determining whether the missingness is MAR. To handle missings that are not MCAR, imputation methods and the Expectation Maximization method can be used in model estimation. A good reference on missing data analysis is Little and Rubin (2002) where historical approaches as well as more recently developed approaches are discussed.

3.1.3 Statistical validation

In general a model is not able to reproduce the exact data underlying the model. As said before, Basel II requires banks to validate the accuracy of the model. To determine the accuracy of the model several statistical tests are available in the literature. The next subsection will discuss the most used tests as a part of statistical validation. We base this section on Harrell (2001), Basel Committee on Banking Supervision (2005) and Engelmann and Rauhmeier (2006). Harrell

(2001) is one of the very few that describes very clear how to validate a logit model with an application to medical science, Basel Committee on Banking Supervision (2005) is a collection of studies on validation methods in general, and Engelmann and Rauhmeier (2006) contains a set of articles about probability of default, loss given default and exposure at default.

3.2 Statistical Model Validation

In the existing literature (Harrell (2001), Basel Committee on Banking Supervision (2005) and Engelmann and Rauhmeier (2006)) models are validated by determining the discrimination and calibration of the model. A model's discrimination is the ability to separate between defaulters and nondefaulters. Calibration is the ability of the model to make unbiased estimates of the outcome. We say that a model is well calibrated when a fraction of p of the events we predict, with a probability p actually occur. Besides discrimination and calibration we consider the following items as part of statistical validation: reproducibility of research, stability of parameters, choice of functional form and influential observations.

The items of statistical validation can be distinguished in validation of the model and validation of the probabilities. In the analysis first the statistical model is estimated, so the estimator $\hat{\beta}$ is determined. Next $\hat{\beta}$ is used to determine estimates of the PDs, \hat{p}_{it} , for observation i in period t . The validity of the model is determined using the items reproducibility of research, stability of parameters and choice of functional form. Discrimination and calibration validate the probabilities. Out-of-sample performance and bootstrap can be used to validate both the model and the probabilities. In the following we discuss the several items of statistical validation.

3.2.1 Reproducibility of Research

Reproducibility of research is defined as the duplication of the results of a former study (McCullough, McGeary, and Harrison (2006)). In the literature reproducibility is also known as replication. Positive and negative replication have a value for the replicated study. A positive reproducibility gives more support to the results of a former study. When a replication is negative it is clear that errors in the research have occurred. Of course the question then remains whether the original study or the reproduced study contains errors. For a researcher to be able to reproduce a study, documentation of the former study must be complete. Documentation is the written information concerning the model. In general, incomplete documentation will make it impossible to reproduce the results of a study. A second problem that makes it difficult to reproduce results is associated with the data. When the data are not recorded and documented correctly and completely they are useless to another researcher, as stated by Dewald, Thursby, and Anderson (1986). Moreover, data are often revised when new information is available. Exact replication will be impossible when a

revised data set is used in a replication. So the researcher has to be sure to use exactly the same data as in the replicated study.

3.2.2 Stability of Parameters

There are two types of stability, stability over time and stability over groups. Often models are intended to be used for predictions, but predictions are only valid if parameters are stable over time. In general we are often interested in stability over time for a subvector of the parameter vector β . For example, interest is in stability over time of the effect of the explanatory variable sex. Divide the parameter vector into two subvectors, $\beta' = (\beta'_1, \beta'_2)$. Let k_i be the length of β_i , $i = 1, 2$, $k_1 + k_2 = k + 1$. Suppose we want to test stability over time of the subvector β_1 . Let T_1 be the potential change point of interest. So we want to test whether the value of the subvector β_1 of β changes after period T_1 . The value of β_2 is assumed to be constant over time. The model to be estimated can be formulated as

$$\Pr(Y_{it} = 1) = \frac{1}{1 + \exp\{-\beta_{1.1}X_{i1t}I_{t \leq T_1} - \beta_{1.2}X_{i1t}I_{t > T_1} - \beta_2 X_{i2t}\}}, \quad (2)$$

where the vector of explanatory variables is divided analogous to the parameter vector. I is an indicator function which equals 1 if the condition is satisfied and 0 elsewhere. In total in the model above we need to estimate $2 \cdot k_1 + k_2$ parameters. The estimator $\hat{\beta}_{1.1}$ of $\beta_{1.1}$ uses the data up to and including period T_1 , the estimator $\hat{\beta}_{1.2}$ of $\beta_{1.2}$ uses the data after period T_1 , and the estimator $\hat{\beta}_2$ of β_2 uses all the data. Now the null hypothesis of stability over time of subvector β_1 can be formulated as

$$H_0 : \beta_{1.1} = \beta_{1.2},$$

this hypothesis will be tested against the two sided alternative

$$H_a : \beta_{1.1} \neq \beta_{1.2}.$$

Let $L(\hat{\beta}_{1.1}, \hat{\beta}_{1.2}, \hat{\beta}_2)$ be the maximum of the likelihood of the model in equation 2 and let $L(\hat{\beta})$ be the maximum of the likelihood of the model

$$\Pr(Y_{it} = 1) = \frac{1}{1 + \exp(-\beta' X_{it})}.$$

The likelihood ratio test can be performed to test H_0 . The test statistics, LR , is defined as

$$LR = 2 \left[\ln L(\hat{\beta}_{1.1}, \hat{\beta}_{1.2}, \hat{\beta}_2) - \ln L(\hat{\beta}) \right].$$

The distribution of LR is $\chi_{k_1}^2$, where the degrees of freedom k_1 is equal to the number of restrictions imposed under H_0 . Of course, this test applies as well if $k_2 = 0$, i.e. all parameters are subject to the test.

In the procedure above we assumed T_1 is known in advance. In general this change point might be unknown, following Andrews (1993) the unknown change point can be estimated in the following way. The likelihood ratio test is performed for each possible value of $T_1 \in \{1, 2, \dots, T\}$, resulting in $T - 1$ values of LR . The change point which

results in the highest value of LR is the estimate of the change point. Let LR^{\max} be the LR test with the highest value. Diebold and Chen (1996) describes two ways to determine the approximate distribution of the test statistic LR^{\max} . The first approximation is the asymptotic distribution, which is the distribution of the supremum of a series of chi-squared distributed statistics. Asymptotically this is correct, but behavior in a finite-sample is unknown. The second approximation is based on the bootstrap method. The bootstrap approximation is performed using the following steps. 1. The test statistic LR^{\max} is calculated. 2. B bootstrap samples are generated using the model parameters estimated under H_0 and disturbances drawn from uniform distribution. The dependent variable is equal to 1 if the probability is larger than a uniform random variable, else it is equal to 0. 3 For each bootstrap sample the test statistic LR^{\max} is calculated, this results in the so-called bootstrap distribution. 4. The p -value is approximated by the fraction of bootstrap LR^{\max} values larger than the LR^{\max} obtained using the observed data. Diebold and Chen (1996) found that the second approximation using the bootstrap distribution outperforms the asymptotic distribution, therefore we use the bootstrap approximation.

As more data becomes available, there might even be multiple change points, Bai and Perron (1998) considers issues related to multiple change points.

To test whether the model is stable over groups the LR test can be performed in an analogous way. Groups can be thought of as different mortgage labels offered by a bank. In order to use the same model for all the labels, the model has to be stable over groups.

DNB (De Nederlandsche Bank N.V. (2005)) asks to take the impact of changing economic conditions into account in determining the PD. Since the time span of the data sets in practice are limited to a few years, economic trends are not part of the model. The best solution for banks at the moment is to check for the stability of the parameters over time, as described above.

3.2.3 Choice of Functional Form

The logit model is used to estimate the PD. An assumption of the model is that a variable X has a linear effect on the logit of $Y = 1$. However, this relation can also be nonlinear. A simple way to describe a nonlinear effect of a variable is to use a transformation of the original variable, for example by taking the logarithm or the squared of the original variable. When the nonlinear effects are too difficult to describe using simple transformations, spline functions can be used (see Harrell (2001)). Restricted cubic spline functions are extremely useful to fit a highly curved function. To explain restricted cubic splines suppose there are two independent variables X_1 and X_2 . The effect of X_1 on the logit of Y is assumed to be linear and the effect of X_2 is assumed to be nonlinear. Therefore the model can be defined as $\text{logit}\{Y_i = 1|X_i\} = \beta_0 + \beta_1 x_{i1} + f(x_{i2})$, where $f(\cdot)$ is a restricted cubic spline. Note that to keep notations simple we omitted the time index. The function $f(\cdot)$

is specified as

$$f(x_{i2}) = \beta_2 x_{i2} + \beta_3 (x_{i2} - t_1)_+^3 + \beta_4 (x_{i2} - t_2)_+^3 + \dots + \beta_{h+2} (x_{i2} - t_h)_+^3,$$

where

$$\begin{aligned} (x)_+ &= \max(0, x), \\ \beta_{h+1} &= \frac{\beta_3(t_1 - t_h) + \beta_4(t_2 - t_h) + \dots + \beta_h(t_{h-2} - t_h)}{(t_h - t_{h-1})}, \\ \beta_{h+2} &= \frac{\beta_3(t_1 - t_{h-1}) + \beta_4(t_2 - t_{h-1}) + \dots + \beta_h(t_{h-2} - t_{h-1})}{(t_{h-1} - t_h)} \end{aligned}$$

and h is the number of knots. The function $f(\cdot)$ is linear before the first knot t_1 and after the last knot t_h and the function is continuous and differentiable at all knots. In practice the number of knots is $h = 3, 4, 5$ or 6 . The variable X_2 is divided into intervals with endpoints t_1, t_2, \dots, t_h . In each interval a cubic polynomial is fitted subject to the restrictions of continuity and differentiability at the knots. Once the parameters $\beta_0, \beta_1, \dots, \beta_h$ are estimated using maximum likelihood, β_{h+1} and β_{h+2} can be calculated. When the effect of X_2 is nonlinear, adding the cubic polynomial terms in the model will give a better fit to the data. In summary, a spline function makes the model more flexible.

3.2.4 Discrimination

Discrimination of a model is the ability to separate subjects' outcomes (Harrell (2001)). Before we discuss several statistics to determine the discrimination of the model we want to ensure discrimination is not confused with calibration. Calibration is the ability of the model to make unbiased estimates of the default probabilities. Table 1 gives an overview of the statistics used by Harrell (2001), Basel Committee on Banking Supervision (2005) and Engelmann and Rauhmeier (2006) to determine the discrimination and calibration of the model.

The Basel Committee's Accord Implementation Group has found that the Accuracy Ratio and the Receiver Operating Characteristic curve are the most meaningful discriminant statistics (Basel Committee on Banking Supervision (2005)). In the practice of banks the coefficient of concordance and Brier score are commonly used to measure the discrimination of a model. In the following the statistics Accuracy Ratio, Receiver Operating Characteristic curve, coefficient of concordance and Brier score will be discussed. For more information on the other discriminant statistics in table 1 we refer to the corresponding sources.

The Accuracy Ratio (AR) is a summary index of the Cumulative Accuracy Profile (CAP). The CAP, also known as Gini curve, Power curve or Lorenz curve, is obtained by first ordering all borrowers on the horizontal axis based on the scores of the model, from the lowest probability to the highest probability. For a given fraction of borrowers on the horizontal axis the percentage of defaulted borrowers with a lower probability than the maximum probability of this fraction is

Table 1: Discrimination and Calibration Statistics

	Discrimination	Calibration
Basel Committee on Banking Supervision (2005)	Cumulative Accuracy Profile (CAP)	Binomial test
	Receiver Operating Characteristic (ROC)	Chi square test
	Coefficient of concordance	Normal test
	Bayesian error rate	Traffic lights approach
	Entropy	
Harrell (2001)	Brier score	
	Coefficient of concordance	α_0 and α_1 refitted model
	Brier score	E_{\max} Generalized R_N^2
Engelmann and Rauhmeier (2006)	Cumulative Accuracy Profile (CAP)	Binomial test
	Receiver Operating Characteristic (ROC)	Chi square test
	Brier score	Normal test
		Traffic lights approach
		Spiegelhalter test Redelmeier test

plotted. The AR is defined as the ratio of the area between the CAP of the model and the CAP of the random model and the area between the CAP of the perfect model and the CAP of the random model. AR has a value between 0.5 and 1, where 0.5 indicates that the model performs equal to the random model and 1 indicates the model performs perfect.

A second graph we can use to determine the discrimination of the model is the Receiver Operating Characteristic (ROC) curve. To plot the ROC curve the borrowers have to be sorted by the score produced by the model in such a way that the defaulters are followed by the non-defaulters. The horizontal axis gives the cumulative percentage of borrowers sorted by the score of the model. The vertical axis gives the cumulative percentage of defaulters. For example, if the point (10%, 50%) lies on the ROC curve, this means that from the first 10% of the sorted borrowers, 50% defaulted on the contract. The aim of the model is to make a good distinction between the defaulters and the non-defaulters based on the borrower characteristics. An ROC curve close to the diagonal, indicates that the model is noninformative. The more the ROC curve lies in the top left corner, the better the model makes the distinction between defaulters and non-defaulters. Or, stated differently, the greater the area under the ROC curve, the better the model. The area under the ROC curve is called coefficient of concordance (c) or Area Under the Curve (AUC). When the value of c is 0.5 the ROC curve is equal to the diagonal and the model makes random predictions. A value of c equal to 1 indicates that the ROC curve lies in the top left corner and the predictions are perfect.

Brier score B is defined as (again omitting the time index for sim-

plicity):

$$B = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - Y_i)^2,$$

where \hat{p}_i is the estimated probability of observation i . B is the average of the difference between the probability and the observed outcome value and can be interpreted as the mean of the sum of squares of the residuals. A value close to 0 indicates the model performs good. Brier score can also be used to determine the discrimination of a rating system with borrower grades (Engelmann and Rauhmeier (2006)), $g = 1, \dots, G$. In this case Brier score is defined as

$$B = \frac{1}{N} \sum_{g=1}^G [n_{1g}(1 - P_g)^2 + n_{0g}(P_g)^2].$$

3.2.5 Calibration

Calibration is the ability of the model to make unbiased estimates of the PD. Calibration is a concept which originates from meteorology, where probability models for weather forecasts are used. In this setting the following definition is given (Seidenfeld (1985)): A set of probabilities are (well) calibrated if p percent of all predictions reported at probability p are true. This definition is general and can also be applied in the setting of default probabilities. Traditionally, the fit of a logit model is often analysed by a classification table. A classification table is a 2×2 table, where the columns are the two predicted values of the dependent variable and the rows are the two observed values of the dependent variable. The predicted values are determined using a cut-off probability which is often equal to 0.5. So the predicted value of the dependent variable is equal to 1 if the predicted probability is above 0.5 and 0 otherwise. The model is perfect if all cases are on the diagonal of the classification table. A classification table gives the percentage of correct predictions. In case of default the data sets are highly unbalanced in the sense that only a small fraction defaulted on their contracts, for example only 2% defaults occur. When a classification table is used to determine the goodness-of-fit one concludes that a model with constant default probability equal to zero will be preferred to a model with several explanatory variables. In case of credit risk, this zero default probability is useless for the calculation of the capital reserve. In other words, in the setting of determining capital reserve a classification table is not a useful calibration tool.

Table 1 shows some tests to determine the calibration of the model. Below we discuss the Binomial test, the chi-square statistic and we describe a refitting method which can be used to determine the calibration.

The first step in calibrating a PD model is often to perform the Binomial test (Engelmann and Rauhmeier (2006)). The Binomial test is for testing a single borrower grade at the time. The number of defaults in grade g , n_{1g} follows a binomial distribution if the assumption

of independent observations is made. So

$$\Pr(n_{1g}) = \binom{n_g}{n_{1g}} P_g^{n_{1g}} (1 - P_g)^{n_g - n_{1g}}.$$

Let the estimated PD in grade g be \hat{P}_g . The null hypothesis that the true PD, P_g , equals \hat{P}_g against the two-sided alternative can now be tested. The test statistic is the number of observed defaults in grade g , n_{1g} . The null hypothesis will be rejected if n_{1g} falls outside the interval $(B(\alpha/2), B(1 - \alpha/2))$, where $B(\cdot)$ is the quantile of the Binomial distribution of n_{1g} with parameters n_g and \hat{P}_g .

The chi-square (or Hosmer-Lemeshow) test statistics compares all borrower grades simultaneously. Define the following variable

$$E_g = n_g \cdot P_g : \text{ the number of expected defaults in grade } g,$$

The chi-square test statistic can now be defined as:

$$\hat{C} = \sum_{g=1}^G \frac{(n_{1g} - E_g)^2}{n_g P_g (1 - P_g)}.$$

The distribution of \hat{C} is approximated by the chi-square distribution with $G - 2$ degrees of freedom, χ_{G-2}^2 .

Harrell (2001) describes a refitting method which can be used to determine the calibration of a logit model. Suppose the original data (Y, X) is splitted in a development set (Y^d, X^d) and a test set (Y^t, X^t) . $\hat{\beta}$ is the maximum likelihood estimator of β based on the development sample. Again omitting the time index, $\hat{\beta}$ is the solution of the following maximum likelihood conditions

$$\sum_{i=1}^{n^d} x_{ij}^d \left(Y_i^d - \frac{1}{1 + \exp(-\beta' X_i^d)} \right) = 0, \quad \text{for } j = 0, 1, \dots, k,$$

where (Y^d, X^d) is the development sample of size n^d . The actual calibration probability and the original predicted probability can be calculated, for the test set (Y^t, X^t) of size n^t , in the following way. The model is refitted

$$p_i^{(c)} = \Pr(Y_i^t = 1 | \hat{\beta}' X_i^t) = \frac{1}{1 + \exp(-\gamma_0 - \gamma_1 \hat{\beta}' X_i^t)},$$

where $p_i^{(c)}$ denotes the actual calibrated probability, $i = 1, \dots, n^t$. Refitting the model means determining the maximum likelihood estimators of γ_0 and γ_1 . The original predicted probability, \hat{p}_i^t , is given by

$$\hat{p}_i^t = \frac{1}{1 + \exp(-\hat{\beta}' X_i^t)},$$

where $\hat{\beta}$ is the maximum likelihood estimator of β based on the development sample, (Y^d, X^d) . Now γ_0 and γ_1 can be estimated using maximum likelihood. Let $\hat{\gamma}_0$ and $\hat{\gamma}_1$ denote the maximum likelihood estimators. If $\hat{\gamma}_0$ is close to zero and $\hat{\gamma}_1$ is close to one, the model is well calibrated. A statistic related to the refitted model is

$$E_{\max} = \max_{\hat{p}} |\hat{p} - \hat{p}_c|,$$

which is the maximum error in the predicted probabilities.

A graphical tool to determine the calibration of a model is the calibration plot (Venables and Ripley (2002)). A calibration plot is obtained in the following way. We look at those loans with predicted probability of default equal to some value, say ω , $0 < \omega < 1$. Next of those loans the proportion p ($0 < p < 1$) of defaulted loans is determined. Then the calibration is obtained by plotting p against ω . A straight line in the calibration plot means the model is well calibrated.

3.2.6 Out-of-sample Performance and Bootstrap

Out-of-sample performance of a model is about how well the model performs on a different data set than the development set. Hence we need two data sets to determine the out-of-sample performance, a development sample and a test sample. First, the model is developed based on the development sample. Second, the test sample is used to determine the out-of-sample performance of the model by means of calculating the discrimination and calibration of the model.

In general we can split the original data into a development and a test sample in two ways. This results in two types of out-of-sample performance, that is out-of-sample performance within the time period and out-of-sample performance outside the time period. These two types of out-of-sample performance are also required by Basel II (section 420). To determine the out-of-sample performance within the time period a subset of the complete data set is used in model development and hence the development set contains observations over T periods. The remaining data also contains observations over T periods and is used to determine the out-of-sample performance of the model. Out-of-sample performance outside the time period means that the data is splitted in the following way. The observations in the first $T - q$ periods are used to develop the model and the observations in the last q periods are used to determine the out-of-sample performance.

The disadvantage of out-of-sample performance is that the size of the sample used to develop the model is smaller than the original sample of size. The bootstrap method overcomes this problem. We first describe the simple bootstrap method. The simple bootstrap method first generates B bootstrap samples. A bootstrap sample is a sample with replacement of size N drawn from the original sample. On each of these bootstrap samples the model is estimated. The B fitted models are applied to the original sample to give B values of a discrimination or calibration measure. The overall accuracy is the average of the B measures. This simple bootstrap method turns out not to work very well. Efron and Tibshirani (1993) describe an enhanced method that works better than the simple method. It is shown that this enhanced method performs better than the simple method (see for example Gong (1986) or Efron (1990)). First B bootstrap samples are drawn and B models are estimated using the bootstrap samples. The fitted models are applied to the original sample to give B measures.

The fitted models are also applied to the bootstrap samples (used to fit the model) to give B measures based on the bootstrap samples used to fit the model. The so-called optimism is calculated for each bootstrap sample by taking the difference between the measure based on the original sample and the measure based on the bootstrap sample. This results in B values of the optimism. The overall optimism is the average of the B values of optimism. To determine the discrimination or calibration of the final model, the overall optimism is subtracted from the measure calculated on the final model which is fitted based on the original sample.

3.2.7 Influential Observations

In developing a model, all the observations have a certain influence on the final model. One or a few observations might have a significant effect on the final model. If these so-called influential observations are left out in the model development, the final model will have significantly different parameters.

Harrell (2001) describes a few reasons of influential observations. The first, most common reason is having too few observations for the complexity of the fitted model, this is the problem of overfitting. Errors in the data is a second reason of influential observations. A third reason can be extreme values of the predictors. A last reason of influential observations is that there might be a disagreement between the predictors and the response. Influential observations must be carefully taken into account as a subject of model validation. If the influential observations are identified the question remains how to deal with them.

There are several statistical measures which can be used to detect influential observations of a logit model (Pregibon (1981)). We describe two measures, $DFBETAS$ and $DFFITs$. For observation i $DFBETA_i$ is defined as

$$DFBETA_i = \hat{\beta} - \hat{\beta}_{(-i)},$$

where $\hat{\beta}$ is the estimate based on the complete data set and $\hat{\beta}_{(-i)}$ is the estimate leaving out observation i . $DFBETA_i$ measures the change in the parameter estimated when observation i is omitted. To obtain $DFBETAS_i$ we standardize $DFBETA_i$ by dividing it with the square root of $((X'WX)^{-1})_{ii}$, where W is a diagonal matrix with elements $w_{ii} = \hat{p}_i(1 - \hat{p}_i)$, estimated probabilities are based on $\hat{\beta}$. In determining the cut off value for $DFBETAS_i$ several things can be taken into account, Belsley, Kuh, and Welsch (1980) provide a clear overview on how to determine cut off values. One possible cut off value for $DFBETAS_i$ is $2/\sqrt{N}$, so observations for which $DFBETAS_i \geq 2/\sqrt{N}$ are influential observations.

The second measure is based on the change in the estimated $\beta'X_i$ due to deletion of observation i , $DFFIT_i = \hat{\beta}'X_i - \hat{\beta}'_{(-i)}X_i$. Analogous to the above standardization $DFFITs_i$ is obtained. Again there are several cut off possibilities, one is to consider observations with $DFFITs_i \geq 2 \cdot \sqrt{(k+1)/N}$ as influential observations.

The statistics are based on omitting one observation to see the influence of this observation on the model. Of course this strategy can be extended by leaving out multiple influential observations. For the formulas of statistical measures we refer to the key paper Pregibon (1981) on diagnostic measures.

4 Application

In the empirical part of this paper we develop a logit model to estimate the probability that a given borrower defaults on his mortgage. The data we use are from Friesland Bank, a bank in the Netherlands. Friesland Bank wants to meet the requirements Basel II stated for the Foundation IRB Approach. To meet the requirements a model has to be developed to predict the probability that a borrower defaults on his contract within 1 year.

4.1 Description of the Data

The data set we use contains information on mortgages. Several variables are available in the data set. Some problems arise with the data. The first problem is that some variables have missing values for some of the cases. The second problem is that some of the variables are not measured correctly. In the data set for some missing values a 0 is inserted, so we can not determine for which case the value is missing and for which case the value is 0. For example, the average number of children is 0.01198. The variables which are not measured correctly can not be used to predict the probability of default. The data set contains yearly information from 2000 till 2003. Note that for a typical observation, the explanatory variables are measured at the beginning of each period and the default variable is measured at the end of each period. So the estimated PD is the probability that default occurs within one year.

Friesland Bank already developed a logit model. The model they developed contains the variables loan to value, loan to value missing, loan to income, expired duration, expired duration missing, mortgage type and overdue payment. A short explanation of the variables can be found in appendix B. The variable mortgage type in the model of the bank is an indicator variable which states whether the loan is of a linear type or of a different type. The model contains two dummy variables, loan to value missing and expired duration missing. Loan to value missing is a dummy for the cases where the loan to value is missing and expired duration missing is a dummy for the cases where the expired duration is missing. The coefficients of the logit model of Friesland Bank are shown in table 10 in appendix C. The coefficient of concordance of this model based on the development set is 0.8898.

4.2 Model Estimation

We first develop univariate logit models. The univariate models predict the probability of default using one explanatory variable. All calculations are done using the program *R* (Copyright 2005, the R Foundation for Statistical Computing, version 2.1.1). The results are used to determine which variables we take into account in the multivariate analysis. For some cases some variables are missing, we use the method of pairwise deletion in the univariate analysis. So, the univariate models are estimated using the cases for which the particular variable is available, the missing cases are deleted. For a number of estimated univariate models the coefficient of the variable is 0. In a univariate model these variables have no effect on the estimated probability of default. We choose not to use variables which have no effect in a univariate model in the multivariate analysis.

The candidate variables for the multivariate analysis are expired duration, credit limit, age, overdue payment and mortgage type. The variable mortgage type we use can take on 4 values, the mortgages types are annuity, life, linear and other mortgages, the reference type is interest-only. The variables loan to value and loan to income are also included as candidate variables because these variables are contained in the model developed by Friesland Bank.

A multivariate logit model is developed containing these variables. The results show that all variables, except the variable age, are significant. Next a model is developed omitting age, the results of this model can be found in tables 2 and 3. Table 2 shows the estimated coefficients of the model and table 3 shows the Wald statistics for the variables. The Wald statistic tests the null hypothesis that the coefficient of the variable is equal to zero. For example the Wald statistic of expired duration (29.92) tests whether the coefficient of the variable expired duration is equal to zero. The p -value is close to zero (< 0.0001), so the coefficient is significantly different from zero. The results show all the variables are significant in predicting the PD. This model is the starting model in the remaining analysis. The model we use as starting model is different from the model estimated by Friesland Bank. When we compare the results of the models we see the signs of the coefficients are the same.

In section 3.1.1 we already pointed out that the assumption of independent observations is violated due to the panel structure in the data. One of the consequences of this violation is that the standard errors are underestimated. This problem can be solved by using what we call the grouped bootstrap. Grouped bootstrap is based on grouped jackknife as described by Therneau and Grambsch (2000), where they leave out one client at the time rather than one observation at the time. Based on this grouped jackknife we performe grouped bootstrap, where the bootstrap method is applied at client level to determine the standard errors. The results can be found in table 11 in appendix C. The first four columns show the results when the assumption of independent observations is made (the same as table 2) and the last three columns show the results when the grouped bootstrap method is

applied. The number of bootstrap samples here is 1000. The results show that the standard errors are indeed higher using the grouped bootstrap, but this does not lead to insignificant coefficients.

Table 2: Fitted Multivariate Logit Model

	coef.	std.err	z	p-value
Intercept	-6.3362194	0.14291493	-44.3356	0.0000e + 00
expired.duration	-0.0051296	0.00093784	-5.4696	4.5101e - 08
credit.limit	0.0068541	0.00060592	11.3119	0.0000e + 00
overdue.payment	2.9610597	0.11031172	26.8427	0.0000e + 00
mortgage.type=annuity	0.6001313	0.11010283	5.4506	5.0188e - 08
mortgage.type=life	0.2690253	0.09577538	2.8089	4.9708e - 03
mortgage.type=linear	0.6567474	0.19549649	3.3594	7.8117e - 04
mortgage.type=other	0.4347098	0.17976977	2.4181	1.5600e - 02
ltv	0.0057829	0.00089986	6.4264	1.3065e - 10
debttoincome	0.0987587	0.02303781	4.2868	1.8126e - 05

Table 3: Wald Statistics for `b2default`

	χ^2	<i>d.f.</i>	<i>P</i>
expired.duration	29.92	1	< 0.0001
credit.limit	127.96	1	< 0.0001
overdue.payment	720.53	1	< 0.0001
mortgage.type	37.35	4	< 0.0001
ltv	41.30	1	< 0.0001
debttoincome	18.38	1	< 0.0001
TOTAL	1443.12	9	< 0.0001

4.3 Statistical Model Validation in Practice

4.3.1 Reproducibility of Research

The bank we consider in the empirical part already developed a logit model to estimate the probability of default. We can not reproduce the exact outcome of this research. One reason is the data we use are different from the data used by the bank. The bank used a data set with measurements on eight different dates, instead of four different dates. A second reason is how missing values are treated. We used the methods of listwise and pairwise deletion to handle missing values. Friesland Bank used some kind of imputation method, so they included some additional information.

4.3.2 Stability of Parameters

To determine whether the parameters are stable over time, we follow perform the test described in section 3.2.2. We test whether the parameter vector β is stable over time, the value of the test statistics is 12.630. We use the bootstrap method with $B = 2000$ to determine

the p -value and find a p -value of 0.372. So, the null hypothesis of an unknown break is rejected. Or stated differently, there is no structural break in the period from 2000 till 2003.

4.3.3 Choice of Functional Form

In the models estimated so far, we assumed the variables have a linear effect on the log odds in the default model. In this part of model validation we test whether some of the variables have a nonlinear effect. We use the method of restricted cubic splines. Most variables appear to have a linear effect on the log odds, except for the variable credit limit. We estimate a model containing nonlinear terms of the variable credit limit using a restricted cubic spline with 5 knots. The results are shown in tables 12 and 13 in appendix C. The coefficient of the nonlinear terms of credit limit is significantly different from zero, so the variable has a nonlinear effect on the log odds.

4.3.4 Discrimination

In the analysis above we developed two models, one is the starting model and the other is the model with a spline function. Next we determine the discrimination of the two models. For now we focus on two measures of discrimination, coefficient of concordance (c) and Brier score (B). The values of the measures are shown in table 4. The results

Table 4: Discrimination

Model	c	B
Starting model	0.914	0.015
Spline function	0.917	0.015

show that the Brier scores of the models are the same and are also very close to zero, which can be interpreted as a small sum of squares of the residuals. The coefficient of concordance of the model with spline function is higher compared to the starting model, so the model with spline function discriminates better than the starting model. In summary, the model with spline function can separate outcomes better than the starting model.

4.3.5 Calibration

Now we determine the calibration of the two models by means of calibration plots. The two plots are shown in pictures 3 and 4. The diagonal line in these plots show the ideal case of perfect calibration. The dotted line shows the apparent calibration of the model. The straight line will be discussed in section 4.3.6. Both calibration plots show similar pattern. For predicted probabilities above 0.4 the models are both not well calibrated. When the focuss is on probabilities below 0.4 we see the model with spline function is better calibrated than the

starting model. Or stated differently, the model with spline function is better in making unbiased estimates of the default probabilities.

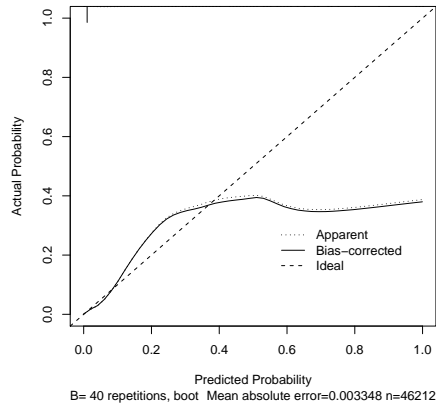


Figure 3: Calibration Plot Starting Model

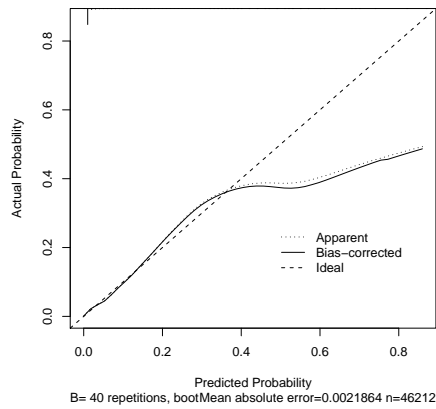


Figure 4: Calibration Plot Model with Spline

The calibration plots show how well the model is calibrated based on the development set. The plots showed the model is better calibrated for lower probabilities. A natural step now is to quantify the calibration for observations with low estimated probability of default. In order to do so we would like to determine calibration slope and intercept for a subset of the observations. The data set can be divided in subsets based on the estimated probability of default. For example the quartiles of the estimated probabilities can be used as cut off values. Next the calibration slope and intercept can be determined for

the subsets. However, in practice this strategy is not useful since the subset with low PD contains very few defaults which makes it very difficult to estimate a logit model.

4.3.6 Out-of-sample Performance and Bootstrap

Above we determined how well the two models perform on the data set which is also used for developing the model. In this section we determine the performance of the model on a different data set. We use the coefficient of concordance (c) and Brier score (B) to determine the discrimination and use calibration intercept and slope (γ_0 and γ_1) for calibration of the models.

First we consider out-of-sample performance within the time period. The data set is divided into two subsets, the development set contains a random sample of 75% of the complete data set. The remaining data are used as test set. The results are shown in table 5. The table also shows the measures calculated on the development set. Note γ_0 and γ_1 estimated on the development set are always equal to 0 and 1, respectively. As we already concluded in the previous sections, results here show the model with spline function discriminates better and is also better calibrated compared to the starting model. This also holds out-of-sample within the time period.

Table 5: Out-of-sample within the time period

Model	Development/Test set	c	B	γ_0	γ_1
Starting model	Development set	0.9191	0.0149	0	1
	Test set	0.9087	0.0153	-0.0955	0.9408
Spline function	Development set	0.9225	0.0146	0	1
	Test set	0.9089	0.0148	-0.0003	0.9734

Second we consider out-of-sample performance outside the time period. Analogous to section 4.3.2 the data set is divided into two subsets, the first containing the years 2000, 2001 and 2002 and the second contains 2003. Results of out-of-sample performance outside the time period are very similar to the results within the time period. So again we see the model with spline function performs better than the starting model.

Table 6: Out-of-sample outside the time period

Model	Development/Test set	c	B	γ_0	γ_1
Starting model	Development set	0.9183	0.0142	0	1
	Test set	0.9053	0.0163	0.0678	0.9919
Spline function	Development set	0.9216	0.0142	0	1
	Test set	0.9115	0.0161	-0.0529	0.9567

Next we use the bootstrap method described in subsection 3.2.6 with 40 bootstrap samples. The calibration plots are shown in figures

3 and 4. The straight lines in the plots show the bias corrected calibration plot using the bootstrap method described in subsection 3.2.6. The plots show both models are not well calibrated for high probabilities. For low probabilities the model with spline function is better calibrated than the starting model. The results of the measures mentioned above are shown in table 7. Again results show the model with

Table 7: Bootstrap

Model	c	B	γ_0	γ_1
Starting model	0.9167	0.0150	-0.0125	0.9963
Spline function	0.9187	0.0147	-0.0184	0.9941

spline function discriminates better and the starting model is better calibrated.

4.3.7 Influential Observations

In this section we try to determine whether there are influential observations. First we determine $DFBETAS$ based on a cut off point of $2/\sqrt{N}$. First we determine for each observation whether it is influential on any of the coefficients of the starting model. Next we check on how many coefficients the observations are influential. The histogram in figure 5 shows how many observations are influential on 1, . . . , 10 out of 10 coefficients. We see, for example, that 184 of the observations are influential on all 10 coefficients.

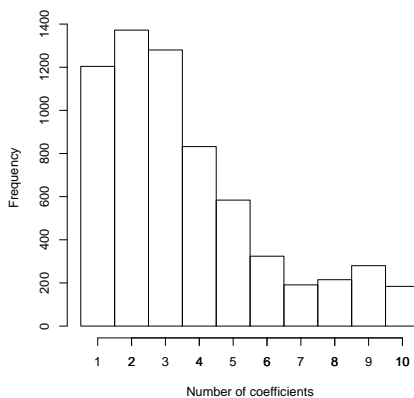


Figure 5: Frequency of influential observations

The next step in the analysis of influential observation is to leave out some variables. The question now raises which observations should be omitted. First we omit the observations which are influential on all

10 coefficients. Table 8 shows the estimation results when the observations which are influential on all 10 coefficients are omitted.

Table 8: Omitting observations influential on 10 coefficients

	coef.	std.err	z	p-value
Intercept	-6.6759146	0.16090702	-41.4893	0.0000e + 00
expired.duration	-0.0048416	0.00106827	-4.5322	5.8362e - 06
credit.limit	0.0066672	0.00069387	9.6086	0.0000e + 00
overdue.payment	2.6438122	0.11511132	22.9674	0.0000e + 00
mortgage.type=annuiteiten	0.9696886	0.12358836	7.8461	4.2188e - 15
mortgage.type=gemengd	0.6641441	0.10799990	6.1495	7.7733e - 10
mortgage.type=lineair	1.0902317	0.21144574	5.1561	2.5217e - 07
mortgage.type=overig	0.8711869	0.19349916	4.5023	6.7229e - 06
ltv2	0.0049836	0.00103006	4.8381	1.3107e - 06
debttoincome2	0.1618302	0.02689704	6.0167	1.7806e - 09

The original data set contains only 1.7% defaults. But the data set used to estimate the model in table 8 is based only 1.3% defaults. Next we estimate the model omitting the observations which are influential on 9 or 10 of the coefficients (see table 9 for the results). The data set used here contains only 0.8% defaults. We see that when omitting even more influential observations results in omitting more defaults compared to nondefaults. So estimating the model will become even more difficult.

Table 9: Omitting observations influential on 9 or 10 coefficients

	coef.	std.err	z	p-value
Intercept	-7.2625164	0.21248761	-34.1785	0.0000e + 00
expired.duration	-0.0059560	0.00142961	-4.1662	3.0976e - 05
credit.limit	0.0072758	0.00090166	8.0693	6.6613e - 16
overdue.payment	1.9882543	0.12965902	15.3345	0.0000e + 00
mortgage.type=annuiteiten	1.7190314	0.17389018	9.8857	0.0000e + 00
mortgage.type=gemengd	1.6206412	0.15427016	10.5052	0.0000e + 00
mortgage.type=lineair	1.8904935	0.26681163	7.0855	1.3853e - 12
mortgage.type=overig	1.5864949	0.25033653	6.3374	2.3360e - 10
ltv2	0.0041693	0.00135253	3.0826	2.0519e - 03
debttoincome2	0.1863634	0.03514124	5.3033	1.1375e - 07

4.3.8 Statistical Validation in Conclusion

Above we applied the methodology of section 3.2 to a data set on mortgages of Friesland Bank. The overall conclusion we can draw from the results is the model with spline function performs better than the starting model. Since we were unable to reproduce the exact results obtained by Friesland Bank we can not compare their model in depth to the model with spline function. However, we can conclude based on the coefficient of concordance the model with spline function performs better.

5 Conclusion

The new Basel Capital Accord forces banks to develop models to estimate the probability of default. These models need to be validated on a continuous basis. However, supervisors like the Dutch central bank, are not very clear what they think constitutes proper validation. In this paper we try to fill this gap. We give an overview of methods used to analyze and validate logit models. Validation is classified into three classes: theoretical validity, data validity and statistical validity. Theoretical validity reviews the theories and assumptions underlying the proposed model, data validity is about the accuracy of the data and statistical validity is concerned with the use and errors of the model.

The main focus of this paper is on statistical model validation. We discuss the items reproducibility of research, stability of parameters, choice of functional form, discrimination, calibration, out-of-sample performance and bootstrapping, and influential observations. We conclude that the literature does provide very useful concepts which can be used in statistical model validation. The classification given in this paper can be used to systematically validate a default model, application will lead to a better model.

In the empirical part of statistical model validation we paid no attention to influential observations. In future research this item can be taken into account. We made several assumptions in our analysis to make the calculations rather simple. Some of these assumption are not very realistic. In future research these assumptions must be reconsidered. We used the methods of listwise and pairwise deletion to handle missing values. Due to this approach some valuable information is not used. Imputation methods can be applied to get better results. We also assumed that the observations are independent. The data set we use contains information on borrowers measured on 4 different dates. So, in principle, a borrower can occur four times in the data set. This dependence is ignored in this paper. In a future research this dependence can be taken into consideration. In the theoretical part of this paper we provided a large number of measurements to use in model validation. In the empirical part we did not calculate all the measurements. In future research the remaining measurements can be used in order to make a better comparison amongst the measurements.

References

- Andrews, Donald W.K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821–856.
- Bai, Jushan and Pierre Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometric* 66, 47–78.
- Basel Committee on Banking Supervision (2001a). The consultative document: The internal ratings-based approach. www.bis.org/publ/bsbca05.pdg (download of August 15, 2005).

- Basel Committee on Banking Supervision (2001b). The consultative document: The standardised approach to credit risk. www.bis.org/publ/bcbsa04.pdf (download of August 15, 2005).
- Basel Committee on Banking Supervision (2005, February). Working paper no. 14: Studies on the validation of internal ratings systems.
- Basel Committee on Banking Supervision (2006, June). International convergence of capital measurements and capital standards: A revised framework comprehensive version.
- Belsley, D.A., E. Kuh, and R.E. Welsch (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.
- Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics Methods and Applications*. Cambridge: University Press.
- De Nederlandsche Bank N.V. (2005). Bazel II: Governance rond modelontwikkeling, -validatie en gebruik.
- Dewald, W.G., J.G. Thursby, and R.G. Anderson (1986). Replication in empirical economics: The journal of money, credit and banking project. *The American Economic Review* 76, 587–603.
- Diebold, Francis X. and Celia Chen (1996). Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. *Journal of Econometrics* 70, 221–241.
- Duffie, D. and K.J. Singleton (2003). *Credit Risk*. Princeton: Princeton University Press.
- Efron, Bradley (1990). More efficient bootstrap computations. *Journal of the American Association* 85, 79–89.
- Efron, B. and R.J. Tibshirani (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Engelmann, B. and R. Rauhmeier (2006). *The Basel II Risk Parameters: Estimation, Validation, and Stress Testing*. Heidelberg: Springer.
- Gass, S.I. and B.W. Thompson (1980). Guidelines for model evaluation: an abridged version of the u.s. general accounting office exposure draft. *Operation Research* 28(2), 431–439.
- Gong, Gail (1986). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of the American Statistical Association* 81, 108–113.
- Harrell, Jr. F.E. (2001). *Regression Modeling Strategies*. New York: Springer.

- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.
- Ibrahim, J.G., M.H. Chen, S.R. Lipsitz, and A.H. Herrin (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association* 100.
- Kalbfleisch, J.D. and R.L. Prentice (1980). *The Statistical Analysis of Failure Data*. New York: John Wiley and Sons.
- Lee, L.-F. (1982). Specification error in multinomial logit models. *Journal of Econometrics* 20.
- Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* 83, 1198–1202.
- Little, R.J.A. and D.B. Rubin (2002). *Statistical Analysis with Missing Data*. New York: Wiley Interscience.
- McCullough, B.D., Kerry Anne McGeary, and Teresa D. Harrison (2006). Lessons from the JMCB archive. *Journal of Money, Credit, and Banking* 38, 1093–1107.
- Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics* 9(4), 705–724.
- Seidenfeld, Teddy (1985). Calibration, coherence, and scoring rules. *Philosophy of Science* 52, 274–294.
- Therneau, Terry M. and Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer: New York.
- Venables, W.N. and B.D. Ripley (2002). *Modern Applied Statistics with S*. New York: Springer.

A Basel II requirements

Basel Committee on Banking Supervision (2006) section III.H states the minimum requirements for banks which opt for the IRB approach. This appendix gives an overview of the most important requirements for the IRB approach. Since in this paper we focus on statistical validation of PD models, only the more statistical requirements for PD models are listed.

389. A rating system must provide a meaningful assessment of borrower and transaction characteristics. The system must be consistent with internal use.

394. The term rating system comprises all of the methods, processes, controls, and data collection and IT systems that support the assessment of credit risk, the assignment of internal risk ratings, and the qualification of default and loss estimates.

395. A bank may use multiple systems. The rationale for assignment must be documented. The system must reflect the risk of the borrowers in best possible way.

397. Separate exposures to same borrower must be assigned to same borrower grade, irrespective of any differences in the nature of each specific transaction.

402. For each pool PD, LGD and EAD must be estimated. At a minimum, the following risk drivers must be considered:

- Borrower risk characteristics
- Transaction risk characteristics
- Delinquency of exposure

403. A bank must have meaningful distribution of exposures across grades with no excessive concentrations, on both its borrower-rating and its facility-rating scales.

404. A bank must have at least 7 borrower grades for non-defaulted borrowers and one for those that defaulted.

411. all relevant and material information must be used in assigning ratings to borrowers and facilities. The less information a bank has, the more conservative must be its assignments.

414. Although the time horizon used in PD estimation is one year, banks are expected to use a longer time horizon.

415. A borrower rating must represent the bank's assessment of the borrower's ability and willingness to contractually perform despite adverse economic conditions or the occurrence of unexpected events.

416. Given the difficulties in forecasting future events, a bank must take a conservative view of projected information.

417. In the estimation of a PD model

- the variables must form a reasonable set of predictors.
- there must be place for vetting the data (accuracy, completeness and appropriateness of the data).
- the underlying data must be representative of the population.
- human judgement must take in to account information not considered by the model.
- a bank must have procedures for human review.
- a bank must have a regular cycle of model validation (monitoring of model performance and stability; model relationships, testing of model outputs against outcomes).

420. In case of a statistical model the following must be documented:

- Outline of the theory, assumptions, mathematical and empirical basis of the assignment, and the data sources used.
- A rigorous statistical process (out-of-time and out-of-sample performance tests) for validation.
- An indicate when the model does not work effectively.

425. The ratings must be refreshed at least on an annual basis.

430. Rating histories must be maintained, including the rating since the borrower/guarantor was assigned an internal grade, dates, methodology and key data used.

434. A bank must have in place sound stress testing processes for the use in assessment of capital adequacy.

435. A bank must perform a credit risk stress test to assess the effect of certain specific conditions on its IRB regulatory capital requirements.

447. PD estimates must be a long-run average of one-year default rates for borrowers in the grade.

448. Internal estimates of PD, LGD, and EAD must incorporate all relevant, material and available data, information and methods.

452. A default occurred when either or both of the two following events have taken place:

- The obligor is unlikely to pay.
- The obligor is past due more than 90 days.

458. The bank must have clearly articulated and documented policies in respect of the counting days past due.

501. Banks must regularly compare realised default rates with estimated PDs for each grade and be able to demonstrate that realised default rates are within the expected range for that grade.

503. Banks must demonstrate that quantitative testing methods and other validation methods do not vary systematically with the economic cycle.

504. Banks must have well-articulated internal standards for situations where deviations in realised PDs, LGDs and EADs from expectations become significant enough to call the validity of the estimates into question.

506. Banks under the foundation IRB approach, which do not meet the requirements for own-estimates of LGD and EAD, must meet the minimum requirements described in the standardised approach (section II.D).

B Explanation of the Variables

Loan to value is the ratio between the original amount of debt and the appraisal value of the house, expressed as a percentage. For the cases where loan to value exceeds 400, we inserted a value 0 and treated the variable as a missing value. Debt to income is the ratio between the original amount of debt and income of the borrower, expressed as a percentage. The cases where debt to income exceeds 5 are truncated at 5. Expired duration is the period between the start of the contract and the snapshot, measured in months. The variable mortgage type used by the bank is an indicator variable which states whether the loan is of a linear type or of a different type. The variable mortgage type we use can take on 4 values, the mortgages types are annuity, life, linear and other mortgages. The reference mortgage type is interest-only, the coefficients of the 4 types indicates the effect on the probability of default when the borrower has mortgage type different from interest-only. Overdue payment is an indicator variable which states whether there was an overdue amount during the 12 months prior to the snapshot. Credit limit is the average percentage of the credit limit that is taken up during the last 3 months prior to snapshot. The age of the borrower is measured in years at the snapshot.

C Default Models

Table 10: Logit Model Frieslandbank

tabel.schattingresultaten	coef.
Intercept	-5.8641597
expired.duration	-0.0070640
expired.duration missing	-0.2574613
overdue.payment	3.2007371
mortgage.type=linear	0.5140892
ltv	0.0044468
ltv missing	0.5440784
debttoincome	0.1255584

Table 11: Grouped Bootstrap Standard Errors

tabel.boot	coef.	std.err	z	p-value	std.boot	z.boot	p.boot
Intercept	-6.3362194	0.14291493	-44.3356	0.0000e + 00	0.1572124	-40.3036	0.0000e + 00
rijping.maanden	-0.0051296	0.00093784	-5.4696	4.5101e - 08	0.0011074	-4.6321	3.6198e - 06
gem.uitnutting.3mnd	0.0068541	0.00060592	11.3119	0.0000e + 00	0.0015091	4.5418	5.5776e - 06
achterstand.in.12mnd	2.9610597	0.11031172	26.8427	0.0000e + 00	0.1381778	21.4293	0.0000e + 00
lening.type=annuiteiten	0.6001313	0.11010283	5.4506	5.0188e - 08	0.1419991	4.2263	2.3756e - 05
lening.type=gemengd	0.2690253	0.09577538	2.8089	4.9708e - 03	0.1168849	2.3016	2.1356e - 02
lening.type=lineair	0.6567474	0.19549649	3.3594	7.8117e - 04	0.2167396	3.0301	2.4446e - 03
lening.type=overig	0.4347098	0.17976977	2.4181	1.5600e - 02	0.2150810	2.0211	4.3265e - 02
ltv2	0.0057829	0.00089986	6.4264	1.3065e - 10	0.0011250	5.1405	2.7407e - 07
debttoincome2	0.0987587	0.02303781	4.2868	1.8126e - 05	0.0279850	3.5290	4.1716e - 04

Table 12: Fitted Multivariate Logit Model

tabel.schattingresultaten	coef.	std.err	z	p-value
Intercept	-6.5301061	0.14719000	-44.3651	0.0000e + 00
expired.duration	-0.0055812	0.00095022	-5.8736	4.2638e - 09
credit.limit	0.0661639	0.00438228	15.0981	0.0000e + 00
credit.limit'	-0.1746259	0.01239340	-14.0902	0.0000e + 00
overdue.payment	2.2121109	0.12350226	17.9115	0.0000e + 00
mortgage.type=annuity	0.5683589	0.11003047	5.1655	2.3984e - 07
mortgage.type=life	0.2327164	0.09591518	2.4263	1.5255e - 02
mortgage.type=linear	0.6841588	0.19573015	3.4954	4.7332e - 04
mortgage.type=other	0.4405753	0.18138247	2.4290	1.5141e - 02
debttoincome	0.0949513	0.02302551	4.1237	3.7276e - 05
ltv	0.0057748	0.00091426	6.3163	2.6786e - 10

Table 13: Wald Statistics for b2default

	χ^2	d.f.	P
expired.duration	34.50	1	< 0.0001
credit.limit	317.85	2	< 0.0001
<i>Nonlinear</i>	198.53	1	< 0.0001
overdue.payment	320.82	1	< 0.0001
mortgage.type	35.01	4	< 0.0001
debttoincome	17.01	1	< 0.0001
ltv	39.90	1	< 0.0001
TOTAL	1606.00	10	< 0.0001