A Practical Approach to Validating a PD Model

Abstract

The capital adequacy framework Basel II aims to promote the adoption of stronger risk management practices by the banking industry. The implementation makes 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.

JEL classification: E42; E58; G21

Keywords: Credit Risk; Probability of default; Basel II; Statistical Validation; Logit Model

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 the 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" (Federal Register (2004)).

Typically, the portfolio on loans consist of loans to business (small, large, retail) and loans to individuals (mortgages). The main difference in the approach to determine PDs for loans to business and loans to individuals, stems from the fact that for business banks make use of external ratings. For loans to business banks use external ratings to determine PDs, for example ratings of a credit bureau or Standard & Poor's or Moody's ratings. Carling, Jacobson, Lind, and Roszbach (2007) base the PDs of firms partly on ratings determined by a credit bureau. For individuals with a loan such external ratings do not exist. Therefore, banks need to estimate the PDs, for example, by means of a logit model. Validation of PD models for loans to

business is concentrated on validating the PD estimates. Traditionally, PDs are validated by measuring the discrimination and calibration (see Dwyer and Stein (2006)). Discrimination and calibration are measures that determine how well the estimated PDs fit the data. Of course, discrimination and calibration can also be used to validate PD models for loans to individuals. However, discrimination and calibration will only provide information on how well a model fits the data. For loans to individuals banks use a logit model to estimate the PDs. In this case validation is not only restricted to the PDs (by means of discrimination and calibration), because in addition the parameter vector can also be validated. By also taking the parameter vector into account, the validation will be more rigorous since information on how the fit can be improved is obtained. Careful examination of the effects of risk drivers on the PD may show that such an effect changes over time, or is different by type of product.

Nowadays banks pay a lot of attention to the validation process, but still a general 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 the portfolio of loans to individuals. In our methodology we focus not only on validation of the PDs, but we specifically pay attention to validation of the parameter vector of the underlying model. This will provide information on how well the model fits and on how the fit may be improved.

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 this paper is organised as follows. Section 2 provides 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 partly 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 mininum 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 have 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 each 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. 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 is the index of clients, $i = 1, \ldots, n_t$. n_t is the number of observations in period t, t is the time index, $t = 1, \ldots, T$. Define the total number of observations as $N = \sum_{t=1}^{T} n_t$. 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.

Let $X_{it} = (1, x_{it,1}, \ldots, x_{it,k})$ be the $(k+1)$-vector of explanatory variables of client i at time t. 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). Let Y_{it} be the dependent variable which equals 1 if client i defaults between time t and $t+1$, and 0 otherwise. Denote the probability that Y_{it} equals y given X_{it} by $\Pr(Y_{it} = y|X_{it}; \beta)$, $y = 0, 1$, and define $p_{it} = \Pr(Y_{it} = 1|X_{it}; \beta)$ as the probability that Y_{it} equals 1, where β is the $(k+1)$-parameter vector of interest.

In general, p_{it} can take on every value between 0 and 1. In practice banks often divide the loans into borrower grades, with a fixed probability of default for each grade. With respect to borrower grades some additional notions are defined. g is the index of borrower grades, $g = 1, \ldots, G$. Let n_{1g} be the number of loans in borrower grade g that defaulted, define $n_{0g} = n_g - n_{1g}$. Let P_g be the 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, the focus is 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 occured 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 of the observed explanatory variables. Both cases of 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

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

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, $G(X_{it}; \beta) = \Phi(\beta' X_{it})$, where $\Phi(\cdot)$ the standard normal distribution function. $\beta' X_{it}$ is sometimes referred to as the index. In practice the logit model is often assumed.

The assumption of a logit model is not restrictive. Equation (1) can be rewritten as $\Pr(Y_{it} = 1|X_{it}; \beta) = G(X_{it}; \beta) = \Lambda(\Lambda^{-1}(G(X_{it}; \beta)))$, 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 separate PD models for each loan class (loans to retail, mortgages, loans to small business and loans to large business). 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 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. 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. 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.

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 data ware 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 1.

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. 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 focuses 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.

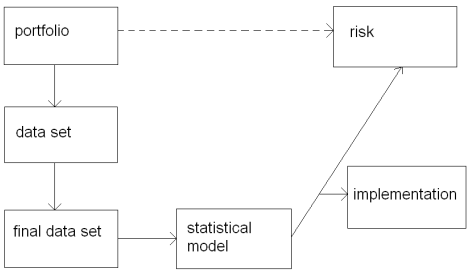


Figure 1: Determine Risk of Portfolio

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 that 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 that 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. A second problem with the variables is that some variables are difficult to measure. For example measurement of default itself is difficult. According to Basel II (section 452) default occured when the obligor is unlikely to pay and/or the obligor is past due more than 90 days. In practice it is difficult to measure when an obligor is unlikely to pay.

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 that 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. This means there is less information available 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. The most commonly used method to handle missing data is complete case analysis. Complete case analysis uses only the complete observations. However, complete case analysis will give, at best, unbiased but inefficient estimates and, at worst, biased estimates. 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 non-defaulters. 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. Discrimination and calibration both compare the estimated probabilities with the observed frequency of default in the data set. So by measuring discrimination and calibration the PDs are validated. However, validation can be more rigorous since the parameters of the model ($\hat{\beta}$) can also be validated. We validate the parameters by means of reproducibility of research, stability of parameters and choice of functional form. Besides we describe out-of-sample performance and bootstrap to validate the PDs as well as the parameteres. Subsequently we discuss the items of statistical validation: reproducibility of research, stability of parameters, choice of functional form, discrimination,

calibration, out-of-sample performance and bootstrap.

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 occured. 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. 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

$$p_{it} = \frac{1}{1 + \exp\left\{-\beta_{1.1}' X_{i1t} I_{t \leq T_1} - \beta_{1.2}' X_{i1t} I_{t > T_1} - \beta_2' X_{i2t}\right\}}, \tag{2}$$

where the vector of explanatory variables is divided analogously to the parameter vector. $\beta_{1.j}$ is the subvector β_1 of β for period $j = 1, 2$. $I_{\cdot}$ 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

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

The likelihood ratio test can be performed to test H_0. The test statistic, 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 $\chi^2_{k_1}$, 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 $T_1 \in \{1, 2, \ldots, T\}$, 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) describe 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 draw from the uniform distribution, 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 in the application.

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)) requires 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 to 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 written 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

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

where

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

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 $t_1, t_2, \ldots, t_h$. In each interval a cubic polynomial is fitted subject to the restrictions of continuity and differentiability at the knots. Once the parameters $\beta_0, \beta_1, \ldots, \beta_h$

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. Several statistics are available to determine discrimination and calibration. Table 1 gives an overview of the statistics proposed by Harrell (2001), Basel Committee on Banking Supervision (2005) and Engelmann and Rauhmeier (2006).

Table 1: Discrimination and Calibration Statistics

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

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. Therefore, we will discuss the Accuracy Ratio, the Receiver Operating Characteristic curve, the coefficient of concordance and Brier score. 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 Ac-

curacy 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 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. Let C be the probability based on which the borrowers are classified into defaulters and non-defaulters. If the estimated probability is above C the borrower is classified as defaulter, else the borrower is classified as non-defaulter. The borrowers classified as defaulters can be split into two groups, borrowers which are correctly classified as defaulters and borrowers which are incorrectly classified as defaulters. The borrowers classified as non-defaulters can also be split into two group, borrowers which are correctly classified and borrowers which are incorrectly classified. The percentage of defaulters which are correctly classified as defaulters is called the hit rate, denoted by $HR(C)$. The hit rate depends on the cut off value C. The percentage of non-defaulters incorrectly classified as defaulters is called false alarm rate, $FAR(C)$. The ROC curve is obtained by plotting HR against FAR for different values of C. 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. In practice ROC curves are not only used to determine the discrimination, but also to determine a cut-off point for granting loans (see Stein (2005)). 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 simplicity):

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

where $\hat{p}_i$ is the estimated probability of observation i. B is the average of the squared 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)), $g = 1, \ldots, G$. In this case Brier score is defined as

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

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

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

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

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

The chi-square test statistic is defined as

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

If the chi-square test is performed on the development set, the distribution of $\hat{C}$ is approximated by the chi-square distribution with $G-2$ degrees of freedom, χ^2_{G-2}. If the test is applied on out-of-sample data the distribution of $\hat{C}$ is chi-squared with G degrees of freedom.

Harrell (2001) describes a refitting method which can be used to calibrate 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

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

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

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

where $p_i^{(c)}$ denotes the actual calibrated probability, $i = 1, \ldots, n^t$. 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

$$\mathrm{E}_{\max} = \max_{\hat{p}} |\hat{p} - \hat{p}^{(c)}|,$$

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

The statistics defined above can be applied to the development set to determine the performance of the model. However, we want to determine the performance of the model for future predictions. Using the same data both to develop the model and to determine the performance of the model will result in an overestimation of the performance for future predictions. For example, the value of Brier score determined on the development set will be lower than the value determined on a different data set. If the performance is determined on the development set the performance will be estimated too optimistic. To correct for this optimism out-of-sample performance and bootstrap methods can be applied.

So, we are interested in how well the model performs on a different 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. The 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 substracted from the measure calculated on the final model which is fitted based on the original sample.

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. Hence, a model has to be developed to predict the probability that a borrower defaults on his contract within 1 year. All calculations are done using the program R (Copyright 2005, the R Foundation for Statistical Computing, version 2.1.1).

4.1 Description of the Data

The data set contains yearly information from 2000 till 2003 on mortgages to individuals. 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. A short description of the variables can be found in appendix A.

Friesland Bank already developed a logit model. Their model contains the variables loan to value, loan to value missing, loan to income, expired duration, expired duration missing, mortgage type and overdue payment. 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 estimated coefficients are shown in table 7 in appendix B. The coefficient of concordance of this model based on the development set is 0.8898.

We estimated a multivariate logit model with expired duration, credit limit, age, overdue payment, mortgage type, loan to value and loan to income as explanatory variables. Here mortgage type can take on 4 values, annuity, life, linear and other mortgages, the reference type is interest-only. All variables, except for age, turn out to be significant. Next a model is developed omitting age, the estimated coefficients of this model can be found in tables 2. The results show that all parameters are significant. Wald statistics (not shown here) show that the four coefficients of mortgage type are jointly significant. This model is referred to as the starting model. 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 that the signs of the coefficients are the same.

Table 2: Estimates starting model.

	coef	std.err	z	p-value
intercept	−6.336	0.143	−44.336	0.000
expired.duration	−0.005	0.001	− 5.470	0.000
credit.limit	0.007	0.001	11.312	0.000
overdue.payment	2.961	0.110	26.843	0.000
mortgage.type=annuity	0.600	0.110	5.451	0.000
mortgage.type=life	0.269	0.096	2.809	0.005
mortgage.type=linear	0.657	0.195	3.359	0.001
mortgage.type=other	0.435	0.180	2.418	0.016
loan.to.value	0.006	0.001	6.426	0.000
debt.to.income	0.099	0.023	4.287	0.000

4.2 Theoretical Validation

The results of the starting model show that expired duration has a negative relationship with the probability of default. This means that when a

mortgage matures the probability of default is lower. The binary variable overdue payment has a positive influence on the PD. So when a mortgage is in arrear the probability of default is higher. The coefficients of mortgage type are positive, so in comparison to the reference category interest-only, the categories annuity, life, linear and other result in a higher PD. Loan to value and debt to income have positive relation with the PD. The signs of the variables are in correspondence with expectations.

4.3 Data Validation

The data are representative for the underlying population since we use the complete portfolio of mortgages.

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 truely is 0. The variables which are not measured correctly can not be used to predict the probability of default.

For some cases the values for certain variables are missing, we use complete case analysis to estimate the models.

4.4 Statistical Validation

4.4.1 Reproducibility of Research

Friesland Bank already developed a logit model to estimate the probability of default. However, 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. A second reason is how missing values are treated. We used complete case analysis to handle missing values. Friesland Bank used some kind of imputation method, so they included some additional information.

4.4.2 Stability of Parameters

To determine whether the parameters are stable over time, we 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.

4.4.3 Choice of Functional Form

In the models estimated so far, we assumed the variables have a linear effect on the logit of $Y = 1$. In this part we use the restricted cubic splines to

test whether the continuous variables have a nonlinear effect. It turns out that credit limit has a nonlinear effect. 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 8 and 9 in appendix B. The coefficient of the nonlinear terms of credit limit is significantly different from zero, so the variable has a nonlinear effect.

4.4.4 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 3.

Table 3: Discrimination of starting and spline model.

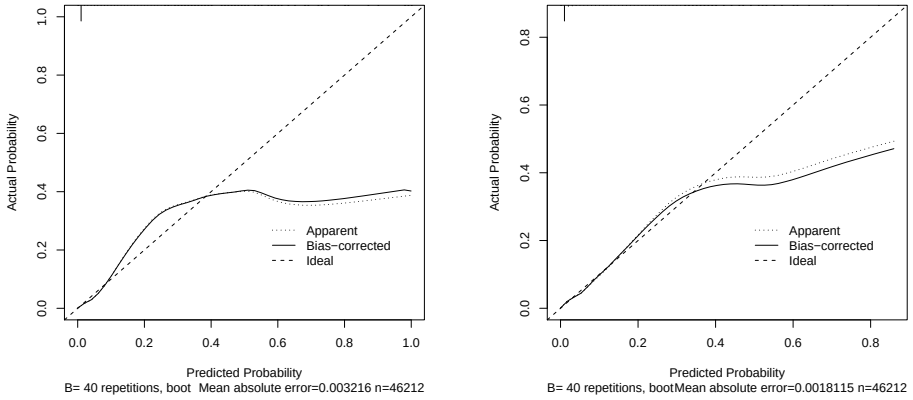
	c	B
starting model	0.914	0.015
spline model	0.917	0.015

The results 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 slightly better than the starting model.

4.4.5 Calibration

The calibration of the two models is analysed by means of calibration plots (see figures 2(a) and 2(b)). The number of observations used in the model development is $n = 46212$. The other information on the horizontal axis will be explained in section 4.4.6. The diagonal line 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 4.4.6. Both calibration plots show similar pattern. For predicted PDs above 0.4 the models are both not well calibrated. When the focus is on PDs 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 slightly better in making unbiased estimates of the PDs.

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



(a) Starting model.

(b) Spline model.

Figure 2: Calibration plots.

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.

4.4.6 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 4. 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 also show the model with spline function discriminates slightly better and is also better calibrated compared to the starting model.

Second we consider out-of-sample performance outside the time period. The data set is divided into two subsets, the first containing the years 2000,

Table 4: Out-of-sample performance within the time period.

	c	B	γ_0	γ_1
starting model - development set	0.916	0.015	0.000	1.000
starting model - test set	0.912	0.016	−0.096	0.941
spline model - development set	0.918	0.014	0.000	1.000
spline model - test set	0.919	0.016	0.000	0.973

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 a little better than the starting model.

Table 5: Out-of-sample performance outside the time period.

	c	B	γ_0	γ_1
starting model - development set	0.912	0.014	0.000	1.000
starting model - test set	0.917	0.016	0.093	0.991
spline model - development set	0.918	0.014	0.000	1.000
spline model - test set	0.921	0.015	0.116	1.006

Next we use the bootstrap method described in subsection 3.2.6 with 40 bootstrap samples. The calibration plots are shown in figures 2(a) and 2(b). The straight lines in the plots show the bias corrected calibration plot using the bootstrap method described in subsection 3.2.6. The error referred to is the mean absolute error below the horizontal axis is the difference between the predicted value and the corresponding bias-corrected value. 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 6.

Table 6: Bootstrap performance.

	c	B	γ_0	γ_1
starting model	0.915	0.015	−0.023	0.992
spline model	0.918	0.015	−0.064	0.978

Again results show that the model with spline function discriminates

little better and the starting model is better calibrated.

4.4.7 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 slightly 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 that the model with spline function performs sligthly 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, there are no clear guidelines as to what 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 and in particular we focus on valiation of the effects of risk drivers. 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 validation. Traditionally validation is focused on PDs by means of discrimination and calibration. In case of a portfolio of mortgages to individuals a bank need to estimate a logit model that forms the basis of the PDs. In this paper we argue that the parameter vector of the model also need to be validated. We validate the parameter vector by determining reproducibility of research, stability of parameters, choice of functional form, out-of-sample performance and bootstrapping.

We conclude that when the model underlying the PDs is estimated within the bank, validation can be more rigorous when it consists of two parts, validation of the PDs and validation of the parameter vector. Validation of the PDs will give information on how well the model fits the data and validation of the parameter vector will provide information on where improvement of the model can be gained. The classification given in this paper can be used to systematically validate a default model, application will lead to a better model.

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 complete case analysis to handle missing values. However, this method is only valid when

the missingness is not related to the data (observed or missing), which might not be a realistic assumption. We also assumed that the observations are independent. The data set contains information on borrowers measured on four 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.

Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics Methods and Applications*. Cambridge: University Press.

Carling, Kenneth, Tor Jacobson, Jesper Lind, and Kasper Roszbach (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking and Finance 31*, 845–868.

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 comparision of analytic and bootstrap procedures. *Journal of Econometrics 70*, 221–241.

Duffie, D. and K.J. Singleton (2003). *Credit Risk*. Princeton: Princeton University Press.

Dwyer, Douglas and Roger M. Stein (2006). Inferring the default rate in a population by comparing two incomplete default databases. *Journal of Banking and Finance 30*, 797–810.

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.

Federal Register (2004, October). Internal ratings-based systems for retail credit risk for regulatory capital. http://www.ots.treas.gov/docs/9/961141.pdf (download of June 26, 2007). p 62765.

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.

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. 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.

Seidenfeld, Teddy (1985). Calibration, coherence, and scoring rules. *Philosophy of Science 52*, 274–294.

Stein, Roger M. (2005). The relationship between default prediction and lending profits: Integrating roc analysis and loan pricing. *Journal of Banking and Finance 29*, 1213–1236.

Venables, W.N. and B.D. Ripley (2002). *Modern Applied Satistics with S*. New York: Springer.

A 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.

B Default Models

Table 7: Logit model Frieslandbank.

	coef.
intercept	−5.864
expired.duration	−0.007
expired.duration missing	−0.257
overdue.payment	3.201
mortgage.type=linear	0.514
loan.to.value	0.004
loan.to.value missing	0.544
debt.to.income	0.125

Table 8: Estimates spline model.

	coef	std.err	z	p-value
intercept	−6.530	0.147	−44.365	0.000
expired.duration	−0.006	0.001	− 5.874	0.000
credit.limit	0.066	0.004	15.098	0.000
credit.limit'	−0.175	0.012	−14.090	0.000
overdue.payment	2.212	0.124	17.912	0.000
mortgage.type=annuity	0.568	0.110	5.165	0.000
mortgage.type=life	0.233	0.096	2.426	0.015
mortgage.type=linear	0.684	0.196	3.495	0.000
mortgage.type=other	0.441	0.181	2.429	0.015
loan.to.value	0.006	0.001	6.316	0.000
debt.to.income	0.095	0.023	4.124	0.000

Table 9: Wald statistics spline model.

	χ^2	df	p-value
expired.duration	34.499	1	0
credit.limit	317.847	2	0
nonlinear	198.534	1	0
overdue.payment	320.822	1	0
mortgage.type	35.014	4	0
loan.to.value	39.896	1	0
debt.to.income	17.005	1	0
TOTAL	1606.001	10	0