



## A practical approach to validating a PD model

Lydian Medema<sup>a,c,\*</sup>, Ruud H. Koning<sup>a,c</sup>, Robert Lensink<sup>b,c</sup>

<sup>a</sup> Department of Economics and Econometrics, University of Groningen, Groningen, The Netherlands

<sup>b</sup> Department of Finance, University of Groningen, Groningen, The Netherlands

<sup>c</sup> Center of International Banking, Insurance and Finance (CIBIF), Faculty of Economics and Business, University of Groningen, P.O. Box 800, 9700 AV, Groningen, The Netherlands

### ARTICLE INFO

#### Article history:

Received 13 February 2008

Accepted 20 November 2008

Available online 7 December 2008

#### JEL classification:

E42

E58

G21

#### Keywords:

Credit risk

Probability of default

Basel II

Statistical validation

Logit 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.

© 2009 Elsevier B.V. All rights reserved.

### 1. Introduction

Since June 1999 the Basel Committee on Banking Supervision (BCBS) has published several proposals for revising the existing Basel I capital adequacy framework. The revised framework, known as Basel II (BCBS, 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.

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 as-

sess their capital adequacy and by supervisors to review such assessments. Since a smaller probability of default (PD) will result in a lower capital reserve, banks have the incentive to underestimate the PDs (Blum, 2007). Therefore 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 internal ratings based (IRB) approach of capital adequacy.

Typically, the portfolio on loans consist of loans to businesses (small, large, retail) and loans to individuals (mortgages). The main difference in the approach to determine PDs for loans to businesses and loans to individuals, stems from the fact that for businesses banks make use of external ratings (for example ratings of a credit bureau or Standard and Poor's or Moody's ratings). Carling et al. (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 businesses is concentrated on validating the PDs by measuring discrimination and calibration (Dwyer and Stein, 2006). For loans to individuals banks use a logit model to estimate the PDs. In this case validation is not only restricted to the PDs, in addition the parameter vector can also be validated. By also taking the parameter vector

\* Corresponding author. Address: Center of International Banking, Insurance and Finance (CIBIF), Faculty of Economics and Business, University of Groningen, P.O. Box 800, 9700 AV, Groningen, The Netherlands. Tel.: +31 503633811; fax: +31 503637337.

E-mail addresses: [l.medema@rug.nl](mailto:l.medema@rug.nl) (L. Medema), [r.h.koning@rug.nl](mailto:r.h.koning@rug.nl) (R.H. Koning), [b.w.lensink@rug.nl](mailto:b.w.lensink@rug.nl) (R. Lensink).

into account, validation will be more rigorous since information on how the fit can be improved is obtained. Discrimination and calibration can also be used to validate PD models for loans to individuals, but will only provide information on how well a model fits the data.

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. Since our methodology focuses also on validation of the parameter vector, it will only be applicable when a statistical model (for example a logit model) is used to estimate the PDs. However, due to data limitations, it is sometimes impossible to estimate a statistical model. For example, when a data set is small and there are only very few defaults, estimation of a logit model will be difficult and banks may use an expert model instead. An expert model is based on knowledge of experts as opposed to a statistical model which is based on historical data. An expert model does not result in an estimated parameter vector, but in an estimation of the probability of default and hence an expert model may require other and additional measures of validation. For example, explicit discussion of the role of the experts in the organization, and the reasons for manual adjustments, if any.

Managerial judgement and a qualitative analysis of the model are also highly important when evaluating a model. 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 organized as follows. Section 2 provides some background information on the Basel II accord and discusses several credit risk models. In Section 3 our proposed validation methodology will be set out. 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

In 1999 the Basel Committee proposed the Basel II accord to replace the existing Basel I accord. 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 (BCBS, 2001b). Fixed risk weights are used and no differentiation is made based on the actual risk.

Under IRB approaches, four inputs are needed for credit risk determination and capital calculations: the PD, an estimate of the loss given default, the exposure at default, and the remaining maturity of the loan (BCBS, 2001a). IRB approaches permit 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.

In the Foundation IRB Approach a bank determines the PD 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.

Banks opting for an IRB approach have to estimate the PD 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 (Basel II, §395). According to the Basel Accord a default takes place when the borrower is past due more than 90 days and/or when the borrower is unlikely to pay.

### 2.2. 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. Disadvantages of duration models are that the data sets are often too limited and that the model does not provide an estimate of the PD directly, which is required by Basel II.

The other main approach in modelling the probability of default is through classification models (an excellent overview is given in Hastie et al. (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 non-defaulters. Based on the borrower characteristics the bank determines to which population the borrower belongs. A disadvantage of this approach is that it does not yield estimated PDs.

In a probability model the PD is modelled as a function of the characteristics of the borrower. Let  $Y_{it}$  be the dependent variable which equals 1 if client  $i$  defaults between time  $t$  and  $t + 1$ , and 0 otherwise, for  $i = 1, \dots, n_t$ ,  $t = 1, \dots, T$ . Let the true model be  $\Pr(Y_{it} = 1 | X_{it}; \beta) = G(X_{it}; \beta)$ , where  $X_{it}$  is the vector of explanatory variables of client  $i$  at time  $t$ , including the intercept, and  $\beta$  is an unknown parameter vector.

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  $\Phi(\cdot)$  is the standard normal distribution function.  $\beta'X_{it}$  is sometimes referred to as the index. In practice the logit model is often assumed, this assumption is not restrictive. The true model 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 nonlinear 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 assump-

tion of a logit model is not restrictive, as long as one allows for enough flexibility in the systematic part of the model (i.e. the index).

Banks are allowed to estimate separate PD models for each loan class. 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 and a so-called expert model. 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 (Basel II, §§417, 428).

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 (Basel II, §404).

### 3. Model validation

#### 3.1. General ideas

The IRB approaches of Basel II requires banks to model the risk associated with their portfolios. Banks are required to use all relevant information to determine the risk of the portfolio (Basel II, §411). 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. Finally, based on this data set a statistical model can be estimated to 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).

Basel II requires the validation of this process (Basel II, §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. The methodology we develop in this section focuses mainly on probability models. We specifically focus on logit models, since in our application we have the task to validate such a model.

##### 3.1.1. Theoretical validation

Theoretical validation requires the review of the theories and assumptions underlying the proposed model. This corresponds with §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 estimate PDs conservatively (Basel II, §411).

Reviewing the assumptions underlying the model is part of the theoretical validation. The use of the logit model to estimate PDs typically assumes the observations to be independent. However, this assumption is violated since the data available for model estimation often contains observations at several points in time.

##### 3.1.2. Data validation

Data validation is about the data underlying the model. The data must be validated (Basel II, §417) and banks must show that

the data used are representative for the underlying population. We distinguish three parts of data validation: representative data, appropriateness of the variables, and completeness of the data set.

*3.1.2.1. Representative data.* Banks can use either internal data or external data to estimate the model. Basel II allows the use of external data (§§448, 463), but it requires banks to demonstrate that the data are representative. 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. The sampling procedure has to be reviewed to determine whether the subset is representative of the underlying population.

*3.1.2.2. Appropriateness of the variables.* At a minimum borrower characteristics, transaction risk characteristics, and delinquency of exposure have to be considered as explanatory variables in a PD model (Basel II, §402). Examples of borrower characteristics are age, income, marital status, and occupation. Transaction risk characteristics are for example mortgage type, loan to value, and payment history. Several problems may arise with the variables. First, the values of a variable can change over time. Second, some variables are difficult to measure. For example measurement of default itself is difficult. According to Basel II (§452) default occurs 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.

*3.1.2.3. Completeness of the data set.* Basel II requires the length of the underlying historical observation period to be at least five years (Basel II, §463). In practice it might be that banks have information on less than five years. 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, §§451, 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 (Basel II, §411). 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. Incomplete cases are removed from the data set. 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.

#### 3.2. Statistical model validation

In general a model is not able to reproduce the exact data underlying the model. To determine the accuracy of the model several statistical tests are available in the literature. We base this section on [Harrell \(2001\)](#), [BCBS \(2005\)](#) 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, [BCBS \(2005\)](#) 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. In the existing literature models are validated

by determining discrimination and calibration of the model. A model's discrimination is its 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 if the fraction of events that actually occur, is estimated unbiasedly by the estimated probability of these events. 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 ( $\beta$ ) can also be validated. We validate the parameters by examining the reproducibility of research, stability of parameters and choice of functional form. Besides we describe out-of-sample performance and we use the bootstrap method to validate the PDs as well as the parameters.

### 3.2.1. Reproducibility of research

Reproducibility, or replication, of research is defined as the duplication of the results of a former study (McCullough et al., 2006). 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. 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 et al. (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.

### 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 interested in stability over time for a subvector of the parameter vector  $\beta$ . For example one may be interested in stability of the effect of the explanatory variable loan to value. The likelihood ratio test can be used to test for stability of the parameters. However, in general the change point is unknown. Andrews (1993) describes how to use the likelihood ratio test to test for stability of parameters over time when the change point is unknown. First the likelihood ratio is determined for each possible change point, resulting in  $T - 1$  values of the likelihood ratio. Next the change point is estimated by the change point with the highest likelihood ratio (denoted by  $LR^{max}$ ). Following Diebold and Chen, 1996 the distribution of  $LR^{max}$  can be approximated using the bootstrap method.

To test whether the model is stable over groups the likelihood ratio test can also be used. 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 too difficult to describe using simple transformations, spline functions can be used (see Harrell, 2001). Restricted cubic spline functions make the model flexible and use piecewise polynomials to fit a highly curved function. That is, they fit polynomials on intervals of the variable which has a nonlinear effect. The splines are fitted such that they are smooth, and the first and second derivatives exist at the knots of the intervals.

An alternative specification test would be obtained by comparing results of the estimated logit model with nonparametric or semiparametric alternatives. If discrimination or calibration of such a model would be better than the one proposed one could argue that the functional form of the proposed model is too restrictive. However, estimation of a non- or semiparametric model may require more data than are typically available. Because defaults are rare events, the effective number of observations is much smaller than the number of loans in a portfolio (Cramer, 2004). For this reason, we are somewhat reluctant to use non- or semiparametric alternatives to a logit model, such as trees. The variance of such classification models tends to be higher (Hastie et al., 2001). Moreover, the supervisor states in De Nederlandsche Bank N.V. (2005): "Validation also consists of a qualitative analysis of the model by means of evaluating the transparency (no black box) and intuition of the model" (p. 7, translation ours) and trees are not transparent or provide guidance on the importance of different risk drivers. For these reasons, we prefer functional form tests based on a flexible specification of the index as discussed in this subsection. By adding nonlinear terms and interactions, most if not all reasonable relations between covariates and the probability of default can be estimated.

### 3.2.4. Discrimination

Discrimination of a model is its ability to separate subjects' outcomes (Harrell, 2001). Several statistics are available to determine discrimination and calibration. Table 1 gives an overview of the statistics proposed by Harrell (2001), BCBS (2005), and Engelmann and Rauhmeier (2006).

We will discuss 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 references.

A graph used to determine the discrimination of the model is the Receiver Operating Characteristic (ROC) curve. Let a borrower

**Table 1**  
Discrimination and calibration statistics.

	Discrimination	Calibration
BCBS (2005)	Cumulative Accuracy Profile, Accuracy 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	$\alpha_0$ and $\alpha_1$ refitted model
	Brier score	$E_{max}$ Generalized $R_N^2$
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

be classified as a defaulter if the estimated PD exceeds  $C$  and as a non-defaulter otherwise (frequently one uses  $C = 0.5$ ). The borrowers classified as defaulters can be split into two groups, defaulters which are correctly classified as defaulters and non-defaulters which are incorrectly classified as defaulters. The borrowers classified as non-defaulters can also be split into two groups, non-defaulters which are correctly classified and defaulters which are incorrectly classified. The percentage of defaulters that 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 (Blöchlinger and Leippold, 2006; 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 (omitting the time index for simplicity):  $B = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - Y_i)^2$ , where  $\hat{p}_i$  is the estimated probability of observation  $i$ .  $B$  is the average of the 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 well. Brier score can also be used to determine the discrimination of a rating system with borrower grades (Engelmann and Rauhmeier, 2006).

### 3.2.5. Calibration

Calibration is the ability of the model to make unbiased estimates of the PDs. A set of probabilities are (well) calibrated if  $p$  percentage of all predictions reported at probability  $p$  are true. Traditionally, the fit of a logit model is often analysed by a classification table. A classification table is a  $2 \times 2$  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 modelling, data sets are highly unbalanced in the sense that only a small fraction defaulted on their contracts. 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.

Instead we describe a refitting method which can be used to determine the calibration (Harrell, 2001). Suppose the original data set is split into a development set  $D$  and a test set  $T$ .  $\hat{\beta}$  is the maximum likelihood estimator for  $\beta$  based on the development set. If the model is well calibrated, defaults in the test set are well estimated using these parameters:  $\hat{p}_i = A(\hat{\beta}'x_i)$ ,  $i \in T$ . Unbiasedness of this predictor for the actual default events in the test set can be assessed by noting  $\Pr(Y_i = 1 | x_i, \hat{\beta}) = A(-\gamma_0 - \gamma_1 \ln(\frac{1}{\hat{p}_i} - 1)) = A(-\gamma_0 - \gamma_1 \hat{\beta}'x_i)$ ,  $i \in T$  with  $\gamma_0 = 0$  and  $\gamma_1 = 1$ .  $\gamma_0$  and  $\gamma_1$  are estimated easily, and the test is straightforward to implement.

A graphical tool to determine the calibration is the calibration plot. In such a plot, the relation between the estimated default probabilities (horizontal) and the actual outcome (vertical) is graphed. This relation can for example be the line connecting the average default frequencies in each decile of the estimated probabilities, or a nonparametric regression of actual default on the estimated probability. The calibration plot is useful to determine in what region of the estimated probabilities the model provides a good fit.

### 3.2.6. Out-of-sample performance

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 observations. 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 (Efron and Tibshirani, 1993).

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 (§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 size. To overcome this disadvantage we use the enhanced bootstrap method of Efron and Tibshirani (1993), which performs better than the simple bootstrap (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. Monitoring the performance of the estimated model

Section 4 will provide an empirical example of the methodology we have set out so far. In line with the main aim of this paper, the

empirical example will primarily focus on the initial statistical validation of the model. However, after the initial validation of the model, the model will be implemented and periodical validations of the developed risk model need to be organized. Although validation of the performance monitoring is outside the scope of this paper, we want to make some general remarks on this issue.

After the implementation of the risk model, the bank needs to determine when and how often periodical validation of the model takes place. The Basel II agreement requires that validation takes place at least once per year, but depending on the type of the model more frequent validations may be necessary. During the year, the bank needs to monitor the performance of the model, and can decide to bring the periodical validation forward. Different organizations within the bank, such as internal audit and risk management, may have important tasks with respect to the performance monitoring and the decision to speed up the periodical validation. The performance monitoring deals with several issues, such as (1) an analysis of the outcomes of the model; (2) a comparison of PDs with realized default rates; (3) the degree to which the model is actually used; (4) the availability of new data; (5) an identification of the part of the portfolio for which the model does not perform well, for instance by using backtests and (6) identification of new risk factors that possibly need to be added to the risk model.

The validation team needs to evaluate to what extent the performance monitoring is adequately organized within the bank, needs to assess the analyses resulting from the performance monitoring, and needs to examine whether the periodical validations are arranged.

## 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. In order to be able to qualify for Foundation IRB, the bank has been developing risk models. The default risk on mortgage model we discuss here is one of their earlier models. In this model, one-year default probabilities are estimated. The bank used the software package SAS, for this empirical analysis we use *R* (Copyright 2008, the *R* Foundation for Statistical Computing, version 2.7.0).

### 4.1. Description of the data

The data set consists of a number of snapshots, taken at different moments in time. 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 description of the variables can be obtained on request.

The bank modelled default risk in this portfolio with a logit model. The model contained the following risk factors: loan to value, loan to income, expired duration, mortgage type and an indicator whether a payment was overdue. These variables are standard risk factors. The variable mortgage type was a dummy variable that distinguishes between linear mortgages and other mortgages. The model contained two additional variables: indicators for loan to value missing and expired duration missing. All variables were statistically significant, and the coefficient of concordance was 0.89.

For our analysis, we estimated a slightly different model, with expired duration, credit limit, age of the borrower, overdue payment, mortgage type, loan to value, and loan to income as explanatory variables. Mortgage type is measured through four dummy variables, representing annuity, life, and linear and other mort-

gages respectively. The reference category is the interest-only mortgage. 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 Table 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 start model. Even though our model specification differs slightly from the one proposed by the bank, qualitatively the effects of the covariates are similar.

### 4.2. Theoretical validation

The results of the start model show that expired duration has a negative relationship with the probability of default. This means that when a mortgage matures the PD is lower. The binary variable overdue payment has a positive influence on the PD. So when a mortgage is in arrear the PD 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 cannot determine for which case the value is missing and for which case the value truly is 0. The variables which are not measured correctly cannot 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. Implicitly, we assume that the mechanism that generates missingness of variables is unrelated to both the probability of default, and the variable itself. To estimate a model with proper allowance for incomplete observations is beyond the scope of this paper.

### 4.4. Statistical validation

#### 4.4.1. Reproducibility of research

We cannot reproduce the outcome of the model of the bank precisely. One reason is that the data we use are different from the data used by the bank. A second reason is how missing values are treated. We used complete case analysis to handle missing

**Table 2**  
Estimation of starting model and spline model.

Independent variables	Start model		Spline model	
	Coefficient	S.D.	Coefficient	S.D.
Constant	-6.336**	(0.143)	-6.530**	(0.147)
Expired.duration	-0.005**	(0.001)	-0.006**	(0.001)
Credit.limit	0.007**	(0.001)	0.066**	(0.004)
Credit.limit – nonlinear			-0.175**	(0.012)
Overdue.payment	2.961**	(0.110)	2.212**	(0.124)
Mortgage.type=annuity	0.600**	(0.110)	0.568**	(0.110)
Mortgage.type=life	0.269*	(0.096)	0.233*	(0.096)
Mortgage.type=linear	0.657*	(0.195)	0.684**	(0.196)
Mortgage.type=other	0.435*	(0.180)	0.441*	(0.181)
Loan.to.value	0.006**	(0.001)	0.006**	(0.001)
Debt.to.income	0.099**	(0.023)	0.095**	(0.023)

Standard deviations of the coefficients (S.D.) in brackets.

\* Variable significant at 5%.

\*\* Variable significant at 1%.

values. The bank used some kind of imputation method, so they included some additional information. These issues are minor and were readily corrected.

4.4.2. Stability of parameters

We test whether the parameter vector  $\beta$  is stable over time, the value of the test statistics is 12.630, with a  $p$ -value of 0.372 (based on 2000 bootstrap samples). So, the null hypothesis of an unknown break is rejected.

4.4.3. Choice of functional form

In the models estimated so far, we assumed that 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. The results are shown in Table 2. It can be seen that the coefficient of the nonlinear terms of credit limit is significantly different from zero.

4.4.4. Discrimination

In the analysis above we developed two models, one is the start 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 ( $B$ ). For the start model we find  $c = 0.914$  and  $B = 0.015$ , and for the spline model we find  $c = 0.917$  and  $B = 0.015$ . 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 slightly higher compared to the start model, so the model with spline function discriminates slightly better than the start model.

4.4.5. Calibration

The calibration of the two models is analysed by means of calibration plots (see Fig. 1a and b). Only a very small fraction of predicted probabilities is above 0.5, therefore the plots show the calibration for probabilities below 0.5. The number of observations

used in the model development is  $n = 46,212$ . The other information on the horizontal axis will be explained in the next subsection. The diagonal line show the ideal case of perfect calibration. The dotted line shows the apparent calibration of the model based on a nonparametric regression of the default event on the estimated probability (loess-smooth). The straight line will be discussed below. Both calibration plots show similar pattern. For predicted PDs above 0.35 the models are both not well calibrated. When the focus is on PDs below 0.35 we see the model with spline function is slightly better calibrated than the start 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. In this example, calibration is better for lower probabilities. Because the estimate of the calibration curve is based on a nonparametric smoothing method, there is no straightforward numerical summary of calibration. Perhaps an appealing ad-hoc measure could be the maximum distance between the calibration curve and the 45° line, perhaps by decile of the predicted probabilities. We prefer just to present the graphs, without numerical summary.

4.4.6. Out-of-sample performance

First 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 3. The table also shows the measures calculated on the development set. Note that  $\gamma_0$  and  $\gamma_1$  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 start 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, 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

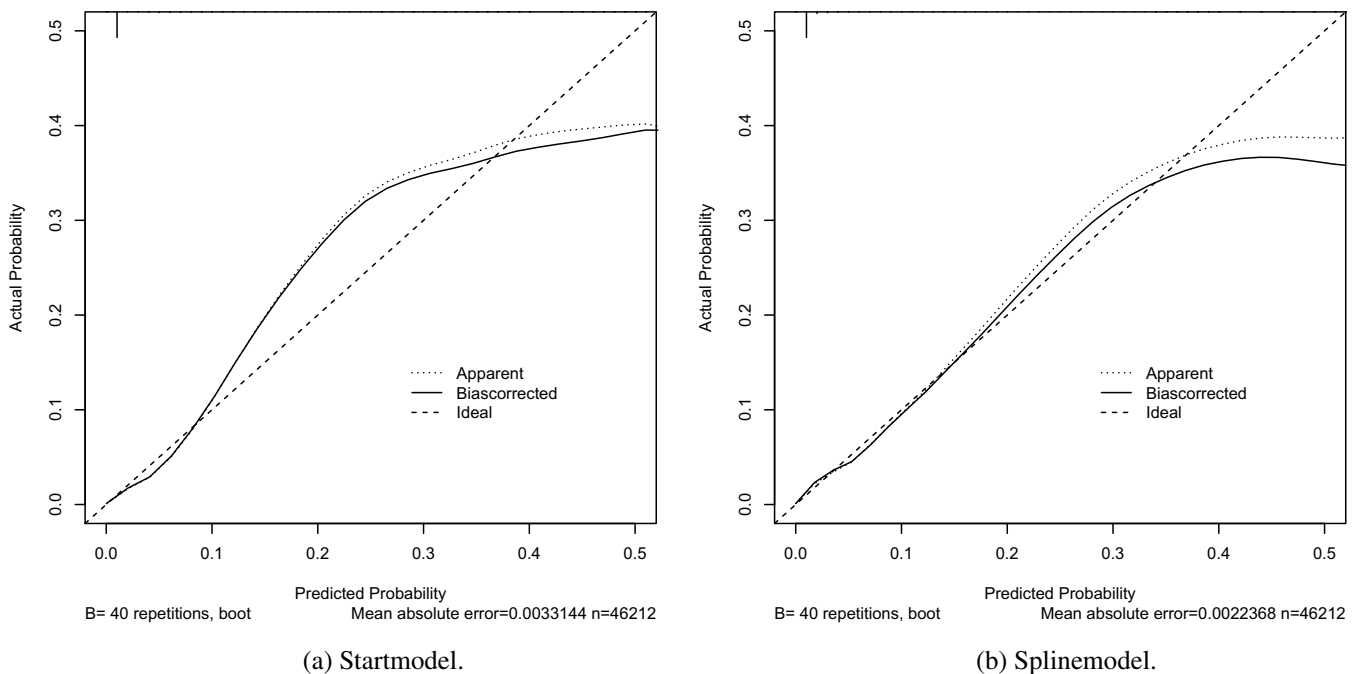


Fig. 1. Calibration plots.

**Table 3**  
Out-of-sample and bootstrap performance.

Model	Data set	$c$	$B$	$\gamma_0$	$\gamma_1$
<i>Within time</i>					
Start model	Development	0.916	0.015	0.000	1.000
	Test	0.912	0.016	-0.096	0.941
Spline model	Development	0.918	0.014	0.000	1.000
	Test	0.919	0.016	0.000	0.973
<i>Outside time</i>					
Start model	Development	0.912	0.014	0.000	1.000
	Test	0.917	0.016	0.093	0.991
Spline model	Development	0.918	0.014	0.000	1.000
	Test	0.921	0.015	0.116	1.006
<i>Bootstrap</i>					
Start model		0.915	0.015	-0.023	0.992
Spline model		0.918	0.015	-0.064	0.978

the model with spline function performs a little better than the start model.

Next we use the bootstrap method described in Section 3.2 with 40 bootstrap samples to see whether these conclusions are biased by too much optimism. The calibration plots are shown in Fig. 1a and b. The straight lines in the plots show the bias corrected calibration plot using the bootstrap method described in Section 3.2. The error refers to 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 start model. The results of the measures mentioned above are shown in Table 3. Again results show that the model with spline function discriminates slightly better and the start model is better calibrated.

The overall conclusion we can draw from the application is that the model with spline function performs slightly better than the start model. Focus on the slope coefficients shows a significant nonlinearity in one of the regressors. Discrimination of the model with the nonlinear term is slightly higher (as judged by the coefficient of concordance) and calibration is better as well. Based on the coefficient of concordance we can conclude that the model with spline function performs slightly better than the model proposed by the bank.

## 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 analyse and validate logit models and in particular we focus on validation 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 keep the calculations 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, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821–856.
- BCBS, 2001a. The Consultative Document: The Internal Ratings-based Approach. <[www.bis.org/publ/bsbca05.pdf](http://www.bis.org/publ/bsbca05.pdf)> (download of August 15, 2005).
- BCBS, 2001b. The Consultative Document: The Standardised Approach to Credit Risk. <[www.bis.org/publ/bcbca04.pdf](http://www.bis.org/publ/bcbca04.pdf)> (download of August 15, 2005).
- BCBS, 2005. Working Paper No. 14: Studies on the Validation of Internal Ratings Systems, February.
- BCBS, 2006. International Convergence of Capital Measurements and Capital Standards: A Revised Framework Comprehensive Version, June.
- Blöchliger, A., Leippold, M., 2006. Economic benefit of powerful credit scoring. *Journal of Banking and Finance* 30, 851–873.
- Blum, J.M., 2007. Why Basel II may need a leverage ratio restriction. *Journal of Banking and Finance* 32, 1699–1707.
- Carling, K., Jacobson, T., Lindé, J., Roszbach, K., 2007. Corporate credit risk modeling and the macroeconomy. *Journal of Banking and Finance* 31, 845–868.
- Cramer, J.S., 2004. Scoring bank loans that may go wrong: A case study. *Statistica Neerlandica* 58, 365–380.
- De Nederlandsche Bank N.V., 2005. Bazel II: Governance rond modelontwikkeling, -validatie en gebruik.
- Dewald, W.G., Thursby, J.G., Anderson, R.G., 1986. Replication in empirical economics: The journal of money credit and banking project. *The American Economic Review* 76, 587–603.
- Diebold, F.X., Chen, C., 1996. Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. *Journal of Econometrics* 70, 221–241.
- Duffie, D., Singleton, K.J., 2003. *Credit Risk: Pricing Measurement and Management*. Princeton University Press, Princeton, NJ.
- Dwyer, D., Stein, R.M., 2006. Inferring the default rate in a population by comparing two incomplete default databases. *Journal of Banking and Finance* 30, 797–810.
- Efron, B., 1990. More efficient bootstrap computations. *Journal of the American Association* 85, 79–89.
- Efron, B., Tibshirani, R.J., 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Engelmann, B., Rauhmeier, R., 2006. *The Basel II Risk Parameters Estimation Validation and Stress Testing*. Springer, Heidelberg.
- Harrell, J.F.E., 2001. *Regression Modeling Strategies*. Springer, New York.
- Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
- Little, R.J.A., Rubin, D.B., 2002. *Statistical Analysis with Missing Data*. Wiley, New York.
- McCullough, B.D., McGeary, K.A., Harrison, T.D., 2006. Lessons from the JMBC archive. *Journal of Money, Credit, and Banking* 38, 1093–1107.
- Stein, R.M., 2005. The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking and Finance* 29, 1213–1236.