Credit scoring and the sample selection bias

Thomas Parnitzke*

Institute of Insurance Economics
University of St. Gallen
Kirchlistrasse 2, 9010 St. Gallen, Switzerland

this version:
May 31, 2005

Abstract

For creating or adjusting credit scoring rules, usually only the accepted applicant's data and default information are available. The missing information for the rejected applicants and the sorting mechanism of the preceding scoring can lead to a sample selection bias. In other words, mostly inferior classification results are achieved if these new rules are applied to the whole population of applicants. Methods for coping with this problem are known by the term "reject inference." These techniques attempt to get additional data for the rejected applicants or try to infer the missing information. We apply some of these reject inference methods as well as two extensions to a simulated and a real data set in order to test the adequacy of different approaches. The suggested extensions are an improvement in comparison to the known techniques. Furthermore, the size of the sample selection effect and its influencing factors are examined.

Keywords: credit scoring, sample selection, reject inference

JEL Classification: C51, G21

*Corresponding author. Tel.: +41-71-243-4090; fax: +41-71-243-4040. *E-mail address*: thomas.parnitzke@unisg.ch

1 Introduction

Credit scoring methods used to separate good from bad risks are commonly used for the scoring of standardized credit products. The aim of creating new selection rules is to find a mechanism that perfectly separates the applicants who will fully repay from defaulters. Probably, a perfect mechanism cannot be found due to certain, however small, default probabilities even for very good risks, but it is worth trying to find or to improve a credit scoring system that is superior to other possible ones. Another important reason for such a search is that a good-working and bias-free rating system is one of the prerequisites required for an internal ratings-based approach by the Basel Committee on Banking Supervision.[1] The use of such an internal approach can lead to a lower capital requirement in comparison with the exclusive application of external rating information, which leaves many unrated positions and requires a higher demand of capital.

In general, the developer of a credit scoring system posseses solely the default information of accepted applicants. Whenever the presently operating scoring system is better than a random assignment of applicants to different groups, this will lead to selection-based bias and to inferior classification results for the next scoring model. Methods for coping with this problem are known as reject inference techniques. These techniques attempt to get additional data for rejected applicants or try to infer the missing default information. Until now there has been little literature that compares the effects of these methods empirically. Examples are Ash and Meester (2002), Banasik et al. (2003), and Crook and Banasik (2004). The most common reject inference methods – enlargement, reweighting, and extrapolation – will be applied here to a simulated and a real data set with the aim of testing the adequacy of these different approaches. Additionally, two new extensions to existing methods will be used for the correction of the sample selection effect. These modifications achieve improved results in comparison to the other techniques for our data samples. Due to unexpected and not yet fully described outcomes arising from setting the sample size and cutoff, the sample selection will be analyzed in the context of influencing factors.

[1] Basel Committee on Banking Supervision (2003), p. 72: "The rating definitions and criteria must be both plausible and intuitive and must result in a meaningful differentiation of risk."; p. 74: "The model must be accurate on average across the range of borrowers or facilities to which the bank is exposed and there must be no known material biases."

Section 2 describes the basic ideas of sample selection and summarizes the most common performance measures for the classification success of a credit scoring system. Different reject inference techniques are introduced in Section 3. These techniques are then applied to the mentioned data sets in Section 4. In Section 5, the determinants of the sample selection effect will be analyzed. The results will be recapitulated in Section 6.

2 Sample selection bias

The term "credit scoring" describes statistical methods used for the classification of credit applicants (see Hand and Henley, 1997). Credit scoring, as applied here, assigns the potential borrower a score S (points) based on individual input data $X_{ij} = (x_{i1}, ..., x_{in})$. For the scoring process we use logistic regression. The calculated score $S_i(X)$ thus should be correlated with the probability of repayment π_i. Low values of S_i are expected to correlate with low probabilities of repayment and vice versa. It is the lender's business to fix a special barrier c, which will determine the applicant's creditworthiness.

The examined credit scoring process consists of two stages illustrated by the questions: Did the applicant obtain credit? Was the credit repaid? The granting decision A depends, as mentioned, on c:

$$A = \begin{cases} 0 & \text{if } S_i < c \text{ credit not granted} \\ 1 & \text{if } S_i \geq c \text{ credit granted.} \end{cases}$$

After issue of the credit and a defined time period t, we will get for the group of accepted applicants ($A = 1$) two possible outcomes:

$$Y = \begin{cases} 0 & \text{default} \\ 1 & \text{nondefault.} \end{cases}$$

For building a new or adjusting the old scoring model we can use only the information for the population ($A = 1$) and their status information Y. Due to sample selection induced by credit scoring we will get biased results, leading to

a scoring rule suited for the population ($A = 1$) but not for the population of applicants as whole ($A = 1$ and $A = 0$). This case is described by Greene (2003) as "incidental truncation".

Another common classification approach is that of Little and Rubin (1987). For missing default information, they define three types of cases:

- MCAR (missing completely at random): The values of Y are missing at random and the missing information does not depend on X. The probability of being selected in group ($A = 1$) is identical for all cases.
- MAR (missing at random): The missing of the Y value and the probability of being accepted depends on X. The fraction of ($Y = 1$) for each subgroup ($A = 1$) and ($A = 0$) remains unchanged.

$$P(Y = 1|A = 1) = P(Y = 1|A = 0)$$

- MNAR (missing not at random): The missing of the Y characteristics depends on X and Y. This is the described case of sample selection. The selection is contingent on the use of a credit scoring model based on Y and X. The fraction of ($Y = 1$) is changed due to sample selection.

$$P(Y = 1|A = 1) \neq P(Y = 1|A = 0)$$

In assessing credit scoring systems, one can distinguish counting measures from separating measures. Separating measures include the Lorenz-curve and the receiver-operating-characteristic, measures based on them such as the Gini-coefficient and the accuracy-ratio, as well as the discriminatory power (see Kraft et al., 2002). The simplest counting measure is the 2×2 contingency table, shown in Table 1. It represents the credit decision A tied to the credit status Y and thus the success of the scoring system. The main diagonal (n_1 and n_4) of this table shows the correctly classified applicants; the secondary diagonal (n_2 and n_3) displays wrongly classified cases. Even though there is only limited use of the 2×2 contingency table in the real world, because of the unobservability of n_1 and n_3, this form of presentation will be useful for the purpose of this examination. For the simulated data used here, the outcomes Y of group ($A = 0$) are fully

	$A = 0$	$A = 1$
$Y = 0$	n_1	n_2
$Y = 1$	n_3	n_4

Table 1: Example contingency table

observable.

3 Reject inference techniques

The simplest approach for coping with sample selection is to grant credit to all applicants for a short time period (see, e.g., Rosenberg and Gleit, 1994). However, this approach is not feasible in the real world because of its high financial risks.

Hand (2002) suggests a soft-accept-reject threshold for the improvement of credit scoring models. On the basis of the applicant's score S_i, a probability $p(S_i)$ for accepting or rejecting the credit request will be computed. Credit applicants with lower values of S_i have a lower chance of being granted credit. This method attempts to improve model results through additional data, but again there exists the possibility of financial risk due to accepting high-risk cases.

The reweighting method, in various realizations, is well known and widely applied. Crook and Banasik (2004) have described a method in which the accepted applicants and their default information are used in the determination of the new model with the inverse of the probability $p(S_i)$, hence $1/p(S_i)$. This method gives cases near the cutoff a higher weight, with the idea that these cases are nearer to the credit situation of a rejected applicant.

Reclassification assigns the status $Y = 0$ to the $x\%$ lowest scores, with the assumption that this group will definitely be defaulters. However, this method can lead to a considerable bias because the $x\%$ lowest scores are not de jure defaulters (see, e.g., Ash and Meester, 2002).

Various types of extrapolation are also used. For example, the method described by Ash and Meester (2002), as well as by Crook and Banasik (2004), is based on posterior probabilities of default, which were extrapolated for the rejected

credit applicants. By setting a cutoff, the population ($A = 0$) will be divided into good and bad risk cases. The information from this division, together with the observed values Y from group ($A = 1$), determines the new model, but leads to only small improvements (see Crook and Banasik, 2004).

Another method suggested by Ash and Meester (2002) is the Heckman approach. Heckman (1979) considers the problem of sample selection as a problem of dropped-down variables. Although this method is designed for continuous, not dichotomous, variables – what harms the assumptions made – the idea can be used for model building by adding a variable representing the selection mechanism. Further, Boyes et al. (1989) used a censored bivariate probit model to determine the probabilities of default, a method based on the work of Poirier (1980). This model understands the problem of sample selection as a case of partial unobservability. As with the Heckman approach, the underlying selection mechanism will be incorporated, but in a bivariate probit estimation of scoring parameters.

Two extensions based on the techniques described above will be presented in this work. The first approach combines the soft-accept-reject threshold suggested by Hand (2002) with the reweighting technique of Crook and Banasik (2004). The second concerns a modified extrapolation. The extrapolated posterior probabilities of repayment were used for simulating the outcomes Y, which will be an additional data basis for the parameter estimation of the new scoring model.

4 Reject inference techniques applied

4.1 Overview

In this section, some reject inference methods are applied to two different data sets. We concentrate on the soft-accept-reject threshold, reweighting, and a combination of both. Additionally, a version of extrapolation is used. The consideration of granting a credit to all applicants and to build a scoring model on this information and so get a model unbiased from sample selection is adopted as a benchmark for performance measurement.

4.2 Simulated data

The analysis of this section is based on generated, normal-distributed data. Each data set consists of 20,000 cases so as to achieve stable results and be fairly realistic. The sample is divided into a training and a test subsample, each consisting of 10,000 cases. For every case i, four characteristics x_j were generated. The values of x_1 represents the true but unobservable score, and the combined true probabilities of repayment π_i as well as the probability of default $1-\pi_i$. The idea applied here is that in real life, the probability of default is unobservable, due to many influencing factors x, such as personal characteristics, which are rarely observable and/or incapable of measurement. The variables $x_{2,3,4}$ represent observable parameters, such as income or fixed expenses, correlated up to a certain degree with the true score x_1. This connection can be displayed by the following correlation-matrix:

$$\rho = \begin{pmatrix} 1 & 0.6 & -0.7 & 0.2 \\ & 1 & -0.6 & 0.3 \\ & & 1 & -0.7 \\ & & & 1 \end{pmatrix}.$$

Using the true score and the logit distribution function

$$\pi_i = \frac{e^{x_1}}{1+e^{x_1}},$$

the real probabilities of repayment are calculated. Then, the real outcome of the credit status information Y is simulated out of π_i with a random generator, assuming that real life is random in a similar way. The random generator assigns approximately 23.5% of the population as defaulters. To score the cases, we start with the following rule:

$$S_i = -0.2x_2 - 0.8x_3 + 0.6x_4.$$

This rule is better than random assignment, but worse in comparison to a logit model determinable from the whole data set. The probability of acceptance $p(S_i)$,

estimable out of the observed variables with the help of S_i, can be defined by

$$p(S_i) = \frac{e^{S_i}}{1 + e^{S_i}} \,.$$

By setting a cutoff c after the 25% lowest scores, the sample will divided as described into rejected ($A = 0$) and accepted ($A = 1$) applicants (see also Figure 1). After the time period t has virtually elapsed, the generated outcome Y for the

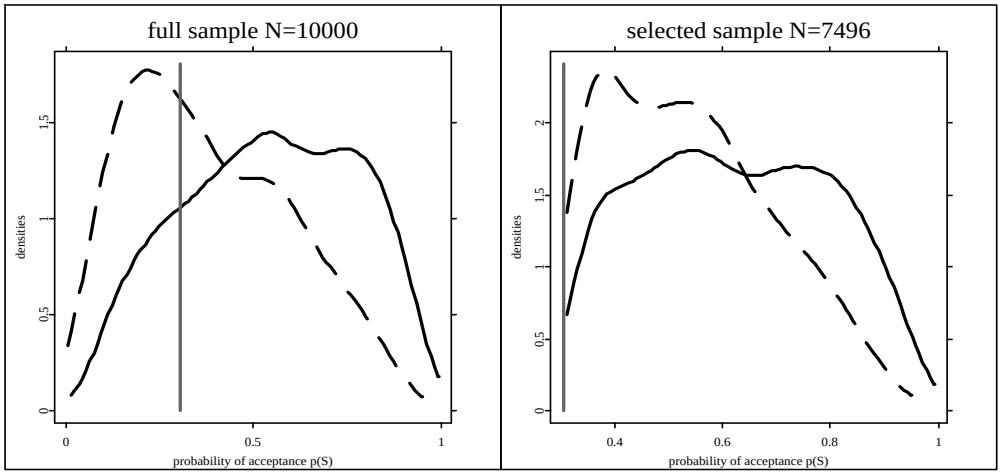

Figure 1: Example for the simulated data $Y = 0$ (dashed), $Y = 1$ (solid), and the cutoff c (grey)

group ($A = 1$) can be used in combination with the variables x_j to find new β_j for the logit-scoring function ($S_i = X_{ij}\beta_j$). These new scoring parameters will be applied to the test data set. The lowest 25% of the score values were again declared as rejected applicants. For stable results, the simulation was run 100 times. The advantage of this simulation is that the correctness of the assignment for all fields of the 2×2 contingency table can be compared to a perfect model (derived from the whole data) simultaneously.

Table 2 shows a classification success of 71.8% for the first scoring function. The term "classification success" is used to mean correctly allocated cases, which means $A = 0 \cap Y = 0$ or $A = 1 \cap Y = 1$. Next, models were built based on group ($A = 1$) data and on the whole data, which is, of course, impossible in reality but feasible for this data set. Scoring the test data set with the different

	$A=0$	$A=1$
$Y=0$	10.22%	13.12%
$Y=1$	15.08%	61.58%

Table 2: Results of the first score function

rules leads to the results represented in Tables 3 and 4.

	$A=0$	$A=1$
$Y=0$	14.54%	8.84%
$Y=1$	10.46%	66.16%

Table 3: Test data results based on the model derived from the full sample

	$A=0$	$A=1$
$Y=0$	13.34%	10.04%
$Y=1$	11.66%	64.96%

Table 4: Test data results based on the model derived from the selected subsample

The full sample model achieved 80.7% correctly classified cases; the model from the truncated data achieved only 78.3%. The results for the two models differ 2.4 percentage points due to the sample-selection-induced loss of information. The classification success of these two models is hereafter used as a benchmark for the applied reject inference methods.

Enlargement

The following section describes the application of Hand's (2002) soft-accept-reject threshold. By using the probabilities of acceptance $p(S_i)$, computed from the score values of the first model, additional applicants from group $(A = 0)$ were selected into group $(A = 1)$. Applying a random generator on $p(S_i)$, additional 19.5% of $(A = 0)$ or, approximately 5% of the whole population, has an access to credit. For a graphical presentation of the changed densities, see Figure 2. The additional data leads to 79.12% correctly classified cases. This is 0.82 per-

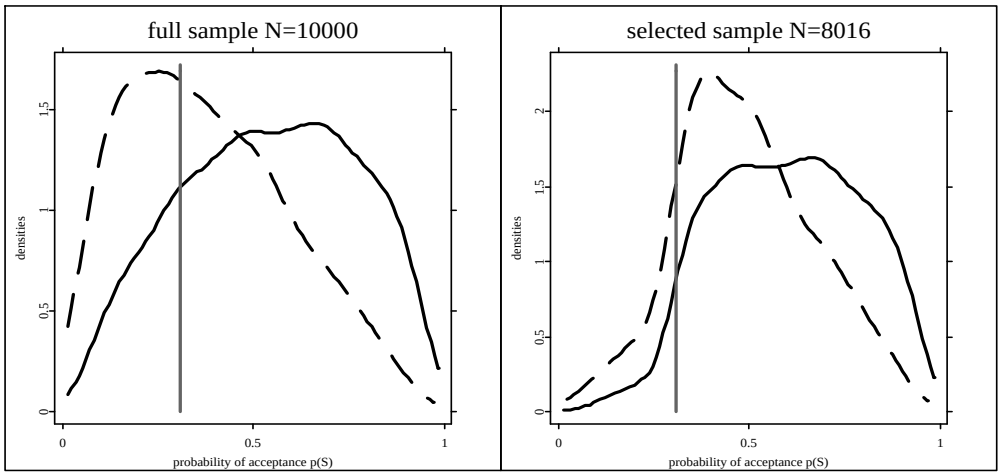

Figure 2: Granting of credit with a soft-accept-reject threshold

centage points better than the result reached by the pure subsample model. An increase or decrease in additional data for the first scoring leads to improvement or impairment of the test data results. For example, with a supplementary 1.95% of the primarily rejected applicants, the results can be enhancend by only 0.18 percentage points.

Reweighting

To test the reweighting model, the method described by Crook and Banasik (2004) was implemented. The selected subsample cases were reweighted by $1/p(S_i)$ the inverse probability of acceptance refined from the first scoring function. This reweighting led to a melioration of 1.07 percentage points for the test data sample.

Enlargement with reweighting

This section describes the combination of the soft-accept-reject threshold proposed by Hand (2002) with the reweighting of $1/p(S_i)$ used in Crook and Banasik (2004). A priori accepted applicants ($A = 1$) are given the weight 1, because they are, dependent on the cutoff, assured of being in group ($A = 1$). The cases ac-

cepted on the basis of $p(S_i)$ were reweighted with the inverse of this probability, on the theory that randomly picked cases can represent many cases around them through reweighting. This will work well if the applicants around the chosen one are highly homogeneous. With the already mentioned 19.5% of previously rejected applicants and applying the reweighting technique, a classification success of 80.66% was reached. This is very little different from the results of the model based on the full sample, which reached a classification success of 80.7%. By applying only 10% of this additional data or 1.95% of the rejected applicants and using the associated weights multiplied by 10, a result of 80.28% was reached. With a few additional cases the results can be considerably improved.

Extrapolation

In the following the concept of extrapolation is used in a modified way. The posterior probabilities of repayment were extrapolated for the rejected applicants. However, instead of setting a cutoff, the status Y was simulated for the rejected applicants. Here we assume that personal probability of default leads to repayment or default by chance. The groups $(A = 1)$ with the observed default status and $(A = 0)$ with the simulated status Y can be used to build a new scoring model.

For the extrapolation, the group $(A = 1)$ was sorted after the score values created by the first score function. The sorted data were then divided into bands for which the observed probability of repayment was calculated. These numbers, in connection with the mean $p(S_i)$ of every band, were used for an OLS-parameter estimation. With the derived parameters, the probabilities of being a good risk were extrapolated for the scores of the rejected applicants. Based on these probabilities, a random generator creates the status Y.

The new model based on the additional information achieves 80.48% correctly classified cases for the test data set, results comparable to the improvements obtained by the enlargement with reweighting method.

4.3 Empirical Data

With the aim of verifying the results achieved so far, another real data set is deployed. The data set, from Fahrmeir and Hamerle (1984), consists of 1,000 cases. The data set is stratified – 700 cases are good risks and are 300 bad – in contrast to a real life accepted-only data set, which would normally have a lower percentage of bad credit risks. However, this larger, more realistic fraction of defaulters (nearer to a population of applicants) is useful for applying reject inference methods. For simplification, only 7 out of 20 explanatory variables were chosen and partly recoded for our purposes. The data set was used in the following way. The sample was separated at random into a trainings sample with 600 cases, leaving the residual 400 cases as the test sample. For the training sample, credit scoring was applied and the 150 cases with the lowest scores were assigned to group ($A = 0$). The new scoring model was built based on the remaining cases, and the reject inference techniques were applied. Through random group composition and prior knowledge of the credit status Y, tests comparable to the ones applied on simulated data could be run, but in contrast to the simulated data set, binomial variables and even more predictor variables are used. However, this data set is rather small and the results are somewhat unstable. The model from the reduced data set works better in some simulation runs than the model from the full data. Applying 100 simulation runs evinced the sample selection effect in mean, but showed considerable variation in results. Raising the number of simulations to 500 mitigated the fluctuation.

At the beginning, a first score function was determined by a logit model based on a randomly chosen 250 elements of the full data sample. This function was used to assign every case of the training data a score and the related probability of acceptance $p(S_i)$. Next, the cases were sorted in order of their scores and the lowest 150 or 25% of the training data were declared as rejected. The first score function leads to 68.11% correctly classified cases. Again logit models were built on the full data as well on the selected subsample and were adopted for the test data set. The new model built on the full sample leads to 69.34% correct assignments (see Table 5) and the function based on the reduced data achieves 68.21% (see Table 6), a difference of about 1.13 percentage points.

	$A = 0$	$A = 1$
$Y = 0$	11.96%	17.62%
$Y = 1$	13.04%	57.38%

Table 5: Test data results based on the model derived from the full sample

	$A = 0$	$A = 1$
$Y = 0$	11.39%	18.18%
$Y = 1$	13.61%	56.82%

Table 6: Test data results based on the model derived from the selected subsample

Enlargement

Cases of the population ($A = 0$) were reallocated to group ($A = 1$) on the basis of the probability of acceptance $p(S_i)$ computed by the first score function. Conditional on higher values of $p(S_i)$, for this data 60% of the formerly rejected applicants were accepted for credit. This method delivers substantial improvements (68.99% correctly classified cases) due to the high amount of supplementary data, but it is unacceptable because of the occurrence of high risk one has to accept. One-fifth of this data amount or 12% of ($A = 0$) leads to 68.40% accurate assignments. Only small improvements are feasible for less extra data. Again, a relationship between supplementary data and classification results can be demonstrated.

Reweighting

According to the procedure for the simulated data, the cases from subgroup ($A = 1$) were reweighted by $1/p(S_i)$. The model computed from the reweighted data reaches a result of 68.20% correctly classified cases. No improvements can be observed from this method. A possible explanation is the high $p(S_i)$ values of the accepted applicants near c, which gives them a smaller weight in comparison to the nearly rejected of the simulated data. Another possibility is that the cases near c have an inferior match to the rejected cases.

Enlargement with reweighting

The supplementary cases selected with help of the probability $p(S_i)$ were reweighted with $1/p(S_i)$. The cases selected in $(A = 1)$, depending on the cutoff c, were again weighted with 1. A slight correction to 69.05% in comparison to the plain enlargement of the selected sample (68.99%) was shown. The use of 12% additional data and, accordingly, multiplying the weight by a factor of 5 leads to 67.73% correctly classified cases. The technique leading to increased classification success for the simulated data leads here to a result that is worse when compared to the model based on the selected subsample. The reason for this might be found in the variable characteristics for the simulated data: one selected case from the simulated data is representative of many others around it. Using instead the 12% additional cases in combination with basic reweighting, improvements (68.44%) can be achieved in comparison to using only the extra data (68.40%).

Extrapolation

The same approach for extrapolation was used with the real data set as was used for the simulated data. Classification success was 68.86%. The real data results are not as intriguing as they were for the simulated data. Potential explanations are a worse fitting linear regression and, again, the small number of cases. To compensate for the small data basis, the variable Y was simulated three times for the rejected applicants and consequently the logit model was computed three times. From these parameter sets, the mean was taken and applied to the test data. The results are an improvement (68.95% correctly assigned cases) compared to the plain extrapolation. The extrapolation approach and enlargement with reweighting (69.05%) differ little in their obtained results. The use of extrapolation is preferred because this approach prevents the need for additional data and the combined risk.

5 The sample selection bias and its influencing factors

During testing, the model based on the reduced sample occasionally produced better classification results for the test data than the model built on the full data sample (see results from the Fahrmeir and Hamerle (1984) data set). This seemingly confusing result depending on sample size and other rationales will be discussed in this section. The basis for the following simulations is the data described in Section 4.2 with one distinction – the starting sample size is 1,000 instead of 10,000 cases. The differences of classification success reached by the models based on the full and the reduced data set were calculated and presented after sorting. Positive values of the differences represent the expected case, meaning that the full data model works better than the one derived from the reduced data set. Negative values mean the opposite.

Sample size

First, the variation in sample size and its influence on the results was analyzed. The other factors were held constant. As can be seen in Figure 3, for small sample sizes (500 or 1,000 cases), the mentioned phenomenon occurs that is, the selection rules from the subsample ($A = 1$) achieve better results. A possible explanation for this pattern is that with small samples the true nature of the dependence between the variables and the outcome Y is not completely revealed. The produced models work better on the test sample by chance. This can be observed in the simulations based on the data set from Fahrmeir and Hamerle (1984), which is rather small. Another finding is that for large data sets, ceteris paribus, a selection effect will occur within a small bandwidth of variation.

Cutoff c

By changing the cutoff c, the amount of data selected in the group of accepted applicants can be changed. The different outputs are represented in Figure 4. Obviously, if all applicants are accepted for credit, there will be no difference in classification success. Also, common knowledge is the displayed circumstance

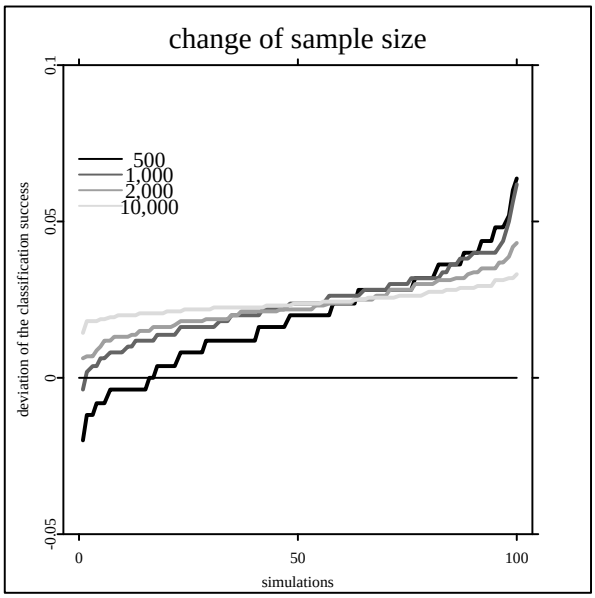

Figure 3: Selection effect and sample size

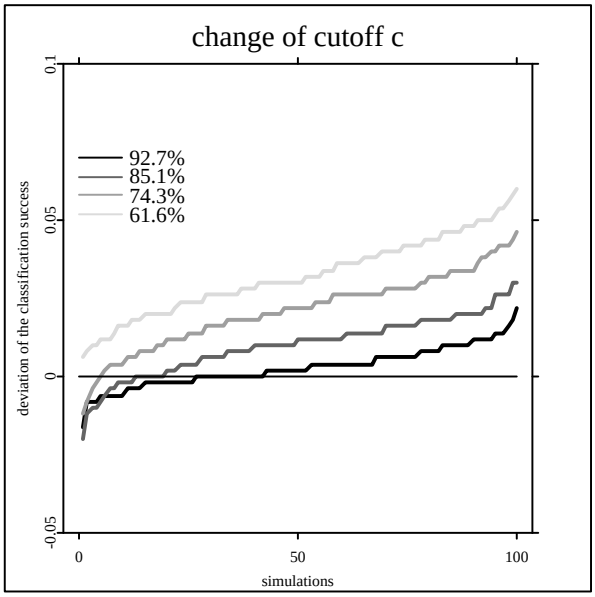

Figure 4: Selection effect and cutoff c

such that with a decreasing selected subsample, the gap between classification results grows, and with an increasing subsample, the gap narrows.

Change of default frequency

In this analysis, the propensity to default was altered, leaving all other variables unchanged. The random generator was adjusted in a way that more or less cases defaulted on the same probabilities of repayment. The described adjustment was deployed on the training and the test data in the same way. Figure 5 shows the

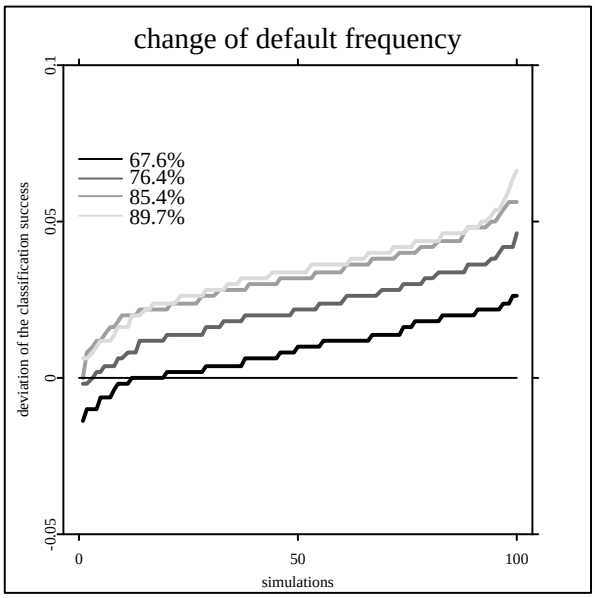

Figure 5: Selection effect and default frequency

differences in classification success, depending on the proportion of good credit risk applicants. For higher proportions of "good" applicants, the differences in classification success increase. The variability of the differences was unaffected. This can be explained as follows. If the proportion of "bad" is very small in the sample, given a fixed cutoff c there is hardly any information about the defaulters in the selected subsample. This leads to an increased bias for the scoring model due to the reduced data set. In conjunction with the cutoff, default frequency can be a considerable problem in real life. If for the creation of a new scoring

model only data with a small proportion of defaulters, dependent on the cutoff, is available, the calculated model will be inferior compared to one computed from a full data set.

A change in default frequency could be used to demonstrate the effects of an economic change over time. Hence, models that were estimated in a good economic environment were tested under inferior or superior conditions. It was assumed that the creditor perceives the economic changes and can adjust the cutoff to reflect the new proportion of defaulters in the whole population. The left side of

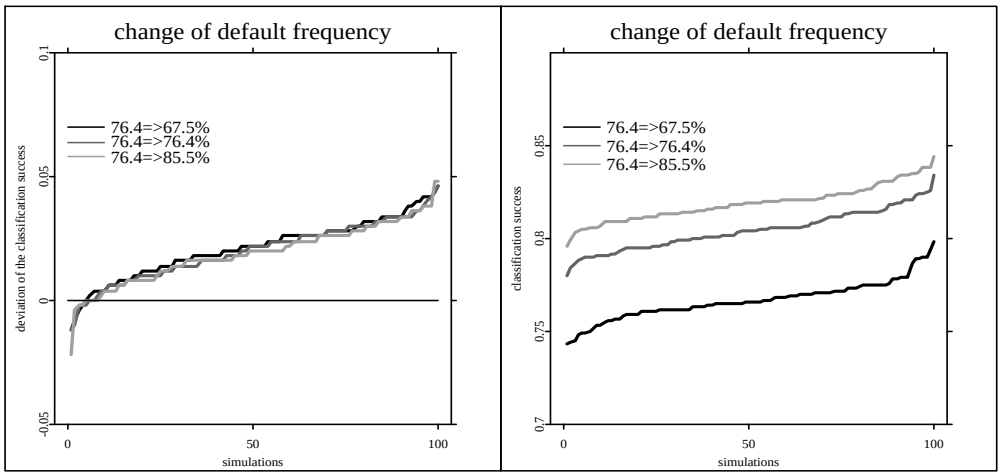

Figure 6: Selection effect and variable default frequency

Figure 6 shows that the differences induced by sample selection are not influenced at all by changing economic conditions. However, such change has an impact on the model from the full and the reduced data set in the same way. The right side of Figure 6 represents this circumstance for the full data model. If the proportion of good risk cases decreases, the results of the model deteriorated. This shows that aside from the sample selection effect, there are other considerations that constrain the efficiency of a scoring system.

6 Conclusions

In this work, the problem of sample selection in the context of building or recalibrating credit scoring systems was analyzed. Methods for coping with this problem were presented and applied to two data samples. As found by others, reweighting appears unsuitable for the purpose of bias reduction. Without an exact knowledge of the default propensity of the rejected applicants, this method leads to improvements only by chance. Granting credit depending on the soft-accept-reject threshold described by Hand (2002) leads to advances subject to the amount of supplementary data. By combining this approach with reweighting, the results were improved. Even though this method leads to better classification, it should not be forgotten that these results were obtained from additional high-risk cases. The question is whether the gains from better case assignment outweigh the losses of additional credit amounts. The extrapolation method as described in this work needs no additional risk and combined default cost and led to a remarkable improvement for the data here used. In future research it would be interesting to test the efficacy of the suggested extensions on further data sets and with different scoring approaches.

Many factors influence the scale of sample selection bias, including the size of the sample, different settings of the cutoff c, or the default frequency. For high cutoffs and small quantities of bad risk in the accepted applicant subsample, in combination with a large sample size, reject inference techniques could be very interesting. Besides these obvious factors there are rationales, such as the change of the economic environment over time, which leave the distortion between the model based on the full and the one based on the reduced data unchanged. Nonetheless, this factor influences the success of both models in the same way. Further, override of the scoring systems by employees of the credit granting organization or changes in the population of credit applicants (see Phillips and Yezer, 1996) are other possible sources of bias.

Although sample selection has only a small influence on classification results, in combination with other factors the distortions can cause considerable effects on the results of a credit portfolio and should not be ignored by credit granting organizations. These effects should be utilized in building or calibrating scoring models.

References

Ash, D. and S. Meester. 2002. Best practices in reject inferencing, Presentation at credit risk modelling and decisioning conference, Wharton Financial Institutions Center, Philadelphia.

Basel Committee on Banking Supervision. 2003. The new Basel capital accord, Bank for International Settlements, http://www.bis.org.

Banasik, J., J. Crook, and L. C. Thomas. 2003. Sample selection bias in credit scoring models, *Journal of the Operational Research Society* 54, 822–832.

Boyes, W. J., D. L. Hoffman, and S. A. Low. 1989. An econometric analysis of the bank credit scoring problem, *Journal of Econometrics* 40, 3–14.

Crook, J. and J. Banasik. 2004. Does reject inference really improve the performance of application scoring models?, *Journal of Banking and Finance* 28, 857–874.

Fahrmeir, L. and A. Hamerle. 1984. *Multivariate statistische Verfahren*, Walter de Gruyter, Berlin.

Greene, W. H. 2003. *Econometric Analysis*, 5th ed., Prentice Hall.

Hand, D. J. 2002. Measurement and prediction models in consumer credit, Presentation at credit risk modelling and decisioning conference, Wharton Financial Institutions Center, Philadelphia.

Hand, D. J. and W. E. Henley. 1997. Statistical classification methods in consumer credit scoring: A review, *Journal of the Royal Statistical Society Series A* 160, 523–541.

Heckman, J. 1979. Sample selection bias as a specification error, *Econometrica* 47, 153–161.

Kraft, H., G. Kroisandt, and M. Müller. 2002. Assessing the discriminatory power of credit scores, Discussion paper, Fraunhofer Institut für Techno- und Wirtschaftsmathematik (ITWM) and Humboldt-Universität zu Berlin.

Little, R. J. A. and D. B. Rubin. 1987. *Statistical analysis with missing data*, John Wiley & Sons, Inc., New York.

Phillips, R. F. and A. M. Yezer. 1996. Self-selection and tests for bias and risk in mortgage lending: Can you price the mortgage if you don't know the process?, *Journal of Real Estate Research* 11, 87–102.

Poirier, D. J. 1980. Partial observability in bivariate probit models, *Journal of Econometrics* 12, 210–217.

Rosenberg, E. and A. Gleit. 1994. Quantitative methods in credit management: A survey, *Operations Research* 42, 589–613.