SVERIGES RIKSBANK
WORKING PAPER SERIES

154

SVERIGES RIKSBANK

Bank Lending Policy, Credit Scoring and the Survival of Loans

Kasper Roszbach

NOVEMBER 2003

WORKING PAPERS ARE OBTAINABLE FROM

Sveriges Riksbank • Information Riksbank • SE-103 37 Stockholm
Fax international: +46 8 787 05 26
Telephone international: +46 8 787 01 00
E-mail: info@riksbank.se

The Working Paper series presents reports on matters in
the sphere of activities of the Riksbank that are considered
to be of interest to a wider public.
The papers are to be regarded as reports on ongoing studies
and the authors will be pleased to receive comments.

Bank lending policy, credit scoring and the survival of loans*

Kasper Roszbach†

Sveriges Riksbank Working Paper Series
No. 154
November 2003

Abstract

To evaluate loan applicants, banks increasingly use credit scoring models. The objective of such models typically is to minimize default rates or the number of incorrectly classified loans. Thereby they fail to take into account that loans are multiperiod contracts for which reason it is important for banks not only to know if but also when a loan will default. In this paper a bivariate Tobit model with a variable censoring threshold and sample selection effects is estimated for (1) the decision to provide a loan or not and (2) the survival of granted loans. The model proves to be an effective tool to separate applicants with short survival times from those with long survivals. The bank's loan provision process is shown to be inefficient: loans are granted in a way that conflicts with both default risk minimization and survival time maximization. There is thus no trade-off between higher default risk and higher return in the lending policy.

JEL classification: C34, C35, D61, D81, G21
Keywords: Banks, lending policy, credit scoring, survival, loans.

*I am indebted to the referees and the editor for detailed and helpful reports. I would also like to thank Marcus Asplund, Kenneth Carling, Luigi Ermini, Lennart Flood, Gerrit-Jan Förch, Dennis Hoffman, Tor Jacobson, Sune Karlsson, Jesper Lindé, Rickard Sandin, Patrik Säfvenblad, Paul Söderlind, Anders Vredin and participants at seminars at the Stockholm School of Economics, Sveriges Riksbank, De Nederlandsche Bank, the ECB, EARIE and at the Computational Finance conference at NYU for helpful comments, and Yngve Karlsson and Björn Karlsson, formerly at UC, for providing and discussing the data. Financial support from the Jan Wallander and Tom Hedelius Foundation is gratefully acknowledged. The views expressed in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank.

†Address: Sveriges Riksbank, Research Department, SE 103 37 Stockholm, Sweden. *Email*: kasper.roszbach@riksbank.se

1. Introduction

Consumer credit has come to play an increasingly important role as an instrument in the financial planning of households. When current income falls below a household's permanent level and assets are either not available or not accessible for dissaving, credit is a means to maintain consumption at a level that is consistent with permanent income. People expecting a permanent increase in their income but lacking any assets, like students, have a desire to maintain consumption at a higher level than their current income allows. Borrowing can assist them in doing that. Those who accumulate funds in a pension scheme but are unable to get access to them when they experience a temporary drop in current income can also increase their welfare by bridging the temporary fall in income with a loan.

The quantitative importance of consumer credit may be illustrated by the fact that total lending, excluding residential loans, by banks and finance companies to Swedish households amounted to SEK 310 bn. (199 bn.), or SEK 34,779 per capita (22,494), by the end of 2001 (1996). That is equivalent to 13.7 (10.9) percent of Swedish GNP or 28.1 (22.3) percent of total private consumption. Viewed from the perspective of financial institutions, household credit also constitutes a significant part of their activities, making up 34.3 (36) percent of total lending to the public. If one includes residential loans, that are often granted by separate subsidiaries, in total lending, this figure drops to 15.5 (12.8) percent. When looking at the risk involved in these loans instead of their volume, their importance is even greater, however. Current BIS rules stipulate an 8 percent capital requirement on consumer credit compared to, for example, 4 percent on residential loans.

From these numbers, it may be clear that a lending institution's decision to grant a loan or not and its choice for a specific loan size can greatly affect households' ability to smooth consumption over time, and thereby even households' welfare. At a more aggregate level, consumer credit makes up a significant part of financial institutions' assets and the effects of any loan losses on lending capacity will be passed through to other sectors of the economy that rely on borrowing from the financial sector. For this reason, the properties and efficiency of banks' credit granting process are of interest not merely because the factors determining the optimal size of financial contracts can be examined. At least as important are the implications these contracts have for the welfare of households and the stability of financial markets.

The starting point of every loan is the application. When lending institutions receive an application for a loan, the process by which it is evaluated and its degree of sophistication can vary greatly. Most continue to use rather naïve, subjective evaluation procedures. This could be a non-formalized analysis of an applicant's personal characteristics or 'scoring with integer numbers' on these characteristics. Some banks, however, use a statistical 'credit scoring' model to separate loan applicants that are expected to pay back their debts from those who are likely to fall into arrears or go bankrupt.

By far the most commonly used methods are discriminant analysis and logistic regression. Altman, Avery, Eisenbeis and Sinkey [1] contains a good review of this literature. Both models have been fit to separate good loans from bad ones among *approved* applications. The estimated parameters are thus subject to sample selection bias when these models are applied to *all* applicants. More recent studies have employed k-nearest-neighborhood [9] and count data models [6], classification trees and neural networks [2]. These methods tend to suffer from problems with either the calibration, estimation or interpretation of their parameters *in addition to* the sample selection bias mentioned earlier. All the above mentioned models, however, fail to account for the multiperiod character of an optimal debt contract and the implications this has for the credit-granting decision. In financial markets with perfect information, any optimal multiperiod financial contract can be obtained by a sequence of one-period loan agreements [13]. Loan applicants will be willing to pay the competitive interest rate that corresponds to their idiosyncratic risk and choose a first-best loan size. Under asymmetric information things become more intricate. In the literature that studies credit markets and the form of optimal financial contracts in the presence of adverse selection or costly state verification, credit rationing - the unequal treatment of ex-ante equal people - is a recurring phenomenon. See for example Stiglitz and Weiss [12] or Williamson [14]. When rationing is the mechanism that equilibrates credit markets, some applicants will be excluded from credit despite being equally creditworthy as those granted a loan. The allocation of resources will thus be inefficient.

Let us assume that *single-period* agreements are optimal[1] and that only the probability of default is unobservable to the lender. Gale and Hellwig [7] show that the optimal one-period debt contract consists of a pair (l, θ) where l is the size of the loan and θ the level of the endowment shock below which the debtor will be declared bankrupt. Under these circumstances traditional credit scoring models - by enabling a lending institution to rank potential customers according to their default risk - could improve the allocation of resources, from a second

[1] Since the default probabilities are not observed, this would be in a second best sense.

best towards the first best equilibrium. In a more general context, however, this does not solve a lender's profit maximization problem because financial contracts typically stretch out over several time periods. Townsend [13] proves that, in the presence of asymmetric information between the borrower and lender, ex-ante optimal contracts can only be created by multiperiod debt contracts because they allow payoffs to be dependent on past and present behavior of the borrower.

A loan, being a multi-period contract, generates a flow of funds until it either is paid off or defaults, in which case a part of the principal may still be recovered. The net present value of a loan is thus not determined by whether it's paid off in full or not, but - if it is not - by the duration of the repayments, amortization scheme, collection costs and possible collateral value. It may, for example, still be profitable to provide a loan, even if the lender is certain that it will default. Since the goal of financial institutions is to maximize profit (or utility), not to rank potential customers according to default risk, credit scoring models leave much room for subjective factors in the loan approval process. In a sense, banks use statistical models to forecast bankruptcy, but - conditional on this forecast - resort to ad-hoc methods to predict profitability.

Boyes, Hoffman and Low [4] address this deficiency and investigate if the provision of credit currently takes place in an efficient way. For this purpose they estimate a bivariate probit model with two sequential events as the dependent variables: the lender's decision to grant the loan or not, and - conditional on the loan having been provided - the borrower's ability to pay it off or not. If the lending institution is minimizing credit risk, one ought to find opposite signs for the parameter of one particular explanatory variable in the two different equations. This would imply that variables that increase the probability of positive granting decision also decrease the likelihood of a default, or vice versa. They find, however, that variables like duration of job tenure, education and credit card ownership carried equal signs, indicative of a policy that conflicts with default risk minimization. As I noted earlier, lenders may nevertheless prefer such a policy of supplying loans with a higher default risk because they have a higher expected rate of return (either the interest rate is higher or the default is expected to occur after a long period with regular installments and interest payments). Moreover, Boyes et al. show that unexplained tendencies to extend credit are positively correlated with default frequencies - another fact consistent with a policy that trades off default risk against profitability.

This paper deals with two issues. First, in order to improve upon the currently available methods for evaluating loan applications, I construct and estimate a

Tobit model with sample selection effects and variable censoring thresholds. This model can be used to predict the expected survival time on a loan to any potential applicant. Using the predicted survival time allows banks to make a more realistic evaluation of the return on a (potential) loan than when they merely employ an estimate of the default risk associated with an individual - as a traditional credit scoring model does.[2]

Secondly, I take up the question about the efficiency of banks' loan provision process that is raised by the results in Jacobson and Roszbach [10] and Boyes et al. [4]. The latter suggest that the fact that some variables increase the probability of a positive granting decision while at the same time increasing the likelihood of a default is a consequence of profit maximizing behavior by the lender. Here, it will be investigated if a similar relationship continues to exist when one models the survival time of a loan instead of the probability of its default. If variables that increase the likelihood of an applicant obtaining a loan also increase the expected survival and vice versa, then this would constitute further evidence of banks' behaving in a way that is consistent with profit-maximization.

The rest of this paper is organized as follows. Section 2 describes the data set and its sources. In Section 3, I derive the econometric model. Section 4 contains the empirical results and Section 5 concludes the paper with a discussion of the results and possibilities for future research.

2. Data

The data set consists of 13,337 applications for a loan that were processed by a major Swedish lending institution between September 1994 and August 1995. All applications were submitted in stores where potential customers applied for instant credit to finance the purchase of a consumer good. Out of 13,337 applications, 6,899 were rejected and 6,438 were approved. The dataset includes 127 second attempts by individuals that had applied once before.

The evaluation of each application took place in the following way. First, the store phoned to the lending institution to get an approval or a rejection. The lending institution then analyzed the applicant with the help of a database with personal characteristics and credit variables to which it has on-line access. The database is maintained by Upplysningscentralen AB, the leading Swedish credit bureau which is jointly owned by all Swedish banks and lending institutions. If

[2] Carling, Jacobson and Roszbach [5] study the consequences of dormancy risk for bank loan profitability for a subsample of granted loans.

approval was given, the store's salesman filled out a loan contract and submitted it to the lending institution. The loan is revolving and administered by the lending institution as any other credit facility. It is provided in the form of a credit card that can only be used in a specific store. Some fixed amount minimum payment by the borrower is required during each month. However, since the loan is revolving, there is no predetermined maturity of the loan. Earnings on the loan come from three sources: a one-time fee paid by the customer; a payment by the store that is related to the total amount of loans granted through it; and interest on the balance outstanding on the card. There is no annual fee, neither for the store nor for credit recipients. For this study, the lending institution provided a data file with the personal number of each applicant, the date on which the application was submitted, the size of the loan that was granted, the status of each loan (good or bad) on October 9, 1996, and the date on which bad loans gained this status.

Although one can think of several different definitions of a 'bad' loan, I classify a loan as bad once it is forwarded to a debt-collecting agency. I do not study what factors determine the differences in loss rates, if any, among bad loans. An alternative definition of the set of bad loans could have been 'all customers who have received one, two or three reminders because of delayed payment. However, unlike 'forwarded to debt-collecting agency", one, two or three reminders were all transient states in the register of the financial institution. Once customers returned to the agreed-upon repayment scheme, the number of reminders was reset to zero. Such a property is rather undesirable if one needs to determine unambiguously which observations are censored and which are not. Upplysningscentralen provided the information that was available on each applicant at the time of application and which the financial institution accessed for *its* evaluation. By exploiting the personal number that each resident of Sweden has, the credit bureau was able to merge these two data sets. In Sweden, each person is assigned a unique 10 digit personal number at birth or at immigration. The first six digits consist of the year, month and day of birth, whereas the remaining four indicate, among other things, gender and (foreign) citizenship. The personal number is used for a variety of purposes, like employment, taxes, social security matters, bank accounts, loans and rental registration. An employer, for example, needs to register its employees by their personal number to deposit payroll taxes, while a bank requires the personal number to open a bank account, among other things to be able to pay base rate capital income taxes to the tax authorities. Typically, tax authorities store data on a number that are relevant for determining tax rates, such as marital status, the city of residence, several income components, real

Table 1: Definition of explanatory variables and descriptive statistics
The table provides definitions and descriptive statistics for all variables and splits up the sample into rejected and granted applicants and further divides granted applicants into defaulted and still performing loans (N=13337).

	Type of applicant							
	Rejected		Granted		Defaulted		Performing	
	Mean	*Std. Dev.*	*Mean*	*Std. Dev.*	*Mean*	*Std. Dev.*	*Mean*	*Std. Dev.*
Variable definition	Min.	*Max.*	Min.	Max.	Min.	Max.	Min.	Max.
PERSONAL								
BIGCITY	0.41	*0.49*	0.37	*0.48*	0.41	*0.49*	0.36	*0.48*
dummy, applicant lives in one of three metropolitan areas								
DIVORCE	0.13	*0.34*	0.14	*0.35*	0.20	*0.40*	0.14	*0.35*
dummy, applicant is a widow(er) or a separated woman								
MALE	0.62	*0.48*	0.65	*0.48*	0.67	*0.48*	0.65	*0.48*
dummy, applicant is male								
MARRIED	0.47	*0.50*	0.47	*0.50*	0.24	*0.43*	0.48	*0.50*
dummy, applicant is divorced								
FINANCIAL								
CAPINC	0.12	*0.32*	0.07	*0.25*	0.04	*0.20*	0.07	*0.26*
dummy, applicant has taxable income from capital								
ENTREPR	0.04	*0.21*	0.02	*0.16*	0.02	*0.13*	0.03	*0.16*
dummy, applicant has taxable income from a business								
HOUSE	0.34	*0.47*	0.47	*0.50*	0.28	*0.45*	0.48	*0.50*
dummy, applicant owns (part of) a house								
INCOME	1.30	*0.70*	1.89	*0.76*	1.65	*0.82*	1.91	*0.75*
annual taxable income from wages (× 100,000 SEK)	0	*7.38*	0	*10.93*	0	*10.93*	0	*10.32*
DIFINC	0.05	*0.34*	0.09	*0.35*	0.04	*0.39*	0.09	*0.34*
$INCOME_t$ - $INCOME_{t-1}$	-4.39	*2.53*	-6.23	*5.01*	-1.35	*4.40*	-6.23	*5.01*
CREDIT								
COAPPLIC	0.16	*0.36*	0.14	*0.35*	0.07	*0.26*	0.14	*0.35*
dummy, appl. has a guarantor								
ZEROLIM	0.15	*0.36*	<0.01	*0.05*	0.04	*0.20*	<0.01	*0.02*
dummy, applicant has no collateral-free loans outstanding								
LIMIT	0.80	*0.94*	0.50	*0.50*	0.41	*0.58*	0.51	*0.49*
collateral-free credit facilities already outstanding (× 100,000 SEK)	0	*17.03*	0	*6.27*	0	*5.12*	0	*6.27*
LIMUTIL	0.64	*0.39*	0.53	*0.34*	0.76	*0.33*	0.52	*0.33*
share of LIMIT that is actually being utilized	0	*2.78*	0	*1.24*	0	*1.24*	0	*1.12*
LOANSIZE	-	-	-	-	0.71	*0.40*	0.71	*0.38*
amount of credit granted (× 10,000 SEK)					0.30	*2.45*	0.30	*3.00*
NRLOANS	2.99	*2.42*	3.65	*2.04*	2.34	*1.64*	3.74	*2.04*
number of collateral-free loans that is registered	0	*18*	0	*16*	0	*11*	0	*16*
NRQUEST	4.69	*2.60*	4.81	*2.68*	6.15	*2.85*	4.72	*2.64*
number of requests for information on the appl. that the credit agency received during the last 36 months	1	*10*	1	*19*	1	*14*	1	*19*
Number of observations	6899		6438		388		6050	

estate property and other assets (exceeding the tax exempt amount). The Swedish *Freedom of Press Act* entitles not only Swedish residents but even aliens to read all official government documents. Examples of public documents are letters written to and by a government institution, but also tax forms and job applications at government institutions. Only drafts of documents, unless filed or sent out by the authority, and individual documents that are classified as secret under the *Secrecy Act*, are exempt from this entitlement. Every individual or

Table 2: Descriptive statistics for survival time.
Mean, standard deviation and percentiles for survival time. For defaulted loans: days between granting of loan and default. For good loans: days between granting and monitoring moment.

				Percentiles							
Sub-sample	μ	σ	min	5	10	25	50	75	90	95	max
t, bad loans	400.1	151.1	130	156	192	278	403	514	606	648	789
t, good loans	632.8	94.0	34	470	497	564	652	704	746	767	795

company can thus access and process personal income statements and the accompanying personal information. However, without permission from the Swedish Data Inspection Board (SDIB), data from different sources can not be merged or stored for more than some predetermined time period. Government authorities can, for example, compare social security payments with income tax data. But unless a bank or credit agency obtains an exemption from the SDIB, it may only store data on payment remarks for up to three years. For the purpose of this study the personal numbers were removed before handing over the data set. Overall, the database includes a total of 49 variables (approximately 60 after transforming qualitative variables into dummies). A full list is provided in Table 7 in Appendix B. The major part consists of publicly available, governmentally supplied information such as sex, citizenship, marital status, postal code, taxable income, taxable wealth, house ownership. The remaining variables, like the total number of inquiries made about an individual, the number of unsecured loans and the total amount of unsecured loans, are reported to Upplysningscentralen by the Swedish banks. Table 1 contains definitions and descriptive statistics for the explanatory variables that are used in the analysis in Section 4. Of the applicants, 6,899, or 51.7 percent, were refused credit. The remaining 6,438 obtained a loan ranging from 3,000 to 30,000 Swedish kronor (approximately US$ 350 - 3500) . The lending institution's policy was that no loans exceeding 30,000 kronor were supplied. Although there is an indicated amortization scheme, the loans have no fixed maturity - they are revolving.

On 9 October 1996, all people in the sample were monitored by the lending institution. On that occasion 388 (6.0 %) of those who had obtained a loan had defaulted and been forwarded to a debt collection agency. All other borrowers still fulfilled their minimum repayment obligations at that time. The survival time in the sample, calculated as the number of calendar days between the date of application and the date of default, ranged from 130 days (a defaulted loan) to 795 days (a censored observation). Descriptive statistics for survival time are provided in Table 2.

3. Econometric model

Under ideal conditions evaluating loan applicants or studying efficiency in the provision of bank loans would entail modelling the revenue on each loan as a function of a set of personal characteristics and macro-economic indicators. However, since few banks store complete time series of interest payments and amortizations on loans, the information presently available and useful for such a study is limited to the current balance and status (good or bad) of each loan. Therefore, I will instead model the survival time of each loan. With some simplifying assumptions imposed on the amortization scheme and cost structure, one can then in principle calculate an estimate of the return on each loan as a function of survival time.

The econometric model consists of two simultaneous equations, the first one for the binary decision to provide a loan or not, y_i, and the second one for the *natural logarithm* of survival time of a loan (in days), for notational simplicity denoted by t_i. Because the bank from which I obtained the data merely considered whether it would accept an application or not, all people who were granted a loan received the amount of credit they applied for at the going rate of interest. The first equation therefore models a binary decision. I do not model how individuals determine the amount of credit they apply for.

I use the superscript $*$ to indicate an unobserved variable and let y_i^* and t_i^* follow

$$\begin{aligned} y_i^* &= \mathbf{x}_{1i}.\boldsymbol{\beta}_1 + \varepsilon_{1i} \\ t_i^* &= \mathbf{x}_{2i}.\boldsymbol{\beta}_2 + \varepsilon_{2i} \quad \text{for } i = 1, 2, ..., N \end{aligned} \tag{1}$$

where the disturbances are assumed to be homoscedastic and bivariate Normal distributed:

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N \begin{pmatrix} 0 & 1 & \sigma_{12} \\ 0 & \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{2}$$

As mentioned, y_i, is a binary choice variable that takes value 1 if the loan was granted and 0 if the application was rejected:

$$y_i = \begin{cases} 0 & if \quad y_i^* < 0 \\ 1 & if \quad y_i^* \geq 0 \end{cases} \tag{3}$$

For loans that turn bad, we can observe the exact survival time. For loans that are still performing on the day of monitoring, survival is censored because we do not know if and when they will turn bad. Because all loans are monitored on October 9, 1996, but are granted anywhere between September 1994 and August 1995, the good loans' survival times will be censored at varying thresholds. For example, a loan granted on September 1, 1994, has a censoring threshold of 768 days. For a loan granted on August 31, 1995, this is 434 days. A loan's censoring threshold for survival time will be denoted as $\overline{t_i}$. The above can be summarized in the following censoring rule:

$$t_i = \begin{cases} t_i^* & if \quad t_i^* < \overline{t_i} \\ \overline{t_i} & if \quad t_i^* \geq \overline{t_i} \end{cases} \tag{4}$$

Due to the fact that one only observes survivals for loans that are actually granted, there is not only a censoring rule for t_i but even an *observation* rule:

Table 3: Observation rule for y_i and t_i.
Entries in the 2×2 table show pairs $(y_i,\ t_i)$ that are observed for all ranges of y_i^* and t_i^*.

	$t_i^* \leq \overline{t_i}$	$t_i^* > \overline{t_i}$	
$y_i^* < 0$	$(0, .)$	$(0, .)$	
$y_i^* \geq 0$	$(1, t_i^*)$	$(1, \overline{t_i})$	

A dummy variable d_i splits up the sample of granted loans into good ones and bad ones. If a loan's survival is uncensored, $t_i^* \leq \overline{t_i}$, it must be a defaulted one. If survival is censored, it must be a good loan.

$$d_i = \begin{cases} 0 & if \quad t_i^* \leq \overline{t_i} \\ 1 & if \quad t_i^* > \overline{t_i} \end{cases} \tag{5}$$

Because three types of observations exist: no loans, bad loans with survival t_i, and good loans with survival $\overline{t_i}$, the likelihood function will take the following form:

$$\ell = \prod_{no\ loans} \text{pr}\left(no\ loan\right) \cdot \prod_{bad\ loans} \text{pr}\left(t_i \cap bad\ loan\right) \times \prod_{good\ loans} \text{pr}\left(\overline{t_i} \cap good\ loan\right) \quad (6)$$

Combining (1), (2) – (5) and Table 3, equation (6) becomes

$$\ell = \prod_{i=1}^{N} \text{pr}\left(y_i^* < 0\right)^{(1-y_i)} \cdot \prod_{i=1}^{N} \text{pr}\left(y_i^* \geq 0,\ t_i\right)^{y_i \cdot (1-d_i)} \times \prod_{i=1}^{N} \text{pr}\left(y_i^* \geq 0,\ t_i^* \geq \overline{t_i}\right)^{y_i \cdot d_i} \quad (7)$$

In Appendix **A.1** it is shown that (7) implies the following loglikelihood:

$$\begin{aligned} \ln\ell = & \sum_{i=1}^{N} (1-y_i) \cdot \ln\left[1 - \Phi\left(\mathbf{x}_{1i}\boldsymbol{\beta}_1\right)\right] + \\ & \sum_{i=1}^{N} y_i \cdot (1-d_i) \left\{ \ln\Phi\left(\frac{\mathbf{x}_{1i}\boldsymbol{\beta}_1 - \frac{\sigma_{12}}{\sigma_2^2}(t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2)}{\sqrt{(1-\rho^2)}}\right) + \right. \\ & \left. -\frac{1}{2}\ln 2\pi + \ln\left(\frac{1}{\sigma_2}\right) - \frac{1}{2}\left(\frac{t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2}\right)^2 \right\} + \\ & \sum_{i=1}^{N} y_i \cdot d_i \ \ln\Phi_2\left(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\ \frac{\mathbf{x}_{2i}\boldsymbol{\beta}_2 - \overline{t_i}}{\sigma_2};\ \rho\right) \end{aligned} \quad (8)$$

where $\Phi(\cdot)$ and $\Phi_2(\cdot, \cdot, \rho)$ represent the univariate and bivariate standard Normal c.d.f., the latter with correlation coefficient ρ.

4. Empirical results

Because no theoretical framework exists to help select candidate explanatory variables from the full data set, the following three step procedure was applied. First, correlations between all relevant variables were calculated to check for potential multicollinearity problems. When a set of variables was highly correlated with each other, I checked for their univariate explanatory power in predicting the probability of obtaining a loan and the probability of default. As a rule, out of each 'set' the variable with the highest correlation with either event was selected for the third (estimation) stage. In addition, plots against empirical default and granting probabilities were studied to investigate the monotonicity of the (univariate) relationship. In the first two stages, most of the income and tax variables were consequently eliminated. As mentioned in Section 2, the original dataset contained many different income components, both gross and net of taxes, plus quite some tax variables. Moreover, for a fair number of these income and tax measures two subsequent annual observations were available. Correlations within both the group of income variables and the group of tax variables - both simultaneous and lagged - were generally high and often around or above .9. In a similar way, *BALANCE* was eliminated in favor of *LIMIT*.

In stage three, the candidate variables were used to estimate equation (8). In principle, all possible model permutations were estimated and the model with the best properties was selected. Criteria that were used for this purpose were: t-statistics, likelihood value, number of parameters and model stability. Some of the candidate variables were omitted from the final model because they were insignificant in both equations for every permutation of the bivariate Tobit model. Age, citizenship (Swedish, Nordic, non-Nordic), the number of months since immigration (and some transformations), the combined value of all real estate a person has (partial) ownership in, shop and chain dummies and *BALANCE/INCOME* were removed for this reason.

It is also worthwhile to make some remarks on the distribution of survival time. QQ normality plots, that compare survival time's sample distribution with a Normal distribution with equal mean and variance, suggest that a logarithmized transformation increases the symmetry of the survival time's distribution slightly. In addition, actual (uncensored) survival time is likely to display large outliers, that can be handled better by a log-transformation.[3]

The parameters of (8) are estimated by the following procedure. First, I calculate starting values for $\boldsymbol{\beta}_1$ from a univariate probit on the first equation in (1). These are consistent although not efficient, because the covariation between ε_1 and ε_2 is not taken into account. The starting values for $\boldsymbol{\beta}_2$ and σ_2 come from a Tobit model with variable censoring threshold on the survival time of the granted loans. This model implicitly assumes that $\rho = 0$. Under the restriction that $\rho = 0$, one can estimate the second equation in (1) separately. Because one ignores the rejected loan applications, these parameter estimates suffer from a sample selection bias and are inconsistent if $\rho \neq 0$ - which is the case here, as we will see below. In all tests of the the model with simulated data, however, these estimates were found to be close (plus minus a decimal) to the true parameter values. The iterative procedure on the full model *with* sample selection converged rather easily when using these estimates as starting values. By comparison, when I let either

[3]PP-plots and QQ-plots are available from the author upon request.

Table 4: Univariate and bivariate MLE of loan granting, survival time and default equations

Univariate estimators of β_1 and β_2 come from separate estimation of the first and the second equation in (1) respectively, so that ρ=0 is assumed. Bivariate estimators come from joint estimation of both equations in (1), so that the that the sample selection effect is taken into account and estimation of ρ is required. The bivariate probit estimator α_2 results when the second equation instead measures default risk.

	Equation				
	Loan granting		Survival time		No default
	β_1		β_2		α_2
Variable	Univariate	Bivariate	Univariate	Bivariate	Bivariate
CONSTANT	**-0.336**	**-0.328**	**8.246**	**9.065**	**2.460**
	(0.051)	(0.051)	(0.156)	(0.193)	(0.113)
BIGCITY	**-0.232**	**-0.222**	**-0.128**	0.233	-0.042
	(0.027)	(0.027)	(0.058)	(0.053)	(0.055)
DIVORCE	**-0.186**	**-0.179**	-0.124	-0.009	-0.071
	(0.040)	(0.039)	(0.078)	(0.069)	(0.074)
MALE	**-0.207**	**-0.196**	**-0.106**	0.024	0.024
	(0.028)	(0.028)	(0.061)	(0.058)	(0.059)
MARRIED	**-0.242**	**-0.233**	**0.187**	**0.344**	**0.253**
	(0.030)	(0.030)	(0.068)	(0.066)	(0.066)
CAPINC	**-0.284**	**-0.272**	-0.057	**0.195**	0.142
	(0.051)	(0.050)	(0.123)	(0.100)	(0.127)
ENTREPR	**0.570**	**0.570**	0.135	0.162	0.147
	(0.064)	(0.064)	(0.181)	(0.160)	(0.157)
HOUSE	**0.110**	**0.102**	0.061	-0.023	-0.011
	(0.028)	(0.028)	(0.061)	(0.062)	(0.060)
INCOME	**0.901**	**0.886**	0.038	**-0.286**	**-0.226**
	(0.018)	(0.018)	(0.042)	(0.050)	(0.050)
DIFINC	**-0.243**	**-0.237**	**0.147**	**0.185**	**0.206**
	(0.035)	(0.035)	(0.075)	(0.079)	(0.073)
COAPPLIC	**0.156**	**0.146**	**0.509**	**0.374**	**0.421**
	(0.034)	(0.034)	(0.109)	(0.104)	(0.098)
ZEROLIM	**-2.253**	**-2.218**	**-2.244**	**-0.328**	**-0.703**
	(0.106)	(0.114)	(0.401)	(0.140)	(0.306)
LIMIT	**-0.861**	**-0.848**	0.006	**0.561**	**0.486**
	(0.019)	(0.021)	(0.059)	(0.050)	(0.056)
LIMUTIL	**-0.747**	**-0.759**	**-1.295**	**-1.223**	**-1.210**
	(0.045)	(0.045)	(0.120)	(0.116)	(0.093)
LOANSIZE	-	-	-0.069	-0.070	-0.067
	-	-	(0.073)	(0.070)	(0.069)
NRLOANS	**0.086**	**0.086**	**0.329**	**0.259**	**0.270**
	(0.007)	(0.007)	(0.026)	(0.023)	(0.020)
NRQUEST	-0.007	-0.004	**-0.116**	**-0.097**	**-0.104**
	(0.005)	(0.005)	(0.012)	(0.011)	(0.010)
σ^2	-	-	**0.919**	**1.096**	-
	-	-	(0.045)	(0.057)	-
ρ	-	-	-	**-0.986**	**-0.911**
	-	-	-	(0.021)	(0.056)
Pseudo R^2	*0.57*	*0.57*	*0.37*	*0.30*	*0.31*
Loglikelihood	*-6510.89*	*-7674.17*	*-2430.19*	*-7674.17*	*-7529.76*

Standard errors are shown in parentheses. Variables significant at a significance level of at least 10 percent are printed in bold style. Observe that the loglikelihood value for the bivariate estimations refers to both equations jointly.

an OLS or a Heckman's two-step procedure generate the starting values for $\boldsymbol{\beta}_2$ and σ_2 - thus taking the sample selection effect into account while ignoring the censoring in t_i - it was more time-consuming or even impossible to find a maximum for the loglikelihood function (8).

With these starting values and letting $\rho^{start} = 0$, I then estimate $\boldsymbol{\beta}_2$, σ_2 and ρ simultaneously by maximizing (8) under the restriction that $\boldsymbol{\beta}_1 = \widehat{\boldsymbol{\beta}}_1^{probit}$. These estimates of $\boldsymbol{\beta}_2$, σ_2 and ρ are consistent and are in their turn used as starting values in the last step. Estimating $\boldsymbol{\beta}_2$, σ_2 and ρ first and then estimating $\boldsymbol{\beta}_2$, σ_2, ρ *and* $\boldsymbol{\beta}_1$ by FIML saves a lot of time compared to doing FIML directly. The FIML iterations provide consistent and efficient estimators of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_2 and ρ and a consistent estimator of the variance-covariance matrix. The FIML parameter estimates, their standard errors and t-statistics are presented in Table 4.

The first two columns of Table 4 contain two sets of parameter estimates for the loan granting decision: the first one from estimation as a single equation and the second from estimation together with the survival equation. There appears to be no clear gain in efficiency in the estimate of β_1 from estimating the two equations in (1) simultaneously. *LOANSIZE* could not be used as an explanatory variable in this equation because no data on this variable were available for rejected applications. The effect of most variables on the probability of obtaining a loan is in line with our expectations. *INCOME* and *HOUSE* confirm their role as important factors that contribute positively, while *LIMIT, LIMUTIL* and *DIVORCE* have the traditional negative effects. The coefficients on *MARRIED, DIFINC* and *CAPINC* are less intuitive, however.

The third and fourth columns of Table 4 compare two different estimators of $\boldsymbol{\beta}_2$ and σ_2. The parameter estimates in column three are obtained from a Tobit model with a variable censoring threshold that ignores the sample selection effect that arises when disregarding the rejected loan applications. This is equivalent to estimating $\boldsymbol{\beta}_2$ and σ_2 in (1) under the hypothesis that $\rho = 0$. One is, in other words, assuming that the likelihood of a survival of a certain length is not affected in any systematic way by the inferences one can make from observing y_i and x_{1i}. If the hypothesis is true, then the parameters in the first and second equation in (1) can be estimated separately from each other. However, if the disturbances ε_1 and ε_2 *are* correlated, these estimators of $\boldsymbol{\beta}_2$ and σ_2 will be biased. The fourth column in Table 4 contains the consistent parameters estimates of $\boldsymbol{\beta}_2$ and σ_2 obtained by estimating the complete model $(1)-(5)$. The purpose of comparing these two estimators is to investigate to what extent any misunderstandings about the relation between people's characteristics and financial discipline may have originated in an incorrect way of sampling data by financial institutions. A comparison of the two estimators will help us to determine if inconsistencies in bank lending policy find their origin in a sample selection bias. If there is any such sample selection bias, we will also want find out whether it is also quantitatively important.

In the final model, all explanatory variables enter the model linearly. I have checked for the presence of non-linear effects by adding quadratic terms of all continuous variables. Their coefficients were never significant, however. Out of 16 explanatory variables, four lose or reduce their significance and three turn significant or increase their level of significance when disregarding the sample selection effect. Of the parameters for the remaining nine variables, four are insignificant while the remaining five are significant and have identical signs in both models. So although accounting for the sample selection effect never reverses the sign of any of the coefficients, it does clearly affect their magnitude. The influence of the variable *ZEROLIM*, for example, would be badly overestimated if one did not account for the sample selection effect. A look at Table 1 may help us understand this phenomenon. Although having no loans outstanding is rather uncommon among the granted loans, it appears to be associated with defaulting rather than with proper repayment behavior. However, this overlooks the fact that 15% of all rejected applicants did not have any previous loan. If rejected applications are not so much different from approved ones, then the actual impact of having a zero limit may well be much smaller than one would expect by merely looking at granted loans.[4] Similarly, *INCOME* is not significant in the (third) column with biased estimators, whereas the consistent parameter estimate in column four ís significant and negative. Although one should be careful not to rationalize each counter-intuitive finding, we can look for a tentative explanation. Table 4, columns 1-2 and Table 1, columns 1-3 clearly show that people with higher incomes are more likely to be granted a loan and that people who are granted loans on average have higher incomes. Consequently, disregarding the rejected applicants (who have low incomes) could well have led us to infer that income does not influence a loan's default risk (within the subsample of accepted applicants), as Table 4, column 3 suggests. Moreover, if in fact other factors than *INCOME* determine a loan's survival, then the selection of applicants may be taking place on the basis of a negative bivariate relation between *INCOME* and defaults (see Table 1) that will disappear in a multivariate setting when one controls for both

[4] Rejected applications will differ very little from approved ones if the lending institution grants loans to applicants on the basis of characteristics that have little impact on survival.

the sample selection effect and the correlation with other variables. In our sample people with higher income actually turn out to constitute a bigger risk than people with lower incomes, as columns 4 and 5 of Table 4 show. Other variables of interest that have significant coefficients are *NRQUEST, ZEROLIM, NRLOANS, LIMIT* and *LIMUTIL*. The number of information requests is considered to reflect a person's efforts to obtain additional credit and as such expected to contribute negatively to survival. Not having any loan at all, as indicated by *ZEROLIM*, is a sign of inexperience with servicing debt and has a negative effect on survival. For debtors/applicants wíth a loan, *NRLOANS* and *LIMIT* are considered to proxy for experience with servicing debt and increase survival time to have a negative influence. *LIMUTIL* captures the extent to which a debtor uses it and reduces decreases survival.

It is also worth commenting the value of the correlation coefficient. Its value of -.98 may create the impression that the algorithm had problems converging. In extensive tests of the model with different sets of explanatory variables and varying sample sizes, rho took values between -.55 and -.98. In tests with the bivariate probit model, the final parameter estimates of which are reported in the last column of Table 4, ρ ranged from approximately -.65 to -.93. For the bivariate probit model Boyes et al. [4] report -.35, Jacobson and Roszbach [10] -.92.. As is the case with most models with limited dependent variables (see Bermann [3]), the computations for the Tobit and probit models did not converge for some configurations of explanatory variables. When the computations broke down, divergence always took place after relatively few iterations, however, with rho breaking its constraint before any of the other parameters had stabilized around a final value. In the estimation of the final model, all parameters settled down around their final values rather quickly.

Overall, the above results clearly illustrate the importance of taking rejected applicants into account when analyzing survival behavior. Ignoring the sample selection effect in an analysis of the duration of loans leads to considerable biases in the parameter estimates. Even though the sign of parameters is never reversed when switching from a biased Tobit model to one that takes the sample selection effect into account, some variables, like income, outstanding loans, and income from assets, would have appeared to be unrelated to survival time had one disregarded the sample selection effect; the importance of others, such as gender, metropolitan residency and lacking experience with credit, would have been overestimated. Biases like these, and the resulting misunderstandings, may well give rise to inefficient lending policies at financial institutions.

In the last column of Table 4, I present parameters of the bivariate probit model, as constructed by Boyes et al. [4] but re-estimated with the data used in this paper.[5] When studying the probit parameters that determine the probability of a loan *not* defaulting ($\boldsymbol{\alpha}_2$), we see that all variables except for one have the same sign as in $\boldsymbol{\beta}_2$ in the Tobit model of logged survival time.[6] Variables that increase (decrease) the probability of a default thus also decrease (increase) the expected survival time of a loan. However, a number of variables, like *MALE, DIVORCE, HOUSE, BIGCITY* and *ENTREPR,* that are significant in the loan granting decision actually affect neither default risk nor survival, while *NRQUEST* does have a significant effect on survival but doesn't play any role in the acceptance process. *MARRIED, INCOME, DIFINC, CAPINC* and *LIMIT* have opposite signs in the loan granting and no-default equations, implying that the bank uses these variables in such a way that they increase (decrease) the likelihood of a loan being granted although they in fact increase (decrease) the risk of default. Because all but one parameters in the survival equation have the same sign as the bivariate probit parameters, these variables also reduce (raise) expected survival and return on the loan. In other words: if the bank is not minimizing default risk in its loan granting policy, it is not doing so because loans with higher default risk have longer expected survival times and therefore higher expected returns.[7] Moreover, the negative values of ρ in Table 4 indicate that any non-systematic propensity to grant loans is associated with shorter survival times and higher default risk. This is consistent with the above observation that the bank does not appear to trade off risk against return. Rather, the loan granting policy appears to be inefficient

[5] Note that the bivariate probit model implicitly assumes that loans, that are still performing at the monitoring occasion, will not default later on.

[6] *CAPINC* is significant in the survival time equation while it's not in the probit equation of default risk.Otherwise, all variables with significant parameters in the survival equation of the Tobit model with sample selection (1) also have significant coeffficients in the 'probability that loan doesn't default' equation of the bivariate probit model. The reverse, however, does not hold! Because the main objective here is to develop and illustrate the properties of the bivariate Tobit model, variables that would have been significant in a bivariate probit model but are not in the Tobit model have been omitted.

[7] It is possible that the expected return *per* loan does not increase with the expected survival time, for example because loans that survive stop paying interest or carry a smaller balance. Here, we can exclude the first possibility because of the bank's definition of default. As far as the second possibility is concerned, the return per Krona will be affected by loan size only due to any fixed costs that the bank incurs when handling existing loans. Because we do no dispose of data on these costs, we cannot determine here at what point a longer survival time will be associated with a lower return due to a concomitant reduction in loan size.

and contain non-systematic components that are strongly negatively correlated with survival.

Unlike Boyes et al., I here control for the size of the loan when estimating $\boldsymbol{\alpha}_2$ and $\boldsymbol{\beta}_2$. As Table 4 shows, neither default risk nor survival is affected by *LOANSIZE.* This has two implications. First, within the size range available for the loans studied here, more credit does not imply greater default risk nor does it imply either shorter or longer survivals. Secondly, the equal parameter signs for $\boldsymbol{\alpha}_2$ and $\boldsymbol{\beta}_2$ demonstrate that greater default risk is associated with shorter survival, not with *longer* survival as Boyes et al. suggest. Within this category of loans, riskier loans are thus very likely to have *lower* expected returns.[8] The lending institution that I study, however, always extended loans with size equal to the amount applied for - independent of the risk associated with the applicant - and thus did not trade off higher default risk against higher expected earnings (coming from bigger loans). Such behavior is compatible with neither return or survival time maximization nor default risk minimization, as the lending institution can increase its earnings by granting bigger loans without increasing their riskiness. One possible explanation could be that the bank attempts to maximize some other objective function than the rate of return on their loan portfolio, e.g., the number of customers, lending volume subject to a minimum return constraint, or total profits from a range of financial products. Any such alternative is, however, not in agreement with the practice reported to me by the lending institution that provided our data. Except for rejecting applicants that are too risky, the lending institution has no explicit policy. In fact, loan officers do not dispose of any information neither the (historical or expected) behavior or return on a loan or the range of products that an applicant already obtains at the lending institution. Thus although we cannot preclude the possibility that another economic explanation does exist for this apparently inefficient behavior, the big discrepancy between the loan granting policy and loans' survival/default behavior provide strong evidence that the lending institution was far from minimizing risk or maximizing returns.[9]

[8] But see the earlier comment in footnote 7.

[9] Additional costs or sources of income, such as annual fees, that are unrelated or not linearly related to duration, could cause deviations from a risk minimizing rule. As was mentioned in Section 2, however, no fixed annual fees were charged by the bank. Discriminatory behavior, either positive or negative, that is not model-based, could also lead to (probably minor) deviations from an optimizing rule. Alternatively, special in-store promotional activities towards special customer groups could have taken place, for example towards women. No such policy or activity was reported to us by the lending institution, however. Moreover, the possibility of

Ultimately, the practical importance of the sample selection bias will depend on the gain in efficiency one can achieve by using the bivariate Tobit model $(1) - (4)$, or, in other words, the model's ability to select the best candidate loans among a set of loan applicants. For this purpose, I compare the selection accuracy of the bivariate Tobit model with that of two other models discussed above: the (unbiased) bivariate probit model of default and the (biased) univariate Tobit model of survival time, and a commonly used, directly related, model that I didn't discuss: the (biased) univariate probit model of default.[10] The latter model's parameter estimates are displayed in Table 6 in Appendix B. To evaluate the model's accuracy, I report seven indicators that shed light on the selective ability in a number of different dimensions in Table 5.

First, I look at each model's ability to single out loans that actually defaulted, because the minimum a good model ought to be able to achieve is to ex-post identify a relatively large share of the defaulted loans as being among the riskiest applicants, when sorting all loan applications according to their predicted survival time $E\left[t_i^* | \mathbf{x}_{2i}, \widehat{\beta_2}\right]$ or probability of default $E\left[PD_i^* | \mathbf{x}_{2i}, \widehat{\alpha}_2\right]$. The first six rows of Table 5 show that the bivariate models are significantly better at sifting out the the defaults than the biased models, with the Tobit models marginally outperforming the probit models. Of all defaults, 51 percent is ranked among the 10 percent with the shortest survival time; another 16 and 11 percent are in the second and third decile, while only 11 percent are considered to be among the best five deciles. Were the bank to use this model to make its loan granting decisions, then only about 10 percent of the actually defaulted debtors would have been given credit. This corresponds to .62 percent of the granted loans, compared with 6.03 percent in the sample portfolio. Although the biased Tobit model does only marginally worse in terms of the share of defaults granted credit (.68 percent), its performance is clearly poorer in the upper deciles. Of all defaults 22 percent would have been identified as belonging to the least promising decile, with the following deciles accounting for 28, 17, 12 and 7 percent. If the bank, for example, would like to grant a larger share of its applicants credit using the biased Tobit model, then its bad performance in the upper deciles would lead it to take more risk than necessary. The two probit models displayed in the last two columns have a

discriminatory behavior can be excluded since the loan granting process is the sole responsibility of the lending institution and handled by telephone at the institution - not in the store.

[10] See Jacobson and Roszbach [10] and Boyes et al. [4] for a comparison between the univariate and bivariate probit models.

Table 5: Selection accuracy of the models

The table presents a comparision of the in-sample selection accuracy of the survival time equation of the bivariate Tobit and univariate Tobit models displayed in Table 4, the default risk equation of bivariate probit model displayed in Table 5 and a univariate probit model of default risk shown in Appendix B. All statistics are based on the assumption that the financial institution will grant 6438 loans, as in the sample.

		Model			
		Tobit		Probit	
	Unit	Bivariate	Univariate	Bivariate	Univariate
Actual defaults identified in the model's k-th riskiest decile [deciles are declining in risk]					
Decile 1	share	0.51	0.22	0.48	0.21
Decile 2	- " -	0.16	0.28	0.16	0.29
Decile 3	- " -	0.11	0.17	0.10	0.17
Decile 4	- " -	0.07	0.12	0.08	0.12
Decile 5	- " -	0.04	0.07	0.06	0.07
Decile 6-10	- " -	0.11	0.14	0.11	0.13
Actual defaults granted or rejected by model i relative to total number of loans					
Defaults selected	percentage	0.62	0.68	0.65	0.68
Defaults rejected	- " -	5.41	5.34	5.37	5.34
Actual defaults	- " -	6.03			
Actual riskiness [according to unbiased model] of loans that would be granted under a biased rule [deciles are declining in risk]					
Decile 1	share		0.00		0.00
Decile 2	- " -		0.01		0.00
Decile 3	- " -		0.02		0.01
Decile 4	- " -		0.05		0.02
Decile 5	- " -		0.09		0.08
Decile 6	- " -		0.13		0.14
Decile 7	- " -		0.15		0.17
Decile 8	- " -		0.18		0.19
Decile 9	- " -		0.19		0.19
Decile 10	- " -		0.19 +		0.20 +
			1.00		1.00
Share of best loans [according to unbiased model] that would be granted under a biased rule					
	share		0.82		0.86
Average quality of loans that would be granted under a biased rule [evaluated using (un)biased model] relative to the quality obtained under an unbiased rule [evaluated using unbiased model]					
Using unbiased model	ratio		0.94		1.39
Using biased model	- " -		0.26		2.09
Average quality of granted applicants relative to rejected ones [evaluated using unbiased model]					
If model based	ratio	6.60		0.07	
In sample	- " -	0.39		0.75	
Average quality of applicants that would be selected using model relative to actual loans [both evaluated using unbiased model]					
	ratio	3.22	3.04	0.15	0.22

Above (higher) quality refers to (longer) survival times for the Tobit model and (smaller) probabilities of default for the probit model. Unbiasedness (biasedness) refers to the bivariate (univariate) models.

selection accuracy that is in line with that of the Tobit models.

The third block of Table 5 confirms our intuition that using a biased model leads to both unnecessary risk-taking and shorter survivals. Of the 6438 applicants that the bank would grant credit with the help of the bivariate Tobit model, only 82 percent would have been so under the univariate model. For the probit models, this figure is 86 percent. Lines 10 to 19 in the table show that the biased models do quite well in identifying the very best applicants, but loose this ability when moving to the applicants in the middle deciles of the survival time and default risk distributions. The biased models would, for example, select only 65-70 of the applicants in the sixth decile - that would all have been accepted by the "unbiased" bank. Instead, the biased banks would grant loans to applicants in the second to fifth deciles, where default risk is much higher and expected survival times much shorter.

Being good at picking out actual defaults among granted loans is not enough, though, for the bivariate Tobit model to pass the test. Another desirable property one would like the model to display is for it to also recognize differences in creditworthiness among applicants who were granted credit and did not default and those were denied credit. The last three lines of Table 5 show that the bivariate Tobit model would allow the bank to select applicants with expected survival times that are more than three times longer than in the sample portfolio. Using the biased model would also result in a significant increase of survival time, but the rise would turn out 18 percent smaller. With the probit models one obtains a similar improvement. Using the bivariate probit model reduces the portfolio's average default risk by 85 percent, compared with 78 percent for the biased probit model.

The bivariate model is not only better than both the bank and the univariate model at sorting applicants according to their survival time, it is also relatively good at picking out the véry best ones: granted loans have a survival time that is 6.6 times as long as that of rejected applicants. In the sample portfolio, the granted applications actually had shorter expected survival times than the rejected ones did. The probit model also sorts applicants in a more efficient way than the bank. People that would be granted credit in a model-based portfolio have an average probability of default risk that is merely 7 percent of that of rejected applicants. In the bank's sample portfolio this figure is only 75 percent. These figures also confirm that survival time is not just a simple transformation of default risk. While the bank at least succeeds in selecting applicants with on average lower default risk than those who are denied credit, it is not able to pick out applicants with longer survival times. The reason for this difference is that, although the debtors who (already) defaulted in the sample portfolio often are those with the shortest expected survivals, big differences exist in the expected survival of rejected applicants and accepted applicants who have not defaulted (yet).[11]

5. Discussion

Traditionally, the objective of credit scoring models used by financial institutions is to minimize default rates or the number of loans that is incorrectly classified as defaulted or non-defaulted. From a profit or utility maximizing perspective, however, it is not only important to know *if* but also *when* a loan will default. Traditional credit scoring models predict default risk and therefore fail to take into account this multiperiod nature of loans contracts. To allow for a more realistic evaluation of the return on a loan, a Tobit model with sample selection and variable censoring thresholds has been constructed and estimated in this paper. This model is shown to be a useful tool to predict the expected survival time on a loan to any kind of applicant. A comparison with a nested model that disregards rejected applications - as has been common in studies of creditworthiness - shows that ignoring the sample selection effect leads to a large bias in the parameters estimates and an inefficient lending policy.

From the empirical results we gain several insights. They confirm the findings in Boyes et al. that financial institutions' lending policies are not compatible with default risk minimization. At the same time, though, the results also conflict with the notion that the financial institution would be trading off higher default risk against higher returns. The lending policy does not favor people that survive longer and thus are likely to have higher rates of return. Firstly, some of the variables that increase (decrease) applicants' odds of obtaining a loan reduce (raise) the expected survival time (and thus return) on a loan and raise (reduce) the likelihood of a default. Secondly, the financial institution is found to be indifferent

[11] Note also that the Tobit model basically implies a non-linear weighting scheme. Observations with exceptionally short survival times get a relatively larger weight in the estimation than defaulted loans do. It is therefore not unlikely that the selection accuracy of the Tobit model relative to the probit model would increase were one to monitor the sample portfolio later on and re-estimate the models. This could be caused by the fact that some still performing loans will default later one. in the probit model.

between loans of different sizes, given its expected survival time. There is thus no evidence of banks' behaving in a way that is consistent with profit-maximization. This impression is strengthened by an analysis of the selection accuracy of the bivariate Tobit model and a comparison with other frequently used models. Such an analysis shows that the bivariate Tobit model has a greater ability to pick out future defaults and select applicants with longer survival times. The model also significantly reduces the riskiness compared to the sample portfolio and increases the average survival time. While a bivariate probit model reduces default risk by up to 85 percent, the Tobit model increases the portfolio's average survival time by a factor two compared.

The lending behavior studied here is thus very likely to be a symptom of an inefficient lending policy. However, I cannot exclude the possibility that the financial institution is aiming at some other objective than profit maximization or default risk minimization, for example a composite objective such as the return on a range of products or revenues from several sources of income. Banks may also be maximizing some other objective like provision income from the turnover on credit cards, the number of customers or lending volume subject to a minimum return constraint. None of these suggestions agree, however, with the practices reported by the lending institution that provided the data. Rather, the results bear strong evidence of a lending institution that has attempted to minimize risk or maximize a simple return function without success.

Censoring of data, as is the case with the non-defaulted loans in the sample, increases the uncertainty in the parameter estimates of the survival function. Appropriate changes in sampling methods can improve their accuracy. A longer period of observation of the loans would reduce the regression error. Even better would be to set up an experiment where a predetermined number of applicants is granted a loan without consideration of their personal characteristics. If each loan is monitored at least at termination of the contract then separate survival time functions for good and bad loans can be estimated. An ideal model of bank profitability or bank efficiency will have to be built on time series data for fees, interest payments and amortizations on loans, personal characteristics, macroeconomic indicators and all costs involved.

References

[1] Altman, Edward I., Robert B. Avery, Robert A. Eisenbeis and Joseph F. Sinkey, (1981), *Application of classification techniques in business, banking*

and finance, JAI Press, Greenwich, CT.

[2] Arminger, Gerhard, Daniel Enache and Thorsten Bonne, (1997), Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis and feedforward networks, *Computational Statistics* 12, 293-310.

[3] Bermann, Georgina, (1993), *Estimation and inference in bivariate and multivariate ordinal probit models*, dissertation, Department of statistics, Uppsala University.

[4] Boyes, William J., Dennis L. Hoffman and Stuart A. Low, (1989), An Econometric Analysis of the Bank Credit Scoring Problem, *Journal of Econometrics* 40, 3-14.

[5] Carling, Kenneth, Tor Jacobson and Kasper F. Roszbach, (2001), Dormancy risk and expected profits of consumer loans, *Journal of Banking and Finance* 25 (4), 717-739.

[6] Dionne, George, Manuel Artís and Montserrat Guillén, (1996), Count data models for a credit scoring system, *Journal of Empirical Finance* 3, 303-325.

[7] Gale, Douglas, and Martin Hellwig, (1985), Incentive-compatible debt contracts: The one-period problem, *Review of Economic Studies* LII, 647-663.

[8] Greene, William, (1993), *Econometric Analysis*, 2nd edition, Macmillan, New York.

[9] Henley, W.E., and David J. Hand, (1996), A k-nearest-neighbor classifier for assessing consumer credit risk, *The Statistician* 45 (1), 77-95.

[10] Jacobson, Tor, and Kasper F. Roszbach, (2003), Bank lending policy, credit scoring and Value-at-Risk, *Journal of Banking and Finance* 27 (4), pp. *615-633.*.

[11] Shumway, Tyler, (2001), Forecasting bankruptcy more accurately: a simple hazard model, *Journal of Business* 74 (1), pp.101-124.

[12] Stiglitz, Joseph E., and Andrew Weiss, (1981), Credit rationing in markets with imperfect information, *American Economic Review* 71, 393-410.

[13] Townsend, Robert M., (1982), Optimal multiperiod contracts and the gain from enduring relationships under private information, *Journal of Political Economy* 90 (61), 1166-1186.

[14] Williamson, Stephen, (1987), Costly monitoring, loan contracts and equilibrium credit rationing, *Quarterly Journal of Economics* 102 (1), 135-145.

A. Likelihood function and gradient

A.1. Likelihood function

The likelihood function

$$\begin{aligned}\ell \;=\; & \prod_{i=1}^{N} \operatorname{pr}\left(y_i^* < 0\right)^{(1-y_i)} \cdot \prod_{i=1}^{N} \operatorname{pr}\left(y_i^* \geq 0,\ t_i\right)^{y_i \cdot (1-d_i)} \times \\ & \prod_{i=1}^{N} \operatorname{pr}\left(y_i^* \geq 0,\ t_i^* \geq \overline{t_i}\right)^{y_i \cdot d_i}\end{aligned} \tag{9}$$

implies that

$$\begin{aligned}\ln\ell \;=\; & \textstyle\sum_{i=1}^{N} (1-y_i) \cdot \ln\left[\operatorname{pr}\left(\varepsilon_{1i} < -\mathbf{x}_{1i}\boldsymbol{\beta}_1\right)\right] + \\ & \textstyle\sum_{i=1}^{N} y_i \cdot (1-d_i)\ \ln\ \left[\operatorname{pr}\left(\varepsilon_{1i} \geq -\mathbf{x}_{1i}\boldsymbol{\beta}_1 \cap \varepsilon_{2i} = t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2\right)\right] + \\ & \textstyle\sum_{i=1}^{N} y_i \cdot d_i\ \ln\left[\operatorname{pr}\left(\varepsilon_{1i} \geq -\mathbf{x}_{1i}\boldsymbol{\beta}_1 \cap \varepsilon_{2i} \geq \overline{t_i} - \mathbf{x}_{2i}\boldsymbol{\beta}_2\right)\right]\end{aligned} \tag{10}$$

If we use that $\varepsilon_{1i}|\varepsilon_{2i} \sim N\left(\frac{\sigma_{12}}{\sigma_2^2}\varepsilon_{2i},\ (1-\rho^2)\right)$ for $\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim \mathbf{N}\left(\mathbf{0},\ \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, then we can simplify by expressing the second line in terms of a univariate Normal cdf and - pdf. As a result

$$\begin{aligned} & \operatorname{pr}\left(\varepsilon_{1i} \geq -\mathbf{x}_{1i}\boldsymbol{\beta}_1 \cap \varepsilon_{2i} = t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2\right) \Leftrightarrow \\ & \operatorname{pr}\left(\varepsilon_{1i} \geq -\mathbf{x}_{1i}\boldsymbol{\beta}_1 \middle|\ \varepsilon_{2i} = t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2\right) \operatorname{pr}\left(\ \varepsilon_{2i} = t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2\right) \Leftrightarrow \\ & \operatorname{pr}\left(\varepsilon_{1i} < \mathbf{x}_{1i}\boldsymbol{\beta}_1 \middle|\ \varepsilon_{2i} = t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2\right) \operatorname{pr}\left(\ \varepsilon_{2i} = t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2\right) \Leftrightarrow \\ & \Phi\left(\frac{\mathbf{x}_{1i}\boldsymbol{\beta}_1 - \frac{\sigma_{12}}{\sigma_2^2}(t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2)}{\sqrt{(1-\rho^2)}}\right) \frac{1}{\sigma_2}\phi\left(\frac{t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2}\right)\end{aligned} \tag{11}$$

Taking natural logarithms we get

$$\ln\Phi\left(\frac{\mathbf{x}_{1i}\boldsymbol{\beta}_1 - \frac{\sigma_{12}}{\sigma_2^2}(t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2)}{\sqrt{(1-\rho^2)}}\right) - \frac{1}{2}\ln 2\pi + \ln\left(\frac{1}{\sigma_2}\right) - \frac{1}{2}\left(\frac{t_i - \mathbf{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2}\right)^2 \tag{12}$$

The last line in (10) can be rewritten in terms of a bivariate Normal c.d.f.:

$$\begin{array}{l} \text{pr}\left(\varepsilon_{1i} \geq -\mathbf{x}_{1i}\boldsymbol{\beta}_1 \cap \frac{\varepsilon_{2i}}{\sigma_2} \geq \frac{\overline{t_i}-\mathbf{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2}\right) \Leftrightarrow {}^{12} \\ \Phi_2\left(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\ \frac{\mathbf{x}_{2i}\boldsymbol{\beta}_2-\overline{t_i}}{\sigma_2};\ \rho\right) \end{array} \tag{13}$$

Consequently, the loglikelihood function can be written as

$$\begin{array}{rcl} \ln\ell & = & \sum_{i=1}^{N}(1-y_i)\cdot\ln\left[1-\Phi\left(\mathbf{x}_{1i}\boldsymbol{\beta}_1\right)\right]+ \\ & & \sum_{i=1}^{N} y_i\cdot(1-d_i)\ \left\{\ln\Phi\left(\frac{\mathbf{x}_{1i}\boldsymbol{\beta}_1-\ \frac{\sigma_{12}}{\sigma_2^2}(t_i-\mathbf{x}_{2i}\boldsymbol{\beta}_2)}{\sqrt{(1-\rho^2)}}\right)+\right. \\ & & \left.\quad -\frac{1}{2}\ln 2\pi+\ln\left(\frac{1}{\sigma_2}\right)-\frac{1}{2}\left(\frac{t_i-\mathbf{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2}\right)^2\right\}+ \\ & & \sum_{i=1}^{N} y_i\cdot d_i\ \ln\Phi_2\left(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\ \frac{\mathbf{x}_{2i}\boldsymbol{\beta}_2-\overline{t_i}}{\sigma_2};\ \rho\right) \end{array} \tag{14}$$

After further simplification, by setting

$$\begin{array}{rcl} \boldsymbol{\alpha}_2 & = & \boldsymbol{\beta}_2/\sigma_2 \\ lng2 & = & \ln\left(\frac{1}{\sigma_2}\right) \\ e_{2i} & = & -\left(\frac{t_i-\mathbf{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2}\right) \\ & = & \mathbf{x}_{2i}\boldsymbol{\alpha}_2-t_i\cdot\exp\left(lng2\right) \\ \overline{e_{2i}} & = & \mathbf{x}_{2i}\boldsymbol{\alpha}_2-\overline{t_i}\cdot\exp\left(lng2\right) \\ \delta & = & \frac{1}{(1-\rho^2)^{1/2}} \end{array}$$

we get that

$$\begin{array}{rcl} \ln\ell & = & \sum_{i=1}^{N}(1-y_i)\cdot\ln\left[1-\Phi\left(\mathbf{x}_{1i}\boldsymbol{\beta}_1\right)\right]+ \\ & & \sum_{i=1}^{N} y_i\cdot(1-d_i)\ \ \{\ln\Phi\left(\delta\left[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\ \rho\cdot e_{2i}\right]\right)+ \\ & & -\frac{1}{2}\ln 2\pi+lng2-\frac{1}{2}e_{2i}^2\}+ \\ & & \sum_{i=1}^{N} y_i\cdot d_i\ \ln\Phi_2\left(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\ \overline{e_{2i}};\ \rho\right) \end{array} \tag{15}$$

The parameters with respect to which we maximize $\ln\ell$ are $\boldsymbol{\beta}_1, \boldsymbol{\alpha}_2$, $lng2$, and ρ.

A.2. Gradients

The gradients corresponding to each observation in the above loglikelihood function are:

[12] See Greene [8] p.661 for a summary of results on the bivariate normal cdf.

$$
\begin{aligned}
\frac{\partial \ln \ell_i}{\partial \boldsymbol{\beta}_1} &= (1-y_i)\,\frac{-\phi(\mathbf{x}_{1i}\boldsymbol{\beta}_1)}{1-\Phi(\mathbf{x}_{1i}\boldsymbol{\beta}_1)}\cdot \mathbf{x}_{1i}+\\
&\quad y_i\,(1-d_i)\cdot \frac{\phi(\delta[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\rho e_{2i}])}{\Phi(\delta[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\rho e_{2i}])}\cdot \delta \mathbf{x}_{1i}+\\
&\quad y_i\cdot d_i\;\frac{\phi(\mathbf{x}_{1i}\boldsymbol{\beta}_1)\Phi\left(\frac{\overline{e_{2i}}-\rho\mathbf{x}_{1i}\boldsymbol{\beta}_1}{(1-\rho^2)^{1/2}}\right)}{\Phi_2(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\ \overline{e_{2i}};\ \rho)}\cdot \mathbf{x}_{1i}\\
\frac{\partial \ln \ell_i}{\partial \boldsymbol{\alpha}_2} &= y_i\cdot(1-d_i)\;\left\{\frac{\phi(\delta[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\rho e_{2i}])}{\Phi(\delta[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\rho e_{2i}])}\delta\,(-\rho)\,\mathbf{x}_{2i}-e_{2i}\mathbf{x}_{2i}\right\}+\\
&\quad y_i\cdot d_i\frac{\phi(\overline{e_{2i}})\Phi\left(\frac{\mathbf{x}_{1i}\beta_1-\rho\overline{e_{2i}}}{(1-\rho^2)^{1/2}}\right)}{\Phi_2(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\ \overline{e_{2i}};\ \rho)}\cdot \mathbf{x}_{2i}\\
\frac{\partial \ln \ell_i}{\partial 1ng2} &= y_i\cdot(1-d_i)\left\{\frac{\phi(\delta[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\rho e_{2i}])}{\Phi(\delta[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\rho e_{2i}])}\delta\,(-\rho)\,(-t_i)\exp(lng2)+\right.\\
&\qquad\qquad \left.+1+e_{2i}\cdot t_i\cdot\exp(lng2)\right\}+\\
&\quad y_i\cdot d_i\left\{\frac{\phi(\overline{e_{2i}})\Phi\left(\frac{\mathbf{x}_{1i}\beta_1-\rho\overline{e_{2i}}}{(1-\rho^2)^{1/2}}\right)}{\Phi_2(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\ \overline{e_{2i}};\ \rho)}\cdot\left(-\overline{t_i}\cdot\exp(lng2)\right)\right\}\\
\frac{\partial \ln \ell_i}{\partial \rho} &= y_i\cdot(1-d_i)\;\frac{\phi(\delta[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\rho e_{2i}])}{\Phi(\delta[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\rho e_{2i}])}\left(-\delta e_{2i}+\rho\delta^3\left[\mathbf{x}_{1i}\boldsymbol{\beta}_1-\ \rho e_{2i}\right]\right)+\\
&\quad y_i\cdot d_i\cdot\frac{\phi_2(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\overline{e_{2i}};\ \rho)}{\Phi_2(\mathbf{x}_{1i}\boldsymbol{\beta}_1,\ \overline{e_{2i}};\ \rho)}
\end{aligned}
\tag{16}
$$

After convergence of the iterative procedure, a consistent estimator of the variance-covariance matrix is obtained by applying the delta method.[13] If we define $\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \sigma_2\\ \rho\end{pmatrix}=\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\alpha}_2\left[\exp(lng2)\right]^{-1}\\ \left[\exp(lng2)\right]^{-1}\\ \rho\end{pmatrix}\equiv \mathrm{f}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}=\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\alpha}_2\\ lng2\\ \rho\end{pmatrix}$, then

$$
\boldsymbol{\Gamma}\equiv\frac{\partial \mathrm{f}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}=\begin{bmatrix}\mathbf{I}_{K_1} & \mathbf{0}_{\mathbf{K}_1\times\mathbf{K}_2} & \mathbf{0}_{\mathbf{K}_1\times 1} & \mathbf{0}_{\mathbf{K}_1\times 1}\\ \mathbf{0}_{\mathbf{K}_2\times\mathbf{K}_1} & \frac{\mathbf{I}_{K_2}}{[\exp(\mathrm{lng2})]} & \frac{-\boldsymbol{\alpha}_2}{[\exp(\mathrm{lng2})]} & \mathbf{0}_{\mathbf{K}_2\times 1}\\ \mathbf{0}_{1\times\mathbf{K}_1} & \mathbf{0}_{1\times\mathbf{K}_2} & \frac{-1}{[\exp(\mathrm{lng2})]} & 0\\ \mathbf{0}_{1\times\mathbf{K}_1} & \mathbf{0}_{1\times\mathbf{K}_2} & 0 & 1\end{bmatrix}
\tag{17}
$$

In some of the iterations, I also made the additional transformation $\rho=\frac{\exp x-1}{\exp x+1}$ to assure that $0\le\rho\le 1$. In those cases the gradient $\partial\mathrm{ln}\ell/\partial\rho$ needs to be multiplied by the term $\partial\rho/\partial x=\frac{2\exp x}{(\exp x\ +1)^2}$ while the 4th diagonal element in $\boldsymbol{\Gamma}$ needs to be set equal to $\partial\rho/\partial x$.

[13] A description of the delta method is provided in Greene [8], pp. 297.

B. Additional tables

Table 6: Univariate MLE of default equation

The univariate estimator of α_2 comes from separate estimation of the second equation in (1) respectively, so $\rho = 0$ is assumed.

	Equation	
	α_2	α_2
Variable	Univariate	Bivariate
CONSTANT	**2.032**	**2.460**
	(0.143)	(0.113)
BIGCITY	**-0.157**	-0.042
	(0.063)	(0.055)
DIVORCE	**-0.177**	-0.071
	(0.084)	(0.074)
MALE	**-0.118**	0.024
	(0.067)	(0.059)
MARRIED	**0.175**	**0.253**
	(0.073)	(0.066)
CAPINC	0.026	0.142
	(0.149)	(0.127)
ENTREPR	0.176	0.147
	(0.218)	(0.157)
HOUSE	0.049	-0.011
	(0.069)	(0.060)
INCOME	0.048	**-0.226**
	(0.051)	(0.050)
DIFINC	**0.189**	**0.206**
	(0.086)	(0.073)
COAPPLIC	**0.582**	**0.421**
	(0.110)	(0.098)
ZEROLIM	**-2.810**	**-0.703**
	(0.423)	(0.306)
LIMIT	0.024	**0.486**
	(0.080)	(0.056)
LIMUTIL	**-1.538**	**-1.210**
	(0.115)	(0.093)
LOANSIZE	**-0.088**	-0.067
	(0.083)	(0.069)
NRLOANS	**0.383**	**0.270**
	(0.026)	(0.020)
NRQUEST	**-0.136**	**-0.104**
	(0.012)	(0.010)
Pseudo R^2	*0.42*	*0.31*
Loglikelihood	*-1029.05*	*-7529.76*

Standard errors in parentheses. Variables significant at a significance level of at least 10 percent are printed in bold style. Estimates for the bivariate model are copied from Table 4.

Table 6 presents the parameter estimates for a probit model of default risk that disregards the sample selection effect. These parameter estimates are used to generate the results in the last column of Table 5. To facilitate comparison with the bivariate probit parameter estimates, the latter have been copied from Table 4. Table 7 contains a list of all variables in the original dataset, both from the credit bureau and the financial institution..

Table 7: Original dataset

GENDER	Gender
AGE	Age
MARISTAT	Unmarried, married man, married woman living together (LT), marr. woman not LT, divorced, widow(er), deceased
MSCHANGE	Months since change in marital status
NAT	Swedish, Nordic, Other, Stateless
IMDAT	Months since immigration
ZIP	Zip code
FIRMA	Owns company
NR-RE	Pieces of real estate that are owned (partially)
RESHARE	Assessed value (AV) of *NR-RE* / *REVAL*
REVAL	Total AV of real estate property in which shares are owned
NRQUEST	No. info. requests credit agency received in past 3 years
NRQUESTP	Idem, but only as a private person
NRQUESTB	Idem, but only as a business
TXYEAR	Taxation year
INCOME	Annual income before taxes
INCATAX	Income after taxes
TAX15	Business income
TAX17	Wage income
TAX18	Capital income
TAX19	Capital costs
TAX21	General deductions
TAX30	Individual's share of the family's assets
TAX31	Family's total assets

Table 7: Original dataset - continued

TAX41	Final taxes
TAX42	Prepaid income tax (as employee)
TAX43	Prepaid income tax (as a business)
RESTTAX	Final taxes minus prepaid taxes
LIMIT	$\sum$ all collateral free credit registered at UC excl. application
BALANCE	Utilized amount of collateral free credit
NRLOANS	Number of loans registered at UC
CODE	Emigrated, deceased, not registered at UC, secret address, ID card lost, declaration of incapacity
PRODUCT	Product type
ACCOND	Code that refers to the contract's conditions.
ACREM	No. reminders since loan was sent to a debt collection agency [can return to zero again]
REMDAT	Date of last change in *ACREM*
ACSTA	Status of an account/loan: rejection, active, bad, cancelled (when unused too long), ended by customer, ended by lending institution. [Useless due to registration delays]
STADAT	Latest change in *ACSTA*: equals *BADDAT* if loan turned bad.
BADDAT	Date at which loan was forwarded to debt collection agency
LOSSDAT	Date at which loss was registered (after collection attempt)
LOANSIZE	Amount of credit granted
UTLOAN	Amount of credit utilized
SHOP	Shop number
MAINSHOP	Main shop number
CHAIN	Chain to which shop belongs
LOSS	Registered loss [long delays in registration]
INQUIRY	Individuals for which an inquiry could be made at UC (0/1)
DECDAT	Date at which decision by bank on application was made
COAPPLIC	Applicant has a guarantor

Earlier Working Papers:

Sveriges Riksbank
Visiting address: Brunkebergs torg 11
Mail address: se-103 37 Stockholm

Website: www.riksbank.se
Telephone: +46 8 787 00 00, Fax: +46 8 21 05 31
E-mail: registratorn@riksbank.se

ISSN 1402-9103 Riksbank Printing Office Stockholm 2003