Journal of Empirical Finance 16 (2009) 457–465

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Estimation of default probabilities using incomplete contracts data

J.M.C. Santos Silva [a,c,*], J.M.R. Murteira [b,c,*]

[a] Department of Economics, University of Essex, UK
[b] Faculdade de Economia, Universidade de Coimbra, Portugal
[c] CEMAPRE, Portugal

ARTICLE INFO

Article history:
Received 4 October 2007
Received in revised form 11 August 2008
Accepted 11 November 2008
Available online 24 November 2008

JEL classification code:
C21
C51
G21

Keywords:
Beta-binomial distribution
Credit scoring
Population drift

ABSTRACT

This paper develops a count data model for credit scoring which allows the estimation of default probabilities using incomplete contracts data. The main advantage of the proposed approach is that it permits a more efficient use of the data, including that for the most recent clients. Moreover, because the probability of default is specified as a function of the age of the contract, the model provides some information on the timing of the defaults. The model is based on the beta-binomial distribution, which is found to be particularly adequate for this purpose. A well-known dataset on personal loans is used to illustrate the application of the proposed model.

1. Introduction

Models for credit scoring are widely used in practice and raise a number of interesting and challenging research questions. Therefore, it is not surprising to find that they have been the subject of a considerable literature (see, among many others, Altman et al., 1981; Maddala, 1996; Hand and Henley, 1997; Hand and Jacka, 1998; Thomas, 2000; Thomas et al., 2002, and the references therein).

Consider a lending institution, hereinafter referred to as the bank, that wants to use information on the characteristics and repayment behavior of its clients to estimate the probability of a prospective borrower to default on a loan.[1] The nature of the statistical methods to use in the estimation of this model will depend on the type of loan being considered. Indeed, different types of loans generate data with different characteristics, and that is critical for the choice of the statistical methodology to use.

In this paper we restrict our attention to a particular, yet important, type of loan. Specifically, we develop a model for the probability of default for the case in which the loan is to be repaid in a number of regular installments.[2] Scoring models for this type of loan are typically estimated using only data on contracts that have reached their maturity. However, models constructed in this way waste the information on clients who are currently repaying their loans. This is inappropriate, not only because it is an inefficient use of the data, but also because the resulting model may be affected by *population drift* problems, caused by changes in

* Corresponding authors. Santos Silva is to be contacted at Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. Murteira, Faculdade de Economia, Universidade de Coimbra, Av. Dias da Silva, 165, 3004-512 Coimbra, Portugal.
E-mail addresses: jmcss@essex.ac.uk (J.M.C. Santos Silva), jmurt@fe.uc.pt (J.M.R. Murteira).

[1] In this paper only the probability of default is modelled. Although this is only a part of the optimization problem faced by the lending institution, it is of critical importance both for its profitability and for the households' welfare (see Carling et al., 2001).

[2] Therefore, the model developed here may be inappropriate to describe defaults when the loans take the form of overdrafts, or are granted through credit cards. Thomas (2000) surveys different techniques that are useful in constructing scoring models for these other types of loans.

0927-5398/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jempfin.2008.11.003

the distribution of the characteristics of the clients (Kelly et al., 1999). To mitigate these problems, in practice, current clients are often included in the sample and are classified according to their present status. However, this procedure will inevitably induce a degree of misclassification because some clients currently classified as non-defaulters may eventually default before the end of their contracts.

This paper shows that, using an appropriate count data model for the number of payments missed by the borrowers, it is possible to use the data on all contracts, including the more recent ones, to estimate the conditional probability of a client becoming a defaulter. This is particularly important in the case of long term contracts, e.g., mortgages, because in these cases the characteristics and repayment behavior of clients with completed contracts may have little to do with those of the prospective borrowers. Moreover, for smaller and newer banks, the number of observations on clients with completed long term contracts may be very small.

The proposed model also has some additional advantages. First, by modelling the actual number of missed payments rather than just an indicator of default, the model explores more of the available information. Second, it allows the probability of default to depend on the age of the contract, thereby providing some information on the way the probability of default varies in time. The timing of defaults has been previously addressed in the literature (e.g., Roszbach, 2004; Duffie et al., 2007), but the particular characteristic of the model developed here is that it allows the researcher to estimate the probability of default for different time horizons, without actually using any duration data on the timing of defaults. Finally, the proposed modelling strategy is interesting because it can be adapted to a variety of circumstances and permits the test of a number of interesting hypothesis, like the possible change in repayment behavior after the client is considered a defaulter.

An issue that has to be addressed in the construction of credit scoring models is the potential sample selection problem caused by the fact that the bank only has information on clients to whom it has decided to grant a loan. This situation is problematic if the decision to accept or refuse the credit applications is made using information on the clients that is not available for the construction of the credit scoring model. In this case the sample is endogenously stratified and there is not much that can be done to solve the problem without relying on very strong assumptions (see Hand and Henley, 1993). On the contrary, if all the information used to decide about the credit applications is available for the construction of the credit scoring model, standard inference methods can be used because in this case the sample is exogenously stratified (see Pudney, 1989; Wooldridge, 1999). This more favorable situation is the one considered here.

The remainder of the paper is organized as follows. The next section introduces a count data model that allows the estimation of default probabilities using data on incomplete contracts. Section 3 presents, purely for illustrative purposes, an application of the proposed model using a well-known dataset on personal loans, and Section 4 concludes the paper.

2. An appropriate count data model

Consider a loan that has to be repaid in N regular installments and let n denote the present age of the contract, measured by the number of installments that should have been paid since the contract began. Furthermore, let Y be the number of payments so far missed by a client. Therefore, we have that $0 \leq Y \leq n \leq N$.

The purpose of the scoring model is to estimate the probability that Y will be larger than the maximum number of payments that a client is allowed to miss without being considered a defaulter, denoted by l. Obviously, this probability is zero for any $n \leq l$.

For clients whose loans have reached their maturity date, i.e., $n = N$, it is possible to construct a binary variable indicating whether or not the client defaulted. Credit scoring models are typically estimated by using appropriate statistical methods to model these indicators. The main drawback of this approach is that is wastes information, non only on the current clients whose final status is yet unknown, but also on the actual number of payments missed by the former clients.

Rather than just modelling the default indicator, the alternative approach we follow here is to model the count variable Y, taking into account that this variate is bounded between 0 and n. These bounds on Y have implications for the type of count data model to use. Indeed, the distributions more often used in applied work, e.g., Poisson and negative binomial, are not suitable in this context because they assume that the counts have no upper bound. In order to account for the peculiar characteristics of Y, we use a beta-binomial model, which explicitly imposes that $Y \leq n$. The beta-binomial regression was first used by Heckman and Willis (1977) in a very different context, but has seen little use in practice.

2.1. Beta-binomial regression

Suppose that, besides Y, N and n, the bank observes a set x of characteristics of the contract and of its clients.[3] The objective is then to estimate the probability that Y will cross the threshold above which the client will be classified as a defaulter, given N, n and x.

In order to take explicitly into account the upper bound on the value of Y, the model developed here has as a starting point the binomial distribution characterized by

$$P(Y = y|n) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \tag{1}$$

[3] Depending on the nature of the problem, macroeconomic indicators may also be useful predictors of default (see, e.g., Bellotti and Crook, 2007).

where p is the probability that the individual will miss any of the n payments and $y=1,2,\ldots,n$. Even if p is parameterized as a function of x, the simple binomial model defined by Eq. (1) is unlikely to be adequate to describe the credit default data due to the presence of unobserved individual heterogeneity.

To account for extra-binomial variation, we assume that p is distributed in the population as a beta random variable with parameters that depend on the value of N and x. In particular, it is assumed that

$$f(p|N,x)=\Gamma\left(\frac{1}{\alpha}+\frac{1}{\alpha\theta}\right)\frac{p^{\frac{1}{\alpha}-1}(1-p)^{\frac{1}{\alpha\theta}-1}}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{1}{\alpha\theta}\right)},$$

where $f(p|N,x)$ denotes the conditional density function of p, and α and θ are positive parameters that may depend on N and x. Therefore, Y follows a beta-binomial distribution defined by

$$P(Y=y|N,x,n)=\frac{n!}{y!(n-y)!}\frac{\Gamma\left(\frac{\theta+1}{\alpha\theta}\right)\Gamma\left(\frac{1+n\alpha\theta-y\alpha\theta}{\alpha\theta}\right)\Gamma\left(\frac{1+y\alpha}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{1}{\alpha\theta}\right)\Gamma\left(\frac{\theta+1+n\alpha\theta}{\alpha\theta}\right)}, \tag{2}$$

with

$$E(Y|N,x,n)=n\frac{\theta}{1+\theta},$$

$$V(Y|N,x,n)=E(Y|N,x,n)\frac{(1+\theta+n\alpha\theta)}{(\theta+1)(1+\theta+\alpha\theta)}.$$

To complete the model specification it is necessary to define how α and θ depend on N and x. Naturally, the particular form of these functions is an empirical issue and no general rules can be established. However, given that they are positive parameters, and in line with standard practice in count data models, it is natural and convenient to specify α and θ as exponential functions of the covariates and N.[4]

In a credit scoring context, the probability of default is the main object of interest. Recalling that l denotes the maximum number of repayments that a client may miss without being considered a defaulter, the probability of default implied by Eq. (2) can be written as

$$P(Y>l|N,x,n)=1-\sum_{j=0}^{l}P(Y=j|N,x,n). \tag{3}$$

Once the relevant parameters are estimated, the creditworthiness of prospective clients can be gauged by evaluating $P(Y>l|N,x,n)$ for the corresponding values of x and for different values of n , for example $n=N$. Notice that in this sort of model the probability of becoming a defaulter will depend on the time horizon considered (see Fig. 1 below). This is important since, from the point of view of the bank, it is not indifferent when the client becomes a defaulter.

The beta-binomial regression accounts for extra-binomial variation by assuming that p is distributed in the population as a beta random variable. An alternative way to account for extra-binomial variation in Eq. (1) is to model the distribution of the unobservables semi-parametrically, as it is done by Johansson and Palme (1996). There are several reasons why we prefer to use a fully parametric specification based on the beta-binomial distribution.

First of all, the beta distribution is well known for its flexibility (see, e.g., Johnson et al., 1995) and can easily accommodate situations in which the distribution of the unobservables depends on the conditioning variables. This is difficult to do, if at all possible, if the semiparametric approach is adopted. Indeed, in this case it is generally assumed that the unobservables are statistically independent of the covariates. Therefore, although the parametric approach is restrictive (in that it requires the specification of the distribution of the unobservables) it is nevertheless quite flexible and has the advantage of allowing the distribution of the unobservables to depend on the conditioning variables.

Additionally, an interesting feature of the beta-binomial distribution is that, besides this interpretation as a binomial distribution with individual heterogeneity, it can be viewed as giving the total number of successes in n Bernoulli trials when both success and failure are contagious (see Johnson et al., 2005). Therefore, this model can accommodate a situation in which the probability that an individual will miss a certain payment depends on his previous repayment behavior.

Finally, a fully parametric model is generally much easier to estimate and to interpret than a semiparametric specification. In the present case, this motive is strengthened by the fact that this simplicity is certainly of great value for the practitioner in charge of the practical application of the model.

[4] See Wooldridge (1992) for more general specifications. It is also interesting to note that when θ is specified as $\theta=\exp(x'\beta-\ln(N))$, the limiting distribution of Y when N passes to infinity is negative binomial with mean $\lambda=\exp(x'\beta)$ and variance $\lambda+\alpha\lambda^2$.

2.2. A beta-binomial hurdle model

As noted before, the beta-binomial regression has many characteristics that make it appropriate to model the number of missed payments. However, in its basic formulation, the beta-binomial regression ignores the fact that the repayment behavior of a client may change after he is classified as a defaulter. Indeed, the bank may put pressure on defaulters to repay their debts, for instance by threatening to take legal action. In more serious cases, defaulted loans may enter special debt-collection procedures and so, after default, the number of missed payments will have a very different nature.

The fact that the number of missed payments before and after default may have different characteristics suggests that a hurdle model (Mullahy, 1986) will be appropriate. In such a model, the probability that a client becomes a defaulter can be obtained from the specification of the first stage, that is, the stage that describes the behavior of the client before being classified as a defaulter.[5] A hurdle model is also convenient because it provides a simple test to check whether or not the repayment behavior actually changes after default.

If it is assumed that the behavior of the clients changes for $y>l$, the parameters of interest can be estimated maximizing a likelihood function with individual contributions of the form

$$L_1(\theta, \alpha) = [P(Y=y|N, x, n)]^{(1-d)}[P(Y>l|N, x, n)]^d, \tag{4}$$

where $d=I(y>l)$, $P(Y=j|N,x,n)$ is defined in Eq. (2) and $P(Y>l|N,x,n)$ is the probability of default, as defined in Eq. (3).

In case the researcher is also interested in the probabilities of the missed payments after the client is considered a defaulter, say $P^*(Y=y|N,x,n)=P(Y=y|N,x,n,y>l)$, the second stage of the hurdle model has to be estimated. Different specifications of the second stage can be considered. Indeed, $P^*(Y=y|N,x,n)$ can be given by any probability distribution with support on the integers $l+1,l+2,\ldots,n$. A convenient specification of $P^*(Y=y|N,x,n)$ is

$$P^*(Y=y|N,x,n) = \frac{P(Y=y|N,x,n)}{1-\sum_{j=0}^{l} P(Y=j|N,x,n)},$$

where $P(Y=y|N,x,n)$ is again given by Eq. (2), but its parameters are estimated from a likelihood function with individual contributions of the form

$$L_2(\theta,\alpha) = \left[P^*(Y=y|N,x,n)\right]^d. \tag{5}$$

The main advantage of this specification is that it facilitates the test of the hypothesis that the repayment behavior changes for $y>l$. Indeed, in this case, a test for the hypothesis that the parameters of the two parts of the hurdle are identical provides a check for the significance of the possible change of behavior for $y>l$.[6] Naturally, the suitability of this specification is an empirical matter that has to be checked in each application.

3. An empirical illustration

3.1. The data

In this section we use the data studied by Dionne et al. (1996) to illustrate the estimation of default probabilities using incomplete contracts data. The original data consisted of a sample of 4691 clients of a Spanish bank that in May 1989 were repaying personal loans in monthly installments. According to the bank criteria, a client is considered to be a defaulter if he misses more than 3 payments, which corresponds to the usual 90 day default rule. Therefore, in this case, $l=3$.

As explained by the authors of the original study, the personal loans considered in this sample are typically small, short-term and destined to finance the purchase of cars or other durables. However, the loans are not necessarily backed by a collateral.

An important feature of the dataset available for this study is that it contains only the sub-sample of 2446 observations used by Dionne et al. (1996), who deleted all the observations with incomplete records, as well as cases with outlying values of the explanatory or of the dependent variable. In particular, these authors deleted 217 observations with $y>11$. This creates a truncation problem that will have to be taken into account in the modelling stage.[7]

Besides containing information on y, the number of missed payments, and on n, the number of months from the beginning of the contract at the sampling date, this dataset also contains information on some characteristics of the loan and of the client. All the information on the client was provided at the application stage. These variables are described in Table 1. Because all the regressors

[5] In their pioneering work, Dionne et al. (1996) used a hurdle model to describe this kind of data. Moffatt (2005) also used hurdle models to model default. These hurdle models are, however, very different from the one considered here.

[6] To simplify the notation, we do not distinguish the parameters of the two parts of the hurdle model. However, of course, these are generally different.

[7] This truncation was neglected by Dionne et al. (1996) but it is likely to be relevant, even if the interest is restricted to the first stage of a hurdle model.

Table 1
Description of the regressors.

DT6	Total number of payments implied by the contract is larger than 48
AGE1	Client's age group is 18–24 years
AGE2	Client's age group is 25–39 years
AGE3	Client's age group is 40 years or more
DESTIN	Credit is used to purchasing a good with a collateral
ETU1	Client has not completed primary education
ETU2	Client has completed primary education
ETU3	Client has completed higher education
ETU4	Client has a university degree
RECSAL	Client receives the salary through the bank
M1	Client is married, non-homeowner and has a salary <$3000
M2	Client is married, non-homeowner and has a salary ≥$3000
M3	Client is married, homeowner and has a salary <$3000
M4	Client is married, homeowner and has a salary ≥$3000
NM1	Client is not married and non-homeowner
NM2	Client is not married and homeowner
CENTRE	Credit is granted by a store
RESID	Client is resident in the city for at least four years
Z1	Client is resident in the south of the country
Z2	Client is resident in the north of the country
Z3	Client is resident in the east of the country
Z4	Client is resident in the centre of the country

Note: The table presents the definition of the regressors used in this study. Because all regressors are dummies, the table only lists the cases for which the variables are equal to one. Descriptive statistics for all the variables and further information on their definition can be found in Dionne et al. (1996).

are dummies, Table 1 just lists the cases for which the variable is equal to one. Descriptive statistics for all the variables and further information on their definition can be found in the original paper. Notice that, although the value of N is not available, DT6 gives some information about the length of the contract.

3.2. Estimation and results

Because they do not account for the truncation induced by the exclusion of the observations with $y>11$, the models proposed in Section 2 are not appropriate to describe the data available for this study. However, the model defined by Eq. (2) can still be estimated if the sample used is restricted to the 677 observations for which $1\leq n\leq 11$. Indeed, for this subsample, y is necessarily smaller than 12 and therefore these observations are not affected by the truncation of the observations with $y>11$.

Besides avoiding the truncation, the use of this sub-sample is interesting because it contains only data on the more recent loans, thereby possibly reducing population drift issues. Indeed, this sample contains only clients whose contracts are so recent that their data would generally be ignored in the construction of a credit scoring model. To emphasize the ability of the model to use this information, this restricted sample is taken as the starting point of the analysis. Naturally, considering only the observations with $1\leq n\leq 11$ considerably reduces the sample size and consequently the results reported here should be interpreted merely as an illustration. However, for applications using data from reasonably large lending institutions, the reduction of the sample size resulting from considering only the more recent loans would not be a problem.

In order to proceed, it is necessary to specify how θ and α depend on x. For this particular exercise, because both α and θ are required to be positive, the following specification was initially adopted: $\theta=\exp(x'\beta)$ and $\alpha=\exp(x'\gamma)$.[8] However, estimation of the model defined by Eq. (2) using this unrestricted parametrization revealed that, for the sample considered, α seems to depend only on DESTIN. Therefore, α was parametrized as $\alpha=\exp(\gamma_0+\gamma_1\text{DESTIN})$. To check the validity of this restriction, the new model was tested against the more general specification, leading to a score test statistic of 13.666, to which corresponds a p-value of 0.6236.[9] Therefore, this test provides no evidence against the validity of the restricted model, whose results are presented in Table 2.

Given the small dataset used, it is not surprising to find that most parameters are estimated with poor precision and only a handful of them are significant at the usual 5% level. Moreover, the dataset contains very little information on the characteristics of the loans and on the financial status of the borrowers, which are likely to be the most important determinants of the default probability. However, it is clear that the variables DESTIN and RECSAL have a significant impact on θ, and therefore on the expected value of the missed payments. This result is not surprising because when either of these variables is equal to 1 it is easier for the bank to put pressure on the client to pay any amount that is due.

[8] More general specifications of θ and α could be used (e.g., Wooldridge, 1992). However, identification of the additional shape parameters would necessarily be very tenuous with such a small sample.

[9] Throughout, score tests are performed using estimates of the information matrix based on the observed Hessian, computed using analytical derivatives. All computations in this section were performed using TSP 5.0 (Hall and Cummins, 2005).

Table 2
Estimation results for the beta-binomial model.

	θ		α	
	Estimates	Std. errors	Estimates	Std. errors
Intercept	**−2.38902**	0.44830	**1.190910**	0.16143
DT6	**0.32232**	0.19347	–	–
AGE1	−0.06700	0.38884	–	–
AGE2	0.23748	0.19953	–	–
DESTIN	**−0.55835**	0.19926	**0.84984**	0.26234
ETU1	0.32796	0.61233	–	–
ETU2	0.48941	0.32390	–	–
ETU3	0.13500	0.32278	–	–
RECSAL	**−0.48391**	0.19117	–	–
M1	0.42638	0.23842	–	–
M2	0.40664	0.64312	–	–
M3	0.29326	0.34608	–	–
NM1	0.27855	0.23968	–	–
CENTRE	−0.28465	0.21158	–	–
RESID	−0.03906	0.19554	–	–
Z2	−0.18915	0.24062	–	–
Z3	−0.29189	0.24123	–	–
Z4	−0.28896	0.33197	–	–
Log-likelihood	−614.970			
Sample size	677			

Note: The table presents maximum likelihood estimates and standard errors for the parameters of model (2). Results are based on observations for contracts less than 1 year old. θ and α are specified as exponential functions of linear combinations of the regressors. The exclusion restrictions in α are supported by a score test of this specification vs. the unrestricted model with all regressors in α. Statistically significant estimates at the 5% level are typed in boldface.

Interestingly, the variable DESTIN also has a positive impact on α. Therefore, the fact that the credit is used to purchase a good with a collateral reduces the expected value of missed payments, but increases its variance. This may be a result of the variable nature of the collaterals in these loans. Thus, depending on the particular nature of the collateral, on which no information is available, the variable DESTIN may indicate very different situations, which is naturally reflected by the larger variance. In any case, the overall effect of the variable DESTIN on the probability of default is negative, as it can be seen in Fig. 1, which displays the estimated probability of default as a function of the age of the contract, for individuals with DESTIN = 0 and DESTIN = 1, when RECSAL = 1 and all the other covariates are equal to zero.

It is now interesting to try to answer some of the questions raised in Section 2. In particular, it is possible to test for a change of behavior after the client is considered a defaulter by testing the estimated model against the hurdle model with individual contributions to the likelihood function given by

$$L_h(\theta,\alpha) = L_1(\theta,\alpha)L_2(\theta,\alpha),$$

where $L_1(\theta,\alpha)$ and $L_2(\theta,\alpha)$ are defined by Eqs. (4) and (5). The statistic for the score test of Eq. (2) against this hurdle model has a value of 29.549, to which corresponds a p-value of 0.0775. Therefore, at the conventional 5% level, there is no evidence that the clients change their repayment behavior after being considered defaulters. Dionne et al. (1996) explain that the observations with $y>11$ were deleted "because strong legal actions are used for these cases". This may suggest that in this particular example, although a client is formally considered a defaulter for $y>3$, the bank is somewhat lenient with the clients with $3<y<11$, which may explain why no change in behavior is detected for $y>3$. Alternatively, the dataset used may just not be rich enough to identify this change in behavior.

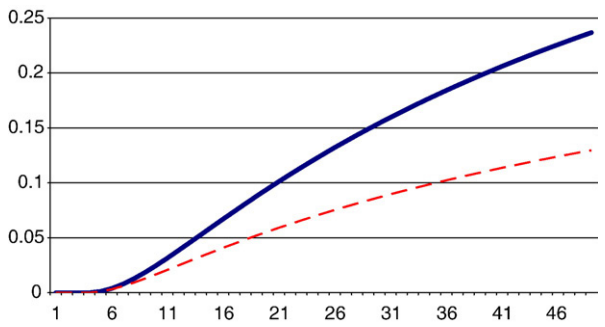

Fig. 1. Estimated probability of default as a function of the age of the contract, for DESTIN = 0 (solid line) and DESTIN = 1 (dashed line), for loans to be repaid in less than 48 months.

Table 3
True and predicted frequencies.

Count	0	1	2	3	>3
Data	0.744	0.126	0.044	0.025	0.061
Model	0.744	0.120	0.054	0.031	0.051

Note: The table presents observed and predicted relative frequencies for different counts of missed payments. Predicted frequencies denote probabilities computed with model (2), evaluated at the estimates presented in Table 2, and summed across all observations in the sample of contracts less than 1 year old.

Next, the possible presence of population drift is investigated using the following truncated model

$$P(Y=y|N,x,n,y<12)=\frac{P(Y=y|N,x,n)}{\sum_{j=0}^{11}P(Y=j|N,x,n)}, \tag{6}$$

where $P(Y=y|N,x,n)$ is defined by Eq. (2). Using the 2437 observations with $n>0$, this model was estimated and the corresponding results can be found in Table A1 in the Appendix A. This model was then tested against an alternative that allows the parameters to be different for observations in the sub-sample with $1\leq n\leq 11$. The resulting score test statistic has a value of 41.892, to which corresponds a p-value of 0.0029. This result indicates that there is strong evidence that the model estimated with older contracts cannot be used to explain the repayment behavior of the clients with $1\leq n\leq 11$. This result may be either the consequence of a population drift or an indication that the repayment behavior changes with the age of the contracts. Unfortunately, the data available for this study do not permit the full clarification of this question. Indeed, to clarify this point we would need to know the number of missed payments in two different moments in time, at least for some clients.

Although that route is not pursued here, for completeness, we note that the hurdle model defined by Eqs. (4) and (5) can also be modified to account for the right truncation of these data. Indeed, estimation of the first stage of the hurdle model accounting for the truncation of the data can be based on a likelihood with individual contributions of the form

$$L_{1t}(\theta,\alpha)=\left[\frac{P(Y=y|N,x,n)}{\left[\sum_{j=0}^{3}P(Y=j|N,x,n)+\sum_{j=4}^{11}P^{*}(Y=j|N,x,n)\right]^{I(n>11)}}\right]^{(1-d)}\left[1-\frac{\sum_{y=0}^{3}P(Y=y|N,x,n)}{\left[\sum_{j=0}^{3}P(Y=j|N,x,n)+\sum_{j=4}^{11}P^{*}(Y=j|N,x,n)\right]^{I(n>11)}}\right]^{d}, \tag{7}$$

where $P(Y=y|N,x,n)$ and $P^{*}(Y=y|N,x,n)$ denote, respectively, the probability of missed payments before and after the client is considered a defaulter. However, because of the truncation, this likelihood depends on the parameters of both stages of the model. Therefore, in presence of truncated data, the likelihood does not factor into two parametrically independent functions, as it is usual in hurdle models. This means that consistent estimation of the first stage parameters depends on the correct specification of the model for the second stage. In contradistinction, the form of Eq. (4) makes clear that, using only the subsample for which $n\leq 11$, it is possible to estimate the probability of default without specifying the second stage of the hurdle model.

Finally, it is interesting to see how the estimated model fits the data. To give an idea of the goodness-of-fit of the model, Table 3 gives the true and predicted frequencies of the number of missed payments. These results suggest that the model fits the data relatively well, being particularly good at predicting the high number of clients with zero missed payments.

A more formal assessment of the fit of the model can be obtained by performing generalized chi-square goodness-of-fit tests, as described by Cameron and Trivedi (1998).[10] These tests check the validity of moment conditions of the form

$$E[I(Y=y)-P(Y=y|N,x,n)]=0,$$

where $P(Y=y|N,x,n)$ is given by Eq. (2). If true, these moment conditions state that the probability of observing y that is specified by the model is equal to the actual probability of that count in the population. This test was performed for the set of moment conditions corresponding to $y=0, 1, 2, 3$, leading to a statistic of 5.477 and to a p-value of 0.24176.

Despite this evidence that the model fits the data well, Table 3 makes clear that the model somewhat under-predicts the probability of default. Given the purpose of the model, it is especially interesting to test if this underprediction is statistically significant. This can be done using again a generalized chi-square goodness-of-fit test. In this case, the relevant moment condition is

$$E[I(Y>l)-P(Y>l|N,x,n)]=0,$$

[10] For details on the implementation of this type of tests and a simulation study on their performance, see Cameron and Trivedi (1998, pp 155-157).

Table 4
Shumway's decile method of forecast accuracy.

Decile of estimated probability of default	1	2	3	4	5
Percentage of total defaulters	0.00	0.00	0.024	0.049	0.073
Decile of estimated probability of default	6	7	8	9	10
Percentage of total defaulters	0.122	0.098	0.222	0.195	0.222

Note: The table presents the percentage of actual defaulters in each decile of the estimated probability of default, for the sample of contracts less than 1 year old. Probabilities of default were computed with model (2), evaluated at the estimates presented in Table 2. E.g., the top two deciles of estimated default probabilities contain 41.7% (=22.2%+ 19.5%) of the total number of defaulters in the sample.

where, as before, $P(Y>l|N,x,n)$ is the probability of default given by Eq. (3). If true, this moment condition states that the probabihlity of default given by the specified model is equal to the actual probability of default in the population. Naturally, if the moment condition is not rejected, then there is some empirical evidence of the appropriateness of the model to predict the probability of default. The value of the corresponding test statistic is 3.837, to which corresponds a p-value of 0.0501. Therefore, continuing to use the standard 5% significance level, the hypothesis that the probability of default is correctly estimated by this model is not rejected.

In order to further explore the predictive accuracy of the model, we used Shumway's (2001) decile method to check the ability of the model to identify the individual defaulters in this dataset. To that end, all the observations were sorted into deciles according to the estimated probability of default. Then, the percentage of the total defaulters in each decile was computed. These results are displayed in Table 4, and show that the top three deciles of the estimated probability of default contain about 65% of all defaulters. Given the limitations of the dataset, and in particular the fact that very few regressors have a statistically significant impact on the probability of default, these results are encouraging.

As mentioned before, due to the shortcomings of the data used, the results in this section should be viewed merely as an illustration of the use of the proposed model. Moreover, the conclusions drawn from this exercise are somewhat fragile as they critically depend on the use of the 5% significance level. Nevertheless, this example shows that the proposed modelling strategy is potentially useful as it can be adapted to a variety of circumstances and permits the test of a number of interesting hypothesis. Of course, it would have been preferable to use richer data, but datasets on credit default are notoriously difficult to obtain.

4. Concluding remarks

This paper shows that by using appropriate count data models it is possible to estimate the conditional probability that a client of the lending institution will default on a loan, using only data from borrowers currently repaying their loans. The advantage of this approach is that it allows the use of data which is both up-to-date and readily available to the lending institution. Moreover, conditional on the characteristics of both the client and the loan, it is possible to see how the probability that a client will default varies with the time horizon considered.

The model used here is based on the beta-binomial distribution. Although this model is easy to estimate and to interpret, it is rarely used in a regression context. For the problem considered here, this model is particularly attractive because it can account for the specific characteristics of the data, and its estimation is not more demanding than that of the logit, which is regularly used in the construction of credit scoring models.

The application of the proposed methodology is illustrated using the data previously studied by Dionne et al. (1996). However, because these data are relatively poor, the results obtained in the empirical part of the paper should be viewed with great caution.

Acknowledgements

We are indebted to two anonymous referees, Montserrat Guillén and Tony Toivonen for many helpful comments on a previous version of this paper. We are especially grateful to Montserrat Guillén for kindly allowing us to use the data studied in Section 4. The usual disclaimer applies. The authors gratefully acknowledge the partial financial support from Fundação para a Ciência e Tecnologia (FEDER/POCI 2010).

Appendix A

Table A1
Results of the estimation of the truncated model.

	θ		a	
	Estimates	Std. errors	Estimates	Std. errors
Intercept	**−2.24063**	0.26956	**1.55797**	0.06539
DT6	**0.43692**	0.13513	–	–
AGE1	**0.59587**	0.26600	–	–
AGE2	**0.34185**	0.13546	–	–
DESTIN	**−0.80669**	0.13079	0.18391	0.10920

Table A1 (*continued*)

	θ		a	
	Estimates	Std. errors	Estimates	Std. errors
ETU1	0.68212	0.37221	–	–
ETU2	**0.40343**	0.18863	–	–
ETU3	0.07301	0.19373	–	–
RECSAL	**−0.89419**	0.12926	–	–
M1	**0.56340**	0.16569	–	–
M2	0.46216	0.45562	–	–
M3	**0.57737**	0.22556	–	–
NM1	**0.45347**	0.17056	–	–
CENTRE	−0.18800	0.15441	–	–
RESID	−0.16492	0.14055	–	–
Z2	−0.18145	0.16234	–	–
Z3	**−0.50307**	0.16675	–	–
Z4	**−0.47613**	0.22249	–	–
Log-likelihood	−2994.40			
Sample size	2437			

Note: The table presents maximum likelihood estimates and standard errors for the parameters of the truncated beta-binomial model defined in Eq. (6). Estimation was performed using all observations with $n>0$ and $y\leq 11$. Estimates that are significant at the 5% level are typed in boldface.

References

Altman, E.I., Avery, R.B., Eisenbeis, R.A., Sinkey, J.F., 1981. Application of Classification Techniques in Business, Banking and Finance. JAI Press, Greenwich (CT).
Bellotti, T., Crook, J., 2007. Credit scoring with macroeconomic variables using survival analysis. The Wharton Financial Institutions Center, Working-paper #07-15.
Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.
Carling, K., Jacobson, T., Roszbach, K., 2001. Dormancy risk and expected profits of consumer loans. Journal of Banking and Finance 25, 717–739.
Dionne, G., Artís, M., Guillén, M., 1996. Count data model for a credit scoring system. Journal of Empirical Finance 3, 303–325.
Duffie, D., Saita, L., Wang, K., 2007. Multi-period corporate failure prediction with stochastic covariates. Journal of Financial Economics 83, 635–665.
Hall, B.H., Cummins, C., 2005. TSP 5.0 User's Guide, TSP International, Palo-Alto (CA).
Hand, D.J., Henley, W.E., 1993. Can reject inference ever work? IMA Journal of Mathematics applied in Business and Industry 5, 45–55.
Hand, D.J., Henley, W.E., 1997. Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society, A 160, 523–541.
Hand, D.J., Jacka, S.D. (Eds.), 1998. Statistics in Finance. Arnold, London.
Heckman, J.J., Willis, R.J., 1977. A beta-logistic model for the analysis of sequential labor force participation by married women. Journal of Political Economy 85, 27–58.
Johansson, P., Palme, M., 1996. Do economics incentives affect work absence: empirical evidence using Swedish micro data. Journal of Public Economics 59, 195–218.
Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. 2nd ed. Continuous univariate distributions, vol. 2. John Wiley & Sons, Inc, New York.
Johnson, N.L., Kemp, A.W., Kotz, S., 2005. Univariate Discrete Distributions, 3rd ed. John Wiley & Sons, New York.
Kelly, M.G., Hand, D.J., Adams, N.M., 1999. The impact of changing populations on classifier performance. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 367–371.
Maddala, G.S., 1996. Applications of limited dependent variable models in finance. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 14. North-Holland, Amsterdam.
Moffatt, P.G., 2005. Hurdle models of loan default. Journal of the Operational Research Society 56, 1063–1071.
Mullahy, J., 1986. Specification and testing in some modified count data models. Journal of Econometrics 33, 341–365.
Pudney, S., 1989. Modelling individual choice. The Econometrics of Corners, Kinks and Holes. Blackwell, Oxford.
Roszbach, K., 2004. Bank lending policy, credit scoring and the survival of loans. Review of Economics and Statistics 86, 946–958.
Shumway, T., 2001. Forecasting bankruptcy more accurately: a simple hazard model. Journal of Business 74, 101–124.
Thomas, L.C., 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting 16, 149–172.
Thomas, L.C., Edelman, D.B., Crook, J.N., 2002. Credit Scoring and its Applications. SIAM, Philadelphia.
Wooldridge, J.M., 1992. Some alternatives to the Box-Cox regression model. International Economic Review 33, 935–955.
Wooldridge, J.M., 1999. Asymptotic properties of weighted M-estimators for variable probability samples. Econometrica 67, 1385–1406.