Estimation of Default Probabilities Using Incomplete Contracts Data*

J. M. C. Santos Silva

ISEG, Universidade Técnica de Lisboa

J. M. R. Murteira

Faculdade de Economia, Universidade de Coimbra

First draft: January 7, 2000

This version: June 1, 2000

Abstract

This paper develops a count data model for credit scoring which allows the estimation of default probabilities using incomplete contracts data. The model is based on the beta-binomial distribution, which is found to be particularly adequate to describe this sort of data. A well known data set on personal loans granted by a Spanish bank is used to illustrate the application of the proposed model.

JEL classification code: C21, C51, G21.

Key Words: Beta-binomial distribution; Credit scoring; Hurdle models.

*We are very grateful to Montserrat Guillén for kindly allowing us to use the data studied in this paper, and for helpful discussions. Address for correspondence: João Santos Silva, ISEG, R. do Quelhas 6, 1200 Lisboa, Portugal. Fax: 351 21 392 27 81. E-mail: jmcss@iseg.utl.pt.

1. INTRODUCTION

Models for credit scoring are widely used in practice, and raise a number of interesting and challenging academic questions. Therefore, it is not surprising to find that they have been the subject of a considerable literature (see, for example, Altman *et al.*, 1981, Maddala, 1996, Hand and Henley, 1997, and the references therein).

The typical situation dealt with in the study of credit scoring models is the case in which data on previous clients of a lending institution are used to define a set of rules that permits the classification of prospective clients as credit worthy or not. The situation considered here is slightly different in that attention is focused on the case in which the credit scoring model is estimated using data on the current clients of the lending institution. In this case there is an obvious observability problem, because the contracts have not been completed.

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, Jacobson and Roszbach, 1998).

The remainder of the paper is organised as follows. In the next section, the problem is presented and an appropriate model is proposed. Section 3 describes the data set used in the study and presents the results obtained. Finally, section 4 concludes the paper.

2. THE PROBLEM AND A MODEL

Consider a lending institution, hereafter referred to as the bank, that wants to use the information available on the characteristics and repayment behaviour of its present clients to construct a credit scoring model to evaluate the probability of a prospective client to become a defaulter.

A first problem that this setup raises is that the bank only has information on clients to whom it has decided to grant a loan. Therefore, there is a potential problem of sample selection. 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 to 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 imposing very strong assumptions (see Hand and Henley, 1997). However, if all the information used to decide about the credit application is available for the construction of the credit scoring model, the sample available to the researcher is exogenously stratified, and standard inference methods can be used (see Pudney, 1989 and Wooldridge, 1999). This more favourable situation is the one considered here.

The second issue that has to be addressed is that the repayment behaviour of a client may change after he is classified as defaulter. In fact, after a client is classified as a defaulter, the bank may put pressure on the client to repay his debt, for instance by threatening to take legal action against him, and that may alter the client's behaviour. In these circumstances a hurdle model will be appropriate. The probability that a client becomes a defaulter can be obtained from the specification of the first stage of the model, that is, the stage that describes the behaviour of the client before being classified as a defaulter. In their pioneering work, Dionne, Artís and Guillén (1996) also used a hurdle model to describe this kind of data. However, the hurdle model considered here differs from the one described by Dionne, Artís and Guillén (1996) in several aspects.

The particular nature of the data being considered imposes a number of restrictions on the type of models that may be adequate. In particular, a model for this kind of data has to account for the fact that once a loan is granted to a client, it is generally repaid in a number of regular instalments. Let N denote the total number of payments implied by the contract. At a given point in time there is an upper bound on the

number of payments the client may have missed. That bound is just the age of the contract measured as the number of payments that should have been made since the contract begun. This upper bound will be denoted by $n \leq N$. Moreover, because the contracts are not completed, the clients that are currently classified as non-defaulters may become defaulters in the future. Therefore, all the analysis must be conditional on n. All these issues will be addressed in the model proposed bellow.

Let Y be the number of payments missed by a bank client, and suppose that, besides N, the bank knows a set x of characteristics of the contract and of its clients. 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, x and n. Notice that in this sort of model the probability of becoming a defaulter will depend on the time horizon considered. This is important since, from the point of view of the bank, it is not indifferent when the client becomes a defaulter (see Roszbach, 1998).

Because n can be very small, models that assume an infinite upper bound for the variate of interest are not appropriate in this situation. Therefore, the count data models more often used in applied work, e.g. Poisson and negative binomial, are not appropriate in this context. 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 model defined by

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

where p is the probability that the individual will miss any of the N payments. Although this model is well known, it is rarely used in applied econometrics (see, however, Johansson and Palme, 1996).

In order to account for individual heterogeneity, it is supposed that p depends on a set of observed and unobserved individual characteristics. In particular, it is assumed 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

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

where $f\left(p|N,x\right)$ 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

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

with

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

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

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, Kotz and Kemp, 1992). Therefore, this model can accommodate a situation in which the probability that an individual will miss a certain payment depends on his previous repayment behaviour.

To complete the model specification it is necessary to define how α and θ depend on N and x. A convenient parametrization is given by $\theta=\exp\left(x'\beta+g\left(N\right)\right)$ and α, for the moment, is assumed to be constant. It is not difficult to verify that when $g\left(N\right)=-\ln\left(N\right)$, the limiting distribution of the total number of non-payments when N passes to ∞ is negative binomial with mean $\lambda=\exp\left(x'\beta\right)$ and variance $\lambda+\alpha\lambda^{2}$.

As noted before, after the client is considered defaulter by the bank his repayment behaviour may change. Therefore, in order to model the total number of missed payments, a hurdle model (Mullahy, 1986) that accounts for the different nature of the two processes should be used. The model described above can be adequate to

describe the number of payments missed by a client that is not a defaulter, and to estimate the probability of default, which is given by

$$P(D|N,x,n) = 1 - \sum_{y=0}^{l} \frac{n!}{y!\,(n-y)!} \frac{\Gamma\left(\frac{\theta+1}{\alpha\theta}\right)\Gamma\left(\frac{1+n\alpha\theta-y\alpha\theta}{\alpha\theta}\right)\Gamma\left(\frac{1+y\alpha}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{1}{\alpha\theta}\right)\Gamma\left(\frac{\theta+1+n\alpha\theta}{\alpha\theta}\right)},$$

where l is the maximum number of repayments that a client may miss without being considered a defaulter by the bank.

In order to estimate the parameters of interest, the following likelihood function is maximized

$$L(\theta,\alpha) = \left[P\left(Y=y|N,x,n\right)\right]^{(1-d)} \left[1 - \sum_{y=0}^{l} P\left(Y=y|N,x,n\right)\right]^{d}, \tag{3}$$

where $P(Y=y|N,x,n)$ is defined by (2), and $d = I(y>l)$. In case the researcher is also interested in the probability of non-payments after the client is considered a defaulter, say $P^*\left(Y=y|N,x,n\right)$, a second model has to be estimated using only the observations with $d=1$.

The model defined by (3) deserves some comments. The simple binomial model defined by (1) is unlikely to be adequate to describe the data, due to the presence of neglected individual heterogeneity. There are two different ways to account for extra-binomial variation. One possibility is to model the distribution of the unobservables semiparametrically, as it is done by Johansson and Palme (1996). Alternatively, a fully parametric approach based on the specification of the distribution of p can be adopted. Here, this latter solution is adopted. There are several reasons for this choice. To start with, 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 using the semiparametric approach with a hurdle model would greatly increase the computational costs. The simple estimation and interpretation of the fully parametric specification are certainly an important advantage for the practitioner in charge of the practical application of the model. Moreover,

the model chosen here 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, as in this case it is generally assumed that the unobservables are statistically independent of the covariates. Therefore, although the parametric approach is restrictive (since it requires the specification of the distribution of the unobservables) it has the advantage of allowing the distribution of the unobservables to depend on the conditioning variables. Furthermore, in contradistinction to what happens with the beta-binomial model, the finite mixture model used by Johansson and Palme (1996) is not suited to situations in which there is true occurrence dependence.

3. DATA AND ESTIMATION RESULTS

The data set used here is the one studied by Dionne, Artís and Guillén (1996). It consists of a sample of clients of a Spanish bank that were repaying loans in May 1989. After excluding observations with incomplete records and the elimination of individuals with outlying values of the explanatory or of the dependent variable, the authors were left with a sample containing 2446 observations. According to the bank criteria, a client is considered to be a defaulter if he misses more than 3 payments. Therefore, in this case, $l = 3$. Besides containing information on y, the number of non-payments, and on n, the number of months from the beginning of the contract at the sampling date, this data set also contains information on some characteristics of the loan and of the client. These variables are described in table 1. Notice that, although the value of N is not available, DT6 gives some information about the length of the contract. Descriptive statistics for all the variables and further information on their definition can be found in the original paper.

Table 1: Description of the regressors used in the study

DT6	1 if total contract duration of return period is more than four years
AGE1	1 if age group is 18-24 years
AGE2	1 if age group is 25-39 years
AGE3	1 if age group is 40 years or more
DESTIN	1 if credit is used to purchase a good with a collateral
ETU1	1 if the client has not completed primary education
ETU2	1 if the client has completed primary education
ETU3	1 if the client has completed higher education
ETU4	1 if the client has a university degree
RECSAL	1 if the client receives the salary through the bank
M1	1 if married, non-owner, salary under $3000
M2	1 if married, non-owner, salary higher or equal to $3000
M3	1 if married, owner, salary under $3000
M4	1 if married, owner, salary higher or equal to $3000
NM1	1 if not married, non-owner
NM2	1 if not married, owner
CENTRE	1 if credit is granted by a store
RESID	1 if client is resident in the city for at least four years
Z1	1 if client is resident in the south
Z2	1 if client is resident in the north
Z3	1 if client is resident in the east
Z4	1 if client is resident in the centre

Because observations with large values of y ($y > 11$) were excluded from the sample by Dionne, Artís and Guillén (1996), the model proposed in section 2 has to be modified to account for the right truncation of the data. This point was neglected by Dionne, Artís and Guillén (1996), but it is likely to be relevant, even if the interest is restricted to the first stage of the hurdle model. For this reason, the model defined by (3) is not applicable in the context of this data set. Therefore, estimation of the first stage of the hurdle model accounting for the truncation of the data, would have to be based on a likelihood of the form

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

where $P^*(Y = y|N, x, n)$ denotes the probability of non-payments after the client is considered a defaulter. Needless to say, this modification of the likelihood function greatly increases the programming and computational cost of estimating the model. Furthermore, this likelihood depends on the parameters of the second stage and, therefore, 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 now depends on the correct specification of the model for the second stage. Despite its lack of robustness, this model is interesting because it can be used to construct a simple specification test. In fact, since the first and second stages depend on all the parameters, the correct specification of the model can be assessed by testing if the parameters separately estimated in each of the stages are equal.

However, the standard hurdle model defined by (3) can still be estimated if the sample used is restricted to the 677 observations for which $1 \leq n \leq 11$, as these

observations are not censored. Of course, this is an enormous waste of information, but by doing this it is possible to estimate the probability of default without specifying the second stage of the hurdle model. Furthermore, the computational burden of estimating (4) is avoided. For the moment, this is the approach adopted here, although the results of the estimation of (4), and of the associated specification test, will be included in a forthcoming version of this work. Because of the limited sample used, the results reported here should be interpreted only as an illustrative example.

For this particular exercise, the following specification was adopted: $\theta = \exp(x'\beta)$ and $\alpha = \exp(x'\gamma)$. However, if no restrictions are imposed on this general specification, and given the limitations of the available data set, the model is likely to be badly over-parametrized. Therefore, as a starting point, it was assumed that $\alpha = \exp(\gamma_0)$. The results obtained with this model correspond to those of Model 1 in table 2. To check if the restriction imposed on γ is acceptable, Model 1 was tested against the unrestricted model. The value of the score test statistic against this alternative is 30.832, to which corresponds a p-value of 0.0209. Therefore, there is some evidence that this restriction is invalid. To overcome this problem, the following parametrization was adopted $\alpha = \exp(\gamma_0 + \gamma_1 \text{DESTIN})$. The results obtained with this more general model are given in table 2 under the label: Model 2. This model was also tested against the unrestricted model and the score test statistic obtained was 20.161, to which corresponds a p-value of 0.2131. Therefore, this test provides no evidence against the validity of Model 2.

Given the small data set used, it is not surprising to find that most parameters are estimated with poor precision. Moreover, the data set 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 non-payments.

Table 2: Estimation results

	Model 1		Model 2	
Parameters in θ	Estimates	Std. Errors	Estimates	Std. Errors
Intercept	−2.17737	0.469965	−2.29277	0.469096
DT6	0.26419	0.211036	0.29889	0.211284
AGE1	−0.04930	0.396941	−0.03910	0.409117
AGE2	0.04635	0.216277	0.06646	0.214923
DESTIN	−0.57567	0.205637	−0.55766	0.213923
ETU1	0.31523	0.663036	0.20977	0.655402
ETU2	0.48803	0.337143	0.48444	0.337265
ETU3	0.07781	0.338691	0.11374	0.337323
RECSAL	−0.65268	0.209392	−0.64052	0.206890
M1	0.33560	0.257524	0.41168	0.255294
M2	0.67648	0.698798	0.62648	0.718005
M3	0.00434	0.394545	0.05425	0.389593
NM1	0.34459	0.256366	0.38753	0.254278
CENTRE	−0.13842	0.228028	−0.14482	0.228553
RESID	−0.08937	0.211300	−0.07577	0.209340
Z2	−0.23332	0.260916	−0.16359	0.260367
Z3	−0.22376	0.256523	−0.14214	0.259326
Z4	−0.13935	0.355619	−0.11413	0.351564
Parameters in α	Estimates	Std. Errors	Estimates	Std. Errors
Intercept	1.53276	0.140841	1.19232	0.181595
DESTIN	—	—	0.91806	0.292061
Log-likelihood	−562.711		−557.811	
Sample size	677		677	

Table 3. True and predicted frequencies

	0	1	2	3	>3
Data	0.744	0.126	0.044	0.025	0.061
Model 2	0.744	0.118	0.054	0.031	0.053

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. It is interesting to notice that 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 non-payments, but increases its variance.

As mentioned above, these results should be viewed only as illustrative. However, it is interesting to see how the model fits the data. To give an idea of the goodness of fit of this model, table 3 gives the true and predicted frequencies of the number of non-payments. It is clear that the model somewhat under-predicts the probability of default, but that it fits the data relatively well. In particular it does very well at predicting the high number of clients with zero non-payments.

4. CONCLUDING REMARKS

This paper shows that using appropriate count data models it is possible to estimate the conditional probability that a client will default on a loan, using only data from present clients of the lending institution. The advantage of this approach is that it allows the use of data that is both up-to-date and readily available to the lending institution. Moreover, it is possible to simulate the probability that a client will default after a given time, conditional on the characteristics both of the client and the loan.

The model used here is based on the beta-binomial distribution. Although this model is easy to estimate and to interpret, it appears that it has not been used in a regression context before. For the problem considered here, this model is particularly attractive because it can account for the specific characteristics of the data.

The data set used to illustrate the application of the proposed methodology is relatively poor, and therefore the results obtained should be viewed with great caution. The true usefulness of the model proposed here can only be judged when a more rich data set is available.

REFERENCES

Altman, E.I.; Avery, R.B.; Eisenbeis, R.A. and Sinkey, J.F. (1981). *Application of Classification Techniques in Business, Banking and Finance*, JAI Press, Greenwich (CT).

Carling, K.; Jacobson, T. and Roszbach, K. (1998). *Duration of Consumer Loans and Bank Lending Policy: Dormancy Versus Default Risk*. Working Paper No 280, Working Paper Series in Economics and Finance, Stockholm School of Economics.

Dionne, G.; Artís, M. and Guillén, M. (1996). "Count Data Model for a Credit Scoring System". *Journal of Empirical Finance*, 3, 303-325.

Hand, D.J. and Henley, W.E. (1997). "Statistical Classification Methods in Consumer Credit Scoring: A Review". *Journal of the Royal Statistical Society*, **A**, 160, 523-541.

Johansson, P. and 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. and Kemp, A.W. (1992) *Univariate Discrete Distributions*, 2nd Ed., John Wiley & Sons, Inc., New York.

Maddala, G.S. (1996). "Applications of Limited Dependent Variable Models in Finance", *in* Maddala, G.S. and Rao, C.R. (eds.) *Handbook of Statistics*, vol. 14, North-Holland, Amsterdam.

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. (1998). *Bank Lending Policy, Credit Scoring and the Survival of Loans*. Working Paper No 261, Working Paper Series in Economics and Finance, Stockholm School of Economics.

Wooldridge, J.M. (1999). "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples". *Econometrica*, 67, 1385-1406.