Estimation of Default Probabilities Using Incomplete Contracts Data*

J. M. C. Santos Silva
Department of Economics, University of Essex and CEMAPRE

J. M. R. Murteira
Faculdade de Economia, Universidade de Coimbra and CEMAPRE

This version: March 16, 2007

Abstract

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. However, models constructed using this type of data may suffer from *population drift* problems caused by the continuous changes in the distribution of the characteristics of the clients. That is, the sample used to estimate these models may not be representative of the population of current bank clients and credit applicants. To mitigate this problem, current clients are sometimes included in the sample and are classified according to their present status. However, this procedure will inevitably induce some degree of missclassification because some clients currently classified as non-defaulters may actually default before the end of the contract.

*Address for correspondence: João Santos Silva, Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. E-mail: jmcss@essex.ac.uk.

We address this issue by considering a count data model for credit scoring, which allows the estimation of default probabilities using data on incomplete contracts, and does not require the classification of the clients as defaulters or not. 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 both of 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 is based on the beta-binomial distribution. Although this model is rarely used in a regression context, it is particularly attractive for the problem considered here because it can account for the specific characteristics of the data and its estimation is as easy as that of the logit, which is regularly used in the construction of credit scoring models.

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; Population drift.

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, Avery, Eisenbeis and Sinkey, 1981, Maddala, 1996, Hand and Henley, 1997, Hand and Jacka, 1998, Thomas, 2000, Thomas, Edelman and Crook, 2002, 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. However, models constructed using this type of data may suffer from *population drift* problems (Kelly, Hand and Adams, 1999) caused by the continuous changes in the distribution of the characteristics of the clients. That is, the sample used to estimate these models may not be representative of the population of current bank clients and credit applicants. We address this issue considering a credit scoring model that is estimated using data on the current clients of the lending institution, thereby mitigating the potential population drift problems. Of course, using data on current clients leads to an obvious observability problem because the contracts have not been completed. However, as it will be shown, using appropriate statistical methods it is possible to deal with this situation.

The remainder of the paper is organized as follows. In the next section we highlight some of the most important features of the problem in hand. Section 3 introduces a count data model that allows the estimation of default probabilities using incomplete contracts data and section 4 illustrates the use of the model. Finally, section 5 concludes the paper.

2. THE PROBLEM

Consider a lending institution, hereafter referred to as the bank, that wants to use the information available on the characteristics and repayment behavior of its present clients to construct a credit scoring model to evaluate the probability of a prospective client to become a defaulter.[1]

A first issue that this setup raises is that there is a potential sample selection problem because 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 imposing very strong assumptions (see Hand and Henley, 1993). On the other hand, if all the information used to decide about the credit application is available for the construction of the credit scoring model, standard inference methods can be used as in this case the sample is exogenously stratified (see Pudney, 1989 and Wooldridge, 1999). This more favorable situation is the one considered here.

Another problem that has to be addressed is that the repayment behavior of a client may change after he is classified as a defaulter because the bank may put pressure on him to repay his debt, for instance by threatening to take legal action. In these circumstances, a hurdle model (Mullahy, 1986) will be appropriate and 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 behavior of the client before being classified

[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, Jacobson and Roszbach, 2001).

as a defaulter.[2] Whether this change of behavior is relevant or not is an issue that can be tested in empirical applications.

Credit scoring models are usually estimated using a sample of clients who the bank has classified either as defaulters or non-defaulters and the results obtained are used to predict the future behavior of the credit applicants. This raises the population drift issue since the sample used to estimate the model may not be representative of the population for which inference is going to be made (see Kelly, Hand and Adams, 1999). A first point that has to be noted is that the severity of the population drift problem depends on the particular statistical method used to select the applicants which will be refused credit. If the applicants are scored using an estimate of the conditional probability of default, the predictions will be robust to changes in the marginal distribution of the covariates. Otherwise, the method may be sensitive even to these changes. Therefore, there is a case for trying to correctly specify the conditional probability of becoming a defaulter.

Of course, if the distribution of unobserved characteristics of the clients that are relevant for their repayment behavior changes, not even the conditional distribution of the missed payments will be stable. Therefore, to minimize the potential population drift problem, current clients are sometimes including in the sample and are classified according to their present status. However, this procedure will inevitably induce some degree of missclassification because some clients currently classified as non-defaulters may actually default before the end of the contract. If the conditional distribution of the missed payments is stable, ignoring data on the current clients is of little consequence. However, in the presence of a population drift problem, these observations are the most informative ones and discarding or missclassifying them can have significant costs. In the next section we present a flexible count data model for the total number of payments missed by the

[2] In their pioneering work, Dionne, Artís and Guillén (1996) used a hurdle model to describe this kind of data. However, the hurdle model described in the next section differs from the one considered by those authors in several aspects.

bank's clients which permits the estimation of the conditional probability of becoming a defaulter using data on all contracts, including the more recent ones.

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 installments. 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 by the number of payments that should have been made since the contract began. 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. 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. Bellow, all these issues are addressed by using a beta-binomial model, first used in a regression context by Heckman and Willis (1977).

3. AN APPROPRIATE COUNT DATA MODEL

Let Y be the number of payments missed by a bank client, and suppose that, besides N and n, the bank observes 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, n and x. 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).

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

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

where p is the probability that the individual will miss any of the N payments. Even if p is parameterized as a function of x, the simple binomial model defined by (1) is unlikely to be adequate to describe the credit default data due to the presence of neglected 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. Given that α and θ are positive parameters, it is convenient to specify them as exponential functions of the covariates and N. Naturally, the particular form of these functions will depend on the application considered. However, it is interesting to note that when θ is specified as $\theta=\exp(x'\beta-\ln(N))$, the limiting distribution of

the total number of non-payments when N passes to ∞ is negative binomial with mean $\lambda = \exp(x'\beta)$ and variance $\lambda + \alpha\lambda^2$.

An alternative way to account for extra-binomial variation in (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 preferred to use a fully parametric specification based on the beta-binomial distribution.

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 this simplicity is certainly of great value for the practitioner in charge of the practical application of the model.

Additionally, 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. Indeed, 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.

Finally, 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 behavior.

As noted before, after the client is considered defaulter by the bank his repayment behavior may change. In any case, the model defined by (2) is appropriate 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.

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

$$L_1(\theta,\alpha) = \left[P\left(Y = y|N,x,n\right)\right]^{(1-d)} \left[1 - \sum_{j=0}^{l} P\left(Y = j|N,x,n\right)\right]^{d}, \tag{3}$$

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 $P^*\left(Y = y|N,x,n\right) = P\left(Y = y|N,x,n,y > l\right)$, the second stage of the hurdle model has to be estimated maximizing a likelihood function with individual contributions of the form

$$L_2(\theta,\alpha) = \left[\frac{P^*\left(Y = y|N,x,n\right)}{1 - \sum_{j=0}^{l} P^*\left(Y = j|N,x,n\right)}\right]^{d}. \tag{4}$$

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 $y > l$.

4. AN EMPIRICAL ILLUSTRATION

In this section we use the data studied by Dionne, Artís and Guillén (1996) to illustrate the estimation of default probabilities using incomplete contracts data. The data consist 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.

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 is not applicable if the entire data set is considered.[3] However, the standard model defined by (2) can still be estimated if the sample used is restricted to the 677 observations for which $1 \leq n \leq 11$, since these observations are not truncated. Besides avoiding the truncation, the use of this sub-sample is interesting because it contains only data on the more recent loans, thereby reducing the population drift problems. 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 and the possible existence of a population drift is later tested by comparing these estimates of the parameters with those obtained with the remaining observations. Naturally, considering only the observations with $1 \leq n \leq 11$ considerably reduces sample size in this particular example 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.

[3]This truncation 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 an hurdle model.

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

In order to proceed, it is necessary to specify how θ and α depend on the covariates. For this particular exercise, the following specification was adopted: $\theta = \exp(x'\beta)$ and $\alpha = \exp(x'\gamma)$. However, estimation of the model using this general 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 adequacy of this specification, the model was tested against the unrestricted model and the score test statistic obtained has a value of 13.666, to which corresponds a p-value of 0.6236. Therefore, this test provides no evidence against the validity of the restricted model, whose results are presented in table 2.[4]

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. 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 also 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. This effect on the variance is not unexpected because in this kind of loans guarantees vary widely, often being just a formality for credit granting.

It is now interesting to try to answer some of the questions raised in section 2. In particular, by testing this model against the hurdle model defined by (3) and (4) it is possible to test for a change of behavior after the client is considered a defaulter. The score test statistic for the test of the estimated model against this hurdle has a value of 29.549, to which corresponds a p-value of 0.0775. Therefore, at the conventional 5 percent level, there is

[4] All computations in this section were performed using TSP 4.5 (Hall and Cummins, 1999).

Table 2: Estimation results

	θ		α	
	Estimates	Std. Errors	Estimates	Std. Errors
Intercept	−2.38902	0.44830	1.19091	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			

no evidence that the clients change their repayment behavior after being considered defaulters.[5]

To test for the so-called population drift, the following truncated model was used

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

where $P\left(Y=y|N,x,n\right)$ is defined by (2). Using the 2437 observations with $n>0$, this model was estimated and tested against an alternative that allows the parameters do 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 $1\leq n\leq 11$.[6]

[5]Notice that the hurdle model defined by (3) and (4) can be modified to account for the right truncation of the 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

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

where $P^{*}\left(Y=y|N,x,n\right)$ denotes the probability of non-payments 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, using only the subsample for which $n\leq 11$ it possible to estimate the probability of default without specifying the second stage of the hurdle model.

[6]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 does 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.

Table 3. True and predicted frequencies

	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

Finally, it is interesting to see how the estimated models fit 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 non-payments. The model fits the data relatively well, being particularly good at predicting the high number of clients with zero non-payments. However, it is 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 evaluated using a test for the moment condition

$$E\left[d - 1 + \sum_{j=0}^{l} P\left(Y = j|N, x, n\right)\right] = 0,$$

where, as before, d is a dummy variable that equals 1 if the client is a defaulter and $P\left(Y = j|N, x, n\right)$ is given by (2).[7] The value of this test statistic is 3.837, to which corresponds a p-value of 0.0501. Therefore, continuing to use the standard 5 percent significance level, the hypothesis that the probability of default is correctly estimated by this model is not rejected.

As mentioned before, due to the limitations 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 percent significance level. Nevertheless, this example shows that the proposed modelling strategy is potentially useful as it can be adapted to a variety of

[7]For details on the implementation of this type of tests and a simulation on their performance see Cameron and Trivedi (1998).

circumstances and permits the test of a number of interesting hypothesis. Of course, it would have been preferable to use richer data, but data sets on credit default are notoriously difficult to obtain.

5. CONCLUDING REMARKS

This paper shows that by 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 which is both up-to-date and readily available to the lending institution. Moreover, conditional on the characteristics both of 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 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.

ACKNOWLEDGEMENTS

We are indebted to 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).

REFERENCES

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

Cameron, A.C. and P.K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Carling, K., T. Jacobson and K. Roszbach (2001). "Dormancy Risk and Expected Profits of Consumer Loans". *Journal of Banking and Finance*, 25, 717-739.

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

Hall, B. H. and C. Cummins (1999). *Time Series Processor Version 4.5 User's Guide*. Palo-Alto (CA): TSP International.

Hand, D.J. and W.E. Henley (1993). "Can Reject Inference Ever Work?". *IMA Journal of Mathematics applied in Business and Industry*, 5, 45-55.

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

Hand, D.J. and S.D. Jacka (eds.) (1998). *Statistics in Finance*. London: Arnold.

Heckman, J.J and 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. and M. Palme (1996). "Do Economics Incentives Affect Work Absence: Empirical Evidence Using Swedish Micro Data". *Journal of Public Economics*, 59, 195-218.

Johnson, N.L., S. Kotz and A.W. Kemp (1992). *Univariate Discrete Distributions* (2nd Ed.).New York: John Wiley & Sons, Inc.

Kelly, M.G., D.J. Hand and N.M. Adams (1999). "The Impact of Changing Populations on Classifier Performance", *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 367-371.

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

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*. Oxford: Blackwell.

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.

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., D.B. Edelman and J.N. Crook (2002). *Credit Scoring and its Applications*. Philadelphia: SIAM.

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