Table III

model	1	2	3	4
Discriminant analysis	65.4	62.2	78.0	8.1
Linear regression model	55.1	47.0	87.5	6.2
Probit model	71.9	76.4	54.1	13.1
Poisson Model	62.4	57.7	81.8	7.3
Negative binomial II model	63.3	58.9	80.6	7.6
two step procedure	64.9	61.1	79.8	7.6

Table II

variables	Poisson	Bin.Neg.II
t.i.	0.398	0.398
	(0.358)	(0.474)
P1	0.086	0.086
	(0.027)	(0.069)
P2	-0.005	-0.005
	(0.034)	(0.088)
P3	-0.640	-0.640
	(0.344)	(1.236)
P4	-0.251	-0.251
	(0.028)	(0.072)
P5	-0.248	-0.248
	(0.026)	(0.065)
P6	0.208	0.207
	(0.042)	(0.105)
P7	0.073	0.073
	(0.047)	(0.116)
P8	0.067	0.067
	(0.057)	(0.146)
P9	0.143	0.143
	(0.066)	(0.165)
E1	-0.074	-0.074
	(0.054)	(0.145)
E2	-0.029	-0.029
	(0.052)	(0.140)
F1	-0.411	-0.411
	(0.032)	(0.080)
F2	1.420	1.420
	(0.031)	(0.092)
F3	-0.138	-0.138
	(0.035)	(0.090)
F4	0.003	0.004
	(0.001)	(0.018)
C2	-0.137	-0.137
	(0.072)	(0.178)
C3	0.531	0.531
	(0.163)	(1.257)
alfa		1.340

Table I

Number of defaults	Absolute frequency	Number of defaults	Absolute frequency
0	3002	14	13
1	502	15	11
2	187	16	4
3	138	17	5
4	233	18	8
5	160	19	6
6	107	20	3
7	80	22	1
8	59	24	1
9	53	28	1
10	41	29	1
11	28	30	1
12	34	31	1
13	10	34	1

Adequate use of count data models without mean-variance restriction is useful to find which are the most influencial variables in the studied process. It has to be noted that estimation required assymptotic approximations for standard errors.

Another important issue is the possibility of estimating truncated negative binomial models in order to study truncated samples which have great interest in this context. Estimation results are not shown in the text to keep it brief, but they show a change in the influence of many variables. Income, for exemple, which was not significant becomes significant in the truncated model, that is to say that once a payment has been missed, higher incomes mean significant smaller expected number of defaults.

The possibility of establishing a two step procedure for prediction has been stressed. In fact, this combines the use of discriminant analysis and modelization and it leads to a good global classificatin rate keeping the number of *bad* clients accepted as *good* well bellow 10%.

Further research is needed to desetangle some obscure points such as model selection or missspecification in truncated models. In this situation although prepayment has not been considered, one should see the way to include duration of repayment at sample collection and its influence in final estimation and classification results.

6 References

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

Boyes, W.J., Hoffman, D.L. y S.A. Low (1989) 'An Econometric Analysis of the Bank Credit Scoring Problem', *Journal of Econometrics,* 40, 3-14.

Cameron, A. C. y P.K. Trivedi (1986) 'Econometric Models Based on Count Data: Comparison and Application of Some Estimators and Tests' *Journal of Applied Econometrics,* 1, 29-53.

Frome, E.L., Kutner, M.H. y J.J. Beauchamp (1973) 'Regression Analysis of Poisson-Distributed Data' *Journal of the American Statistical Association,* 68, 344 935-940.

Gourieroux, C., Monfort, A, y A. Trognon (1984a) 'Pseudo-Maximum Likelihood Methods: Theory' *Econometrica,* 52, 681-700.

Gourieroux, C., Monfort, A, y A. Trognon (1984b) 'Pseudo-Maximum Likelihodd Methods: Applications to Poisson Models' *Econometrica,* 52, 701-720.

Grogger, J.T. y R.T. Carson (1991) 'Models for Truncated Counts' *Journal of Applied Econometrics,* 6, 225-238.

Guillén, M. (1992) *Análisis econométrico del credit scoring. Modelos Count Data.* Ph. D. Dissertation. University of Barcelona, Spain.

Hausman, J., Hall, B.H. y Z. Grilliches (1984) 'Econometric Models for Count Data with an Application to the Patents-R&D Relationship' *Econometrica,* 52, 909-938.

Lee, L-F. (1986) 'Specification Test for the Poisson Regression Models ' *International Economic Review,* 27, 689-706.

McCullagh, P. y J.A. Nelder F.R.S. (1983) *Generalized Linear Models.* Monographs on Statistics and Applied Probability. Chapman and Hall. London.

Mullahy, J. (1986) 'Specification and Testing of Some Modified Count Data Models' *Journal of Econometrics,* 33, 3, 341-366.

Myers, J.H. y W. Forgy (1963) 'The Development of Numerical Credit Evaluation Systems' *Journal of the American Statistical Association,* 58, 303, 779-806.

Steenackers, A. y M.J. Goovaerts (1989) 'A Credit Scoring Model for Personal Loans' *Insurance: Mathematics and Economics,* 8, 31-34.

4 Results

In 1990, a Spanish financial institution provided a data set containing almost 5000 clients that had been granted credit in the previous four years. The type of credit of interest is known as personal loan. Personal loans are characterized by the fact that the amount of money granted is moderate. Usually, the loan is returned in a short period of time and it is often repaid monthly with instalments that are constant along the repayment period and not very large compared to individual income. Normally, this kind of loans are related respond, most commonly automobiles.

The variables included in the models are divided in different groups according to the source of the information provided, whether they are items responded by the individual applying for credit, or if those items are information that, although it may be provided by the applicant, the financial institution is able to check in its own files.

Variables in the models may be found in one of the following three groups:

Personal variables (date of birth, marital status, number of children,...)

Socio-economic variables (net monthly income, housing ownership,...)

Financial variables (monthly mortgage, availability of credit card, amount requested...)

The main innovation about the variables that are used in the model for credit scoring is that the above variables provide the information that is needed to create new variables finnally used in the model. Modifications are made in two different senses and have two well established objectives:

a) On the one hand, new combinations may preserve confidentiality of the discriminant functions or model that are being implemented.

b) On the other hand, by using some new variables, the model can cope with interactions.

Table I shows the absolute frequencies for the variable of interest, that is the number of unpaid instalments. Note that 64 % of clients have no defaults.

Table II shows estimates and standard errors for three models. For estimation purposes, some individuals were eliminated from the original sample. Individuals with repayment lasting less than six months at sample collection were excluded from the estimation process on the grounds that there was not enough information about their repayment behaviour and that posterior classification could be misleading.

It is interesting to see that parameter estimates are the same, except for variable F4, but note how estimation of a Poisson model leads to distorted standard errors due to the fact that heterogeneity is not taken into account.

Classification results are shown in Table III. Finally, a two step procedure for prediction is proposed. Firstly, one can use discriminant analysis to predict future behaviour. Secondly, when the score obtained lies within a critical frame (around 50 %), a truncated negative binomial model is used to predict the number of defaulted instalments. If this number is large, the individual is classified to the *bad* group. On the other hand, if it is small, the prediction is *good*. The original criterium given by the bank is used to define the concept of large and small.

The first column represents the total correct classification, the second column is correct classification of *good*, the third is correct classification of *bad* and the forth is the percentage of *bad* accepted into the *good* group.

5 Final remarks

Classification problems in the context of credit granting decisions may use count data models due to the chracteristics of the dependent variable. In fact, the number of defaulted payments is the variable used to define whether a client is *good* (repaying) or *bad* (defaulter).

For the Poisson model, eliminating irrelevant terms, the objective log-likelihood function is:

$$\ell(y,\beta) = \sum_{i \leq n} \left[y_i X_i \beta - \exp(X_i \beta) \right].$$

For Type II negative binomial model, the expression is:

$$\sum_{i=1}^{n} \log \ell(y_i, X_i, \beta, \alpha) = \sum_{i=1}^{n} \log \frac{\Gamma(y + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i + 1)} (\frac{1}{\alpha e^{X_i \beta}})^{\frac{1}{\alpha}} \frac{1}{(1 + \frac{1}{\alpha e^{X_i \beta}})^{y_i + \frac{1}{\alpha}}}$$

The estimation of standard errors, asymptotic approximations through pseudo maximum likelihood were used (Gourieroux and Monfort, 1984).

For the truncated at zero negative binomial model the log-likelihood is:

$$\ln L = \sum_{i=1}^{m} \left[\ln \Gamma(y_i + \frac{\phi_i^c}{\alpha}) - \ln \Gamma \left(\frac{\phi_i^c}{\alpha} \right) - \ln \Gamma(y_i + 1) + y_i \ln \alpha - y_i (c-1) \ln \phi_i \right.$$

$$\left. -(y_i + \frac{\phi_i^c}{\alpha}) \ln(\alpha \phi_i^{(1-c)} + 1) - \ln(1 - P(Y=0)) \right]$$

where m is the number of individuals in the truncated sample, that is for which $y_i > 0$.

For estimation purposes, a particular value for c was fixed and α was estimated in a previous step (for details, see Guillén, 1992).

3 Prediction in credit scoring models

Usually, the performance of credit scoring models is evaluated through the percentage of correct classification for the individuals who already applied for credit, according to their subsequent behaviour. Nevertheless, the percentage of *bad* clients that would be classified as *good* by the scoring is a very important issue. It is this measure that is to be minimized since the smaller it is, the smaller the risk of granting credit to potencial defaulters.

Eventually, studies in this area take a part of the sample for estimation purposes and another part is used to check the preditive performance of the estimated models. In this work, we do not set part of the sample apart since an initial discriminant analysis was done with the whole sample and we did not want to alter this in the view of a final comparison of different approaches.

Definition of *good* and *bad* clients was based on the number of monthly instalments that were defaulted.

When using discriminant analysis, a score was associated to each individual. The score is a transformation of the probability of having been drawn from each of the two populations under study. If the estimated probability of being a *good* client is greater that the estimated probability of being *bad*, the prediction for the individual is that it belongs to the *good* group (and conversely, for a smaller probability). This prediction is compared to the actual client behaviour. When this is done for all individuals in the sample, an estimation of classification rates is obtained.

With the estimation of count data models, prediction has to be performed in two steps. Firstly, the number of expected defaulted monthly instalments is found. Afterwards, the definition of *predicted good* or *predicted bad* is assigned to the individual following the same criterium that is used to define *good* and *bad* clients in the sample. At the end, predicted and real behaviour are compared to obtain estimated classification rates that may be used to evaluate the performance of this methodololy to traditional approaches.

where $y_0 \in \{0, 1, 2, \ldots\}$ y $x_0 \in R^k$ and the λ parameter is related to the covariant variables through the expression bellow, thus preserving positivity

$$\lambda = \exp(X\beta). \tag{1}$$

This model is a generalized linear model (McCullagh and Nelder, 1983) and was presented in an equivalent form by Gourieroux and Monfort (1984a and 1984b).

It is known that the Poisson model does not account for heterogeneity since it has a mean-variance restriction. Negative binomial models are one possible generalization of the previous one, allowing a flexible relationship in terms of mean and variance.

Negative binomial models appear when a disturbance term, ϵ is introduced in relationship (1), so that

$$\ln(\lambda) = X\beta + \epsilon.$$

If ϵ is assumed to have a Gamma distribution, this leads to a conditional negative binomial distribution for the dependent variable Y.

In fact, the negative binomial distribution requires two parameters, unless a constant term is present in X and ϵ is taken with mean equal to one. Thus, both parameters are related to X in the following way:

$$\phi = \exp(X\beta) \tag{2}$$

and

$$\nu = (1/\alpha)(\exp(X\beta))^c \text{ for } \alpha > 0 \text{ and } c \text{ a fixed constant.} \tag{3}$$

Different possibilities for constant c offer different possible negative binomial models and types of (conditional) mean-variance relationships. For $c = 1$, taking fixed X, $\mathrm{Var}(Y) = (1 + \alpha)\mathrm{E}(Y)$, which was called Type I negative binomial model by Cameron and Trivedi (1986). Likely, for $c = 0$ Type II is obtained and the relationship is $Var(Y) = \mathrm{E}(Y)(1 + \alpha\mathrm{E}(Y))$.

When truncated data need to be studied, particularly when the truncation point is zero, modelization is based on the fact that

$$P(Y = y_i | Y > 0) = \frac{P(Y = y_i)}{P(Y > 0)}, \quad y_i = 1, 2, \ldots.$$

To get the truncated negative binomial model we first note that

$$P(Y = y_i) = \frac{\Gamma(y_i + \nu)}{\Gamma(\nu)\Gamma(y_i + 1)}(\frac{\nu}{\phi})^{\nu}(1 + \frac{\nu}{\phi})^{-(y_i+\nu)} \quad y_i = 0, 1, 2, \ldots.$$

then using (2) and (3) the following expression is obtained

$$P(Y = y_i) = \frac{\Gamma(y_i + \frac{\phi_i^c}{\alpha})}{\Gamma\left(\frac{\phi_i^c}{\alpha}\right)\Gamma(y_i + 1)}\alpha_i^y(\alpha\phi_i^{(1-c)} + 1)^{-(y_i+\frac{\phi_i^c}{\alpha})}\phi_i^{(1-c)y_i}.$$

Finally,

$$P(Y = y_i | Y > 0) =$$

$$= \frac{\Gamma(y_i + \frac{\phi_i^c}{\alpha})}{\Gamma\left(\frac{\phi_i^c}{\alpha}\right)\Gamma(y_i + 1)}\alpha_i^y(\alpha\phi_i^{(1-c)} + 1)^{-(y_i+\frac{\phi_i^c}{\alpha})}\phi_i^{(1-c)y_i}(1 - P(Y = 0))^{-1}.$$

Log-likelihood functions for the untruncated models are stated bellow.

classes: *good* and *bad*. *Good* clients would return the money completely, where as *bad* clients would be defaulters.

A data base was available, having information from almost 5000 clients. Firstly, by means of the covariant variables, a discriminant analysis was performed, in order to produce a discriminant rule that would serve as a basis for the decision to grant credit. This procedure has been suggested by many authors in the same situation. For example, Myers and Forgy (1963) worked on a data base from an institution in this way. More recently, Steenackers and Goovaerts (1989) provided a similar approach for a Belgian bank.

Afterwards, looking more carefully to the available data, we tried to find alternative ways to produce an estimation of the expected level of debt for potential creditors. The crucial point was to note that the behaviour of clients was represented exclusively by a variable counting the number defaulted instalments, that is the number of times the client did not pay the money as it was agreed when credit was granted. In fact, this variable was the basis for the definition of the two populations: *good* and *bad* clients.

So, instead of using techniques to classify individuals into populations, in this work we suggest that a sensible approach is to modelize the variable counting the number of defaulted instalments, which is a way to get a model to predict the expected level of debt for new applicants. Moreover, we want to find an adequate model for the creditors having already defaulted at least one payment, in order to see whether there exists a structural change in the model. This would lead to the interpretation that the process generating the transition from no unpayment to one is different from the process from whatever different from zero number of defaults to one more. Mullahy (1986) proposed *hurdle models* to cope with this situation.

Here, our purpose is to use count data models to predict the number of times that an applicant for credit will not pay the accorded amount to return the credit. So, two different issues are addressed: modelization and prediction.

Poisson models are typically used in situations where the dependent variable is discrete, for example, the number of patents applied for by firms (Hausman, Hall and Grilliches, 1984). Alternatively, negative binomial distribution models, that take into account the observed heterogeneity, have also been used (e.g., Cameron and Trivedi, 1986). Other discrete distributions are also useful to model count data, but in this paper they are not stated.

Here, we are going to use Poisson models, negative binomial models and their truncated versions to model our dependent variable. In the next section, we present the model specification and estimation procedure. Afterwards, we explain how prediction performance will be evaluated and finally, the results for the particular data set are presented.

In the last part, we draw some conclusions and we suggest possible ways to improve our approach.

2 Model specification and estimation

The basic models for count data have been studied by many authors such as El Sayyad (1973), Frome, Kutner and Beauchamp (1973), Terza (1985), Cameron and Trivedi (1986) and Mullahy (1986).

Let Y a dependent discrete non-negative variable. Assume that Y is the variable of interest in a population from which we have a random sample of size n. Assume that X is a vector of k explanatory variables that will be used for the modelization of Y Let $y = (y_1, ..., y_n)'$, the vector $n \times 1$ observations of Y, and $X = (x_1, ..., x_k)$, the matrix $n \times k$ for the observations of the explanatory variables where $x_j = (x_{j1}, ...x_{jn})'$ is the vector corresponding to the n observations of the j-th variable, $j = 1, ...k$.

The Poisson model is obtained when the conditional probability takes the following form:

$$P(Y = y_0 \mid X = x_0) = \frac{\exp(-\lambda)\lambda^{y_0}}{y_0!},$$

Count Data Models for a credit scoring system

Montserrat Guillén *
Departament d'Econometria, Estadística i Economia Espanyola
Universitat de Barcelona

Manuel Artís
Departament d'Econometria, Estadística i Economia Espanyola
Universitat de Barcelona

Paper presented at the Third Meeting on the European Conference Series in Quantitative Economics and Econometrics on *Econometrics of Duration, Count and Transition Models.*
Paris, December, 10-11, 1992

Abstract

Credit scoring systems created for the evaluation of new applications are based on the available statistical information which is related to the behaviour of former clients with credit. Usually, financial institutions apply discriminant analysis techniques to create these systems but they lack of good properties due, for example, to the presence of non-normal variables.

As an alternative, the future repayment behaviour is predicted by means of the expected number of unpaid instalments. The use of this latter variable suggests that appropriate models might be of interest, in which some covariant exogenous variables are included in order to specify the expected level of debt. At this point, prepayment is not explicitly considered. These models should be used as explanatory tools when evaluating the level of risk involved in personal credit transactions.

Negative Binomial Distribution models show particularly useful when heterogeneity is taken into account. Some results related to prediction performance are shown for different model specifications in the case of data from a Spanish bank.

Keywords: *count data, NBD models, credit scoring*
AMS Classification: 90A19
Abbreviated title: *Count Data Models*

1 Introduction

The application of statistical techniques for the analysis of decision problems entailing classification has experimented a development in the last decades. Specially, in the context of economics and finance, some particular problems have offered a wide range of possible applications for theoretical results. For a review, see Altman, Avery, Eisenbeis and Sinkey (1981).

The work presented here was motivated as we wished to develop a system for credit scoring. This means that a financial institution desired to find a way to classify new clients applying for credit into two different

*Mailing address: Montserrat Guillén. Departament d'Econometria, Estadística i Economia Espanyola, Universitat de Barcelona, Diagonal, 690, E-08034 Barcelona, Spain. (e-mail: guillen@riscd2.ub.es)