ELSEVIER Japan and the World Economy 10 (1998) 299–316

Japan and the WORLD ECONOMY

Sample selection in credit-scoring models[1]

William Greene*

Department of Economics, Stern School of Business, New York University, 44 West 4th Street, Mec 7-80, New York, NY 10012, USA

Received 30 November 1995; accepted 31 March 1998

Abstract

We examine three models for sample selection that are relevant for modeling credit scoring by commercial banks. A binary choice model is used to examine the decision of whether or not to extend credit. The selectivity aspect enters because such models are based on samples of individuals to whom credit has already been given. A regression model with sample selection is suggested for predicting expenditures, or the amount of credit. The same considerations as in the binary choice case apply here. Finally, a model for counts of occurrences is described which could, in some settings also be treated as a model of sample selection. © 1998 Elsevier Science B.V.

Keywords: Poisson regression; Sample selection; Probit

1. Introduction

In evaluating an application for a large loan, such as a mortgage or a construction loan, a lender will rely on direct, individual scrutiny by a loan officer or committee. But, a credit-card vendor, such as American Express, Visa, or Mastercard, might have to examine millions of applications over the course of a year. Rather than examine each one in detail, vendors usually rely on fairly simple models to assign scores to applications. A high enough score merits acceptance of the application. This process is completely statistical. The evaluator is using a profile of a successful borrower as a yardstick against which to measure the individual applications. Thus, this process sorts applications based on residuals and outliers from statistical models.

There is a potential problem with credit scoring as it is usually done. The statistical models used to evaluate applicants are constructed from historical data. In order to enter the sample used to build the model, an individual must have already been 'accepted.' Thus, the

* Corresponding author. Tel.: +1 212 9980876; fax: +1 212 9954218; e-mail: wgreene@stern.nyu.edu

[1] Working Paper No. 194 presented at The Japan–U.S. Technical Symposium in Japan, November 13, 1995.

PII S0922-1425(98)00030-9

model is a description of some aspect of the behavior of individuals who have already received loans. The scoring model is to be used to evaluate applicants who are drawn, arguable randomly, from the entire population. The individuals whose applications were accepted to begin with were qualitatively different from individuals whose applications were rejected. Since, an application which arrives randomly at the vendor could be of either type, ex post, it is not certain that the model being used is appropriate for the population being measured. To consider a concrete example, the vendor is interested in assessing the likelihood of loan default if an individual's application is accepted. Mathematically, they are interested in Prob[Default|Acceptance], so this is the model that is required for the evaluation. But, because of the selective nature of the data used to build the model, it is possible that using data selected 'ex post' produces a biased estimate of this probability. We find a similar observation applies to a model which attempts to predict expenditures and one which describes an important explanatory variable used in the credit-scoring model, itself, the number of derogatory reports in an applicant's credit history.

This paper will describe three applications in this general area of study. We are interested in how sample selection affects the measurement of some variables of interest to credit-card vendors. The three studies analyze different types of response variables, each of which requires a different type of statistical model and a different estimation technique. The outline of the paper is as follows: The sample selection model is briefly described in general terms in Section 2. Section 3 presents the most important application, the prediction of loan default. Section 4 presents a relatively standard model for expenditure. Section 5 examines the (apparently) most influential measure in the observed credit-scoring model, the number of derogatory reports observed in individuals' credit histories. In each of these applications, a model which accounts for the sample selection problem produces predictions that are quite different from one which does not. Some conclusions and a summary are given in Section 6. The three applications are illustrated with a large sample of observations generated by a major credit-card vendor in 1991. Appendix A describes the data set.

2. The sample selection problem

Applications for credit-card accounts are handled universally by a statistical process of 'credit scoring.' The scorers (who, in many cases, are not the credit-card vendors themselves) use historical data on loan performance to build a profile of a successful loan recipient. An applicant is then measured against successful recipients by constructing their score and comparing it to the norm. The underlying model that is producing the score can be viewed as a predictor of some response (e.g., default), conditional on the sampling rule, in this case, on acceptance of the application. The potential flaw in the model is that if there are factors which enter the acceptance decision but do not appear explicitly in the rule, and these same factors influence (or are correlated with) the response in the performance equation, then the latter equation may produce biased predictions. Thus, to continue the example, a predictor of default risk in a given population of applicants can be systematically biased because it is constructed from a nonrandom sample of past applicants, that is, those whose applications were accepted.

This is a straightforward application of what is known as the problem of sample selection. In general terms, the 'sample selection problem' can be viewed as follows: It is desired to build a model of an economic response, denoted 'y', (default, expenditure, credit history, etc.), for purposes of predicting the behavior of individuals in a specific population denoted A. We might denote the model generally as $E[y|A] = f_A(x, \beta)$, where x denotes a set of attributes which are assumed to explain the variation in $E[y|A]$. Under normal conditions, data would be drawn randomly from population A and used to fit the model, which could then be used to make predictions. But, suppose that an individual's presence in population A is determined by some process that is correlated with, if not necessarily a function of y, itself. Then, the model-building process could be tainted by this latent effort. To consider an example, suppose we desire to explain the incomes of individuals who attend college, so as to compare them to the incomes of individuals who do not attend college. The direct approach that intuition might suggest, just comparing average incomes of the two groups neglects the possibility that individuals who attend college (population A) select themselves into that population on the basis of traits (e.g., motivation, endurance) that will ultimately affect their incomes, whether or not they attend college. This fact will distort the comparison of the two groups. If it is left unaccounted for, the analyst might attribute to college attendance differences which are at least partially explained by differences in the college attending individuals, themselves.

In this paper, we will consider three applications that bear similarity to this example. In the first, we are interested in the prediction of default on a credit-card loan. Loan default as measured here is a binary response; either the individual defaults (D=1) or they do not (D=0). The problem is to build a satisfactory model of Prob$[D = 1|$individual has a credit card$]$ when there are factors which explain default behavior which also enter the vendor's decision to grant a card. The second application is a direct extension. In order to predict the expenditures of a credit-card applicant, one must consider that spending behavior is clearly tied to the same behavior that induces default. This application involves a continuous variable and is very similar to the income example given above. Finally, we examine a variable which is apparently of great interest to credit scorers, the number of derogatory reports in an applicant's credit history. (We infer this from the strong significance of this variable in a model of the acceptance decision.) This application is similar to the expenditure model. The interesting difference for our purposes is the nature of the response variable, which is a count. This requires a newly developed approach to the selected regression problem.

3. Estimating the probability of default

Although one might expect the evaluation of a credit-card application to be directed toward profitability of a loan, default risk appears to be the major focus of credit scoring. The credit-scoring process is typically used simply to estimate, essentially, the probability that an individual will service the debts incurred with the card. In this section, we describe briefly the process used by most vendors and some alternative models that might usefully be considered.

Most vendors use outside agencies for the credit-scoring function. Applications are forwarded from the vendor (e.g., American Express or Visa issuing bank) to an agency which evaluates them and returns them with the scores. The formulas used for credit scoring are closely guarded trade secrets, though the data analyzed here are suggestive of the underlying process.

The most common technique used for credit scoring is linear discriminant analysis. The technique of discriminant analysis rests on the assumption that there are two populations of individuals, which we denote '1' and '0,' each characterized by a multivariate distribution of a set of attributes, x, including such factors as age, income, family size, credit history, occupation, and so on. An individual with attribute vector x_i, is drawn from one of the two populations, and it is desired to determine which. The analysis is carried out by assigning to the application a 'Z' score, computed as

$$Z_\mathrm{i} = a + b' x_\mathrm{i}.$$

Given a sample of previous observations on y_i and x_i, the vector of weights, (a,b), can be obtained as a multiple of the vector of regression coefficients in the linear regression of $d_\mathrm{i}=P_0 D y_\mathrm{i} - P_1(1-y_\mathrm{i})$ on a constant and the set of attributes, where P_1 is the proportion of 1s in the sample and $P_0=1-P_1$. The scale factor is $(n-2)/e'e$ from the linear regression[2]. The individual is classified in group 1 if their 'Z' score is greater than some Z^* (usually 0) and 0 otherwise. The linearity (and simplicity) of the computation is a compelling virtue. Thus, in the current context, we use this 'Z score' method to classify applicants as defaulters (D=1) or nondefaulters (D=0). The applications of individuals assigned to the D=1 class are rejected.

This method divides the universe of loan applicants into two types, those who *will* default and those who *will not*. The crux of the analysis is that at the time of the application, the individual is as if preordained to be a defaulter or a nondefaulter. In point of fact, the same individual might be in either group at any time, depending on a host of attending circumstances and random elements in their own behavior. Thus, prediction of default is not a problem of classification the same way as is, say, determining the sex of prehistoric individuals from a fossilized record.

Index function based models of discrete choice, such as the probit and logit models, assume that for any individual, given a set of attributes, there is a definable probability that they will actually default on a loan. This interpretation places all individuals in a single population. The observed outcome, default/no default, arises from the characteristics and random behavior of the individuals. Ex ante, all that can be produced by the model is a probability. The observation of y_i ex post is the outcome of a single Bernoulli trial.

This alternative formulation does not assume that individual attributes, x_i, are necessarily normally distributed. The probability of default arises conditionally on these attributes and is a function of the inherent randomness of events and human behavior and the unmeasured (and unmeasurable) determinants which are not specifically included in the

[2] See Maddala (1983), pp. 18–25.

model[3]. The core of the formulation is an index function model with a latent regression,

$$D_{\mathrm{i}}^{*} = \beta' x_{\mathrm{i}} + \varepsilon_{\mathrm{i}}$$

The dependent variable might be identified with the 'propensity to default.' An intuitively appealing interpretation of D_{i}^{*} is as a quantitative measure of 'how much trouble the individual is in.' Conditioning variables, x_{i}, might include income, credit history, the ratio of credit-card burden to current income, and so on. If D_{i}^{*} is sufficiently large relative to the attributes, that is, if the individual is in trouble enough, they default. Formally,

$$D_{\mathrm{i}} = 1 \text{ if } D_{\mathrm{i}}^{*} \geq 0 \text{ and } 0 \text{ otherwise,}$$

so the probability of interest is

$$P_{\mathrm{i}} = \mathrm{Prob}[D_{\mathrm{i}} = 1 | x_{\mathrm{i}}]$$

Assuming that ε is normally distributed with mean 0 and variance 1, we obtain the default probability

$$\mathrm{Prob}[D_{\mathrm{i}} = 1|x_{\mathrm{i}}] = \mathrm{Prob}[D_{\mathrm{i}}^{*} > 0|x_{\mathrm{i}}] = \mathrm{Prob}[\varepsilon_{\mathrm{i}} \leq \beta' x_{\mathrm{i}}|x_{\mathrm{i}}] = \Phi(\beta' x_{\mathrm{i}})$$

where $\Phi(\bullet)$ is the standard normal CDF[4]. The classification rule is

$$\text{Predict } D_{\mathrm{i}} = 1 \text{ if } \Phi(\beta' x_{\mathrm{i}}) > P^{*},$$

where P^{*} is a threshold value chosen by the analyst.

Whether one uses discriminant analysis, a probit model, or some other, the quantity ultimately of interest in this exercise is the probability of default that would apply, if the individual were issued a credit card, which we denote $\mathrm{Prob}[D = 1|C = 1, x]$. We denote rejection of the application by 'C=0.' But, the preceding construction, it is unclear whether that is what is actually estimated. Recall the underlying structure of the model,

$$D_{\mathrm{i}} = \beta' x_{\mathrm{i}} + \varepsilon_{\mathrm{i}}$$

The probability we seek is

$$\mathrm{Prob}[D_{\mathrm{i}} = 1|C = 1] = \mathrm{Prob}[D_{\mathrm{i}}^{*} > 0|C = 1] = \mathrm{Prob}[\varepsilon_{\mathrm{i}} < \beta' x_{\mathrm{i}}|C = 1]$$

which we have assumed thus far is simply the normal probability. But, we must now account for the sample selection rule. The individuals who are in the sample are those who have already been granted a card. The selection rule can be written

$$\mathrm{Prob}[C = 1] = \text{some function of}(x, z)$$

where that function remains to be determined. The nature of the selection rule is now critical. In order for the simple model given above,

$$\mathrm{Prob}[D_{1} = 1|C = 1] = \Phi(\beta' x_{\mathrm{i}})$$

[3] Our discussion of this modeling framework will also be brief. More details can be found in Ref. (Greene, 1997, ch. 19).

[4] One might question the normality assumption. But, the logistic and alternative distributions rarely bring any differences in the predictions of the model. For our data, these two models produced virtually identical results at the first stage. However, only the probit form is tractable in the integrated model to follow.

to be correct, it must be true that the events $C=1$ and $D=1$ are independent. But, of course, this is exceedingly unlikely, since cardholder status is explicitly granted based on some kind of assessment of the default probability. This is precisely the selection problem discussed in Section 2. There are two kinds of failures that can arise:

1. z does not actually enter Prob[$C=1$], but x does. Then, even if the normal distribution assumed is correct (which is unlikely), β is not the right coefficient vector. Thus, estimation of the probit model produces a biased estimate of β.
2. If z does enter Prob[$C=1$], then it enters the joint probability and hence the conditional probability. In that event, another source of bias is the omission of z from the estimated conditional probability. The analogy drawn earlier would be our omission of some measure of motivation or endurance in our explanation of the incomes of individuals who attend college.

Both of these produce the possibility that the simple model produces a biased set of coefficient estimates and, therefore, a biased estimate of the default probability.

We will proceed to define and estimate a model of the default probability that accounts for the sample selection. We will use a bivariate probit specification to model this. The structural equations are

$$\textbf{Default equation} : D_{\mathrm{i}} = \beta' x_{\mathrm{i}} + \varepsilon_{\mathrm{i}}$$

$D_{\mathrm{i}}=1$ if and only if $D_{\mathrm{i}}{*}>0$, and 0 else.

$$\textbf{Cardholder equation} : C_{\mathrm{i}}^* = \gamma' v_{\mathrm{i}} + u_{\mathrm{i}}$$

$$C_{\mathrm{i}} = 1 \text{ if and only if } C_{\mathrm{i}}^* > 0, \text{ and 0 else.}$$

$$D_{\mathrm{i}} \text{ and } x_{\mathrm{i}} \text{ are only observed if } C_{\mathrm{i}} = 1$$

$$C_{\mathrm{i}} \text{ and } v_{\mathrm{i}} \text{ are observed for all applicants.}$$

$$\textbf{Selectivity} : [\varepsilon_{\mathrm{i}} U_{\mathrm{i}}] \sim \mathrm{N2}[0, 0, 1, \rho_{\varepsilon u}]$$

The vector of attributes, v_{i}, are the factors used in the approval decision. The probability of interest is the probability of default given that a loan is accepted, which is

$$\mathrm{Prob}[D_{\mathrm{i}} = 1 | C_{\mathrm{i}} = 1] = \frac{\Phi_2[\beta' x_{\mathrm{i}}, \gamma' v_{\mathrm{i}}, \rho]}{\Phi[\gamma' v_{\mathrm{i}}]}$$

where Φ is the bivariate normal cumulative probability. If ρ equals 0, the selection is of no consequence, and the unconditional model described earlier is appropriate. This model was developed and recently applied to an analysis of consumer loans.[5] Note, once again, this is a considerably more involved expression than the simple model. Estimation of the model is described by Greene (1992).

[5] Boyes et al. treated the joint determination of cardholder status and default as a model of partial observability in the sense of Poirier (1980). Since cardholder status is generated by the credit scorer while the default indicator is generated later by the cardholder the observations are sequential, not simultaneous. As such, the model of Abowd and Farber (1982) might apply. But, the simpler censoring interpretation seems more appropriate. It turns out that the difference is only one of interpretation. The log-likelihood functions for Boyes et al.'s model (see their page 6) and ours are the same.

Table 1
Estimated cardholder equation joint with default equation

	Coeff.	Std. error	*t*-ratio
Basic cardholder specification			
Constant	−1.2734	0.1563	−8.1500
AGE	0.0000	0.0039	−0.0060
MTHCURAD	0.0015	0.0006	2.4650
DEPNDNTS	−0.1314	0.0487	−2.7000
MTHMPLOY	0.0003	0.0006	0.4910
MAJORDRG	−0.8230	0.0442	−18.6340
MINORDRG	0.0082	0.0462	0.1780
OWNRENT	0.0129	0.0765	0.1680
MTHPRVAD	0.0003	0.0004	0.6980
PREVIOUS	0.1185	0.1283	0.9240
INCOME	0.0156	0.004	3.8670
SELFEMPL	−0.5651	0.1307	−4.3250
TRADACCT	0.0850	0.0064	13.3520
INCEPER	−0.0550	0.0513	−1.0730
Credit bureau			
CREDOPEN	−0.0096	0.0109	−0.8760
CREDACTV	0.0060	0.0223	0.2700
CRDDEL30	−0.3167	0.1197	−2.6470
CR30DLNQ	−0.0965	0.0317	−3.0480
AVGRVBAL	0.0049	0.005	0.9740
AVBALINC	−0.0014	0.0008	−1.9060
Credit reference			
BANKSAV	−0.4708	0.1731	−2.7190
BANKBOTH	0.5074	0.0694	7.3100
CRDBRINQ	−0.1743	0.0176	−8.3930
CREDMAJR	0.3663	0.0807	4.5410
Correlation between disturbances			
Puε	0.1178	0.258	1.7360

The preceding analysis was applied to a large sample of credit-card application. Of 13 444 applications received, 10 499 were accepted[6]. A full description of the data set is given in the Appendix A. Tables 1 and 2 present estimates of the parameters of a model of loan default based on this sample. The cardholder equation is largely consistent with expectations. The most significant explanatory variables are the number of major derogatory reports and the credit bureau inquiries (negative) and the number of open trade accounts (positive). What Table 1 reveals most clearly is the credit-scoring vendor's very heavy reliance upon credit reporting agencies such as TRW. There is one surprising result. Conventional wisdom in this setting is that the own/rent indicator for home ownership is

[6] The sample used was 'choice based'. At the time the data were generated, the true acceptance rate was closer to 60%. The credit-card vendor provided the choice based sample so as to facilitate analysis of the very low default rate. The Lerman and Manski (1981) WESML procedure was used to correct the bias introduced by the sample design.

Table 2
Default models

Variable	Unconditional			Conditional		
	Coeff.	Std. error	*t*-ratio	Coeff.	Std. error	*t*-ratio
Basic default specification						
Constant	−1.1350	0.0984	11.5330	−1.3752	0.3945	−3.4860
AGE	−0.0031	0.0023	−1.3420	0.0054	0.0094	−0.5820
MTHCURAD	0.0003	0.0003	1.0690	0.0002	0.0013	0.1530
DEPNDNTS	0.0445	0.0294	1.5120	−0.0217	0.1114	−0.1950
MTHMPLOY	0.0007	0.0003	2.3310	0.0007	0.0013	0.5660
MAJORDRG	0.0592	0.0408	1.4480	−0.2969	0.1985	−1.4950
MINORDRG	0.0764	0.0296	2.5860	0.1780	0.0993	1.7930
OWNRENT	−0.0010	0.4312	−0.0230	0.0908	0.1706	0.5330
MTHPRVAD	0.0004	0.0002	1.8170	0.0002	0.0009	0.2740
PREVIOUS	−0.1507	0.0792	−1.9020	−0.1112	0.3103	−0.3580
INCOME	−0.1507	0.0033	−5.6080	−0.0072	0.0151	−0.4760
SELFEMPL	0.0788	0.0850	0.9270	−0.1969	0.3565	−0.5520
TRADACCT	0.0004	0.0044	0.1090	0.0207	0.0205	1.0090
INCPER	0.0228	0.0323	−0.7060	−0.0545	0.1058	−0.5150
EXP_INC	0.4761	0.1717	−2.7740	−0.5790	0.5033	−1.1500
Credit bureau						
CREDOPEN	0.0138	0.0063	2.1950	0.0199	0.0272	0.7320
CREDACTV	−0.1218	0.0126	−9.6570	−0.1500	0.0557	−2.6950
CRDDEL30	0.2841	0.0712	3.9910	0.2829	0.2766	1.0230
CR30DLNQ	0.0806	0.0177	4.5590	0.0446	0.0757	0.5890
AVGRVBAL	0.0011	0.0024	0.4390	0.0156	0.0123	1.2680
AVBALINC	0.0039	0.0004	9.1920	0.0008	0.0021	0.3980
Expenditure						
FITEXP	0.0014	0.0044	3.1030	0.0006	0.0019	0.3360

the single most powerful predictor of whether an applicant will be given a credit card. We find no evidence of this in these data. Rather, as one might expect, what explains acceptance best is a higher income, fewer dependents, and a ‘clean’ credit file with numerous accounts in the reporting agency. Surprisingly, being employed longer at one’s current job appears not to increase the probability of approval, though being self-employed appears significantly to decrease it. We should note, the market descriptive data are interesting for revealing patterns in the default data. But, because they do not relate specifically to the individual, they cannot be used in a commercial credit-scoring model. Before leaving this discussion, we note that this cardholder equation is only a model of the true model. The binary dependent variable, C, is produced by a deterministic rule. However, we know neither the exact functional form nor the precise list of variables which enter the function. We are confident, however, that the number of derogatory reports, the number of credit bureau inquiries and the presence of bank accounts do enter the true equation.

Table 2 gives the probit estimates of the default equation. Predicted monthly expenditure, FITEXP, is computed using the model described in the Section 4. The ‘selection’

Table 3
Estimated default probabilities

Group	Conditional	Unconditional
All observations	0.1498	0.1187
Cardholders	0.1056	0.0947
Noncardholders	0.3090	0.2061
Defaulters	0.1632	0.1437
Nondefaulters	0.0997	0.0895

variable, λ_i, is computed using the coefficients from a single equation estimates of a cardholder equation (not reported here). The single equation, unconditional model is given in the first three columns. The results agree with a conjecture that default rates might be related to expenditures; the idea of cardholders 'getting in over their heads' comes to mind. Table 2 also presents the full information, conditional estimates of the default equation[7].

Table 3 shows the average of the predicted default probabilities computed with the models in Tables 1 and 2 for some subgroups of the data set.

The results show strongly that the model is sharpened by the addition of a control for the selection problem. It also appears that both our and the underlying credit-scoring model are sorting the data appropriately. The conditional model predicts a much higher average default rate for the population as a whole (0.1498 vs. 0.1187), largely because of its assignment of much higher likelihoods of default to the individuals whose applications were ultimately rejected. Also, the conditional model appears to distinguish somewhat more sharply those individuals who actually did default on their loans (average probability of 0.1632 for the conditional model vs 0.1437 for the unconditional one).

4. Predicting expenditure

For a credit-card vendor interested in more than just default risk, consumer expenditures would be another important quantity to study. (Patterns of repayment would also be of interest, but are beyond the scope of this study.) The variable of interest is expenditure, denoted S, conditioned on cardholder status, C=1. Given observed data on expenditures for a sample of cardholders, we can fit an equation for predicting monthly expenditure for cardholders. But, the same selection issue as before arises, and once again, it becomes a question as to whether the familiar technique (in this case, linear regression) can produce an unbiased estimate of the desired regression.

The model is $(S_i|C_i = 1) = \beta' x_i + \varepsilon_i$.

Does linear regression of average monthly expenditure on a set of attributes produce unbiased estimates of β and, thereby, allow unbiased prediction of expenditure for an

[7] A reader has observed that one might suspect that the probability of loan default is also affected by more general characteristics of the economy. Default and bankruptcy do tend to increase during recessions. Since our study is cross sectional in nature – the data were drawn in November 1991 – we have no evidence on this issue one way or the other.

applicant? Once again, we postulate a selection equation,

$$\begin{aligned} C_i^* &= \alpha' v_i + u_i \\ C_i &= 1, \quad \text{if } C_i^* > 0, \\ &= 1 \quad \text{if } u_i > -\alpha' v_i \end{aligned}$$

By implication, then,

$$E[S_i|C_i = 1] = \beta' x_i + E[\varepsilon_i|C_i = 1] = \beta' x_i + (\rho_{\in u}\sigma)M_i,$$

Table 4
Expenditure equations

Variable	Selection model		Uncorrected model	
	Parameter	*t*-ratio	Parameter	*t*-ratio
Constant	−3.3561	0.0200	51.0224	0.3200
AGE	−1.4929	−4.3200	−1.4420	−4.2000
ADEPCNT	−1.3982	−0.5000	1.6088	0.5900
OWNRENT	−5.4236	−0.7100	−10.3701	−1.3700
INCOME	54.1945	26.5800	51.8008	25.9400
SELFEMPL	−28.5441	−2.0100	−17.7309	−1.2600
TRADACCT	0.5302	0.9100	−1.1348	−2.1400
PROF	72.6506	0.4600	60.4627	0.3800
MGT	61.7839	0.3900	51.0549	0.3200
MILITARY	9.4426	0.0600	−4.4128	−0.0300
CLERICAL	26.4140	0.1700	13.2977	0.0800
SALES	113.6751	0.7200	103.0752	0.6500
OTHERJOB	54.4613	0.3500	42.6444	0.2700
BUYPOWER	383.6338	1.0100	375.4135	0.9800
PCTCOLL	1.7577	3.8000	1.7143	3.7000
MEDAGE	−0.0869	−0.1400	−0.0521	−0.0800
MEDINC	14.2060	3.5900	13.4840	3.4000
PCTOWN	−0.5252	−3.9400	−0.5155	−3.8500
PCTBLACK	0.5197	2.9000	0.6157	3.4300
PCTSPAN	0.6288	2.4200	0.6921	2.6500
GROWTH	0.0060	0.3800	0.0054	0.3400
PCTEMPL	−0.0166	−0.5000	−0.0172	−0.5200
APPAREL	0.8221	0.5500	0.8073	0.5400
AUTO	−4.7966	−1.8700	−4.8683	−1.8900
BUILDMTL	1.4894	0.5600	1.2959	0.4900
DEPTSTOR	−6.9411	−0.5000	−6.5342	−0.4700
EATDRINK	−1.2276	−1.4900	−1.2646	−1.5300
FURN	0.9885	0.8500	1.0487	0.9000
GAS	−1.7548	−0.8800	−1.8450	−0.9200
LAMBDA	108.3403	7.2500		
Rho	0.3379			
σ-corrected	320.6213			
R^2	0.0970		0.0925	
s.d. ε_i			316.8383	

Table 5
Average predicted monthly expenditures

	Corrected	Uncorrected
All observations	\$263.29	244.88
Cardholders	\$251.03	251.30
Noncardholders	\$307.03	221.97

where

$$M_\mathrm{i} = \phi(\alpha' v_\mathrm{i})/\Phi(\alpha' v_\mathrm{i}),$$

because of the joint normality. M_i is the inverse Mills ratio, or selectivity correction, familiar in the literature on modeling sample selection. If, in fact, the cardholder decision is correlated with the disturbance in the expenditure equation, that is, if ρ is nonzero, then linear regression of S on x will not produce unbiased estimates. The same results as before are obtained here. If x appears in the cardholder equation, then directly, coefficients on x are biased. Second, if z contains variables not in x, then these variable have been omitted from the equation, and, once again, a bias is imparted to the extent that the included variables are imperfect predictors of the excluded ones. The upshot is that simple linear regression of S on x will not estimate β without bias.

Heckman (1979) proposed the following two step approach. Step 1 – estimation of the cardholder equation using probit analysis, as we did above. Step 2 – linear regression of S on x and M will produce consistent (albeit not unbiased) estimates of β and $\theta{=}\rho\sigma$. The estimated standard errors must be corrected to account for the fact that parameters from the first step appear in the regression at the second. Formulas appear in Heckman (1979), Greene (1997).

It remains to be seen whether the predictions from the model without accounting for selectivity are systematically biased. They may not be, if the biased coefficients systematically offset one another, which is certainly possible. Results for the linear regressions are given below. Table 4 gives the estimates of an expenditure equation estimated with and without the sample selection correction. The highly significant estimate of θ strongly suggests that the selection is influential in the results.

Table 5 gives the average predicted expenditures for the selection corrected regression and for a simple linear regression which ignores the selectivity. The models give the same predictions for the cardholders. (This is to be expected, since this is a linear model.) But, they differ sharply for the observations ultimately rejected. The unconditional model predicts that rejectees would spend somewhat less, on average per month than cardholders, while the conditional model predicts that they would spend substantially more. Which of these is a better explanation of behavior is left for further research, but, as noted earlier, the latter seems more consistent with intuition.

5. Predicting MDRs

By far the most significant variable in the cardholder equation is MDRs, the number of major derogatory reports. One might argue that there is a kind of simultaneity at work, in that although the cardholder equation is conditioned on MDRs, it is very likely that there is a kind of selectivity at work in the determination of MDRs, much the same as in the default equation. In this section, we will examine a sample selection model for MDRs. We leave for later research the search for a more appropriate model of the joint determination of MDRs and cardholder status.

Neither of the previous modeling frameworks applies to the number of MDRs. The response variable in this study is discrete and nonnegative. The typical value is 0 or 1, but values range up to 14. Neither a probit style model for binary choice nor a linear-regression model applies. An appropriate modeling framework for a variable such as the count of MDRs would be a count data model such as the Poisson-regression model. For variable y which takes values 0, 1,$\cdots$, the Poisson-regression model specifies that

$$\mathrm{Prob}[y = j|x] = \exp(-\lambda_\mathrm{i})\lambda_\mathrm{i}^{j_\mathrm{i}}/j_\mathrm{i}!$$

The mean and variance of the distribution are both λ_i. (Although an interesting issue in its own right, we leave the matter of overdispersion to other work.) In order to ensure a nonnegative mean and variance, it is usually assumed that λ_i=exp ($\beta' x_\mathrm{i}$). Given observed data on y and x, maximum likelihood estimation of the Poisson model is extremely straightforward. (See Greene (1997).)

It seems clear that a model for the number of derogatory reports based on cardholder data would be tainted by selectivity in the same fashion as the previous two. It is tempting to modify the Poisson model straightaway by appending a Mills ratio term to the conditional mean, mimicking the Heckman model we used earlier for expenditures;

$$\log\lambda_\mathrm{i} = \beta' x_\mathrm{i} + \theta M_\mathrm{i}.$$

A two step estimator could then be used, in the same fashion as in the regression model of the Section 4: (1) Fit the cardholder equation by probit MLE; and (2) For the cardholders, compute M_i, then use the Poisson model to fit, by MLE, the nonlinear regression with the estimated M_i included as an additional regressor. This is the approach suggested by Greene (1994).

Terza (1995) argues persuasively that this is an inappropriate approach. First, there is no obvious reason for the Mills ratio term to enter the conditional mean function linearly as assumed above. Second, if the original model were a Poisson regression, the Poisson distribution would surely not apply in the selected subpopulation (though it is not obvious at all what distribution would apply.) Terza suggests, instead, a direct approach based on the introduction of heterogeneity in the conditional mean function. Continuing the formulation we have used previously, Terza's results are as follows: The conditional mean in the Poisson model is

$$\log\lambda_\mathrm{i} = \beta' x_\mathrm{i} + \varepsilon_\mathrm{i}$$

The selection equation is the same one specified in the expenditure model. If u and ε are correlated, then

$$E[y|x, C = 1] = \lambda\Phi(\theta + \alpha' w_\mathrm{i})/\Phi(\alpha' w_\mathrm{i}),\ \text{where}\ \theta = \rho\sigma.$$

The two important implications are that this conditional mean is not log-linear and that if the original distribution, conditioned on ε is Poisson, then the conditional distribution, given C=1, surely is not. This does not preclude estimation, however. Since the conditional mean function is known, the parameters of Terza's model can be estimated by nonlinear ordinary least squares (Terza (1995) gives details).

An alternative approach is suggested by Greene (1995c). If $\lambda_i\Phi(\theta+\alpha' w_i)/\Phi(\alpha' w_i)$ is expanded in a linear Taylor series around the point $\theta=0$ (the case of no selectivity bias), the resulting conditional mean is exactly what was suggested at the outset. This validates Greene's approach with an important qualification. The result does not reinstate the Poisson distribution, so nonlinear least squares remains the preferred estimator. Once obtained, the results of Murphy and Topel (1985) are used to obtain an appropriate asymptotic covariance matrix. (see Greene (1995c) for the mathematical results.)

Table 6 presents two sets of estimates for the Poisson model, the simple Poisson model and the selection corrected model described above. It is clear that there are some substantial

Table 6
Estimates of the MDR equation

		Number of major derogatory reports			
	Cardholder			Selection Corrected	
	Probit	Poisson	Poisson	Greene	Terza
Constant	0.542	−3.616	−4.594	−5.345	−4.068
	(−0.184)	(−0.422)	(−0.521)	(−0.74)	(−0.596)
Age	−0.00857	0.0188	0.0162	0.0128	0.0142
	(−0.00498)	(−0.00872)	(−0.00996)	(0.0110)	(0.0106)
Income	0.092	0.134	0.183	0.191	0.136
	(0.0532)	(−0.0543)	(−0.0613)	(−0.0596)	(0.0586)
Exp._Inc.		1.986	1.878	1.775	1.734
		(1.265)	(1.296)	(0.943)	(1.075)
Avg._Exp.		0.0000483	−0.0000236	−0.0000268	−0.0000362
		(0.000395)	(0.000419)	(0.000308)	(0.000405)
Major	0.212	0.242	0.572	1.376	0.811
	(0.103)	(0.268)	(0.316)	(0.590)	(0.491)
Mills ratio			1.788	1.989	3.465
			(0.431)	(0.296)	(30.689)
Sum of squared deviations				165.319	168.262
Own_Rent	0.349				
	(0.101)				
Depndt.	−0.131				
	(0.069)				
Inc._Per	−0.0150				
	(0.0714)				
Self_Empl.	−0.201				
	(0.163)				
	−0.286				
	(0.0245)				
Cur._Add.	−0.000409				
	(0.0007000)				
Active	−0.230				
	(0.0214)				
Log-likelihood		−407.944	−394.157		

Table 7
Averages of predicted numbers of major derogatory reports

	Major derogatory reports		Fitted poisson	Fitted conditional model	
All obs.	0.4564		0.1218		0.2030
C=0	1.5878	0.0967		0.4763	
C=1	0.1290		0.1290		0.1238

differences in the estimated models. The predicted counts shown in Table 7 confirm this finding. The model that ignores the selectivity appears to pick up almost none of the between group differences. The predictions from the conditional model are much more in line with expectations, given that this variable virtually dominates the cardholder decision.

6. Conclusions

The overall characteristics of the results reported here are to be expected. In this application, the effects of the sample selection are likely to be substantial, by construction. The implication for the credit-scoring process is that model builders might want to be wary of results that are based on selected samples in order to be used to make predictions for the larger population (as a whole).

One might be interested in extending the credit-scoring models. Ultimately, simple default is not what interests the vendor; profitability is. Thus, an integrated model involving default, expenditure, and costs might be of interest. (A rudimentary model is given by Greene (1992).) Our results suggest that accounting for the selected nature of the historical data will be important. Acceptance/rejection decisions are based on extremely simple, easily explained and justified criteria. The elaborate equations presented in the preceding could only be illustrative. Consumer laws and the practicalities of this market would make a large integrated model problematic. But, there is another aspect of this market in which one might be useful. Banks are increasingly interested in narrowly targeting their promotional efforts[8].

In this instance, a model which allows a sharper distinction between different types of customers will have some utility. Once again, for predictive purposes, it is important to account for the nature of the observed data in constructing the models.

Appendix A

Data used in the applications

The models described earlier were estimated for a well known credit-card company. The data set used in estimation consisted of 13 444 observations on credit-card applications

[8] The author recently received an offer from Visa promoting a card specifically targeted (somehow) to physicists.

Table 8
Variable used in analysis of credit-card default

Indicators	
CARDHLDR	1 for cardholders, 0 for denied applicants
DEFAULT	1 for defaulted on payment, 0 if not
Expenditure	
EXP1, EXP2, EXP3,···, EXP12	monthly expenditure in most recent 12 months
Demographic and socioeconomic, from application	
AGE	age in years and twelfths of a year
DEPNDTS	dependents, missing data converted to 1
OWNRENT	indicator=1 if own home, 0 if rent
MTHPRVAD	months at previous address
PREVIOUS	1 if previous card holder
ADDLINC	additional income, missing data coded as 0
INCOME	primary income
SELFEMPL	1 if self employed, 0 otherwise
PROF	1 for professional (airline, entertainer, other, sales, tech)
UNEMP	1 for unemployed, alimony, disabled, or other
MGT	1 for management services and other management
MILITARY	1 for noncommissioned and other
CLERICAL	1 for clerical staff
STAFF	1 for sales staff
OTHERJOB	1 for all other categories including teachers, railroad, retired, repair workers, students, engineers, dress makers, food handlers, etc
Constructed variables	
INCOME	labor income+additional income
AVGEXP	(1/12) $\sum_i \mathrm{EXP}_i$
INCPER	income per family member=(income+additional income)/(1+dependents)
EXP_INC	average expenditure for 12 months/average monthly income
Miscellaneous application data	
MTHCURAD	months at current address
CRDBRINQ	number of credit bureau inquiries
CREDMAJR	1 if first credit card indicated on application is a major credit card
CREDDEPT	1 if first credit card indicated is a department store card
CREDGAS	1 if first credit card indicated is a gasoline company
CURTRADE	number of current trade item accounts (existing charge accounts)
MTHEMPLOY	months employed
Types of bank accounts	
BANKSAV	1 if only savings account, 0 otherwise
BANKCH	1 if only checking account, 0 else
BANKBOTH	1 if both savings and checking, 0 else
Derogatories and other credit data	
MAJORDRG	count of major derogatory reports (long delinquencies) from credit bureau
MINORDRG	count of minor derogatories from credit bureau
TRADACCT	number of open, active trade lines
Credit bureau data	
CREDOPEN	number of open and current trade accounts
CREDACTV	number of active trade lines
CRDDEL30	number of trade lines 30 days past due at the time of the report

Table 8
(Continued)

CR30DLNQ	number of 30 day delinquencies within 12 months
AVGRVBAL	dollar amount of average revolving balance
AVBALINC	average revolving balance divided by average monthly income
Market data	
BUYPOWER	buying power index
PCTCOLL	percent college graduates in 5 digit zip code
MEDAGE	median age in 5 digit zip code
MEDINC	median income in 5 digit zip code
PCTOWN	percent who own their own home
PCTBLACK	percent black
PCTSPAN	percent Spanish
GROWTH	population growth rate
PCTEMPL	1987 employment percent
Commerce within 5 digit zip code	
APPAREL	apparel stores percent of retail sales in 5 digit zip code of residence
AUTO	auto dealer stores, percent
BUILDMTL	building material stores, percent
DEPTSTOR	department stores, percent
DRUGSTOR	drug stores, percent
EATDRINK	eating and drinking establishments, percent
FURN	furniture stores, percent
GAS	gas stations, percent

Table 9
Descriptive statistics for variables

Variable	Mean	Std. dev.	Minimum	Maximum	Case
CARDHLDR	0.78094000	0.41362	0.0	1.000	13444
DEFAULT	0.09486600	0.29304	0.0	1.000	10499
EXP1	268.26000000	542.39000	0.0	24650.000	10499
EXP2	252.60000000	537.20000	0.0	24030.000	10499
EXP3	238.89000000	460.30000	0.0	7965.000	10499
EXP4	247.32000000	507.61000	0.0	14240.000	10499
EXP5	253.24000000	504.53000	0.0	17870.000	10499
EXP6	266.46000000	509.99000	0.0	10310.000	10499
EXP7	256.41000000	500.52000	0.0	9772.000	10499
EXP8	248.62	494.10000	0.0	9390.000	10499
EXP9	245.06000000	472.39000	0.0	8377.000	10499
EXP10	228.60000000	441.28000	0.0	6926.000	10499
EXP11	273.66000000	520.60000	0.0	16820.000	10499
EXP12	233.26000000	458.15000	0.0	18970.000	10499
ADDLINC [a]	0.41262000	0.91279	0.0	10.000	13444
BANKSAV	0.03369500	0.18045	0.0	1.000	13444
BANKCH	0.29753000	0.45719	0.0	1.000	13444
BANKBOTH	0.66877000	0.47067	0.0	1.000	13444
AGE	33.47200000	10.22600	0.0	88.670	13444
MTHCURAD	55.31900000	63.09000	0.0	576.000	13444
CRDBRINQ	1.40800000	2.28910	0.0	56.000	13444

Table 9
(Continued)

Variable	Mean	Std. dev.	Minimum	Maximum	Case
CREDMAJR	0.81308000	0.38986	0.0	1.000	13444
DEPNDNTS	1.01730000	1.27910	0.0	9.000	13444
MTHMPOLY	60.64800000	72.24000	0.0	600.000	13444
PROF	0.11537000	0.31948	0.0	1.000	13444
UNEMP	0.00052068	0.02281	0.0	1.000	13444
MGT	0.07430800	0.26228	0.0	1.000	13444
MILITARY	0.02246400	0.14819	0.0	1.000	13444
CLERIICAL	0.08814300	0.28351	0.0	1.000	13444
SALES	0.07832500	0.26869	0.0	1.000	13444
OTHERJOB	0.62087000	0.48519	0.0	1.000	13444
MAJORDRG	0.46281000	1.43270	0.0	22.000	13444
MINORDRG	0.29054000	0.76762	0.0	11.000	13444
OWNRENT	0.45597000	0.49808	0.0	1.000	13444
MTHPRVAD	81.28500000	80.35900	0.0	600.000	13444
PREVIOUS	0.07334100	0.26071	0.0	1.000	13444
INCOME [a]	3.42410000	1.77750	0.1300	20.000	13444
SELFEMPL	0.05794400	0.23365	0.0000	1.000	13444
TRADACCT	6.42200000	6.10690	0.0000	50.000	13444
INCPER [a]	2.17200000	1.35910	0.0363	15.000	13444
EXP_INC	0.07097400	0.10392	0.0001	2.038	13444
CREDOPEN	6.05520000	5.24050	0.0000	43.000	13444
CREDACTV	2.27220000	2.61370	0.0000	27.000	13444
CRDDEL30	0.05556400	0.26153	0.0000	3.000	13444
CR30DLNQ	0.36581000	1.24940	0.0000	21.000	13444
AVGRVBAL	5.28050000	7.59040	0.0000	190.000	13444
AVBALINC	46.57000000	42.72800	0.0000	2523.000	13444
BUYPOWER	0.01396300	0.00909	0.0000	0.113	13444
PCTCOLL	10.72900000	8.51040	0.0000	54.900	13444
MEDAGE	33.18100000	5.42320	0.0000	65.000	13444
MEDINC [a]	2.83410000	1.04370	0.0000	7.500	13444
PCTOWN	53.98300000	28.54900	0.0000	100.000	13444
PCTBLACK	11.77700000	20.55700	0.0000	100.000	13444
PCTSPAN	7.78170000	13.18600	0.0000	96.600	13444
GROWTH [b]	0.00224620	0.00188	−0.0617	0.707	13444
PCTEMPL	40.99300000	108.01000	0.0000	5126.000	13444
APPAREL	2.43980000	2.43120	0.0000	33.300	13444
AUTO	1.49720000	1.32350	0.0000	33.300	13444
BUILDMTL	1.12930000	1.23350	0.0000	33.300	13444
DEPTSTOR	0.15870000	0.25209	0.0000	12.500	13444
EATDRINK	6.66570000	3.95700	0.0000	100.000	13444
FURN	1.84600000	2.51650	0.0000	100.000	13444
GAS	1.76540000	1.79580	0.0000	100.000	13444

[a] Income, Addlinc, Incper, and Medinc are in $10 000 units and are censored at 10.

[b] Population growth is growth/population.

received in a single month in 1988. The observation for an individual consists of the application data, data from a credit-reporting agency, market descriptive data for the 5 digit zip code in which the individual resides, and, for those applications that were accepted, a 12 month history of expenditures and a default indicator for the 12 month period following initial acceptance of the application. Default is defined as having skipped payment for 6 months. A full summary of the data appears in Tables 8 and 9.

References

Maddala, G., 1983. Limited Dependent and Qualitative Variables in Econometrics, Cambridge University Press, New York.

Greene, W., 1992. Credit scoring, Working Paper No. EC-92-29, Department of Economics, Stern School of Business, New York University.

Greene, W., 1994. Accounting for excess zeros and sample selection in poisson and negative binomial regression models, Department of Economics, Stern School of Business, New York University, Working Paper No. EC-94-10.

Greene, W., 1997. Econometric Analysis, 3rd ed., Prentice Hall, Englewood Cliffs, New Jersey.

Greene, W., 1995c. Sample selection in the poisson regression model, Department of Economics, Stern School of Business, New York University, Working Paper No. EC-95-6.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica, 47, 153–161.

Poirier, D., 1980. Partial observability in bivariate probit models. Journal of Econometrics, 12, 209–217.

Abowd, J., Farber, H., 1982. Job queues and union status of workers. Industrial and labor Relations Review, 35, 354–367.

Lerman, R., Manski, C., 1981. On the use of simulated frequencies to approximate choice probabilities. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge.

Murphy, K., Topel, R., 1985. Estimation and inference in two step econometric models. Journal of Business and Economic Statistics, 3, 370–379.

Terza, J., 1995. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects, Department of Economics. Penn State University, Working Paper No. 4-94-14, forthcoming, Journal of Econometrics.