A Statistical Model for Credit Scoring

Article · November 1992

DOI: 10.1017/CBO9780511754197.002 · Source: RePEc

1 author:

William H Greene
New York University
238 PUBLICATIONS **43,352** CITATIONS

A Statistical Model for Credit Scoring

William H. Greene

Department of Economics
Stern School of Business
New York University
100 Trinity Place
New York, NY 10006

April 8, 1992

Abstract. We derive a model for consumer loan default and credit card expenditure. The default model is based on statistical models for discrete choice, in contrast to the usual procedure of linear discriminant analysis. The model is then extended to incorporate the default probability in a model of expected profit. The technique is applied to a large sample of applications and expenditure from a major credit card company. The nature of the data mandates the use of models of sample selection for estimation. The empirical model for expected profit produces an optimal acceptance rate for card applications which is far higher than the observed rate used by the credit card vendor based on the discriminant analysis.

I am grateful to Terry Seaks for valuable comments on an earlier draft of this paper and to Jingbin Cao for his able research assistance. The provider of the data and support for this project has requested anonymity, so I must thank them as such. Their help and support are gratefully acknowledged. Participants in the applied econometrics workshop at New York University also provided useful commentary.

1. Introduction

Prediction of loan default has an obvious practical utility. Indeed, the identification of default risk appears to be of paramount interest to issuers of credit cards. In this study, we will argue that default risk is overemphasized in the assessment of credit card applications. In an empirical application, we find that a model that incorporates the expected profit from issuance of a credit card in the approval decision leads to a substantially higher acceptance rate than is present in the observed data and, by implication, acceptance of a greater average level of default risk.

A major credit card vendor must evaluate tens or even hundreds of thousands of credit card applications each year. These obviously cannot all be subjected to the scrutiny of a loan committee in the way that, say, a real estate loan might. Thus, statistical methods and automated procedures are essential. Banks and credit card issuers typically use `credit scoring models.' In principle, the credit score could incorporate any amount of relevant business information. In practice, credit scoring for credit card applications appears to be focused fairly narrowly on default risk and on a rather small set of attributes.[1] This study will develop an integrated statistical model for evaluating a credit card application which incorporates both default risk and the anticipated profit from the loan in the calculation. The model is then estimated using a large sample of applications and followup expenditure and default data for a major credit card company. The models are based on standard techniques for discrete choice and linear regression, but the data present two serious complications. First, observed data on default and expenditure used to fit the predictive models are subjected to a form of censoring that mandates the use of models of sample selection. Second, our sample used to analyze the approval decision is systematically different from the population from which it was drawn. This nonrepresentative nature of the data is remedied through the use of choice based sampling corrections.

[1]We say `appears to be' because the actual procedures used by credit scoring agencies are not public information nor, in fact, are they even necessarily known by the banks that use them. The small amount of information that we have was provided to us in conversation by the supporters of this study. We will return to this issue below.

Boyes, et. al. (1989) examined credit card applications and account performance using data similar to ours and a model that, with minor reinterpretation, is the same as one of the components of our model. They and we reach several similar conclusions. However, in one of the central issues in this study, we differ sharply. Since the studies are so closely related, we will compare their findings to ours at several points.

The paper is organized as follows: Section 2 will present models which have been used or proposed for assessing probabilities of loan default. Section 3 will describe an extension of the model. Here, we will suggest a framework for using the loan default equation in a model of cost and projected revenue to predict the profit or loss from the decision to accept a credit card application. The full model is sketched here and completed in Section 5. Sections 4 and 5 will present an application of the technique. The data and some statistical procedures for handling its distinctive characteristics are presented in Section 4. The empirical results are given in Section 5. Conclusions are drawn in Section 6.

2. Models for Prediction of Default

Individual i, with vector of attributes $\mathbf{x}_i$, applies for a loan at time 0. The attributes include such items as: personal characteristics including age, sex, number of dependents, and education; economic attributes such as income, employment status and home ownership; a credit history including the number of previous defaults, and so on. Let the random variable y_i indicate whether individual i has defaulted on a loan (y_i=1) or has not (y_i=0) during the time which has elapsed from the application until y_i is is observed. We consider two familiar frameworks for predicting default. The technique of discriminant analysis is considered first. We will not make use of this technique in this study. But, one of the observed outcome variables in the data that we will examine, the approval decision, *was* generated by the use of this technique. So it useful to enumerate its characteristics. We then consider a probit model for discrete choice as an alternative.

2.1. 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 normal distribution

of the attributes, **x**. An individual with attribute vector $\mathbf{x}_i$, is drawn from one of the two populations, and it is needed to determine which. The analysis is carried out by assigning to the application a `*Z*' score, computed as

$$Z_i = b_0 + \mathbf{b}\mathbf{x}_i. \tag{2.1}$$

Given a sample of previous observations on y_i and $\mathbf{x}_i$, the vector of weights, $\mathbf{b} = (b_0, \mathbf{b}_1)$, can be obtained as a multiple of the vector of regression coefficients in the linear regression of $d_i = P_0 y_i - P_1(1-y_i)$ 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 $(n-2)/\mathbf{e}'\mathbf{e}$ from the linear regression.[2] The individual is classified in group 1 if their `*Z*' score is greater than *Z* (usually 0) and 0 otherwise.[3] The linearity (and simplicity) of the computation is a compelling virtue.

The assumption of multivariate normality is often held up as the most serious shortcoming of this technique.[4] This seems exaggerated. Techniques which rely on normality are often surprisingly robust to violations of the assumption, recent discussion notwithstanding.[5] The superiority of the discrete choice techniques discussed in the next section, which are arguably more appropriate for this exercise, is typically fairly modest.[6] Since the left hand side variable in the aforementioned linear regression is a linear function of y_i, $d_i = y_i - P_1$, the *calculated*[7] discriminant function can be construed as nothing more (or less) than a linear probability model.[8] As such, the comparison between discriminant analysis and, say, the probit model could be reduced to one between the linear

[2]See Maddala (1983, pp. 18-25).

[3]We forego full details on the technique since we shall not be applying it to our data nor will we be comparing it to the other methods to be described.

[4]See Press and Wilson (1978), for example.

[5]See Greene (1983), Goldberger (1983), and Manski (1989).

[6]See, for example, Press and Wilson (1978).

[7]We emphasize `calculated' because there is no underlying counterpart to the probability model in the discriminant function.

[8]For a detailed and very readable discussion, see Dhrymes (1974, pp. 67-77).

probability model and the probit or logit model.[9] Thus, it is no surprise that the differences between them are not great; this has been observed elsewhere.[10]

Its long track record notwithstanding, one could argue that the underpinning of discriminant analysis is naive. The technique 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 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 attendant 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.

2.2 Discrete Choice Models

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, $\mathbf{x}_i$, 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 model.[11] The core of this formulation is an index function model with a latent regression,

$$D = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i. \tag{2.2}$$

The dependent variable might be identified with the `propensity to default.' In the present context,

[9]See Press and Wilson (1976) for discussion.

[10]See Aldrich and Nelson (1984) or Amemiya (1985), for example.

[11]Our discussion of this modelling framework will also be brief. Greater detail may be found in Greene (1993, chapter 21).

an intuitively appealing interpretation of D^* is as a quantitative measure of `how much trouble the individual is in.' Conditioning variables, $\mathbf{x}_i$, might include, income, credit history, the ratio of credit card burden to current income, and so on. If D is sufficiently large relative to the attributes, that is, if the individual is in trouble enough, they default. Formally,

so the probability of interest is

$$P_i = \text{Prob}[D_i = 1 \,|\, \mathbf{x}_i]. \tag{2.4}$$

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

$$\begin{aligned} \text{Prob}[D_i = 1 \,|\, \mathbf{x}_i] &= \text{Prob}[D > 0 \,|\, \mathbf{x}_i] \\ &= \text{Prob}[\varepsilon_i \le \boldsymbol{\beta}'\mathbf{x}_i \,|\, \mathbf{x}_i] \\ &= \Phi(\boldsymbol{\beta}'\mathbf{x}_i), \end{aligned} \tag{2.5}$$

where $\Phi(\cdot)$ is the standard normal CDF.[12] The classification rule is

$$\text{Predict } D_i = 1 \text{ if } \Phi(\boldsymbol{\beta}'\mathbf{x}_i) > P^*, \tag{2.6}$$

where P^* is a threshold value chosen by the analyst. The value 0.5 is usually used for P^* under the reasoning that we should predict default if the model predicts that it is more likely than not. For our purposes, this turns out to be an especially poor predictor. Indeed, in applications such as this one, with unbalanced data sets (that is, with a small proportion of ones or zeros for the dependent variable) this familiar rule may fail to perform as well as the naive rule `always (or never) predict $D = 1$.'[13] We will return to the issue in detail below, since it is crucial in our analysis. The vector of marginal effects in the model is

$$\boldsymbol{\theta} = \frac{\partial \text{Prob}[D_i = 1 \,|\, \mathbf{x}_i]}{\partial \mathbf{x}_i} = \varphi(\boldsymbol{\beta}'\mathbf{x}_i)\boldsymbol{\beta}, \tag{2.7}$$

[12]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.

[13]For discussion, see Amemiya (1981).

where $\varphi(\cdot)$ is the standard normal density.[14] If the discriminant score function can be viewed as a `model' (rather than as merely the solution to an optimization problem), the coefficients b_k would be the counterparts. The usefulness of θ is in determining which particular factors would contribute most to a rejection of a credit application. An example is given in Section 5.7.

2.3. Censoring in the Default Data

Regardless of how the default model is formulated, in practice, it must be constructed using data on loan recipients. But, the model is to be applied to a broader population, some (possibly even most) of whom are applicants who will ultimately be rejected. The underlying logic of the credit scoring problem is to ascertain how much an applicant resembles individuals who have defaulted in the past. The problem with this approach is that mere resemblance to past defaulters may give a misleading indication of the individual default probability for an individual who has not already been screened.

The model is to be used to assign a default probability to a random individual who *applies* for a loan, but the only information that exists about default probabilities comes from previous loan *recipients*. The relevant question for this analysis is whether, in the population at large, Prob[D=1$\,|\,$**x**] equals Prob[D=1$\,|\,$**x** *and* C=1] in the subpopulation, where `$C$ = 1' denotes having received the loan, or, in our case, `card recipient.' Since loan recipients have passed a prior screen based, one would assume, on an assessment of default probability, Prob[D=1$\,|\,$**x**] must exceed Prob[D=1$\,|\,$**x**,C=1] for the same **x**. For a given set of attributes, **x**, individuals in the group with C = 1 are, by nature of the prior selection, less likely to default than otherwise similar individuals chosen randomly from a population that is a mixture of individuals who will have C = 0 and C = 1. Thus, the unconditional model will give a downward biased estimate of the default probability for an individual selected at random from the full population. This describes a form of censoring. To be applicable to the population at large, the estimated default model should condition specifically on cardholder status.

[14]While the coefficients in logit and probit models often differ markedly, estimates of θ in the two models tend to be similar, indeed, often nearly identical. [See Greene (1993) and Davidson and Mackinnon (1993, Chapter 15).]

We will use a bivariate probit specification to model this. The structural equations are

Default equation: $D = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i$ (2.8)

$D_i = 1$ if and only if $D > 0$, and 0 else.

Cardholder equation: $C = \boldsymbol{\gamma}'\mathbf{v}_i + w_i$ (2.9)

$C_i = 1$ if and only if $C > 0$, and 0 else.

Sampling rule: D_i and $\mathbf{x}_i$ are only observed if $C_i = 1$ (2.10)

C_i and $\mathbf{v}_i$ are observed for all applicants.

Selectivity: $[\varepsilon_i, w_i] \sim N_2[0,0,1,1,\rho_{\varepsilon w}]$ (2.11)

The vector of attributes, $\mathbf{v}_i$, are the factors used in the approval decision. The probability of interest is the probability of default given that a loan application is accepted, which is

$$\text{Prob}[D_i=1 \mid C_i=1] = \frac{\Phi_2[\boldsymbol{\beta}'\mathbf{x}_i, \boldsymbol{\gamma}'\mathbf{v}_i, \rho]}{\Phi[\boldsymbol{\gamma}'\mathbf{v}_i]}, \tag{2.12}$$

where Φ_2 is the bivariate normal cumulative probability. If ρ equals 0, the selection is of no consequence, and the unconditional model described earlier is appropriate.

The counterparts to the marginal effects noted earlier are

$$\frac{\partial \Phi_2(\boldsymbol{\beta}'\mathbf{x}_i, \boldsymbol{\gamma}'\mathbf{v}_i, \rho)/\Phi(\boldsymbol{\gamma}'\mathbf{v}_i)}{\partial \mathbf{x}_i} = \boldsymbol{\theta}_{|C_i=1}. \tag{2.13}$$

The detailed expression for this derivative is given in Section 5. This model was developed by Wynand and Bernard (1981) and recently applied to an analysis of consumer loans by Boyes, et. al. (1989).[15]

[15]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

3. A Model for Evaluating an Application

Expenditure of a credit card recipient might be described by a linear regression model:

$$S_i = \boldsymbol{\alpha}'\mathbf{z}_i + u_i. \tag{3.1}$$

Expenditure data are drawn *conditionally* on C_i=1. Thus, with the cardholder data, we are able to estimate only

$$E[S_i \,|\, \mathbf{z}_i, C_i=1] = \boldsymbol{\alpha}'\mathbf{z}_i + E[u_i \,|\, C_i=1, \mathbf{z}_i]. \tag{3.2}$$

This may or may not differ systematically from

$$E[S_i \,|\, \mathbf{z}_i] = \boldsymbol{\alpha}'\mathbf{z}_i. \tag{3.3}$$

The statistical question is whether the sample selection into cardholder status is significantly related to the expenditure level of the individuals sampled. The equations of the sample selection model [see Heckman (1979)] used here are

Expenditure $$S_i = \boldsymbol{\alpha}'\mathbf{z}_i + u_i. \tag{3.4}$$

Cardholder Status $$C = \boldsymbol{\gamma}'\mathbf{v}_i + w_i \tag{3.5}$$

$$C_i = 1 \text{ if and only if } C > 0, \text{ and 0 otherwise.}$$

Sample Selectivity $$[u_i, w_i] \sim N_2[0,0,\sigma,1,\rho_{uw}\sigma_u]. \tag{3.6}$$

Selectivity Corrected Regression

$$\begin{aligned} E[S_i \,|\, C_i=1] &= \boldsymbol{\alpha}'\mathbf{z}_i + \mathrm{E}[u_i \,|\, C_i=1] \\ &= \boldsymbol{\alpha}'\mathbf{z}_i + (-\rho_{uw}\sigma_u)\lambda_i \\ &= \boldsymbol{\alpha}'\mathbf{z}_i + \alpha_\lambda\lambda_i, \end{aligned} \tag{3.7}$$

where $\lambda_i = \varphi(\boldsymbol{\gamma}'\mathbf{v}_i) / \Phi(\boldsymbol{\gamma}'\mathbf{v}_i)$.

Estimation techniques are discussed in Section 5.

Finally, it seems likely that even controlling for other factors, the probability of default is related to expenditures. The extension to (2.12) that we will examine is

$$- \qquad \Phi_2[\boldsymbol{\beta}'\mathbf{x}_i + \delta S_i, \boldsymbol{\gamma}'\mathbf{v}_i, \rho]$$

---

ours [see (5.1)] are the same.

$$\text{Prob}[D_i = 1 \mid C_i=1,\mathbf{x}_i,S_i] = \frac{\Phi(\boldsymbol{\gamma}'\mathbf{v}_i)}{\Phi(\boldsymbol{\gamma}'\mathbf{v}_i)} \tag{3.8}$$

where

$$\bar{S}_i = E[S_i \mid C_i=1].$$

Expenditure, like the default probability, is only an intermediate step. Ultimately, the expected profitability of a decision to accept a loan application is a function of the default probability, the expected expenditure, and the costs associated with administering the loan. Let

$$P_D = \text{Prob}[D_i=1 \mid C_i=1].$$

Then

$$\begin{aligned} E[\Pi(\mathbf{x}_i,\mathbf{v}_i,\mathbf{z}_i) \mid C_i=1] &= E[S_i \mid C_i=1]m && \text{(merchant fee)} \\ &+ E[S_i \mid C_i=1](1 - P_D)(f-t) && \text{(finance charge - t bill rate)} \\ &- E[S_i \mid C_i=1]P_D[1 - r(1 + q)] && \text{(losses from default)} \\ &+ \text{fixed fees paid by cardholder} \\ &- \text{overhead expenses for the account.} \end{aligned} \tag{3.9}$$

The merchant fee, m, is collected whether or not the consumer defaults on their loan. This term would also include any float which is accrued before the merchant is reimbursed. The second term gives the finance charges from the consumer, which are received only if default does not occur. The third term includes the direct loss of the defaulted loan minus any ultimate recovery. The term denoted `r' is the recovery rate and 'q' is the penalty assessed on recovered funds.

This is a simple model which involves spending, costs, and the default probability. Obviously, there are elements missing. Finance charges paid by the cardholder are the most complicated element. Specific treatment would requires a subsidiary model of timing of repayment and how the consumer would manage a revolving charge account.[16] For the present, we assume that the finance charge component, if any, is simply included in the term `f' in (3.9). Variations of this value could be used to model different repayment schedules. The model estimated later is for a monthly expenditure, so the applicable figure could range from 0 to 1.5 percent depending on what is assumed about the repayment schedule. The figure is then net of the opportunity cost of the funds, based, for example, on the return on a treasury bill. Admittedly, the model is crude. It is important to emphasize that the preceding model applies to purchases, not to revolving loans. That is, the consumer might well make their purchases, then take years to repay the loan, each month making a minimum payment. The preceding is much simpler than that; it is a single period model which assumes that all transactions occur, either full repayment or default, within the one year period of

[16]Of course, if the finance charges, themselves, were influential in the default rate, this would also have to be considered. This seems unlikely, but either way, this complication is beyond the scope of this study. Our data contain no information about finance charges incurred or paid. We have only the expenditure levels and the default indicator.

observation. Nonetheless, even in this simple formulation, a clear pattern emerges. Based on observed data and the description of the cost structure, consideration of the censoring problem and use of an integrated model produce a prescription for considerably higher acceptance rates for loan applicants than are seen in our observed data.

4. Data Used in the Application

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 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 twelve month history of expenditures and a default indicator for the twelve month period following initial acceptance of the application. Default is defined as having skipped payment for six months. A full summary of the data appears in Tables 1 and 2.

Table 1. Variables 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.
DEPNDNTS = 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.
SALES = 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 = income+aadlinc.
AVGEXP = $(1/12)\Sigma_i$**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 acounts).
MTHMPLOY = months employed.

Table 1. (Continued)

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 delinquincies) 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.
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 2 Descriptive Statistics for Variables

Variable	Mean	Std. Dev.	Minimum	Maximum	Cases
CARDHLDR	.78094	.41362	0.0	1.000	13444
DEFAULT	.094866	.29304	0.0	1.000	10499
DB1	268.26	542.39	0.0	24650	10499
DB2	252.60	537.20	0.0	24030	10499
DB3	238.89	460.30	0.0	7965	10499
DB4	247.32	507.61	0.0	14240	10499
DB5	253.24	504.53	0.0	17870	10499
DB6	266.46	509.99	0.0	10310	10499
DB7	256.41	500.52	0.0	9772	10499
DB8	248.62	494.10	0.0	9390	10499
DB9	245.06	472.36	0.0	8377	10499
DB10	228.60	441.28	0.0	6926	10499
DB11	273.66	520.60	0.0	16820	10499
DB12	233.26	458.15	0.0	18970	10499
ADDLINC[1]	.41262	.91279	0.0	10.000	13444
BANKSAV	.033695	.18045	0.0	1.000	13444
BANKCH	.29753	.45719	0.0	1.000	13444
BANKBOTH	.66877	.47067	0.0	1.000	13444
AGE	33.472	10.226	0.0	88.67	13444
MTHCURAD	55.319	63.090	0.0	576.0	13444
CRDBRINQ	1.4080	2.2891	0.0	56.00	13444
CREDMAJR	.81308	.38986	0.0	1.000	13444
DEPNDNTS	1.0173	1.2791	0.0	9.000	13444
MTHMPLOY	60.648	72.240	0.0	600.0	13444
PROF	.11537	.31948	0.0	1.000	13444
UNEMP	.00052068	.022813	0.0	1.000	13444
MGT	.074308	.26228	0.0	1.000	13444
MILITARY	.022464	.14819	0.0	1.000	13444
CLERICAL	.088143	.28351	0.0	1.000	13444
SALES	.078325	.26869	0.0	1.000	13444
OTHERJOB	.62087	.48519	0.0	1.000	13444
MAJORDRG	.46281	1.4327	0.0	22.00	13444
MINORDRG	.29054	.76762	0.0	11.00	13444
OWNRENT	.45597	.49808	0.0	1.000	13444
MTHPRVAD	81.285	80.359	0.0	600.0	13444
PREVIOUS	.073341	.26071	0.0	1.000	13444
INCOME[1]	3.4241	1.7775	0.1300	20.00	13444
SELFEMPL	.057944	.23365	0.0	1.000	13444
TRADACCT	6.4220	6.1069	0.0	50.00	13444
INCPER[1]	2.1720	1.3591	0.03625	15.00	13444
EXP_INC	.070974	.10392	0.00009	2.038	13444
CREDOPEN	6.0552	5.2405	0.0	43.00	13444
CREDACTV	2.2722	2.6137	0.0	27.00	13444
CRDDEL30	.055564	.26153	0.0	3.000	13444
CR30DLNQ	.36581	1.2494	0.0	21.00	13444
AVGRVBAL	5.2805	7.5904	0.0	190.0	13444
AVBALINC	46.570	42.728	0.0	2523.	13444
BUYPOWER	.013963	.0090948	0.0	.1134	13444
PCTCOLL	10.729	8.5104	0.0	54.90	13444
MEDAGE	33.181	5.4232	0.0	65.00	13444

MEDINC[1]	2.8341	1.0437	0.0	7.500	13444
PCTOWN	53.983	28.549	0.0	100.0	13444
PCTBLACK	11.777	20.557	0.0	100.0	13444
PCTSPAN	7.7817	13.186	0.0	96.60	13444
GROWTH[2]	.0022462	.001877	-0.06172	.7068	13444
PCTEMPL	40.993	108.01	0.0	5126.	13444

Table 2. Descriptive Statistics (Continued)

Variable	Mean	Std. Dev.	Minimum	Maximum	Cases
APPAREL	2.4398	2.4312	0.0	33.30	13444
AUTO	1.4972	1.3235	0.0	33.30	13444
BUILDMTL	1.1293	1.2335	0.0	33.30	13444
DEPTSTOR	.15870	.25209	0.0	12.50	13444
EATDRINK	6.6657	3.9570	0.0	100.0	13444
FURN	1.8646	2.5164	0.0	100.0	13444
GAS	1.7654	1.7958	0.0	100.0	13444

[1]Income, Addlinc, Incper,and Medinc are in $10,000 units and are censored at 10.
[2]Population growth is growth/population.

4.1. The Choice Based Sampling Problem

The incidence of default among our sample of cardholders mimics reasonably closely the incidence of default among cardholders in the population. But, the proportion of cardholders in the sample is, by design, considerably larger than the proportion of applications that are accepted. That is, the rejection rate for applications in the population is much higher than our sample suggests. The sampling is said to be `choice based' if the proportional representation of certain outcomes of the dependent variable in the model is deliberately different from the proportional representation of those outcomes in the population from which the sample is drawn. In our sample, 10499 of 13444 observations are cardholders, a proportion of .78094. But, in the population, the proportion of card applications which are accepted is closer to 23.2%. In view of the fact that we are using `Cardholder' as a selection rule for the default equation, the sample is `choice based.' This is a type of nonrandom sampling that has been widely documented in other contexts, and has been modelled in a counterpart to the study by Boyes, et. al. (1989).

Choice based sampling induces a bias in the estimation of discrete choice models. As has been shown by Manski and Lerman (1977) possible to mitigate the induced bias if one knows the

true proportions that should apply in the sampling. These are listed in Table 3

Table 3. Sampling Weights for Choice Based Sampling

Event	w=sample	W=Population	$\Omega=W/w$
D=1, C=1	996/13444	.232 × .103	.32255
D=0, C=1	9503/13444	.232 × .897	.29441
C=0	2945/13444	.768	3.50594

The `Weighted Endogenous Sampling MLE' or `WESML' estimator is obtained by maximizing where the subscript `*i*' indicates the *i*th individual. There are *J* possible outcomes, indexed by `*j*,' the indicator I_{ij} equals 1 if outcome or choice *j* is occurs for or is chosen by individual *i*, P_{ij} is the theoretical probability that individual *i* makes choice *j*, Ω_j is the sampling weight,

$$\Omega_j = W_j/w_j \tag{4.2}$$

and W_j = the 'true' or population proportion of occurrences of outcome *j* (4.3)

w_j = the sample counterpart to W_j.

(See Table 3.) Note that in our application, this would give smaller weight to cardholders in the sample and larger weight to rejects than would the unweighted log-likelihood.

After estimation, an adjustment must be made to the estimated asymptotic covariance matrix of the estimates in order to account for the weighting. The appropriate asymptotic covariance matrix is

$$\mathbf{V} = \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}, \tag{4.4}$$

where **B** = the Berndt et. al. (1974) estimator

H = inverse of the estimated expected Hessian of the log-likelihood. Both matrices in the expression are computed using the sampling weights given above.

5. Empirical Results

5.1. Cardholder Status

Table 4 presents univariate probit estimates of the cardholder equation both with and without the correction for choice based sampling. We also show the results of applying the familiar prediction rule. The effect of the reweighting is quite clear in these tables. As might be expected,

with the choice based sampling correction, the predictions are more in line with the population proportions than with the distorted sample.

Table 4. Weighted and Unweighted Probit Cardholder Equations

	Choice based sampling		Unweighted	
Variable	Coefficient	t-ratio	Coefficient	t-ratio
ONE	-1.1175	-9.090	0.1070	1.390
AGE	-0.0021	-0.806	-0.0012	-0.672
MTHCURAD	0.0010	2.547	0.0011	3.943
DEPNDNTS	-0.0947	-2.623	-0.0957	-4.079
MTHMPLOY	-0.0002	-0.410	-0.0002	-0.694
MAJORDRG	-0.7514	-13.922	-0.7796	-34.777
MINORDRG	-0.0609	-1.554	-0.0471	-2.005
OWNRENT	0.0514	0.947	-0.0042	-0.119
MTHPRVAD	0.0002	0.626	0.0001	0.767
PREVIOUS	0.1781	1.843	0.2089	2.967
INCOME	0.1153	4.353	0.1362	7.001
SELFEMPL	-0.3652	-3.711	-0.3634	-5.804
TRADACCT	0.0995	19.447	0.1099	25.573
INCPER	-0.0167	-0.476	-0.0007	-0.027
CREDOPEN	-0.0276	-3.550	-0.0227	-4.194
CREDACTV	0.0443	2.825	0.0341	3.074
CRDDEL30	-0.2720	-2.658	-0.2740	-4.776
CR30DLNQ	-0.0947	-3.773	-0.0891	-6.732
AVGRVBAL	0.0095	2.949	0.0094	3.560
AVBALINC	-0.0019	-1.616	-0.0010	-2.573
BANKSAV	-0.5018	-4.012	-0.5233	-7.305
BANKBOTH	0.4630	9.579	0.4751	14.692
CRDBRINQ	-0.1559	-13.907	-0.1719	-23.469
CREDMAJR	0.3033	5.407	0.3092	8.652

	Predicted				Predicted		
Actual	0	1	TOTAL	Actual	0	1	TOTAL
0	.208	.011	**2945**	0	.110	.109	**2945**
1	.420	.361	**10499**	1	.020	.761	**10499**
TOTAL	**8448**	**4996**	**13444**	**TOTAL**	**1748**	**11696**	**13444**

The cardholder equation is largely consistent with expectations. The most significant explanatory variables are the number of major derogatory reports and credit bureau inquiries (negative) and the number of open trade accounts (positive). What Table 7 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 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 at 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 could not be used in a commercial credit scoring model.

5.2. Expenditure

The expenditure equation is estimated using Heckman's sample selection correction and adjustment for the estimated standard errors of the coefficients. The selection mechanism is the univariate probit model for cardholder status. The equations of the model are given in (3.4) to (3.7). Details on the estimation method may be found in Heckman (1979) and Greene (1981, 1993). Parameter estimates and estimated asymptotic standard errors are given in Table 5. Note that the dependent variable in this equation is average monthly expenditure, computed as the simple average of the 12 months beginning with when the credit card was issued.

As might be expected, INCOME is the single most significant explanatory variable in the expenditure equation. The market variables which appear to be very significant are puzzling. Three, PCTOWN, PCTBLACK, and PCTSPAN, given their relationship to average income, would seem to have the wrong sign. But, since MEDINC is already in the equation, as well as the individual income, one must conclude that these variables are picking up some other effect.

The last variable in the equation is the selectivity correction described earlier. Its large t-statistic suggests that the sample selection correction is, indeed, warranted. The coefficient on LAMBDA estimates $-\rho_{uw}\sigma_u$. An estimate of σ_u is given at the top of the results, 319.68, so the implied estimate of ρ_{uw} is -.204. The negative value is surprising given the criteria that are probably used to determine cardholder status. But, since, INCOME, OWNRENT, etc. are already in the equation, it is unclear just what sign should have been expected.

Table 6 displays the average predicted expenditures for three groups of observations. The predicted expenditure is substantially higher for those whose applications were denied.

5.3. Default Probability

Table 7 gives the probit estimates of the default equation. Predicted expenditure, FITEXP, is computed using (3.7). The `selection' variable, λ_i, is computed using the leftmost coefficients in Table 4. The coefficients used in computing the linear function in (3.7) are given in Table 5. The single equation, unconditional model is given in the first three columns. The results agree with our conjecture that default rates might be related to expenditures, the idea of cardholders "getting in over their heads" comes to mind. Table 8 presents the full information, conditional estimates of the default equation based on (2.8) to (2.11) and (4.2) to (4.4) with the reestimated cardholder equation. Estimates of the cardholder equation are given in Table 8.

Table 5. Estimated Expenditure Equation

Dependent Variable = AVGEXP in $ per month
Observations = 10499
Mean of LHS = 251.03
StdDev of residuals = 315.60
Corrected Std. error = 319.68
(This is a consistent estimate of σ_u)
R-squared = 0.0977
Adjusted R-squared = 0.0952
Correlation of disturbance in regression and selection equation = -0.204

Variable	Coefficient	Std. Error	t-ratio
Constant	-44.249	160.270	-0.276
AGE	-1.487	0.34655	-4.291
DEPNDNTS	-2.0829	2.79774	-0.744
OWNRENT	-1.9733	7.71648	-0.256
INCOME	55.0379	2.05561	26.774
SELFEMPL	-33.4684	14.3173	-2.338
TRADACCT	1.5301	0.63709	2.402
PROF	71.8808	157.985	0.455
MGT	60.3144	158.096	0.382
MILITARY	9.0472	159.241	0.057
CLERICAL	25.8032	158.121	0.163
SALES	112.145	158.118	0.709
OTHERJOB	53.4139	157.770	0.339
BUYPOWER	375.513	380.930	0.986
PCTCOLL	1.7967	0.46231	3.886
MEDAGE	-0.0889	0.61771	-0.144
MEDINC	14.3057	3.95810	3.614
PCTOWN	-0.5333	0.13336	-3.999
PCTBLACK	0.5094	0.17949	2.838
PCTSPAN	0.6271	0.25991	2.413
GROWTH	0.00564	0.015846	0.356
PCTEMPL	-0.01769	0.033207	-0.533
APPAREL	0.78475	1.49578	0.525
AUTO	-4.89992	2.56277	1.912
BUILDMTL	1.48865	2.63996	0.564
DEPTSTOR	-6.61155	13.9866	-0.473
EATDRINK	-1.24421	0.82499	-1.508
FURN	0.97996	1.15843	0.846
GAS	-1.77288	1.99177	-0.890
LAMBDA	65.4875	8.52960	7.678

Table 6. Average Predicted Expenditures

All Observations	$263.29
Cardholders	$251.03
Noncardholders	$307.03

Table 7. Default Model

	Unconditional			Conditional			
Variable	Coeff.	Std.Err.	t-ratio	Coeff.	Std.Err.	t-ratio	Partial
Basic Default Specification							
Constant	-1.1350	0.0984	-11.533	-1.3752	0.3945	-3.486	———
AGE	-0.0031	0.0023	-1.342	-0.0054	0.0094	-0.582	-.0018
MTHCURAD	0.0003	0.0003	1.069	0.0002	0.0013	0.153	-.0001
DEPNDNTS	0.0445	0.0294	1.512	-0.0217	0.1114	-0.195	.0073
MTHMPLOY	0.0007	0.0003	2.331	0.0007	0.0013	0.566	.0002
MAJORDRG	0.0592	0.0408	1.448	-0.2969	0.1985	-1.495	.0033
MINORDRG	0.0764	0.0296	2.586	0.1780	0.0993	1.793	.0488
OWNRENT	-0.0010	0.4312	-0.023	0.0908	0.1706	0.533	.0236
MTHPRVAD	0.0004	0.0002	1.817	0.0002	0.0009	0.274	.00002
PREVIOUS	-0.1507	0.0792	-1.902	-0.1112	0.3103	-0.358	-.0434
INCOME	-0.0168	0.0033	-5.608	-0.0072	0.0151	-0.476	.0062
SELFEMPL	0.0788	0.0850	0.927	-0.1969	0.3565	-0.552	-0017
TRADACCT	0.0004	0.0044	0.109	0.0207	0.0205	1.009	-.0028
INCPER	-0.0228	0.0323	-0.706	-0.0545	0.1058	-0.515	-.0094
EXP_INC	-0.4761	0.1717	-2.774	-0.5790	0.5033	-1.150	-.1614
Credit Bureau							
CREDOPEN	0.0138	0.0063	2.195	0.0199	0.0272	0.732	.0066
CREDACTV	-0.1218	0.0126	-9.657	-0.1500	0.0557	-2.695	-.0424
CRDDEL30	0.2841	0.0712	3.991	0.2829	0.2766	1.023	.1120
CR30DLNQ	0.0806	0.0177	4.559	0.0446	0.0757	0.589	.0225
AVGRVBAL	0.0011	0.0024	0.439	0.0156	0.0123	1.268	.0038
AVBALINC	0.0039	0.00042	9.192	0.0008	0.0021	0.398	.0004
Expenditure							
FITEXP	0.0014	0.0044	3.103	0.00064	0.0019	0.336	

Table 8. Estimated Cardholder Equation
Joint with Default Equation

	Coeff.	Std Error	t-ratio
Basic Cardholder Specification			
Constant	-1.2734	0.1563	-8.150
AGE	-0.00002	0.0039	-0.006
MTHCURAD	0.0015	0.0006	2.465
DEPNDNTS	-0.1314	0.0487	-2.700
MTHMPLOY	0.0003	0.0006	0.491
MAJORDRG	-0.8230	0.0442	-18.634
MINORDRG	0.0082	0.0462	0.178
OWNRENT	0.0129	0.0765	0.168
MTHPRVAD	0.0003	0.0004	0.698
PREVIOUS	0.1185	0.1283	0.924
INCOME	0.0156	0.0040	3.867
SELFEMPL	-0.5651	0.1307	-4.325
TRADACCT	0.0850	0.0064	13.352
INCPER	-0.0550	0.0513	-1.073
Credit Bureau			
CREDOPEN	-0.0096	0.0109	-0.876
CREDACTV	0.0060	0.0223	0.270
CRDDEL30	-0.3167	0.1197	-2.647
CR30DLNQ	-0.0965	0.0317	-3.048
AVGRVBAL	0.0049	0.0050	0.974
AVBALINC	-0.0014	0.0008	-1.906
Credit Reference			
BANKSAV	-0.4708	0.1731	-2.719
BANKBOTH	0.5074	0.0694	7.310
CRDBRINQ	-0.1473	0.0176	-8.393
CREDMAJR	0.3663	0.0807	4.541
Correlation Between Disturbances			
$\rho_{w\varepsilon}$	0.4478	0.2580	1.736

Maximum likelihood estimates for the conditional model are obtained by maximizing

$$\log\text{-}L = \Sigma_{C_i=0}\,\Omega_i \log(\text{Prob}[C_i=0]) + \Sigma_{C_i=1,D_i=0}\,\Omega_i \log(\text{Prob}[D_i=0 \mid C_i=1]\text{Prob}[C_i=1])$$

$$+ \Sigma_{C_i=1,D_i=1}\,\Omega_i \log(\text{Prob}[D_i=1 \mid C_i=1]\text{Prob}[C_i=1])$$

$$= \Sigma_{C_i=0}\,\Omega_i \log(1-\Phi(\boldsymbol{\gamma}'\mathbf{v}_i)) + \Sigma_{C_i=1,D_i=0}\,\Omega_i \log\Phi_2[-(\boldsymbol{\beta}'\mathbf{x}_i+\delta\bar{S}_i),\boldsymbol{\gamma}'\mathbf{v}_i,-\rho]$$

$$+ \Sigma_{C_i=1,D_i=1}\,\Omega_i \log\Phi_2(\boldsymbol{\beta}'\mathbf{x}_i+\delta\bar{S}_i,\boldsymbol{\gamma}'\mathbf{v}_i,\rho).^{17}$$

Optimization and construction of the asymptotic covariance matrix for the estimates can be based on the following results: Let $\Phi_2(d,c,\rho)$ and $\varphi_2(d,c,\rho)$ denote the cdf and density, respectively, of the bivariate normal distribution. Then,

$$\begin{aligned}
\partial\Phi_2/\partial c &= \varphi(c)\Phi[(d-\rho c)/(1-\rho^2)^{1/2}] = g_c,\\
\partial\Phi_2/\partial\rho &= \varphi_2,\\
\partial^2\Phi_2/\partial c^2 &= -cg_c - \rho\varphi_2 - g/\Phi_2,\\
\partial^2\Phi_2/\partial c\partial d &= \varphi_2 - g_c g_d/\Phi_2, \qquad (5.1)\\
\partial^2\Phi_2/\partial c\partial\rho &= \varphi_2([\rho/(1-\rho^2)^{1/2}](d-\rho c) - c - g_c/\Phi_2),\\
\partial^2\Phi_2/\partial\rho^2 &= \varphi_2\{[\rho/(1-\rho^2)](1 - (c^2+d^2-2\rho cd)/(1-\rho^2)) + \rho cd - \varphi_2/\Phi_2\}.
\end{aligned}$$

Terms that are symmetric in c and d are omitted.

Partial effects in the single equation model are obtained by multiplying the coefficients by $\partial\Phi(d)/\partial d = \varphi(d)$, which is roughly .13 for these data. By this calculation, the most important behavioral variables in the equation appear to be MAJORDRG (.0077), MINORDRG (.0099), CRDDEL30C (.0369), and CR30DLNQ (.0104). These are counts, so the marginal effects are obtained directly. Note, in particular, the number of trade lines past due at the time of the application. An increase of one in this variable alone would be sufficient to raise the estimated

[17]This is the same log likelihood as maximized by Boyes, et. al. (1989). The second term in their formulation would be $\log[\Phi(d)-\Phi_2(d,c,\rho)]$, but this equals $\log[\Phi_2(-d,c,-\rho)]$, so the two are the same.

default probability from an acceptable level (say .095) to well beyond the threshold (roughly .11). CPTF30, the number of 30 day delinquencies, is similarly influential. The marginal effects in the conditional probability, account for the selection equation. Let the joint probability be denoted

$$\text{Prob}[D=1,C=1] = \Phi_2[d,c,\rho], \tag{5.2}$$

where $$d = \boldsymbol{\beta}'\mathbf{x}_i + \delta[\boldsymbol{\alpha}'\mathbf{z} + \alpha_\lambda\varphi(\boldsymbol{\gamma}'\mathbf{v})/\Phi(\boldsymbol{\gamma}'\mathbf{v})] \tag{5.3}$$

and $$c = \boldsymbol{\gamma}'\mathbf{v}.$$

[See (3.7) and (3.8). Note that the term in square brackets is expected expenditure given cardholder status.] Let **w** denote the union of the variables in **x** [see (2.6)], **v** [see (2.9)], and **z** [see (3.1)]. Then, reconfigure $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha}$ conformably, with zeros in the locations where variables do not actually appear in the original equation. Thus,

$$\frac{\partial \text{Prob}[D=1 \mid C=1,S]}{\partial \mathbf{w}} = \frac{1}{\Phi(c)}\left[g_d\frac{\partial d}{\partial \mathbf{w}} + g_c\frac{\partial c}{\partial \mathbf{w}}\right] - \frac{\Phi_2(d,c,\rho)\varphi(c)}{(\Phi(c))^2}\frac{\partial c}{\partial \mathbf{w}}. \tag{5.4}$$

The outer derivatives, g_d and g_c were defined earlier. The inner derivatives are

$$\partial c/\partial \mathbf{w} = \boldsymbol{\gamma} \tag{5.5}$$

and

$$\partial d/\partial \boldsymbol{\omega} = \boldsymbol{\beta} + \delta[\boldsymbol{\alpha} - \alpha_\lambda\lambda(c+\lambda)\boldsymbol{\gamma}]. \tag{5.6}$$

Inserting the sample means of the variables where required for the computations gives an estimate of approximately +0.0033. The rightmost column in Table 7, labelled 'Partial,' gives a complete set of estimates of the marginal effects for the conditional default equation. It is clear that the coefficients, themselves, are misleading. In particular, the apparent effect of MAJORDRG turns out to be an effect of selection; increases in this variable appear to decrease default only because increases so heavily (negatively) influence the approval decision.

5.4. Predicted Default Probabilities

Table 9 shows the average of the predicted default probabilities computed with the models in Tables 7 and 8 for some subgroups of the data set.

Table 9. Estimated Default Probabilities

Group	Conditional	Unconditional
All observations	.1498	.1187
Cardholders	.1056	.0947
Non-cardholders	.3090	.2061
Defaulters	.1632	.1437
Nondefaulters	.0997	.0895

The standard predictive rule, `predict $y_i = 1$ if $\hat{P}_i > .5$' predicts only 11 defaults, 6 of them incorrectly, in the sample of 10,499 observations which includes 996 defaults. Obviously, this is not likely to be useful. The problem is that the sample is extremely unbalanced, with only 10 percent of the observations defaulting. Since the average predicted probability in the sample will equal the sample proportion, it will take an extreme observation to produce a probability as high as .5. Table 10 shows the effect with three alternative choices of the threshold value. The value .09487 is the sample proportion.

Table 10. Predictions for Different Thresholds

	Predict D=0			Predict D=1			
	.09487	.12	.15	.09487	.12	.15	
Actual							Total
0	5225	6464	7675	4278	3039	1828	9503
1	214	329	494	782	667	502	996
Total	5439	6793	8169	5060	3706	2330	10499

5.5. Expected Profit

The final step in this part of the analysis is to construct the equation for expected profit from

approving an application. The basis of the model is equation (3.9). We used the following specific formulation:

m	= 2% + 10%/52	(merchant fee)	
f	= 1.25%	(finance charges)	
t	= 1%	(opportunity cost of funds)	
r	= 50%	(recovery rate)	(5.7)
q	= 2%	(penalty rate)	
fee	= $5.25	(fee for card(s))	
o	= .2%	(overhead rate on loans)	

This assumes a 2.00 percent merchant fee, 1.25 percent finance charge, plus one week's float on repayment and an interest rate of 10 percent. The net return on finance charges is only 3% per year, but the merchant fees are quite substantial. We assume a 50 percent ultimate recovery rate on defaulted loans and a 2 percent penalty rate. As before, we acknowledge the simplicity of the preceding. Nonetheless, it captures most of the important aspects of the calculation. Based on the estimated expenditure equation and conditional default model, Table 11 lists the sample averages for E[Π] for several subgroups.

Table 11. Sample Average Expected Profits

All Observations	-$4.41
Cardholders	$4.27
Defaulters	-$3.15
Nondefaulters	$5.06
Noncardholders	-$35.32

The values in Table 11 are striking. It is clear that the results are being driven by the default probability. Figure 1 is a scatter plot of the estimated profits against the predicted default probability for 3,000 randomly chosen observations from the full sample.

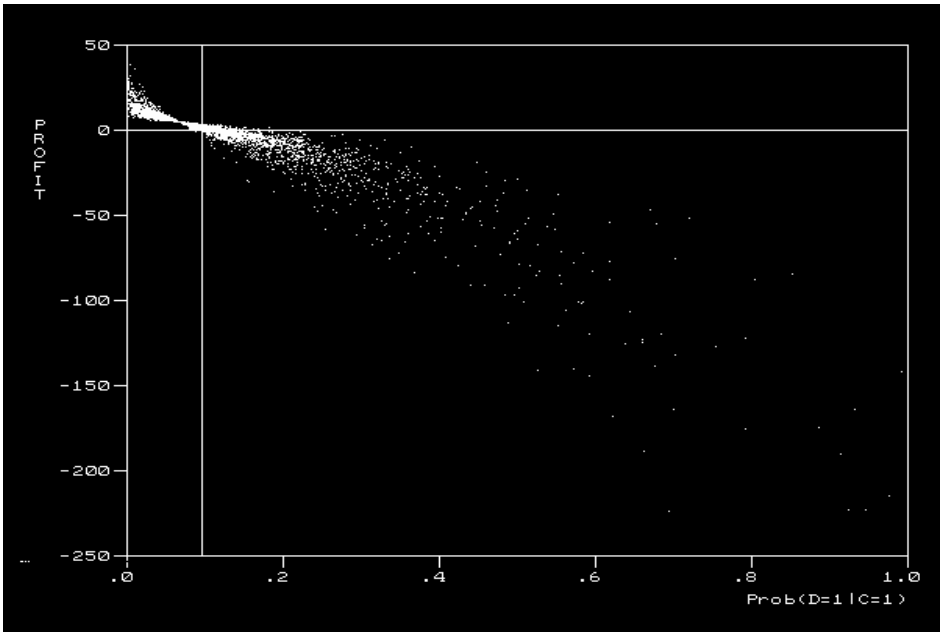


Figure 1. Expected Profit vs. Default Probability

The vertical line in the figure is drawn at the sample average default rate of slightly under 10 percent. The horizontal line is drawn at zero. The figure clearly shows that the model predicts negative profits for most individuals whose estimated default probability exceeds roughly ten percent. The familiar rule of 0.5 for the threshold for predicting default is obviously far too high to be effective in this setting.

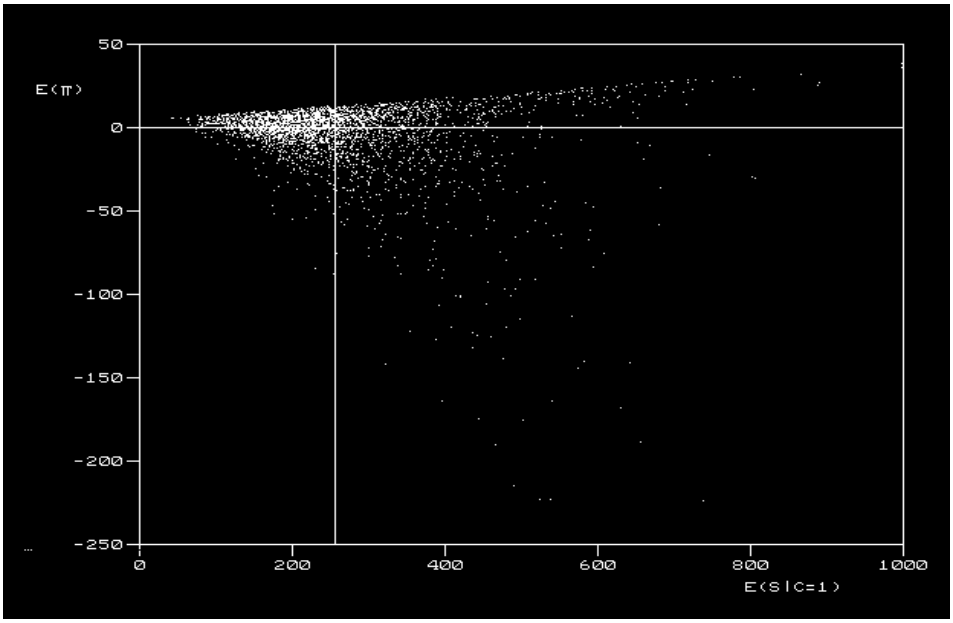


Figure 2. Expected Profit vs. Expected Expenditure

Figure 1 agrees strongly with Boyes, et. al.'s finding that applicants whose default

probability exceeded 9 percent were generally associated with negative profits. We find exactly the same result. But, they suggest at several points that higher balances are likely to be associated with higher expected earnings. Our results strongly suggest the opposite. Figure 2 shows the expected profits plotted against expected expenditure for the same 3,000 observations used to produce Figure 1. Clearly, beyond a surprisingly modest expenditure level, higher expenditures are generally associated with lower, not higher profits. Our own results are easily explained. The expenditure level strongly influences the default probability in our model, and the profit equation is, in turn, heavily dependent on the default probability. The result is explored further in the next section.

5.6. Aggregate Decision Rules for Approving or Denying Credit

Consider a pool of applicants within which default probabilities will be widely distributed. For each individual in the pool, we can compute an expected profit, as in the preceding section, which will depend on both predicted default rate and predicted expenditure. The expected profit of a decision rule can then be obtained by summing the expected profits of those in the pool who are accepted by this rule. An equivalent procedure is to compute the `normalized expected profit,'

$$E^*[\Pi] = E_{P^*}[\{E[\Pi_i] \,|\, P^*\} \times AR(P^*)] \tag{5.7}$$

where $AR(P^*)$ is the acceptance rate associated with a particular threshold probability. Obviously, $AR(P^*)$ increases monotonically with P^*. However, $E[\Pi_i] \,|\, P^*$ falls with P^*. Because the acceptance rate is falling with P^*, the profit that will be obtained from a given pool need not rise with falling P^*. In short, a rule which decreases P_* attracts fewer and fewer better and better loans. Thus, the total, average loans times number of loans, may not rise.

In order to estimate the function in (5.7), we use the following steps: Compute for every individual in the pool (1) probability of acceptance, Prob[C_i=1] = $\Phi[\boldsymbol{\gamma}'\mathbf{v}_i]$, (note that this is only for purposes of dealing with our censoring problem; it is not part of the structure of the model) (2) expected expenditure from (3.7), (3) probability of default from (3.8), and (4) expected profit from (5.7). For different values of P^*, we compute the average value of $E^*[\Pi_i]$ for those individuals whose

estimated default probability is less than P^*. We then multiply this sample mean by the acceptance rate. Table 12 gives the result of this calculation. The last column shows that by this calculation, there is an optimal acceptance rate. Figures 3 and 4 show the relationship between acceptance rate and normalized expected profit.

Table 12. Normalized Expected Profits

P^*	Acceptance Rate	Sample Mean $E^*[\Pi_i]\,\vert\,P^*$	Normalized Profit
0.00000	0.00000	0.00000	0.00000
0.00500	0.00885	21.89900	0.19384
0.01000	0.02581	20.29800	0.52391
0.02000	0.07461	17.41600	1.29933
0.03000	0.13292	15.54800	2.06667
0.04000	0.19154	14.19900	2.71961
0.05000	0.25082	13.12700	3.29249
0.06000	0.30861	12.22200	3.77187
0.07000	0.36180	11.45900	4.14583
0.08000	0.40970	10.79700	4.42353
0.09000	0.45425	10.19100	4.62931
0.10000	0.49636	9.62100	4.77543
0.11000	0.53689	9.07600	4.87285
0.11500	0.55437	8.83700	4.89900
0.12000	0.57200	8.59900	4.91865
0.12500	0.58710	8.38800	4.92460
0.13000	0.60257	8.17170	4.92405
0.13500	0.61871	7.94200	4.91383
0.14000	0.63262	7.74310	4.89850
0.15000	0.66096	7.32700	4.84288
0.16000	0.68826	6.91500	4.75933
0.17000	0.71259	6.52260	4.64791
0.18000	0.73408	6.18000	4.53663
0.19000	0.75268	5.85800	4.40919
0.20000	0.76986	5.42200	4.17418

Table 12 suggests that a rule $P^* = .125$, or an acceptance rate of about 59% is optimal. This is a rule that allows a fairly high default rate, in exchange for higher expected profits. It also accepts some individuals with negative expected profits, since the default rate is not, alone, sufficient to insure positive expected profit. This acceptance rate is noticeably higher than the value actually observed,

which was roughly 25 percent during the period in which these data were drawn.

Figure 3. Profits vs. Default Probability

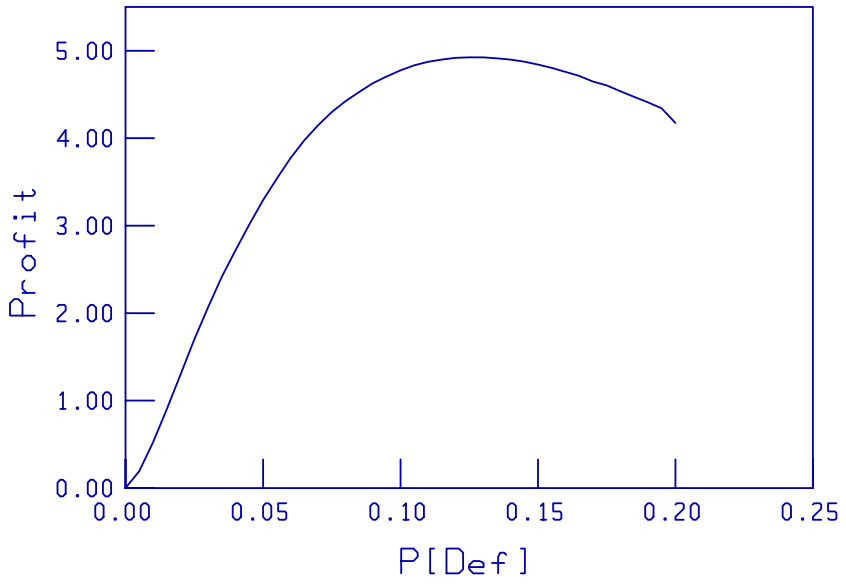
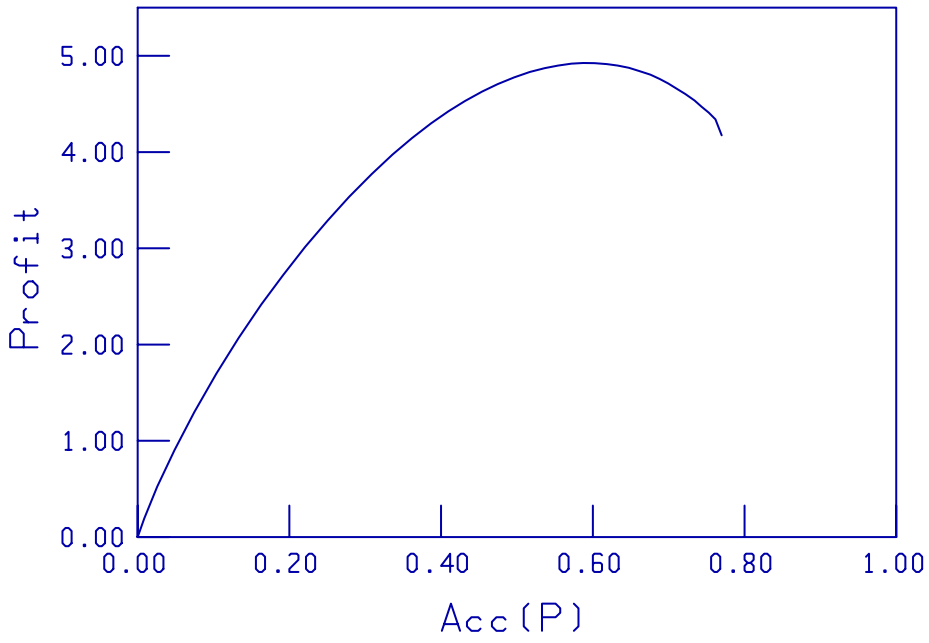


Figure 4. Profits vs. Acceptance Rate



5.7. Ranking Attributes Which Contribute to a Denial of Credit

Denote by R^* the criterion, or `rule' that has been used for the decision whether to approve or deny an application and by $R(\mathbf{w}_i)$ the value of the criterion for a particular individual `i,' where $\mathbf{w}$ is the full vector of attributes and characteristics used in the calculation. In order to establish which factor contributed to an individual's failure to meet the benchmark, we need to determine the values of the factors which are consistent with meeting it. We can do so by sampling individuals which meet the benchmark and empirically determining sample means. We will do so by obtaining for a set of individuals, *all of whom are at or close to the benchmark*, the sample means of the attributes. This estimates $E[\mathbf{w} \mid P=P^*]$. Denote the set of sample means $\bar{\mathbf{w}}^*$. If the sample is large enough (by which we would surmise a few thousand observations), then it will be the case that $R^* \approx R(\bar{\mathbf{w}}^*)$. Now, approximate the rule function evaluated at the particular $\mathbf{w}_i$ with a linear Taylor series, expanding around the point of means that we have obtained;

$$R(\mathbf{w}_i) - R^* \approx \Sigma_k [\partial R(\bar{\mathbf{w}}^*)/\partial \bar{w}](w_{ik} - \bar{w})$$
$$= \Sigma_k \psi_k (w_{ik} - \bar{w}).$$

Thus, the deviation of the individual's `score' from the benchmark is expressed as a linear function of the deviations of their attributes from the benchmark attributes. If the decision rule is the default probability, then the elements of $\boldsymbol{\psi}$ are the marginal effects in (5.3). Some of the numeric values are given in the last column of Table 7. If the expected profit is used, the calculation is only slightly more difficult. By combining terms, the expected profit may be written as

$$E[\Pi] = \pi_0 + E[S](\pi_1 + \text{Prob}[D=1 \mid C=1]),$$

so the extension to this function would be straightforward using results already given.

We will use the default probability for an illustration. For the example, we take as a cutoff our earlier described optimal default probability rule of $R^* = P^* = .125$. Using the model presented in the previous sections, observation number 4805 in our sample has a predicted default probability of .165, so they would be rejected. (They were.) In order to obtain the means for the calculation, we

use observations which have predicted default rates between .115 and .135. (With more data we could use a narrower range.) This leaves about 800 observations from the original 13,444. The set of calculations listed above produces a default probability at the means of roughly $R(\bar{\mathbf{w}}^*) = .116$. The sample mean predicted default probability for these 800 observations is $\bar{R} = .127$. (Recall, we have attempted to match .125, so this is quite close.) The difference between the computed default probability and the benchmark is .165-.125 = 0.040. The decomposition obtained as the sum of the terms gives a value of .0414. The difference of -.0014 would be the remainder term in the Taylor series approximation. The largest single term is associated with CPTF30, the 30 day delinquency count in the last 12 months. The average in the sample for this variable is .242. This individual had 4. The second largest contributor was the number of credit bureau inquiries, for which, once again, this individual (4) was well above the mean (1.2358).

6. Conclusions

The preceding has described a methodology for incorporating costs and expected profits into a credit scoring model for loan approvals. Our main conclusion is the same as Boyes, et. al.'s (1989). When expected return is included in the credit scoring rule, the lender will approve applications that would otherwise be rejected by a rule that focuses solely on default probability. Contrary to what intuition might suggest, we find that when spending levels are included as a component of the default probability, which seems quite plausible, the optimal loan size is relatively small.

The model used for profit in this study is rudimentary. More detailed data on payment schedules would allow a more elaborate behavioral model of the consumer's repayment decisions. Nonetheless, it seems reasonable to expect similar patterns to emerge in more detailed studies. Since, in spite of our earlier discussion, we continue to find that default probability is a crucial determinant of the results, it seems that the greatest payoff in terms of model development would be found here. For example, with better and finer data, it would be possible to examine the timing of

default, rather than simply its occurrence. The relationship between default probability and account size could also be further refined. Finally, our objective function for the lender, expected profit, is quite simple. The preceding is best viewed as merely a simulation. A more elaborate model which makes use of the variation in expenditures from month to month or used the second moment of the distribution of profits might more reasonably characterize the lender's objectives.

Much of the modelling done here is purely illustrative. The equations are somewhat unwieldy. Credit scoring vendors would still be required to manipulate the models with convenience, which would make a more critical specification search necessary. The obvious use of models such as ours is for processing initial applications, which can, in principle, be done at a leisurely pace. But, an equally common application is the in store approval for large purchases. For relatively small purchases this has been automated, and focuses simply on whether the account is already in arrears. But, for very large purchases which often require human intervention, credit card companies often rely on a decidedly ad hoc procedure, the gut reaction of an individual based on a short telephone call. A simple enough behavioral model which incorporates up to date information and behavioral characteristics might be of use in this situation.

References

Abowd, J. and Farber, H., "Job Queues and the Union Status of Workers," *Industrial and Labor Relations Review*, 35, 1982, pp. 354-367.

Aldrich, J. and F. Nelson, *Linear Probability, Logit, and Probit Models*, Beverly Hills: Sage, 1984.

Altman, E., "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance*, 1986, pp. 589-609.

Amemiya, T., *Advanced Econometrics*, Cambridge: Harvard University Press, 1985.

Berndt, E., B. Hall, R. Hall, and J. Hausman, "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 3/4, 1974, pp. 1263-1278.

Boyes, W., D. Hoffman, and S. Low, "An Econometric Analysis of the Bank Credit Scoring Problem," *Journal of Econometrics*, 40, 1989, pp. 3-14.

Davidson, R., and J. Mackinnon, *Estimation and Inference in Econometrics*, New York: Oxford University Press, 1993.

Dhrymes, P., *Econometrics: Statistical Foundations and Applications*, New York: Springer Verlag, 1974.

Doukas, J., "Bankers versus Bankruptcy Prediction Models: An Empirical Investigation, 1979-1982," *Applied Economics*, 1986, pp. 479-493.

Goldberger, A., "Abnormal Selection Bias," in S. Karlin, T. Amemiya, and L. Goodman, eds., *Studies in Econometrics, Time Series, and Multivariate Statistics*, Stamford: Academic Press, 1982.

Greene, W., "Sample Selection as a Specification Error: Comment," *Econometrica*, 49, 1981, pp. 795-798.

Greene, W., *LIMDEP, Version 6.0, Reference Guide*, Bellport, New York: Econometric Software, 1991.

Greene, W., *Econometric Analysis*, 2nd. ed., New York: Macmillan, 1993.

Heckman, J., "Sample Selection as a Specification Error," *Econometrica*, 47, 1979, pp. 153- 161.

Jones, C., and I. Swary, "An Analysis of Risk and Return Characteristics of Corporate Bankruptcy Using CapitalMarket Data," *Journal of Finance*, 1980, pp. 1001-1016.

Maddala, G., *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Makeever, D., "Predicting Business Failures," *Journal of Commercial Bank Lending*, 1984, pp. 14-18.

Manski, C., "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 1989, pp. 341-360.

Manski, C. and S. Lerman, "The Estimation of Choice Probabilities for Choice Based Samples," *Econometrica*, 45, 1977, pp. 1977-1988.

Meng, C. and Schmidt, P., "On the Cost or Partial Observability in the Bivariate Probit Model," *International Economic Review*, 26, 1985, pp. 71-85.

Poirier, D., "Partial Observability in Bivariate Probit Models," *Journal of Econometrics*, 12, 1980, pp. 210-217.

Press, J., and S. Wilson, "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association*, 73, pp. 699-705.

Wynand, P. and M. Bernard, "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection," *Journal of Econometrics*, 17, 1981, pp. 229-252.