NBER WORKING PAPER SERIES

AN EMPIRICAL STUDY OF THE CREDIT MARKET WITH UNOBSERVED CONSUMER TYPERS

Li Gan
Roberto Mosquera

Working Paper 13873
http://www.nber.org/papers/w13873

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2008

We would like to thank participants of the Texas Econometrics Camp 2008, Brit Grosskopf, Liqun Liu, Rajiv Sarin and Guoqiang Tian for their helpful comments. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

An Empirical Study of the Credit Market with Unobserved Consumer Typers
Li Gan and Roberto Mosquera
NBER Working Paper No. 13873
March 2008
JEL No. C81,D12

ABSTRACT

This paper proposes an econometric model to identify unobserved consumer types in the credit market. Consumers choose different amounts of loan because of differences in their time or risk preferences (types). Thus, the unconditional probability of default is modeled using a mixture density combining a type-conditioning default variable with a type-determining random variable. The model is estimated using individual-level consumer credit card information. The parameter estimates and statistical tests support this kind of specification. Furthermore, the model produces better out-of-sample predictions on the probability of default than traditional models; hence, it provides evidence of the existence of types in the consumer credit market.

Li Gan
Department of Economics
Texas A&M University
College Station, TX 77843-4228
and NBER
gan@econmail.tamu.edu

Roberto Mosquera
Department of Economics
Texas A&M University
College Station, TX 77843-4228
rmosquera@tamu.edu

1. Introduction

Several economic theories have long suggested the existence of different yet unobserved types of individuals. In this line of research, theorists focus on mechanisms that may be useful in distinguishing these types. For example, in the classical signaling game of the job market (Spence 1973), there are multiple types of workers, some of which are more productive than others. Since the types can not be observed, each worker's level of education is used as a signal to reveal his/her productivity. However, few empirical studies have either documented the existence of the unobserved types among individuals and/or the economic significance of identifying these types. To our knowledge, this is the first paper that provides an econometric method directly derived from a theoretical equilibrium model to empirically identify the unobserved individual types. Moreover, it shows that identifying the types results in significant economic benefits.

This paper presents two main contributions. First, the paper extends Milde and Riley's (1988) model of the consumer loan market, by assuming that consumers have different, unobserved types characterizing their time discount rates (or their risk preference). At equilibrium, consumers with different types will borrow different amounts of loans. As a result, the probability of default is a function of the parameter that characterizes the types (for example, time preference or risk aversion parameter). In other words, the unconditional probability of default is a function of the types, and it can be written as the sum of the conditional probabilities corresponding to each type. Therefore, ignoring the types would lead to biased estimates of the marginal effect of observed characteristics. Accordingly, this paper proposes a mixture density econometric model that can be used to estimate simultaneously the distribution of the types and the conditional probabilities of default.

Second, this paper estimates the empirical model using individual-level consumer credit information. We find that there are unobserved types in the population, and that having types in the model produces better out-of-sample predictions of the default probabilities. Hence, this paper shows that this econometric model, that takes into account the existence of unobserved types, improves upon traditional credit scoring models.

Using mixture densities to empirically model unobserved types is not new in the literature. For instance, Stahl and Wilson (1994) assume that individuals can be classified into different rational archetypes. Some choose strategies randomly, while others can figure out the best responses predicted by rational models. Stahl and Wilson (1994) use a mixture density model on a set of experimental data, and find evidence supporting the existence the types. On a distinct field, Feinstein (1990) proposes and estimates a mixture density model that considers simultaneously the unobserved violations and the observed detections of violations of laws and regulations. Helland (1998) applies Feinsten's method in studying the effect of pollution control laws. More recently, Knittel and Stango (2003) estimate a mixture density model using state-mandated price ceilings as focal points for unobserved tacit collusions of credit card companies.

The rest of this paper is organized as follows. Section 2 derives an equilibrium model of consumer credit cards with unobserved consumer types, and proposes an econometric method to identify the types. Section 3 applies the econometric model to individual-level data of a consumer credit market, and empirically tests the significance of the types previously defined. It also briefly discusses the application of this model to credit scoring, and shows how this model improves upon models that are only driven by statistical predictability. Section 4 concludes.

2. An Equilibrium Model of Consumer Credit Markets

This theoretical model is based on the work of Jaffee and Russell (1976), and of Milde and Riley (1988). These papers focus on type identification by characterizing the theoretical conditions for the existence of separating equilibrium credit bundles. Nonetheless, in this research we intend to describe the economic effects on consumer credit of the existence of different types in the population, and how these effects motivate an empirical approach to test the significance of these types.

In this model, there are two sets of participating agents: the applicants/borrowers and the issuer banks. We are interested in the potentially different types of borrowers in a population, as their distinct behavior leads to possibly diverse equilibrium actions. So we define $\boldsymbol{\Theta}$ as the set of possible types of borrowers. Here it is assumed that there are two different types, so $\boldsymbol{\Theta} = \{\theta_1, \theta_2\}$. The borrower applies for a credit card. In his application, he discloses personal information (sex, age, income, number of children, credit history, etc), that potentially reveals his type, and may be useful in determining the probability of default. The bank receives the application, and determines the appropriate combination of interest rate and loan amount for the borrower in order to maximize its profit. Bierman and Hausman (1970) indicate that accurate estimations of the probability of default lead to an effective credit policy, in the sense that the amount of credit granted will not constrain sales and gains, nor it will allow increased losses due to uncollectible accounts. Thus, the probability of default is a critical variable for maximizing the bank's profit. It is important to notice that in a credit card the loan amount really is a credit limit. The actual amount loaned may be lower than this limit. It depends on the preferences of the borrower. While the bank does not know to which type a particular borrower belongs, it may know that there is some probability that the borrower belongs to type θ_1. From the economic theory about signals (see Spence, 1974), this probability is a function of the information revealed by the borrower.

A. Indifference Curves for the Borrowers

Following the setup of Milde and Riley (1988), we formulate a simple two-period consumption model to characterize the behavior of the borrowers. Let the intertemporal utility function $U(c_1, c_2)$ be expressed as

$$U(c_1, c_2) = V(c_1) + \alpha_\theta c_2 \tag{1}$$

In this model $V(\cdot)$ is assumed to be a concave function. Now, a first difference with Milde and Riley's (1988) setup is the inclusion of time preference parameter α_θ. For this analysis, we assume[2] that there are two types of borrowers depending on their intertemporal preference for consumption: those who give the same weight to future consumption as present consumption, that is α_{θ_1} is normalized to be 1, and those who prefer present consumption, that is $0 \leq \alpha_{\theta_2} < 1$. This assumption can easily be relaxed to admit the existence of a continuum of types defined by $0 \leq \alpha_\theta \leq 1$. Loosely speaking, one can say that there is a "responsible" type of borrower in the sense the he gives relatively more weight to

[2] Actually this specification is robust to whatever characteristics the different types have; as long as those traits lead to different preference maps for each type. In the Appendix we solve for the case where the types represent different degrees of risk aversion. Since differences in both time and risk preferences can generate similar predictions, our empirical model cannot be used to distinguish the factors that define the unobserved types.

the future consumption, and an "irresponsible" type of borrower in the sense that he put relatively more weight to the present, without worrying too much about the future.

The budget constraint is expressed in terms of consumption in the two periods. The borrower uses amount L of his credit limit in the first period, hence his first-period consumption is $c_1 = y_1 + L$. Loan L must be repaid in the second period, so the second-period consumption is defined as

$$c_2 = \max\left[y_2 - (1+r)L, 0\right], \tag{2}$$

where the minimum consumption level for survival (or guaranteed by some law) is normalized to zero. The second-period income y_2 is random, but we do not make Milde and Riley's (1988) assumption that individuals differ in their expected second period income; hence, they define the types of borrowers as those with a high expected income, and those with a low expected income. In this research we assume that the randomness of second period income comes from exogenous shocks to the borrower, and thus it is independent of the borrower's type, α_θ.

Let D be a random variable describing whether a borrower defaults or not. $D = 1$ if the borrower defaults; otherwise, $D = 0$. From equation (2), $D = 1$ if $c_2 = 0$ or

$$D = 1 \text{ if } y_2 - (1+r)\, L \leq 0. \tag{3}$$

With this setup, for a given value of the utility function, say $U(c_1,c_2) = U$, one can use the implicit function theorem to determine the shape of the indifference curves in a given space. In this research we are interested in the space spawned by the loan amount L and the interest rate, expressed in terms of R=1+r. We restate equation (1) introducing the budget constraint:

$$\begin{aligned} U(c_1, c_2) &= U\left(y_1 + L, \max\left[y_2 - RL, 0\right]\right) \\ &= V(y_1 + L) + \alpha_\theta \int_{RL}^{\infty} (y_2 - RL)\, dF(y_2) \end{aligned} \tag{4}$$

Borrowers maximize this utility function to determine the optimal amount of loan L they demand. For a given interest rate R, the first order condition (FOC) for utility maximization is:

$$V'\left(y_1 + L^*_{\theta_s}\right) = \alpha_\theta R\left(1 - F\left(RL^*_{\theta_s}\right)\right), \tag{5}$$

where $L^*_{\theta_s}$ is the optimal loan for a given interest rate R.

To obtain the indifference curves in the space of *L* and *R*, we fix the value of the utility in (4). Applying the implicit function theorem to (4), we have:

$$[V'(y_1 + L) - \alpha_\theta R(1 - F(RL))]dL = \alpha_\theta L(1 - F(RL))dR \tag{6}$$

Rewriting (6), we have

$$\frac{dR}{dL} = \frac{V'(y_1 + L) - \alpha_\theta R(1 - F(RL))}{\alpha_\theta L(1 - F(RL))} \tag{7}$$

The numerator is the FOC of utility maximization of (4) for a given interest rate *R*. When $L = L^*_{\theta_s}$, the numerator of (7) is zero, we have that *dR*/*dL*=0, which means that the indifference curves' slope has a turning point at $L = L^*_{\theta_s}$. To determine the sign of the slope for $L \neq L^*_{\theta_s}$, let's consider the second derivative with respect to *L*:

$$\frac{dR^2}{dL^2} = \frac{V''(y_1 + L) - \alpha_\theta R^2 f(RL)}{\alpha_\theta L(1 - F(RL))} - \frac{dR}{dL} \cdot \frac{\alpha_\theta [1 - F(RL) - RLf(RL)]}{\alpha_\theta L(1 - F(RL))} \tag{8}$$

As stated before, *dR/dL*=0 if and only if $L = L^*_{\theta_s}$. This implies that at point $L^*_{\theta_s}$, $d^2R/dL^2<0$ which means that the indifference curve is at a maximum. Then, the concavity of $V(\cdot)$ allows one to conclude that

(a) For $L > L^*_{\theta_s}$, $V'(y_1 + L) < \alpha_\theta R(1 - F(RL)) \Rightarrow dR/dL < 0$; and,

(b) For $L < L^*_{\theta_s}$, $V'(y_1 + L) > \alpha_\theta R(1 - F(RL)) \Rightarrow dR/dL > 0$.

To evaluate the effect of α_θ we differentiate *dR/dL* with respect to α_θ:

$$\frac{d^2R}{dLd\alpha_\theta} = -\frac{V'(y_1 + L)}{\alpha_\theta^2 L(1 - F(RL))} < 0.$$

Both the numerator and the denominator are strictly positive for any *L*. Then, a higher value of α_θ implies a lower marginal willingness to accept increments in the interest rate. This means that the marginal willingness to accept an increase in the interest rate is greater for type θ_2 individuals, with higher preference for present consumption (lower α_θ), than for type θ_1 individuals. Graphically this relationship is showed in Figure 1.

B. Iso-Profit Curve of the Bank and Market Equilibria

Now, we turn our attention to the bank. Define i as a rate that captures the economic cost of lending (including opportunity costs). Since there is the possibility of default, the bank actually maximizes its expected profit. From equation (3), for a given level of R and L, the probability of default is $F(RL)$. The expected profit function is

$$
\begin{aligned}
E(\pi) &= \Pr(D = 0)\, RL - (1+i)\, L \\
&= (1 - F(RL))\, RL - (1+i)L \qquad (9)
\end{aligned}
$$

In a competitive market setup, the expected economic profit is bound to be zero. Thus, we derive the iso-profit curve for $E(\pi) = 0$. Similarly to the borrower's indifference curves, the iso-profit curve is characterized by[3]

$$
\begin{aligned}
\frac{dR}{dL} &= \frac{(1+i) - R + R\left[F(RL) + RLf(RL)\right]}{L\left[1 - F(RL) - RLf(RL)\right]} \\
&= \frac{R^2 f(RL)}{\left[1 - F(RL) - RLf(RL)\right]}
\end{aligned}
\qquad (10)
$$

The second equality is obtained by applying $E(\pi) = 0$ to the numerator, which is strictly positive. Now, it is easy to see that the numerator is strictly positive, so the sign of (10) depends only on the sign of the denominator; which in turn depends on the difference between the probability of payment, $1-F(RL)$, and $RLf(RL)$. Thus, the sign of the denominator depends on the characteristics of $F(\cdot)$, the distribution of second period income. If $F(\cdot)$ is an unimodal distribution, then dR/dL will be positive and increasing up to a turning point, $\overline{RL}$, where $dR/dL = \infty$. After this point, dR/dL will be negative, and the iso-profit curve turns back to the left. With this characterization of the behavior of both borrowers and banks, one can analyze the equilibrium in the consumer credit market *via* credit cards.

Figure 2 presents the equilibrium analysis for credit card loans. The bank defines a credit limit L_{CL}[4], and offers a continuum menu of loan amounts, such that $L^S \in [0, L_{CL}]$. The borrower uses the supplied credit up to point where one of his indifference curves is tangent to the zero iso-profit curve. As showed above, the shape of the indifference curves is affected by the type of the individual. Therefore, type θ_2 individuals, with a higher

[3] This result comes directly from applying the implicit function theorem to equation (9).
[4] The credit limit is defined by the total amount of funds available to loan, capital requirements and other regulations.

preference for present consumption (lower α_θ), have a higher equilibrium loan and interest rate than type θ_1 individuals. It is important to notice that for each type the other's equilibrium bundle lies in a lower utility indifference curve. In other words, neither type has an incentive to switch to the other's equilibrium bundle. Hence, we recover Milde and Riley (1988)'s result that one can identify the type of the individual after he chooses the amount of credit he wants to use. For two otherwise identical individuals, the one that uses the highest amount of credit belongs to type θ_2. The equilibrium loan amount for type θ_s, denoted as $L^e_{\theta_s}$, is typically different from the optimal amount of loan $L^*_{\theta_s}$ in (5), since $L^*_{\theta_s}$ is obtained for a given interest rate R, while $L^e_{\theta_s}$ is obtained that allowing R to be changeable.

Now, suppose that there is some legislation that fixes the interest rate, or that the bank does not charge differential interest rates based on the loan amount. The constant rate $\overline{R}$ has to be greater than $1+i$, otherwise the bank will not participate in the market. As showed in Figure 3, the same conclusions as in the unregulated case apply. The only difference is that the equilibrium amounts of loan will be the optimal amounts of loan that maximize the borrower's utility given $\overline{R}$ for each type.

C. Equilibrium Default Rates and the Empirical Model

In this setup, the probability of default depends on the type of each individual. We let α_θ be a random variable with a Bernoulli distribution such that $\Pr(\alpha_\theta=\theta_1) = p$ and $\Pr(\alpha_\theta=\theta_2) = 1-p$. When $\alpha_\theta=\theta_s$, the conditional probability of default is given by

$$\Pr(D=1 \mid \alpha_\theta = \theta_s) = \Pr\left(y_2 - R^e_{\theta_s} L^e_{\theta_s} \le 0\right) = F\left(R^e_{\theta_s} L^e_{\theta_s}\right), \qquad (11)$$

Under the case illustrated in Figure 2, both $L^e_{\theta_s}$ and $R^e_{\theta_s}$ will be functions of the time discount rate α_θ (i.e. the type of a person), first period income y_1, and second period income equation (2) through the cumulative distribution function $F(\cdot)$. In the case of a constant interest rate R, as illustrated in Figure 3, the optimal loan amount $L^e_{\theta_s}$ still is a function of α_θ, y_1, and $F(\cdot)$. Let X be a random vector of the observable characteristics of individuals that may affect probability of default. A simple way to approximate $R^e_{\theta_s} L^e_{\theta_s}$ is

$R^e_{\theta_s} L^e_{\theta_s} \approx \alpha_s y_1 + X\beta_s$. Therefore, $F\left(R^e_{\theta_s} L^e_{\theta_s}\right) \approx F\left(\alpha_s y_1 + X\beta_s\right)$ and the unconditional probability of default can be written as:

$$\begin{aligned}\Pr(D=1) &= \Pr(D=1, \alpha_\theta = \theta_1) + \Pr(D=1, \alpha_\theta = \theta_2) \\ &= \Pr(D=1 \mid \alpha_\theta = \theta_1)\Pr(\alpha_\theta = \theta_1) + \Pr(D=1 \mid \alpha_\theta = \theta_2)\Pr(\alpha_\theta = \theta_2) \\ &= pF\left(R^e_{\theta_1} L^e_{\theta_1} \mid \alpha_\theta = \theta_1\right) + (1-p)F\left(R^e_{\theta_2} L^e_{\theta_2} \mid \alpha_\theta = \theta_2\right) \\ &= pF(\alpha_1 y_1 + X\beta_1) + (1-p)F(\alpha_2 y_1 + X\beta_2).\end{aligned} \tag{12}$$

Similarly,

$$\Pr(D=0) = p(1 - F(\alpha_1 y_1 + X\beta_1)) + (1-p)(1 - F(\alpha_2 y_1 + X\beta_2)).$$

Empirically, we can let $\Pr(\alpha_\theta = \theta_1) = p$ be a function of a set of individual characteristics Z_i, i.e., $\Pr(\alpha_\theta = \theta_1) = G(Z_i\gamma)$. The likelihood function of the model is:

$$\begin{aligned}L = \sum_{i=1}^{n} D_i \ln\left[G(Z_i\gamma)F(\alpha_1 y_1 + X_i\beta_1) + (1 - G(Z_i\gamma))F(\alpha_2 y_1 + X_i\beta_2)\right] + \\ (1 - D_i)\ln\left[1 - G(Z_i\gamma)F(\alpha_1 y_1 + X_i\beta_1) - (1 - G(Z_i\gamma))F(\alpha_2 y + X_i\beta_2)\right]\end{aligned} \tag{13}$$

Since the model in (13) incorporates the unobserved types into the model, we refer this model as the *type-consistent* model.

Feinstein (1990) discusses the identification issues related to this type of models. He argues that these models are basically identified *via* the functional form. Especially, if the same continuous variables appear in both *Z* and *X*, the coefficients *β* and *γ* may not be identifiable. To avoid this problem, we include in *Z* only socio-demographic variables, such as gender, age, and marital status; while *X* only includes economic variables such as income, measures of wealth, and credit history. This distinction follows from signaling theory as personal characteristics may reveal the individual's type. Nevertheless, it is somewhat arbitrary which variables should be included in *Z* or *X*. In the next section, we estimate the model using individual-level credit information.

3. An Empirical Study of the Default Probability

A. The Data

This research uses the information of all main cardholders of a major credit card issuer bank from Ecuador as of February 2006. This data set comprises information of the age of the debt (number of days in default), the demographic characteristics of each individual, some economic variables, and the credit worthiness of each individual.

One cannot simply use the fact that if an individual is late in his payment to classify each person into two groups: default or not default. Being late in payment may simply indicate that the card holder has forgotten to pay. Thus, it is necessary to define how many days of not payment signify that the cardholder has opted to default. Ecuadorian law states that if a consumer credit account is in arrays for 90 or more days, then full provisions must be made and legal collection action must be taken. Therefore, if an individual is in default for 90 days or more, it is a clear sign that he has chosen to default. For this study, clients that not default will be those who are not in default, or those who are in default for less than 90 days (92.5% of the sample). Clients in default will be those in default for 90 or more days (7.5% of the sample).

The data set was randomly partitioned into a design sample (60% of the observations) used to estimate the model, and a test sample (40% of the observations) for further analysis. Both samples maintain the population proportions of default and not default actions. For the estimation of the model, Z_i will comprise variables that represent only individual socio-demographic characteristics. X_i will include income, variables proxy for the wealth of each individual, a set of indicator variables that represent if the individual's residence is in one of the nine largest cities of Ecuador, and credit worthiness indicators in the form of ratings provided by *"Superintendecia de Bancos"*, the government agency that oversees the operations of financial institutions in Ecuador. A description of these variables is provided in table 1.

The data for estimating default probabilities should come from a random sample of all historic applicants (accepted and rejected). However, it is often only possible to identify defaults and not defaults in the clients who were accepted. Therefore, usually there is a bias in the data generating process for these models that reflects the procedures used to accept/reject applicants. The estimators will echo this bias, and it is not possible to determine its direction (Capon, 1982). In this case we have that this particular bank has a very lenient applicant acceptance procedure. It basically relies on confirmation of the information presented in the application form (especially the reported income), and on requesting a guarantor for those cases where the assigned credit officer believes there is a risk. If the guarantor does not qualify, then it asks for another guarantor. Thus, for this bank the rejection rate is negligible, and any bias in the data set will be minimal.

B. Specification Test and the Estimation Results

For identification reason, we let *Z* be different from *X*. More specifically, *Z* is not a subset of *X*, nor is *X* is a subset of *Z*. Therefore, the type-consistent model does not nest the logit model. A specification test of the two models requires a non-nested procedure. Consider the following likelihood function:

$$L_{non-nested} = aL_{\log it} + (1-a)L_{type-consistent}\,, \tag{14}$$

where the likelihood L_{logit} represents the standard likelihood function of a logit model, and the likelihood $L_{type\text{-}consistent}$ is given in equation (13). The nesting parameter, $a \in [0,1]$, tests which model is correct. If $a = 0$, then the type-consistent model is supported; when $a = 1$, the logit model is supported. To estimate the type-consistent model, we approximate $G(\cdot)$ and $F(\cdot)$ in the likelihood function (13) with a logistic[5] CDF. Our estimation of (14) finds that $a = 0.0001$ with a large standard error. Therefore, this non-nested test supports the type-consistent model stated in equation (13). From table 2, the likelihood for the logit model is -4028.95 while the likelihood for the type-consistent model is -3644.29. Given the results of the non-nested test, it is not surprising to see that the type-consistent model has a much higher likelihood value than the logit model.

On average, the probability of default conditional on $\alpha_\theta = \theta_1$ (6.5%) is lower than the probability of default conditional on $\alpha_\theta = \theta_2$ (12.63%). These results are consistent with the hypothesis that one of the types is more "responsible" than the other. Consider first the coefficients for the probability of type θ_1. Of all the socio-demographic variables included in *Z* to describe this probability, only age is statistically significant. Additionally, there is evidence that age has a quadratic effect over the type probability. This result indicates that older people have a higher probability of belonging to type θ_1, but this effect decreases as age increases. Looking back to the theoretical characterization of the types, in the context of either the time preference or the risk aversion definitions of the types, this result implies that older people tend to belong more to the "responsible" type than younger people. In addition, it is worthwhile to note that neither gender nor marital status is statistically

[5] We also estimated the model using a normal CDF and recovered the main results.

significant to determine the type of an individual. Unfortunately, the data available does not include information regarding the level of education of the cardholder. Signaling theory argues that a person's education could carry important signal about his type in a series of economic setups. It would have been interesting to test education's signaling power in the credit market.

In terms of the conditional probabilities of default, it is important to notice that parameter estimates are different for each type of consumers. For example, for the parameter that *Times rated C past 3 months*, the coefficient for type θ_2 consumer is 17.19 (5.61) while the coefficient for type θ_1 consumer is only 0.217 (.107)[6]. It is also true that some parameters are statistically significant for one type of consumers but not for the other type. Another important observation is that the parameter estimates from the simple logit model are different from both type θ_1 and type θ_2 coefficients. Therefore, this result indicates that different types of consumers do behave differently, as predicted by the theoretical model.

In terms of marginal effects, we present two examples to illustrate the differences between the two models. In the first example, we consider an increase of 10 years for a person at age 20, while everything else is taken at their respective sample mean. For the type-consistent model, the probability that the person belongs to type θ_1 increases by 0.034 and the overall probability of default decreases by 0.00130. However, this change based on logit model causes the default probability to increase by 0.000888. In the second example, we consider an increase in 1 in the number of times a person is rated *E* in the past three months, while taking everything else at their mean. For the type-consistent model, the probability of default increases by 0.1573. For the logit model, the probability of default increases by 0.0823. In both examples, the two models produce significantly different marginal effects.

In the next section, we will show that modeling types in the probability of default yields better out-of-sample predictions of the probability of default.

C. An Application to Credit Scoring

[6] Standard errors showed in parenthesis.

In a more mundane field, the model described and estimated in the previous sections provides the theoretical structure for a credit scoring model. Typically, credit card scoring models are derived only from statistical considerations; with the sole objective of achieving greater levels of predictability of the probability of default. Our model provides an economic structure that considers the estimation of the probability of default within an equilibrium setup and unobserved consumer types. As it will be showed, this allows us to define a credit scoring system that improves on traditional techniques. Thus, we verify the existence of types in the credit market in the sense that a model that includes types in its specification produces better out-of-sample predictions of the probability of default.

Even though credit scoring systems are commonly employed by banks and retail companies, one can not find in the published literature much investigation on the subject, especially on specific estimation techniques. Hand and Henley (1997) argue that this is a consequence of the confidentiality lenders maintain on their data and procedures due to security issues and the competitive advantage given by more accurate estimation techniques. The estimation of the probability of default relies on techniques such as discriminant analysis, linear regression, logistic regression, decision trees, expert systems, neural networks, and dynamic programming. Hand and Henley (1997), Rosenberg and Gleit (1994), and Reichert, Cho and Wagner (1983) give a succinct description of these techniques. Wiginton (1980) presents one of the first uses of the logistic regression in credit scoring. His results show that the logit model predictability dominates the linear discriminant results. In the nonparametric field, Hand and Henley (1996) derive a k-nearest neighbor estimator for estimating the probability of default. In addition, Zhu, Beling and Overstreet (2001) build on the notion of second order stochastic dominance to determine the conditions where combining two credit scores leads to a better model in the sense that it estimates probabilities of default more accurately than its components.

We use the logit model for the probability of default to contrast the out of sample performance of the type-consistent model. Logit models are commonly used in the credit industry, so they are a well known benchmark. Several simple statistics and the Lorenz curve will be used to compare the models. Of course, any validation of the predictive power of the model must be assessed out of sample, so in what follows we use the test sample.

(1) *Model Accuracy*

Table 2 reports the within sample pseudo R^2 for the type-consistent model and the logit model. The pseudo-R^2 = 0.8047 for the type-consistent model, higher than the pseudo-R^2 = 0.7841 for the logit model. Since we have more than 60,000 observations in the estimation sample, the (degrees of freedom) adjusted pseudo-R^2 changes little: the type consistent model still has a higher pseudo-R^2 than the logit model. However, for the within-sample average default probabilities, the logit model (0.075119) is closer to the observed default probability (0.075120) than the type-consistent model (0.07506).

In table 3, we calculate out-of-sample statistics for the mean of the probabilities of default, pseudo-R^2, adjusted pseudo-R^2, and the mean squared prediction errors for both models. The type-consistent model produces a mean default probability closer to the observed sample mean, a higher pseudo-R^2, a higher adjusted pseudo-R^2, and a lower mean squared prediction error than the logit model. Therefore, there is evidence that a model that includes types in its specification produces better out-of-sample predictions of the probability of default.

(2) *Lorenz Curves*

Before further comparing the models, it is necessary to establish how the probability of default will be computed using the type–consistent model. In practice the bank can use the computed unconditional probability of default, or either the conditional probability of default for type θ_1 and the conditional probability of default for type θ_2, depending on each individual's probability of being of type θ_1. Thus, different mixtures for estimating the probability of default can be employed. For a credit card issuer, it is important to minimize the number of "default" ("bad" clients) classified by the model as "not default" ("good" clients), as issuers tend to not reject those applicants with high probabilities of default, but to reduce their credit limit. Taking into account this fact, this research compares two different alternatives with the results of the logit model:

(i) A "naïve" approach that only uses the unconditional probability of default from the type-consistent model such that:

$$\Pr(D_i = 1) = G(Z_i\gamma)F(\alpha_1 y_{1i} + X_i\beta_1) + (1 - G(Z_i\gamma))F(\alpha_2 y_{1i} + X_i\beta_2)$$

(ii) A "risk averse" approach based on the distribution of the estimated probability of type θ_I for each individual, to minimize the number of "bad" clients classified by the model as "good".

$$\Pr(D_i = 1) = \begin{cases} F(\alpha_1 y_{1i} + X_i\beta_1) & \text{if } p_i > 0.86 \\ G(Z_i\gamma)F(\alpha_1 y_{1i} + X_i\beta_1) + (1 - G(Z_i\gamma))F(\alpha_2 y_{1i} + X_i\beta_2) & \text{if } 0.8340 < p_i \le 0.8618 \\ F(\alpha_2 y_{1i} + X_i\beta_2) & \text{if } p_i \le 0.8340 \end{cases}$$

The bounds are the upper quantile and the mean of the distribution of the estimated probability of type θ_I at a 95% confidence level.

Typically, the performance of a credit score model is to evaluate the number of "bad" clients the model rates as "good" (Type II error) and the number of "good" clients the model rates as "bad" (Type I error), for a specific probability of default cut off value. This value should be determined to maximize profit, or relative to a particular approval policy if any exists. In the absence of these criteria, one can calculate the proportion of "good" and "bad" clients accepted by the model for a series of probability of default cut off values. Then one can plot the proportion of "good" clients rejected and the proportion of "bad" clients accepted against the cut off values forming Lorenz curves (see Figure 4 and Figure 5). The perfect model would the lower and right axes for both "good" and "bad" applicants. The area between the curves and the axes is used as a measure of the model's discriminatory power (Hand and Henley, 1997).

As stated before, for the credit card industry, the main interest is to identify the "bad" clients in the application process. Extending this idea, this is so because the amount of credit extended and the interest rate can be defined to take into account the level of risk of each individual (see Section 2). Therefore, in practice a "bad" applicant will not be rejected (unless his or her risk level is really high), but will receive a lower credit limit. For "good" applicants identified as "bad" this poses a minor inconvenience, since is common to periodically adjust (increase) the credit limit granted according to consumption and payment behavior. Thus, this research will focus on comparing the performance of the type model versus the logit model regarding the identification of "bad" applicants. The relevant Lorenz curves for the three cases are in Figure 4.

The "risk averse" approach yields a smaller proportion of "bad" clients accepted than the logit or the "naïve" approach for practically the entire range of probability of default cut off values. This means that the bank will do better using the “risk averse” approach for any acceptance rule. Also, the “naïve” model outperforms the logit for cut off values lower then 0.25. This means that if the banks automatically accepts all clients with a probability of default of 0.25 or lower, it will do better using the “naïve” model than the logit. These results indicate that the type-consistent model identifies more accurately the “bad” from the “good”. This is a consequence of identifying the type of each client in a probabilistic sense.

For the “good” clients (Figure 5), obviously the performance of the “risk averse” approach is lower than both the logit model and the “naïve” approach, since this approach was designed to detect the “bad” clients without considering the effect on the “good” ones. However, its acceptance rate of “good” clients is over 90% for practically the whole range. On the other hand, the “naïve” approach (which is neutral between “good” and “bad”) performs better than the logit model for cut off values of 0.26 and larger. This indicates that we could find a mixture similar to the “risk averse” approach, only that it would minimize rejection of “good” clients”. All this indicates that a model that includes types in its specification produces better out-of-sample predictions, and in this sense confirms the existence of types in the population.

4. Conclusions

This paper builds up an equilibrium model of consumer credit markets. The model has two types of consumers, one with a higher time discount rate than the other. Although the types are unobserved, different types of consumers will choose different equilibrium amount of loan. As a result, the probability of default depends on the type of each individual. Subsequently, the paper derives an econometric model for the probability of default that incorporates the unobserved types. Hence, the choice of defaulting, viewed as a random variable, has a mixture density.

The paper then applies the model to the consumer credit market to test the existence of two types of applicants: those individuals who are naturally inclined to fulfill their credit obligations (the “responsible” type), and those individuals who are naturally inclined to default (the “irresponsible” type). We find that including types in the model for the probability of default leads to better out of sample predictions. In this sense, the paper tests and confirms the existence of types in the population. An interesting result is that older people have a higher chance of belonging to the "responsible" type than younger, while gender statistically has no effect in determining the type of a person. This indicates that responsibility is not part of the genetic code, but is a consequence of the experiences a person has in his life. Unfortunately, the data available does not include variables that carry information about a person's experiences (for example, education, type of work, travels to foreign countries, etc.). It would be interesting to analyze which factors lead to higher levels of responsibility for a series of economic setups.

Appendix

Unobserved Types in Terms of Risk Aversion

Now we assume that there are two types of borrowers: those high risk aversion and those with low risk aversion. To introduce risk aversion in the model, we explicitly define $V(\cdot)$ as a power utility function with constant relative risk aversion coefficient α_θ.

In this context type θ_1 individuals have a higher coefficient α_θ than type θ_2 individuals. The two-period utility function is restated as:

$$U(c_1, c_2) = \frac{(y_1 + L)^{1-\alpha_\theta}}{1-\alpha_\theta} + \frac{\beta}{1-\alpha_\theta}\left(\int_{RL}^{\infty}(y_2 - RL)dF(y_2)\right)^{1-\alpha_\theta} \quad \text{(A1)}$$

The borrower chooses L to maximize his utility given R. The FOC is:

$$(y_1 + L)^{-\alpha_\theta} - \beta R\left(\int_{RL}^{\infty}(y_2 - RL)dF(y_2)\right)^{-\alpha_\theta}(1 - F(RL)) = 0 \quad \text{(A2)}$$

Define $A = \int_{RL}^{\infty}(y_2 - RL)dF(y_2)$. The effect of a change of α_θ of the optimal amount of loan $L^*_{\theta_s}$ is calculated using the implicit function theorem.

$$\frac{dL^*_{\theta_s}}{d\alpha_\theta} = -\frac{A^{\alpha_\theta}(y_1 + L)^{-\alpha_\theta}\ln(y_1 + L) - \beta R(1 - F(RL))\ln(A)}{\alpha_\theta A^{\alpha_\theta}(y_1 + L)^{-(1+\alpha_\theta)} + \beta R^2\left[f(RL) + \alpha_\theta(1 - F(RL))^2 A^{-1}\right]} < 0\,, \quad \text{(A3)}$$

if $A^{\alpha_\theta}(y_1 + L)^{-\alpha_\theta}\ln(y_1 + L) > \beta R(1 - F(RL))\ln(A)$.

Thus the amount of loan demanded is lower for those individual with higher risk aversion. One can also use the implicit function theorem to derive the form of the indifference curves.

$$\frac{dR}{dL} = \frac{(y_1 + L)^{-\alpha_\theta} - \beta R A^{-\alpha_\theta}[1 - F(RL)]}{\beta L A^{-\alpha_\theta}[1 - F(RL)]} \tag{A4}$$

dR/dL is zero if and only if $L = L^*_{\theta_s}$. To determine the sign of the slope for $L \neq L^*_{\theta_s}$, let's consider the second derivative with respect to L:

$$\begin{aligned}\frac{d^2R}{dL^2} = &-\frac{\alpha_\theta (y_1 + L)^{-(1+\alpha_\theta)} - \beta R^2 A^{-\alpha_\theta}\left[f(RL) + \alpha_\theta (1 - F(RL))^2 A^{-1}\right]}{\beta L A^{-\alpha_\theta}[1 - F(RL)]} \\ &+ \frac{dR}{dL}\cdot\frac{RL\left[f(RL) - \alpha_\theta (1 - F(RL))^2 A^{-1}\right] - 1}{L[1 - F(RL)]}\end{aligned} \tag{A5}$$

$dR/dL = 0$ if and only if $L = L^*_{\theta_s}$. This implies that at point $L^*_{\theta_s}$, $d^2R/dL^2 < 0$, which means that the indifference curve is at a maximum. Then, the concavity of $V(\cdot)$ allows one to conclude that:

(a) For $L > L^*_{\theta_s}$: $(y_1 + L)^{-\alpha_\theta} < \alpha_\theta R A^{-\alpha_\theta}(1 - F(RL)) \Rightarrow dR/dL < 0$; and,

(b) For $L < L^*_{\theta_s}$: $(y_1 + L)^{-\alpha_\theta} > \alpha_\theta R A^{-\alpha_\theta}(1 - F(RL)) \Rightarrow dR/dL > 0$.

To evaluate the effect of α_θ, we take the second derivative with respect to α_θ:

$$\frac{d^2R}{dLd\alpha_\theta} = \frac{(y_1 + L)^{-\alpha_\theta}\ln\left(\frac{A}{y_1 + L}\right)}{\beta L A^{-\alpha_\theta}[1 - F(RL)]} \tag{A6}$$

Recall that $A = \int_{RL}^{\infty}(y_2 - RL)dF(y_2)$, which means that it is the expected second period consumption. Thus, there are three scenarios:

- $E(c_2) < y_1 + L = c_1$, in this case, $d^2R/dLd\alpha_\theta < 0$ so type θ_1 individuals, with a higher coefficient α_θ, have a lower marginal willingness to accept increases in R than type θ_2 individuals. This implies the same results as developed in Section 2.
- $E(c_2) > y_1 + L = c_1$, in this case $d^2R/dLd\alpha_\theta > 0$, thus type θ_1 individuals, with a higher coefficient α_θ, have a higher marginal willingness to accept increases in R than type θ_2 individuals. This implies that still one can find differences in the

preference map between the two types, but in this case type θ_1 individuals will have a higher equilibrium loan amount.

- $E(c_2) = y_1 + L = c_1$, in this case the expected second period consumption is the same as the first period consumption. Now, $d^2R/dLd\alpha_\theta = 0$ and the existence of different types has no effect on the preference maps. Thus, the results in Section 2 would not hold.

In conclusion, as long as expected second period consumption is different from the first period consumption, the existence of types, defined in terms or different degrees of risk aversion, will differentiate the preference map of type θ_1 individuals from the preference map of type θ_2 individuals. In turn, these differences will lead to different equilibrium loan amounts and different probability of default for each type.

References

Bierman, H. and Hausman, W. H. (1970). "The Credit Granting Decision." *Management Science*, Vol. 16, No. 8, *Application Series*, pp. B519-B532.

Capon, N. (1982). Credit Scoring Systems: A Critical Analysis. *Journal of Marketing*, Vol. 46, No. 2, pp. 82-91.

Edelstein, R. H. (1975). "Improving the Selection of Credit Risks: An Analysis of a Commercial Bank Minority Lending Program." *The Journal of Finance*, Vol. 30, No. 1, pp. 37-55.

Feinstein, Jonathan S. (1990). "Detection Controlled Estimation." *Journal of Law and Economics*, Vol 33, No. 1. (April 1990), pp. 233-276.

Hand, D. J. and Henley, W. E. (1996). "A k-Nearest-Neighbor Classifier for Assessing Consumer Credit Risk." *The Statistician*, Vol. 45, No. 1, pp. 77-95.

Hand, D. J. and Henley, W. E. (1997). "Statistical Classification Methods in Consumer Credit Scoring: A Review." *Journal of the Royal Statistical Society*. Series A (Statistics in Society), Vol. 160, No. 3, pp. 523-541.

Helland, Eric (1998). "The Enforcement of Pollution Control Laws: Inspections, Violations, and Self-Reporting." *Review of Economics and Statistics*, Vol. 89, No 1(February), pp 141-153.

Jaffee, D. M. and Russell, T. (1976). "Imperfect Information, Uncertainty, and Credit Rationing." *The Quarterly Journal of Economics*, Vol. 90, No. 4, pp. 651-666.

Knittel, Christopher, and Victor Stango (2003). "Price Ceilings as Focal Points for Tacit Collusion: Evidence from Credit Cards." *American Economic Review*, Vol. 93, No. 5 (Dec., 2003), pp. 1703-1729

Milde, H. and Riley, J. G. (1988). "Signaling in Credit Markets." *The Quarterly Journal of Economics*, Vol. 103, No. 1, pp. 101-129.

Reichert, A. K., Cho, C., and Wagner G. M. (1983). "An Examination of the Conceptual Issues Involved in Developing Credit-Scoring Models." *Journal of Business & Economic Statistics*, Vol. 1, No. 2, pp. 101-114.

Rosenberg, E. and Gleit, A. (1994). "Quantitative Methods in Credit Management: A Survey." *Operations Research*, Vol. 42, No. 4, pp. 589-613.

Spence, Michael (1973). "Job Market Signaling." *The Quarterly Journal of Economics*, Vol. 87, No. 3, pp. 355-374.

Stahl, Dale and Paul Wilson (1994). "Experimental Evidence on Players' Models of Other Players." Journal of Economic Behavior and Organization, Vol 25, pp309-327.

Wiginton, J. C. (1980). "A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior." *Journal of Financial and Quantitative Analysis*, Vol. 15, No. 3, pp. 757-770.

Zhu, H., Beling, P.A., and Overstreet, G.A. (2001). A Study in the Combination of Two Consumer Credit Scores. *The Journal of the Operational Research Society*, Vol. 52, No. 9, Special Issue: Credit Scoring and Data Mining, pp. 974-980.

Table 1: *Description of the Variables*

Variable	Description	Mean	Std. Dev.	Min	Max
Default	1 if in default for 90 or more days.	0.08	0.26	0	1
Sex	1 if women, 0 if men.	0.36	0.48	0	1
Number of Children	The number of persons the client is financially responsible for.	1.14	1.30	0	12
Age	Age of the client measured in years.	46.57	11.68	20	95
Marital Status	Single	0.19	0.40	0	1
	Married	0.72	0.45	0	1
	Widow	0.02	0.14	0	1
	Divorced	0.06	0.24	0	1
Income	Monthly income (in US$)	1982.98	1997.63	100	65000
Vehicles	Number of vehicles	1.08	0.69	0	21
Properties	Number of properties (houses, apartments, or offices) the client has.	1.23	0.76	0	10
# of cell phones	# of cell phones.	0.93	0.89	0	3
Cell Phone	1 when the client has a cell phone.	0.66	0.47	0	1
Cell Phone and Telephone	1 if has a land line phone and a cell phone.	0.57	0.50	0	1
Times rated A	Credit ratings of each individual in other financial institutions. For each category (A, B, C, D or E), the number of times rated in each category for the last three months (from 0 to 3).	2.49	0.94	0	3
Times rated B		0.18	0.47	0	3
Times rated C		0.06	0.29	0	3
Times rated D		0.02	0.16	0	3
Times rated E		0.17	0.67	0	3
City dummies	City 1	0.48	0.50	0	1
	City 2	0.23	0.42	0	1
	City 3	0.07	0.26	0	1
	City 4	0.03	0.17	0	1
	City 5	0.01	0.10	0	1
	City 6	0.01	0.08	0	1
	City 7	0.02	0.13	0	1
	City 8	0.02	0.13	0	1
	City 9	0.03	0.16	0	1

Table 2 - Estimation Results

Variable	Logit Model	Type-Consistent Model		
		Probability of Type θ_1	Default Prob Type θ_1	Default Prob Type θ_2
Sex	-0.0847	-0.0885	-	-
	(0.0723)[a]	(0.1025)	-	-
Number of children	-0.0176	0.0135	-	-
	(0.0287)	(0.0413)	-	-
Age	0.0206	0.0482**	-	-
	(0.0208)	(0.0199)	-	-
Square Age	-0.000186	-0.000522**	-	-
	(0.000208)	(0.000205)	-	-
Single	0.2494	0.7446	-	-
	(0.4852)	(0.4704)	-	-
Married	0.3713	0.5668	-	-
	(0.4786)	(0.4746)	-	-
Widower	0.5106	0.3620	-	-
	(0.5286)	(0.5551)	-	-
Divorced	0.3708	0.6071	-	-
	(0.4936)	(0.5114)	-	-
Income	-0.000030	-	-0.000024	0.000047
	(0.0000199)	-	(0.000039)	(0.000054)
Number of vehicles	-0.0997**	-	-0.1316	0.0361
	(0.0505)	-	(0.0939)	(0.2742)
Properties	0.0632	-	-0.1093	0.6648**
	(0.0482)	-	(0.0884)	(0.1943)
Number of cell phones	0.4204**	-	0.5415**	0.6955
	(0.1449)	-	(0.2452)	(0.5767)
Has Cell phone	-1.1601**	-	-1.3903*	-2.1888
	(0.4094)	-	(0.7144)	(1.7014)
Has Land line and cell phone	0.6054**	-	0.6269	1.4854
	(0.2626)	-	(0.4736)	(1.1389)
Times rated A past 3 months	-0.9902**	-	-1.7841**	[b]
	(0.0593)	-	(0.1351)	-
Times rated B past 3 months	0.1530**	-	-0.0945	1.0693
	(0.0644)	-	(0.1078)	(1.3951)
Times rated C past 3 months	1.0985**	-	0.2170**	17.1934**
	(0.0653)	-	(0.1070)	(5.6111)
Times rated D past 3 months	2.3029**	-	2.1249**	30.2062
	(0.0913)	-	(0.1184)	(50.3460)
Times rated E past 3 months	2.3737**	-	2.3192**	6.68673**
	(0.0693)	-	(0.1038)	(1.7785)
City 1	-0.1316	-	0.0642	-0.9488*
	(0.1157)	-	(0.2175)	(0.5269)
City 2	0.1185	-	0.1258	0.2418
	(0.1224)	-	(0.2228)	(0.4908)
City 3	-0.5254**	-	-0.8388**	-1.4157
	(0.1811)	-	(0.3730)	(0.9575)
City 4	-0.1676	-	-0.1325	-0.7568
	(0.2245)	-	(0.4434)	(1.1096)
City 5	-0.0562	-	1.5563**	-33.6897
	(0.3439)	-	(0.4210)	(40.6929)
City 6	0.4841	-	1.1057*	0.6668
	(0.3323)	-	(0.6640)	(1.2261)

City 7	-0.4859*	-	-0.0691	-5.4389
	(0.2842)	-	(0.4889)	(10.0657)
City 8	0.4759*	-	0.7500*	-1.5327
	(0.2478)	-	(0.4356)	(2.4561)
City 9	-0.0164	-	0.1838	-9.4377**
	(0.2322)	-	(0.4341)	(3.0267)
Constant	-3.4030**	-	-2.0522**	-5.9895
	(0.7110)	-	(0.3033)	(4.1805)
Log Likelihood	-4028.95		-3644.29	
Probability of Type θ_1		83.4013%		
Predicted Default Probability	7. 5119%		6.4978%	12.6266%
			7.5061%	
Pseudo R^2	0.7841		0.8047	

Notes:

(a) Standard errors are in parentheses.

* Significant at 90% confidence level. ** Significant at 95% confidence level.

(b) The unconstraint estimate of *Times rated A* has a very high standard deviation for type θ_2. The coefficient is set to zero in this estimation.

Table 3 Comparisons of Out-of-Sample Predictions

	type-consistent model	logit model
Mean Default Probabilities*	0.07470	0.07455
Pseudo-R^2	0.7912	0.7791
Adjusted Psuedo-R^2	0.7911	0.7790
Mean Squared Predicted Errors	0.0148	0.0153

* The observed default probability for this sample is 0.07512.

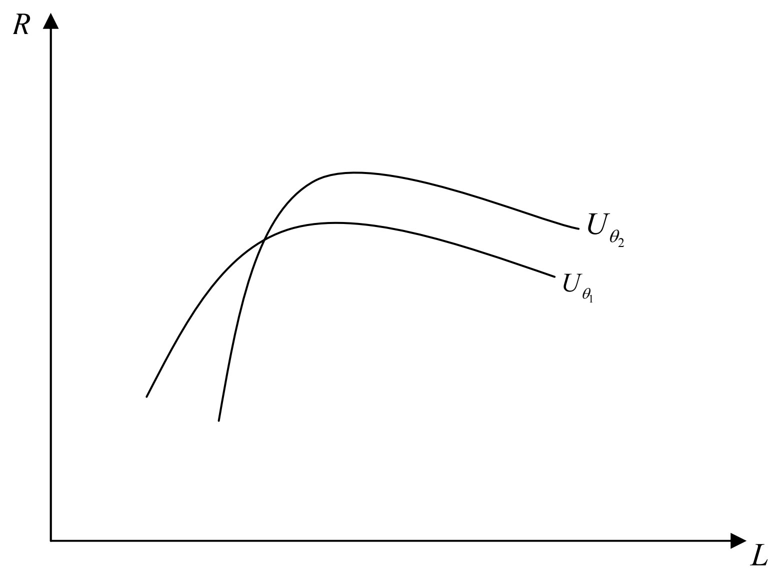

Figure 1 Indifference Curves for two types of consumers

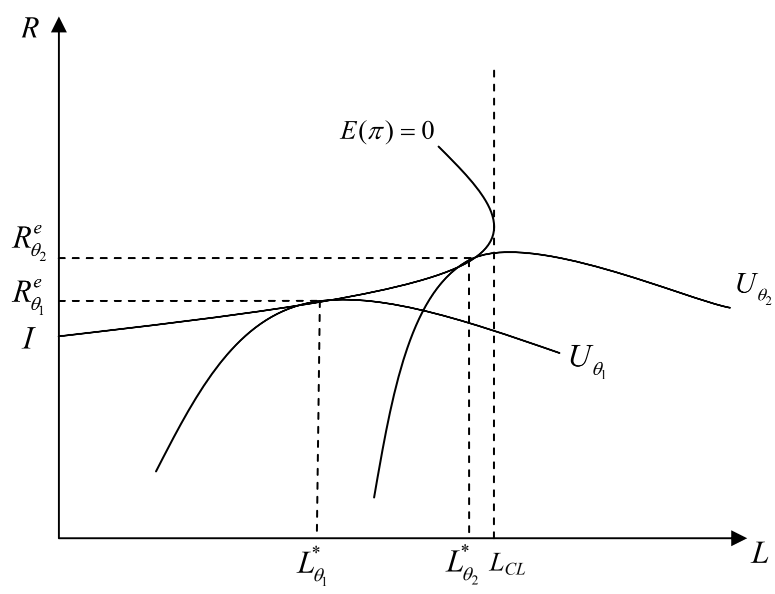

FIGURE 2: Iso-profit Curve and the Indifference Curves

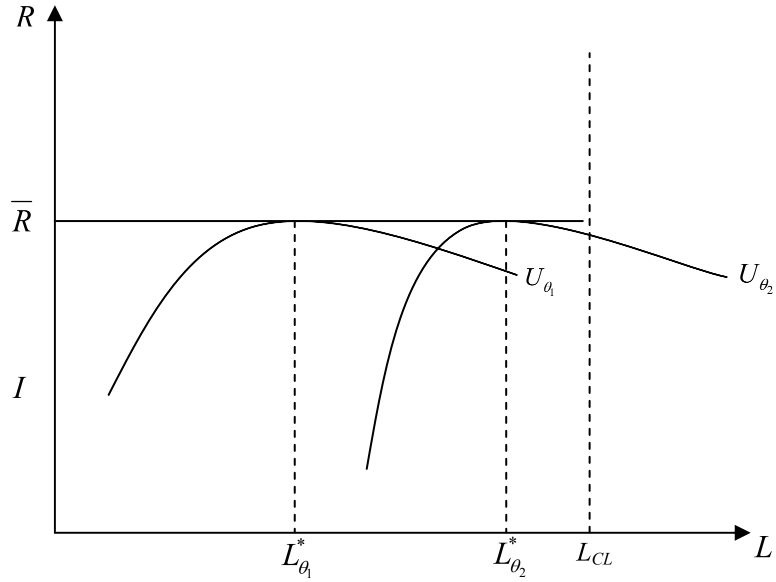

FIGURE 3: Iso-profit Curve and Indifference Curves when *R* is constant

FIGURE 4 - Comparison of Type II Errors

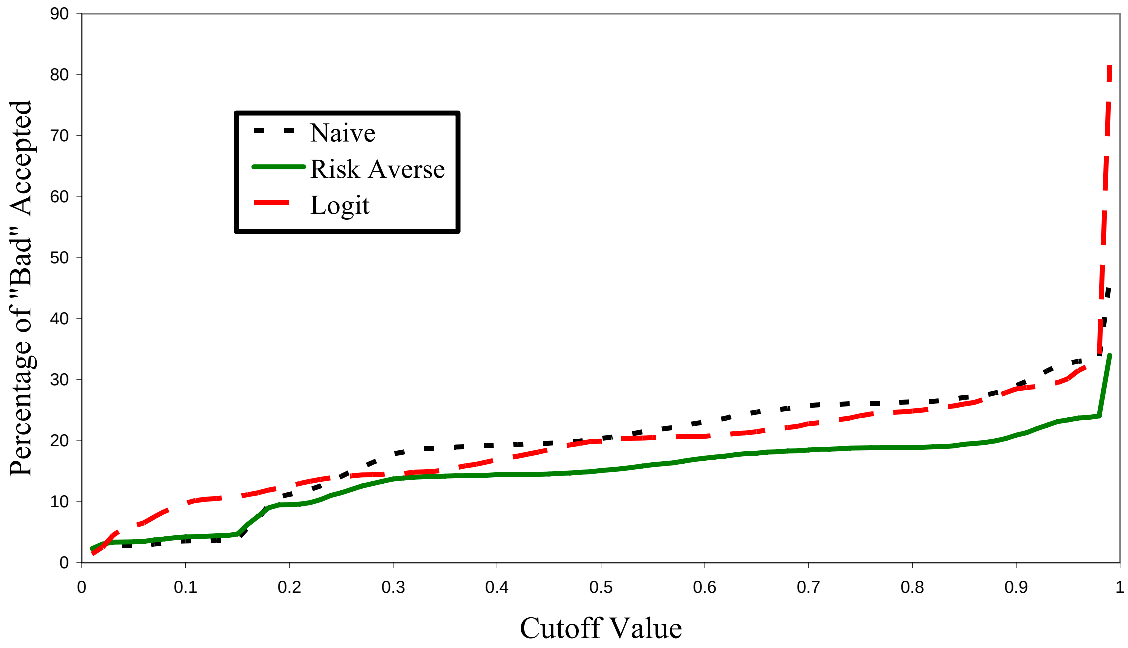

FIGURE 5 - Comparison of Type I Errors