



Munich Personal RePEc Archive

Iterative Least Squares Estimator of Binary Choice Models: a Semi-Parametric Approach

Wang, Weiren and Zhou, Mai

University of Kentucky

July 1995

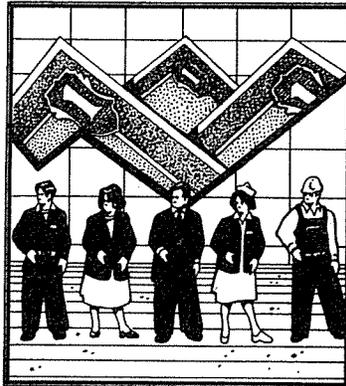
Online at <https://mpra.ub.uni-muenchen.de/46981/>

MPRA Paper No. 46981, posted 16 May 2013 06:31 UTC

Center for Business and Economic Research

Iterative Least Squares Estimator
of Binary Choice Models:
a Semi-Parametric Approach

Weiren Wang and Mai Zhou
University of Kentucky



College of Business and Economics
Lexington, Kentucky 40506-0047

(606) 257-7675

Iterative Least Squares Estimator
of Binary Choice Models:
a Semi-Parametric Approach

Weiren Wang and Mai Zhou
University of Kentucky

Weiren Wang
Department of Economics
University of Kentucky
Lexington, KY 40506-0034
Phone: (606) 257-4149
E-mail: weiren@convex.cc.uky.edu

Mai Zhou
Department of Statistics
University of Kentucky
Lexington, KY 40506-0034

Working Paper #E-180-95
(Revision of E-175-94)

July 1995

The Center for Business and Economic Research
College of Business and Economics
University of Kentucky
Lexington, Kentucky

Iterative Least Squares Estimator of Binary Choice Models:
a Semi-Parametric Approach

Weiren Wang*

Department of Economics
University of Kentucky
Lexington, KY 40506-0034

Phone: (606) 257-4149

E-mail: weiren@convex.cc.uky.edu

Mai Zhou

Department of Statistics
University of Kentucky
Lexington, KY 40506-0027

July 1995

Field: C2

This research is supported by the NSF grant SBR-9321103.

Earlier version of this paper was presented at summer joint meeting at San Francisco,
Aug. 1993.

*Please send correspondence to Weiren Wang.

Abstract

Most existing semi-parametric estimation procedures for binary choice models are based on the maximum score, maximum likelihood, or nonlinear least squares principles. These methods have two problems. They are difficult to compute and they may result in multiple local optima because they require optimizing nonlinear objective functions which may not be unimode. These problems are exacerbated when the number of explanatory variables increases or the sample size is large (Manski, 1975, 1985; Manski and Thompson, 1986; Cosslett, 1983; Ichimura, 1993; Horowitz, 1992; and Klein and Spady, 1993).

In this paper, we propose an *easy-to-compute* semi-parametric estimator for binary choice models. The proposed method takes a completely different approach from the existing methods. The method is based on a semi-parametric interpretation of the Expectation and Maximization (EM) principle (Dempster et al, 1977) and the least squares approach. By using the least squares method, the proposed method computes quickly and is immune to the problem of multiple local maxima. Furthermore, the computing time is not dramatically affected by the number of explanatory variables. The method compares favorably with other existing semi-parametric methods in our Monte Carlo studies. The simulation results indicate that the proposed estimator is, 1) easy-to-compute and fast, 2) insensitive to initial estimates, 3) appears to be \sqrt{n} -consistent and asymptotically normal, and, 4) better than other semi-parametric estimators in terms of finite sample performances. The distinct advantages of the proposed method offer good potential for practical applications of semi-parametric estimation of binary choice models.

Keywords: Binary choice models, EM algorithm, least squares method, semi-parametric estimation.

1. Introduction

Binary choice models are widely used in economics, marketing and other social sciences (Maddala, 1983; Greene, 1993). Recent studies have shown that popular parametric methods, such as the Probit and Logit methods, will cause inconsistency in parameter estimation and misleading predictions of behavior when the error distribution is not normal or logistic (Huber, 1967; Horowitz 1992). As a result, many researchers have tried to avoid restrictive distributional assumptions and advocated semi-parametric estimation methods for binary choice models in recent years (Manski, 1975, 1985; Cosslett, 1983; Klein and Spady, 1993; Ichimura, 1993; Ichimura and Thompson, 1993; Sherman, 1993; Matzkin, 1992; Horowitz, 1992; and Gabler, Laisney and Lechner, 1993). Horowitz (1992, 1994) and Gabler, Laisney and Lechner (1993) also provide excellent summary of these methods.

There are two major disadvantages associated with the existing semi-parametric estimation methods. First, they are difficult to compute, and second, they may generate multiple local optima because they require optimizing nonlinear objective functions which may not be unimode. These problems are exacerbated when the number of explanatory variables increases or the sample size is large (Manski, 1975, 1985; Manski and Thompson, 1986; Cosslett, 1983; Ichimura, 1993; Horowitz, 1992; and Klein and Spady, 1993). Existing semi-parametric methods also need a very large sample size to “obtain the benefits promised by asymptotic theory” (Horowitz, 1992).

In this paper, we propose an *easy-to-compute* semi-parametric estimator for binary choice models. The methodology is completely different from the existing semi-parametric estimation methods for binary choice models. The proposed method is based on a semi-parametric interpretation of the Expectation and Maximization (EM) principle (Dempster et al, 1977) and the least squares approach. By using the least squares method, the estimator computes quickly and is immune to the problem of multiple local maxima. Furthermore, the computing time is not dramatically affected by the number of explanatory variables. Other semi-parametric methods require more (exponentially increasing) computing time or become computationally intractable as the number of explanatory variables increases. We are not aware of any simulation study which deals with models with more than two explanatory variables in the literature. The method compares favorably with other existing semi-parametric methods in our Monte Carlo studies. The simulation results indicate that the estimator is, 1) easy-to-compute and fast, 2) insensitive to initial estimates, 3) appears to be \sqrt{n} -consistent and asymptotically normal, and, 4) better than other semi-parametric estimators, such as the Maximum Score (MS), Smoothed Maximum Score (SMS)¹, Semi-parametric Max-

¹The MS and SMS methods are designed for binary choice models with *symmetric* (can be *het-*

imum Likelihood (Cosslett, 1983), and Klein and Spady (1993) estimators in terms of finite sample performances.

The approach is not completely general, since it requires *homoskedastic* errors. This is a common requirement in many existing methods. Nevertheless, the proposed method is very promising and offers great potential for practical applications of semi-parametric estimation of binary choice models.

The paper is organized as follows. We describe the proposed method and the estimator in section 2. Large sample properties of the estimator are discussed heuristically in section 3. Section 4 presents the results of Monte Carlo experiments with various designs to illustrate as well as to compare the proposed method with other methods for finite sample sizes. We also discuss some important issues such as the consistency, rate of convergence, asymptotic normality of the estimator, relative efficiency of the proposed estimator to other estimators, the choice of initial estimates and the computing times of the estimator. The concluding remarks are given in section 5. Some technical remarks are in the Appendix. Tables and figures are attached to the end of the paper.

2. The Estimator

In this section, we outline the semi-parametric method for estimating binary choice models. The approach is very intuitive and relies on the least squares method.

The binary choice model we want to estimate is of the following form:

$$y_i = \begin{cases} 1 & \text{if } y_i^* = x_i\beta + \epsilon_i > 0, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \dots, n \quad (1)$$

where x_i is a row vector of explanatory variables, y_i is a binary response indicator, y_i^* is a latent variable, β is a vector of unknown parameters to be estimated from data, and the ϵ_i 's, which follow the continuous cumulative distribution function $F(\cdot)$, are independent and identically distributed with zero mean and finite variance.

In equation (1), the indicator y_i , rather than the underlying response variable y_i^* , is observed. However, if we are able to obtain a “good” estimate of y_i^* , then we can apply the powerful least squares method to estimate β . The conditional expectation of y_i^* , $E(y_i^*|x_i\beta, y_i)$ is used as the estimate of y_i^* in this paper. To estimate β , the following steps are performed:

eroskedastic) errors, while the proposed model is designed for binary choice models with *mean zero* and *homoskedastic* errors. The comparison between the proposed method and the MS or SMS method is based on *symmetric and homoskedastic* errors with *zero mean* in the simulation studies.

Step 1: Obtain an initial estimate of β , using the linear probability, Probit, or Logit method. Denote the initial estimate as $\hat{\beta}$ ($|\hat{\beta}_1|$ is normalized to 1)².

Step 2: From observations $(x_i, y_i)_1^n$ and $\hat{\beta}$, use Ayer et al.'s (1955) algorithm to obtain the non-parametric maximum likelihood estimate of $F(\cdot)$, the cumulative distribution of ϵ_i in model (1). Denote this estimate as $\hat{F}(\cdot)$. This algorithm only defines $\hat{F}(\cdot)$ at n points. $\hat{F}(\cdot)$ can be any value besides these n points as long as it is nondecreasing. For convenience, we adopt linear interpolation between any two adjoining points. The details of the algorithm are provided in the Appendix.

Step 3: Compute $\hat{y}_i^* = E(y_i^* | x_i \hat{\beta}, y_i)$ as follows:

$$\hat{y}_i^* = E(y_i^* | x_i \hat{\beta}, y_i) = \begin{cases} x_i \hat{\beta} + \frac{\int_{-x_i \hat{\beta}}^{\infty} \epsilon d\hat{F}(\epsilon)}{1 - \hat{F}(-x_i \hat{\beta})} & \text{if } y_i = 1 \\ x_i \hat{\beta} + \frac{\int_{-\infty}^{-x_i \hat{\beta}} \epsilon d\hat{F}(\epsilon)}{\hat{F}(-x_i \hat{\beta})} & \text{if } y_i = 0, \end{cases} \quad (2)$$

where $\hat{F}(\cdot)$ is from the previous step³.

Step 4: Fit ordinary least squares by regressing \hat{y}_i^* on x_i to obtain a new $\hat{\beta}$.

Step 5: Repeat steps 2-4 until the iterative process converges according to a pre-specified criterion.

The proposed estimator is a semi-parametric estimator, since $F(\cdot)$ is not specified in the method. The method is also an application of the Expectation and Maximization (EM) algorithm. The E-step is performed in a nonparametric fashion and the M-step is performed by the least squares method⁴.

²Equation (1) is invariant through multiplication of β and ϵ_i by a scalar, some normalization is necessary. Without loss of generality, we choose $|\beta_1| = 1$ in this paper.

³Since $\hat{F}(\cdot)$ is a piece-wise linear function by construction, the integration term $\int_{-x_i \hat{\beta}}^{\infty} \epsilon d\hat{F}(\epsilon)$ in (2) can be computed easily as

$$\sum_{t_j = -x_{(j)} \hat{\beta} \geq -x_{(i)} \hat{\beta}} [\hat{F}(t_{j+1}) - \hat{F}(t_j)][t_{j+1} + t_j]/2.$$

⁴The proposed estimator is in fact the iterative solution to the following self-consistent equation:

$$\beta = \left(\sum x_i' x_i \right)^{-1} \sum \hat{y}_i^*(\beta) x_i'$$

where $\hat{y}_i^*(\beta) = E(y_i^* | x_i \beta, y_i) = x_i \beta + y_i \frac{\int_{-x_i \beta}^{\infty} \epsilon d\hat{F}(\epsilon)}{1 - \hat{F}(-x_i \beta)} + (1 - y_i) \frac{\int_{-\infty}^{-x_i \beta} \epsilon d\hat{F}(\epsilon)}{\hat{F}(-x_i \beta)}$.

3. Large Sample Properties of the Proposed Estimator and Identification

In this section, we will discuss the asymptotic properties of $\hat{\beta}$ using heuristic arguments and then support our conjectures using simulations with various designs in the next section.

Consistency: Consider the simple case where only one explanatory variable is included in model (1), i.e., $y_i^* = \alpha + \beta x_i + \epsilon_i$. The estimate of the intercept α is normalized to be 1. The proposed slope estimator $\hat{\beta}$ is, in fact, the solution of the following “normal” equation:

$$\frac{1}{n} \sum_{i=1}^n [E(y_i^* | \hat{\beta} x_i, y_i) - 1 - \hat{\beta} x_i] x_i = 0.$$

Substituting $E(y_i^* | y_i, \hat{\beta} x_i) = 1 + \hat{\beta} x_i + y_i \frac{\int_{-\hat{\beta} x_i}^{\infty} \epsilon d\hat{F}(\epsilon)}{1 - \hat{F}(-\hat{\beta} x_i)} + (1 - y_i) \frac{\int_{-\infty}^{-\hat{\beta} x_i} \epsilon d\hat{F}(\epsilon)}{\hat{F}(-\hat{\beta} x_i)}$, the normal equation becomes

$$\frac{1}{n} \sum_{i=1}^n \left[y_i \frac{\int_{-\hat{\beta} x_i}^{\infty} \epsilon d\hat{F}(\epsilon)}{1 - \hat{F}(-\hat{\beta} x_i)} + (1 - y_i) \frac{\int_{-\infty}^{-\hat{\beta} x_i} \epsilon d\hat{F}(\epsilon)}{\hat{F}(-\hat{\beta} x_i)} \right] x_i = 0 \quad (3)$$

Since the mean of ϵ is zero, we have

$$\int_{-\hat{\beta} x_i}^{\infty} \epsilon d\hat{F}(\epsilon) = - \int_{-\infty}^{-\hat{\beta} x_i} \epsilon d\hat{F}(\epsilon) > 0.$$

The normal equation can be written as

$$\frac{1}{n} \sum_{i=1}^n \hat{w}_i(-\hat{\beta} x_i) [y_i - (1 - \hat{F}(-\hat{\beta} x_i))] x_i = 0,$$

where $\hat{w}_i(t) = \frac{\int_t^{\infty} \epsilon d\hat{F}(\epsilon)}{\hat{F}(t)[1 - \hat{F}(t)]}$, which is always positive.

To study the properties of the above normal equation, we will first consider a simplified version of the normal equation

$$S(\hat{\beta}, \hat{F}) = \frac{1}{n} \sum_{i=1}^n [y_i - (1 - \hat{F}(-\hat{\beta} x_i))] x_i = 0. \quad (4)$$

This can be done by adding weights $1/\hat{w}_i(-\hat{\beta} x_i)$ in Step 4 (linear regression) in the iterative scheme described in the last section.

The solution to the above equation, the estimate $\hat{\beta}$, is likely to be consistent for the following reasons.

First, for the true value, β_* , $\hat{F}(t) \xrightarrow{P} F(t)$ for any given t (Pollard, 1984).
 Second, for the true value, β_* , we have

$$S(\beta_*, \hat{F}) = S(\beta_*, F) + S(\beta_*, \hat{F}) - S(\beta_*, F),$$

and notice that

$$\begin{aligned} S(\beta_*, \hat{F}) - S(\beta_*, F) &= \frac{1}{n} \sum [\hat{F}(-\beta_* x_i) - F(-\beta_* x_i)] x_i \\ &= c_0 \cdot \int_{-\infty}^{+\infty} t[\hat{F}(t) - F(t)] dG_n(t), \end{aligned}$$

where c_0 is some constant, $G_n(t)$ is the empirical CDF of $-\beta_* x_i$. By the theory of uniform consistency of empirical CDF (see, eg. Pollard, 1984. Page 24-36),

$$S(\beta_*, \hat{F}) - S(\beta_*, F) = c_0 \cdot \int_{-\infty}^{+\infty} t[\hat{F}(t) - F(t)] dG(t) + o_p(1).$$

By the general theory on differentiable functionals (see e.g., van der Vaart; 1991) and Groeneboon and Wellner (1992; Theorem 5.5, Page 114), we have

$$\sqrt{n} \int_{-\infty}^{\infty} t[\hat{F}(t) - F(t)] dG(t) \xrightarrow{D} N(0, \sigma^2), \quad (5)$$

where σ^2 is some variance. Therefore,

$$S(\beta_*, \hat{F}) = S(\beta_*, F) + o_p(1).$$

By standard arguments for parametric models (Amemiya; 1985), the solution of $S(\beta, F) = 0$ is consistent (e.g., converges to the true value, β_*). In view of the above approximation, we expect that $S(\beta, \hat{F}) = 0$ also has a solution which is consistent.

Asymptotic Normality: In the iterative scheme we described in Section 2, the estimated $\hat{F}(\epsilon)$ is obtained using Ayer's et al (1955) algorithm, and the resulting function is a piece-wise linear function. This $\hat{F}(\epsilon)$ is convenient in computing $E(y_i^* | \beta x_i, y_i)$, but it is not very convenient in deriving asymptotic properties of $\hat{\beta}$, because it is not continuously differentiable. In principle, we can use any smooth $\hat{F}(\epsilon)$, for example, the kernel based $\hat{F}(\epsilon)$ which is continuously differentiable (Ichimura, 1993) in the iterative scheme. The proof of the asymptotic normality of $\hat{\beta}$ using the smoothed $\hat{F}(\epsilon)$ can proceed as follows. Using Taylor expansion, we can expand $S(\hat{\beta}, \hat{F})$ around β_* :

$$S(\hat{\beta}, \hat{F}) = \sum [y_i - (1 - \hat{F}(-\beta_* x_i))] x_i - \sum \hat{f}(-\tilde{\beta} x_i) x_i^2 (\hat{\beta} - \beta_*) = 0,$$

where $\hat{f}(t) = \partial \hat{F}(t) / \partial t$, and $\tilde{\beta}$ lies between $\hat{\beta}$ and β_* . Thus, we have

$$\sqrt{n}(\hat{\beta} - \beta_*) = \left[\frac{1}{n} \sum \hat{f}(-\tilde{\beta} x_i) x_i^2 \right]^{-1} \frac{1}{\sqrt{n}} \sum [y_i - 1 + \hat{F}(-\beta_* x_i)] x_i.$$

Under certain regularity conditions (such as, \hat{f} being bounded continuous functions converging uniformly to f), the law of large number will guarantee that

$$\frac{1}{n} \sum \hat{f}(-\hat{\beta}x_i)x_i^2 \xrightarrow{P} \lim_{n \rightarrow \infty} \frac{1}{n} \sum E[f(-\beta_*x_i)]x_i^2 = K \quad (\text{say}),$$

where $f(t) = \partial F(t)/\partial t$ (assuming that $F(t)$ is continuously differentiable).

The second part of $\sqrt{n}(\hat{\beta} - \beta_*)$ can be decomposed into two parts:

$$\frac{1}{\sqrt{n}} \sum [y_i - 1 + \hat{F}(-\beta_*x_i)]x_i = C1 + C2,$$

where $C1 = \frac{1}{\sqrt{n}} \sum [y_i - 1 + F(-\beta_*x_i)]x_i$, and $C2 = \frac{1}{\sqrt{n}} \sum [\hat{F}(-\beta_*x_i) - F(-\beta_*x_i)]x_i$.

By the central limit theorem, it is easy to show that $C1 \xrightarrow{D} N(0, \sigma_1^2)$ where σ_1^2 is some variance. By Groeneboom and Wellner (Theorem 5.5, 1992), $C2 \xrightarrow{D} N(0, \sigma_2^2)$, where σ_2^2 is some variance. Then the Cramer-Wold device may be used to show that $\sqrt{n}(\hat{\beta} - \beta_*) \xrightarrow{D} N(0, \sigma_*^2)$, where σ_*^2 is some variance which involves K , σ_1^2 , σ_2^2 and the covariance between $C1$ and $C2$.

Efficiency: Since ϵ_i has mean zero, we have $\int_{-\hat{\beta}x_i}^{\infty} \epsilon d\hat{F}(\epsilon) = -\int_{-\infty}^{-\hat{\beta}x_i} \epsilon d\hat{F}(\epsilon)$, and (3) becomes

$$\frac{1}{n} \sum_{i=1}^n \left[y_i \frac{\int_{-\hat{\beta}x_i}^{\infty} \epsilon d\hat{F}(\epsilon)}{1 - \hat{F}(-\hat{\beta}x_i)} + (1 - y_i) \frac{-\int_{-\hat{\beta}x_i}^{\infty} \epsilon d\hat{F}(\epsilon)}{\hat{F}(-\hat{\beta}x_i)} \right] x_i = 0.$$

In particular, if the error terms ϵ_i follow the standard normal distribution, then $F(\epsilon) = \Phi(\epsilon)$ and $\int_{-\hat{\beta}x_i}^{\infty} \epsilon dF(\epsilon) = f(-\hat{\beta}x_i)$. Therefore in this case the only difference between the above equation and the first order condition for the Probit estimator is that one uses \hat{F} and the other uses F .

This also leads us to conjecture that our estimator is efficient when the error terms are normally distributed.

Identification: This estimation method requires at least one explanatory variable to be continuous and have support on all of R (the real line), otherwise, $\hat{F}(\epsilon)$ will not converge to $F(\epsilon)$. By construction, $\hat{F}(\epsilon)$ is only defined at n points, $(-x_i\beta)_1^n$ (Ayer et al., 1955) and can be any value between these n points as long as it is not decreasing (we use linear interpolation between any two adjoining points in this paper). If x_i 's are continuous, then the number of distinct $-x_i\beta$'s will increase with the sample size n . As n goes to infinity, $\hat{F}(\epsilon)$ will converge to the underlying $F(\epsilon)$. If x_i 's are discrete, then the number of distinct $-x_i\beta$'s will be the same regardless the sample size n . Consequently, $\hat{F}(\epsilon)$ is fixed and independent of the sample size. Therefore, we would not expect that $\hat{F}(\epsilon)$ to converge to $F(\epsilon)$.

4. Monte Carlo Study

In this section, Monte Carlo experiments with various designs are used to illustrate as well as to compare the proposed method with other methods for finite sample sizes. The first Monte Carlo design follows that of Horowitz (1992) with only one parameter β to be estimated. Symmetric error terms are used by Horowitz. In the second experiment, Cosslett's (1986) designs with both symmetric and asymmetric error distributions are used. In the third experiment, we estimate equation (1) with eight explanatory variables to demonstrate the potential of the proposed approach for empirical applications. We also present the variance estimate of the estimator using the bootstrap method (Efron, 1982) for the sample in the third experiment.

Experiment 1 (random design, symmetric errors): we want to estimate β in the following model:

$$y_i = \begin{cases} 1 & \text{if } y_i^* = \alpha x_{i1} + \beta x_{i2} + \epsilon_i > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $(\alpha, \beta) = (1, 1)$, $x_{i1} \sim N(0, 1)$, and $x_{i2} \sim N(1, 1)$. The random error ϵ_i is generated from the following distributions with each distribution corresponding to one experiment⁵. They are:

L: $\epsilon_i \sim$ logistic with median 0 and variance 1;

U: $\epsilon_i \sim$ uniform with median 0 and variance 1;

T3: $\epsilon_i \sim$ Student's t with 3 degrees of freedom normalized to have variance 1.

To compare our results with Horowitz's, the mean bias and the variance of different estimators are computed for three different sample sizes 250, 500 and 1000. There are 1000 replications per experiment (random design). The model is estimated without the intercept. The coefficient of x_{i1} , α , is normalized to 1 for identification purposes. The estimate of β via the linear probability method is used as the starting value of the proposed estimator. Since there is only one coefficient to be estimated, $|\hat{\beta}_t - \hat{\beta}_{t-1}| < 10^{-4}$ is used as convergence criterion. The results of Experiment 1 are summarized in Table 1.

—Insert Table 1 here—

Table 1 compares the variance and bias of the proposed estimator (W-Z in Table 1) with that of three other estimators, including the Logit, the Maximum Score (MS), the

⁵Horowitz (1992) also has another error distribution which is heteroskedastic. We do not include this error distribution in the simulation, since the proposed estimator is based on homoskedastic errors

Smoothed Maximum Score (SMS) estimators. Results of Logit, MS and SMS estimators are cited from Horowitz (1992) when sample sizes are 250, 500 and 1000. Horowitz reported the mean squared errors (MSE), not the variances of the Logit, MS and SMS estimators. To facilitate the comparisons among various estimators, variances of the estimators are reported here. The variance can be easily reproduced by $MSE - Bias^2$ for the Logit, MS and SMS estimators. The results of the proposed estimator is added in Table 1 for the sample size of 2000. Due to different random number generators and possibly different convergence criteria, our results cannot be directly compared to Horowitz's. Since there are 1000 trials in each simulation, we expect the difference resulting from different random number generators to be negligible. To compare the proposed estimator and other estimators, several issues, including the consistency, rate of convergency, asymptotic normality, relative efficiency, choices of the initial estimate and the computing times of the proposed estimator are discussed below. •

Consistency: The proposed estimator exhibits a much smaller finite sample bias (ranging from 1.1% to 2.7%) than the SMS estimator (bias varies between 3.8% and 13.3%) for three different sample sizes and three error distributions. When the sample size is 500, the proposed estimator has smaller bias than the MS estimator for three error distributions. It also has smaller bias than the MS estimator with the uniform distribution, but larger bias with the logistic and $T3$ distributions when the sample size is 250 or 1000. The bias of the estimator also declines as the sample sizes increase for the logistic and $T3$ distributions. For the uniform error distribution, the bias of the estimator first decreases from 0.0176 to 0.0101 as n increases from 250 to 500, then increases a little bit to 0.0107 when n reaches to 1000, and then quickly reduces to 0.0084 as n goes to 2000.

Rate of convergence: To compare the rates of convergence among various estimators, we study the change of variance for each estimator as the sample size doubles by constructing Table 2 from Table 1.

—Insert Table 2 here—

Table 2 confirms that the Logit estimator is \sqrt{n} -convergent, because its variance is approximately halved as the sample size doubles. As shown in Table 2, the variance of the proposed estimator is also approximately halved as the sample size doubles for three error distributions. Furthermore, variance ratios (Var_{500}/Var_{250} , Var_{1000}/Var_{500} , or Var_{2000}/Var_{1000}) are quite stable for the suggested estimator for three error distributions. For example, the variance ratios are 0.5337, 0.5350 and 0.5471 for L, U and $T3$ distributions, respectively when the sample size doubles from 250 to 500 (Table 2).

This supports the conjecture that the proposed estimator is \sqrt{n} -consistent and enjoys the “benefits promised by asymptotic theory” for relatively small sample sizes, such as 250.

The MS estimator is $\sqrt[3]{n}$ -convergent (Kim and Pollard, 1990), which means that as the sample size doubles, the new variance of the MS estimator will be 0.63 ($=2^{-\frac{2}{3}}$) of the original variance. According to Horowitz (1992), the rate of convergence for the SMS estimator in Experiment 1 is $n^{\frac{4}{9}}$. In other words, as the sample size doubles, the new variance of the SMS estimator will be 0.54 ($=2^{-\frac{8}{9}}$) of the original variance. The finite sample results of MS and SMS do not agree with their asymptotic properties for sample sizes up to 1000. Furthermore, variance ratios are not stable for either MS or SMS estimator for three error distributions as the sample size doubles. For example, the variance ratios (for MS) are 0.6523, 0.4352 and 0.5835 for L, U and T3 distributions, respectively when the sample size doubles from 250 to 500. This means that the asymptotic has not “kicked in” yet. As noted in Horowitz (1992, p.518), “large samples are needed to obtain the benefits promised by asymptotic theory” for MS and SMS estimators. It is also noted that the variances of MS and SMS estimators are much larger than that of the proposed estimator in all the situations reported in Table 1.

Asymptotic normality: There are 1000 estimates of β 's in Experiment 1 for each error distribution. The attached Figure 1 shows the histograms of these 1000 estimates of β 's when the sample size is 1000. These histograms closely resemble the distribution of normal variables.

—Insert Figure 1 here—

Relative efficiency: The relative efficiencies of various semi-parametric estimators with the Logit error distribution are presented in Table 1. The Logit model is fully (100%) efficient with the logistic error distribution. The suggested estimator has better finite sample performances than the MS and SMS estimators do. For example, the estimator is 79% ($= 0.0152/0.0193$) efficient when the sample size is 250. However, MS and SMS estimators are only 20% and 28% efficient when $n = 250$, respectively. When the sample size is 1000, the estimator is 86% efficient, while MS and SMS are 15% and 44% efficient, respectively.

To further investigate the efficiency of the suggested estimator, we compare the performances between the proposed and Klein and Spady (K-S) estimators. The K-S estimator is supposed to attain the semi-parametric efficiency lower bound (Klein and

Spady, 1993, p.405). The experiment in Klein and Spady (1993) is reviewed as follows:

$$y_i = \begin{cases} 1 & \text{if } y_i^* = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i > 0 \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \dots, 100$. x_{i1} is a chi-squared variable with 3 degrees of freedom truncated at 6 and standardized to have zero mean and unit variance; x_{i2} is a standard normal variable truncated at ± 2 and similarly standardized. x_{i1} and x_{i2} are independently and identically distributed. The ϵ_i 's are from the standard normal distribution. The number of replication is 1,000. The normalization is set to be $|\beta_1| + |\beta_2| = 2$. The convergence criterion is $\|\hat{\beta}_t - \hat{\beta}_{t-1}\|^2 < 10^{-8}$. Results for β_1 are displayed in Table 3.

—Insert Table 3 here—

Results for the Probit and K-S estimators are directly cited from Klein and Spady (1993, p. 406). As illustrated in Table 3, the proposed estimator has better sample performances than the K-S estimator does. The estimator is 88% efficient while the K-S estimator is only 78% efficient when the sample size is 100.

Initial estimates: It would be the best to use a \sqrt{n} -consistent initial estimate for the proposed method. However, based on the simulations reported here, the estimator is very *insensitive* to initial estimates. For example, in Experiment 1, the true value of β is 1. Initial values ranging from -28 to $+28$ are used to investigate the sensitivity of the estimator. After three or four iterations, the estimate $\hat{\beta}$ becomes very close to 1. The same phenomena were also observed in Experiments 2 and 3, which will be reported shortly.

The proposed approach is an iterative scheme. It may face the problem of nonconvergence, especially for small sample sizes, because the estimated $\hat{F}(\cdot)$ is not continuous in β . Based on the simulation, the proposed algorithm may oscillate between two values (both of which are very close to the true value). First, $\hat{\beta}$ monotonically goes to one of the two oscillating points. Then the oscillation takes place. As the sample size increases, $\hat{F}(\cdot)$ becomes “less discontinuous” and the oscillation is less severe. Asymptotically, $\hat{F}(\cdot)$ is continuous as long as the true $F(\cdot)$ is continuous. For practical purpose, if the sample size is small and the proposed estimator swings between two values, we can take the average of these two values, or any value in between. This is also the strategy adopted in Buckley and James (1979) for censored regression models.

In Experiment 1, the Logit estimator performs well for the symmetric error distributions, although the Logit model is misspecified with the U and $T3$ distributions. This may due to the fact that the explanatory variables in Experiment 1 are generated

from normal distributions. Ruud (1983) shows that the Probit estimator is consistent when explanatory variables are jointly normal. It is also known that the Probit estimator is very close to the Logit estimator. In the next experiment, we will use designs in Cosslett (1986) to compare the performances among the proposed estimator, Probit, Maximum Score, Maximum Rank Correlation (MRC) (Han, 1987), Semi-parametric Maximum Likelihood (SML) (Cosslett, 1983) estimators.

Experiment 2: Consider a binary choice model with two explanatory variables:

$$y_i = \begin{cases} 1 & \text{if } y_i^* = \alpha_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The true parameter values are $\alpha_0 = 0$, $\beta_1 = 1$, and $\beta_2 = -2$.

To generate the explanatory variables x_1 and x_2 , two cases were considered:

- (A) standard normal, $x_1, x_2 \sim N(0,1)$;
- (B) standardized exponential, $x_1, x_2 \sim \exp(1) - 1$.

The error distributions were normal mixtures with zero mean. Three cases were included in the experiment:

- (1) $\epsilon_i \sim N(0, 1)$;
- (2) $\epsilon_i \sim 0.75 * N(0, 1) + 0.25 * N(0, 5)$
(standard deviation 2.65; skewness 0; kurtosis 6.61);
- (3) $\epsilon_i \sim 0.75 * N(-0.5, 1) + 0.25 * N(1.5, 5)$
(standard deviation 2.78; skewness 1.29; kurtosis 6.29).

Two different sample sizes, 250 and 1000, were used in the experiment. There are 1000 replications. The coefficient β_1 is normalized to 1. Since only one parameter β_2 has to be estimated, $|\hat{\beta}_{2(t)} - \hat{\beta}_{2(t-1)}| < 10^{-4}$ is used as the convergence criterion. We report the mean bias and the root mean squared error (RMSE) of the estimates of β_2 for various estimators mentioned above in Tables 4.

—Insert Table 4 here—

In Table 4, results for Probit, MS, MRC, SML and SML-1 (one-step iterative improvement over SML, see Cosslett, 1986) are cited from Cosslett (1986). Because there are 1000 trials in the simulation, we expect that the difference resulting from different random number generators to be negligible.

When x_1 and x_2 are both $N(0, 1)$, the Probit estimator is consistent even the errors are not from $N(0, 1)$. The proposed estimator, SML-1 and the Probit estimator produce very similar results. They compare favorably with other semi-parametric estimators. When the two explanatory variables x_1 and x_2 are both standardized exponentials,

the Probit estimator is inconsistent when the errors are misspecified. The proposed estimator performs better than other estimators, especially when the errors are from M2, the mixture (of normals) with thicker tails and skewness.

The power of the proposed estimator lies on its ability to handle models with many explanatory variables and large sample sizes. The next experiment with eight explanatory variables is a more realistic example for empirical applications. Existing semi-parametric methods become computationally intractable for this experiment. This experiment is the first one in literature to our knowledge.

Experiment 3 (fixed design, 8 explanatory variables):

$$y_i = \begin{cases} 1 & \text{if } y_i^* = \beta_0 + \sum_{j=1}^8 \beta_j x_{ij} + \epsilon_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

where $(\beta_j, j = 0, \dots, 8) = (0, 1, 1, 1, -1, 0.5, 1.3, 1, -0.5)$, $x_1 \sim N(0, 1)$, $x_2 \sim N(1, 1)$ and x_3 is a binary variable, taking 1 and 0 with equal probabilities. x_4 is also a discrete random variable, taking 0, 1, 2, 3, 4 and 5 with equal probabilities. x_5 is a binary variable, taking 0 with probability 0.9 and 1 with probability 0.1. x_6 is a discrete variable, taking 0, 1 and 2 with probabilities 0.5, 0.375 and 0.125, respectively. x_7 and x_8 are generated from the uniform distribution $u[0, 2]$ and the Chi-square distribution with 3 degrees of freedom, respectively. The sample size is 1000. The explanatory variables are generated only once (fix design). The random error is generated 1000 times from $N(0, 2)$. Since there are eight coefficients to be estimated, $\|\hat{\beta}_t - \hat{\beta}_{t-1}\|^2 < 10^{-8}$ is used as the convergence criterion for the iteration process. Table 5 summaries the simulation results for the Logit estimator and the proposed estimator.

—Insert Table 5 here—

Since the error distribution is normal, the “slightly misspecified” Logit model produces good results as expected. However, the suggested estimator performs even better. The estimator has smaller bias than the Logit estimator for all the eight coefficients. The variances of the estimates are either smaller than or comparable to the variances of the Logit estimates. Table 5 also includes the results of the proposed estimator for sample size 2000. As the sample size increases from 1000 to 2000, the bias of the estimator drops substantially. The variance of the estimator is also approximately halved. The attached Figure 2 shows the histograms of the 1000 estimates of β_i ’s when the sample size is 2,000. The histograms closely resemble the distribution of normal variables. This again suggests that the proposed estimator is \sqrt{n} -consistent and asymptotically normal.

—Insert Figure 2 here—

Computing times of the proposed estimator: The proposed method computes $\hat{\beta}$ very quickly, since the computation time of Step 2 is almost independent of the number of explanatory variables. Step 3 computes simple summations. And Step 4 uses OLS, which is minimal in terms of computing time. The mean computing times (in seconds, on a HP730 workstation) for one replication in Experiments 1 and 3 are documented in Table 6.

—Insert Table 6 here—

The reported computing times include the times for generating random numbers, input and output, so the actual computing times are even less.

The variance of the proposed estimator: It is difficult to obtain the analytical solution for the variance of the proposed estimator. Different resampling methods (Efron, 1982; Wu, 1986), can be used to obtain the variance estimate for $\hat{\beta}$. They are commonly used methods for estimating variances of semi-parametric estimators (Horowitz, 1992). Resampling methods are appropriate for the proposed method, since the estimator computes very fast as shown earlier. Table 7 reports the bootstrap (Efron, 1982) variance estimate of $\hat{\beta}$ for one sample ($n = 1,000$) used in Experiment 3.

—Insert Table 7 here—

For the same sample size ($n = 1000$), the bootstrap variance estimates of $\hat{\beta}_i$'s are slightly smaller than the variances of $\hat{\beta}_i$'s reported in Table 5. This may due to the particular sample selected.

5. Conclusions

In this paper, we proposed an *easy-to-compute* semi-parametric estimation method for binary choice models. The methodology is significantly different from the existing methods. The proposed estimator computes very fast and compares favorably with other semi-parametric estimators, including the MS, SMS, SML, MRC, and Klein-Spady estimators, in Monte Carlo studies. The method is especially useful when multiple explanatory variables are included in binary choice models or the sample size is large. This proposed approach opens a new avenue for empirical applications of the semi-parametric estimation of binary choice models.

The proposed method can be extended to ordered multiple choice models as long as there are some “good” semi-parametric estimates of $F(\epsilon)$. This is the topic for future research.

Bibliography

- [1] AMEMIYA, T.: *Advanced Econometrics*, Cambridge, Mass: Harvard University Press, 1985.
- [2] AYER, M. H. D. BRUNK, G. M. EVING, W. T. REID, AND E. SILVERMAN (1955): "An Empirical Distribution Function for Sampling with Incomplete Information," *Annals of Mathematical Statistics*, **26**, 641-647.
- [3] BUCKLEY, J. AND I. JAMES (1979): "Linear Regression with Censored Data," *Biometrika*, **66**, 429-436.
- [4] COSSLETT, S. R. (1983): "Distribution-free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica* **51**, 765-782.
- [5] ——— (1986): "Efficiency of Semiparametric Estimators for the Binary Choice Model in Large Samples: A Monte Carlo Comparison," manuscript.
- [6] ——— (1987): "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models," *Econometrica*, **55**, 559-585.
- [7] DEMPSTER, A. P. N. M. LAIRD, AND D. R. RUBIN (1977): "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," (with Discussion) *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.
- [8] EFRON, B. (1982): "The Jackknife, the Bootstrap and Other Resampling Plans," *CBMS-NSF Regional Conference Series in Applied Mathematics*, **38**.
- [9] GABLER, S. LAISNEY, F. AND M. LECHNER (1993): "Seminonparametric Estimation of Binary-Choice Models with an Application to Labor-Force Participation," *Journal of Business & Economic Statistics*, **11**, 61-80.
- [10] GREENE, W. H. : *Econometric Analysis*, New York: MacMillan Publishing Company, Second Edition, 1993.
- [11] GROENEBOON, P. AND J. A. WELLNER: *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Basel: Birkhäuser, 1992.
- [12] HAN, A. K. (1987): "Non-Parametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, **35**, 303-316.
- [13] HOROVITZ, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, **60**, 505-531.
- [14] ——— (1994): "Advances in Random Utility Models," Working Paper Series No.94-02, Department of Economics, University of Iowa.
- [15] HUBER, P. J. (1967): "The behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistical and Probability*. Berkeley: University of California.

- [16] ICHIMURA, H. (1993): "Semiparametric Least Square (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, **58**, 71-120.
- [17] ICHIMURA, H. AND T. S. THOMPSON (1993): "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution," memo, September 1993.
- [18] KIM, J., AND D. POLLARD (1990): "Cube Root Asymptotics," *Annals of Statistics* **18**, 191-219.
- [19] KLEIN, R. L. AND R. H. SPADY (1993): "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica*, **61**, 387-421.
- [20] LEURGANS, S. (1982): "Asymptotic distributions of slope-of-greatest-convex-minorant estimators," *Annals of Statistics*, **10**, 287-296.
- [21] MADDALA, G. S.: *Limited Dependent and Qualitative Variables in Econometrics*, New York: Cambridge University Press, 1983.
- [22] MANSKI, C. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, **3**, 205-228.
- [23] ——— (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, **27**, 313-334.
- [24] MANSKI, C. AND T. S. THOMPSON (1986): "Operational Characteristics of Maximum Score Estimation," *Journal of Econometrics*, **32**, 65-108.
- [25] MATZKIN, R. L. (1992): "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica*, **60**, 239-270.
- [26] PINKSE, C. A. P. (1993): "On the Computation of Semiparametric Estimates in Limited Dependent Variable Models," *Journal of Econometrics*, **58**, 185-205.
- [27] POLLARD, D.: *Convergence of Stochastic Processes*, New York: Springer-Verlag, 1984.
- [28] RUUD, P. A. (1983): "Sufficient Conditions for the Consistency of Maximum Likelihood Estimation despite Misspecification of Distribution," *Econometrica*, **49**, 505-514.
- [29] SHERMAN, R. P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, **61**, 123-138.
- [30] VAN DER VAART, A. W. (1991): "On Differentiable Functionals," *Annals of Statistics*, **19**, 178-205.
- [31] WU, C. F. J. (1986): "Jackknife, Bootstrap and Other Resampling methods in Regression Analysis," *Annals of Statistics*, **14**, 1261-1295.

Appendix: Non-Parametric Maximum Likelihood Estimator of $F(\cdot)$

Ayer et al. (1955) provides an algorithm, to estimate $F(\cdot)$ based on the *ordered* (ascending order in $-x_i\beta$) observation pairs $(y_{(i)}, -x_{(i)}\beta)_1^n$ for any given value of β . The reader is referred to Cosslett (1983, p.772) for a heuristic explanation of this procedure and Amemiya (1985, p.347) for a simple numerical example. The algorithm proceeds as follows:

(i) Rank-ordering: rearrange pairs $(y_i, -x_i\beta)$ such that $-x_i\beta$ is in ascending order. Denote the new sequence as $(y_{(i)}, -x_{(i)}\beta)$.

(ii) Isotonizing: group the new sequence such that there is a group boundary between observations j and $j + 1$ if and only if $y_{(j)} = 0$ and $y_{(j+1)} = 1$. In other words, every group consists of only one non-increasing (in $y_{(i)}$) sequence. Assume there are K groups.

(iii) Computing the proportion of ones in each group: there are K values, (p_1, \dots, p_K) , for K groups. The empirical counterpart of $F(\cdot)$ is defined as $\hat{F}(-x_{(i)}\beta) = p_k$ for the (i) th observation in k th group.

(iv) If $p_k < p_{k-1}$ for some k , then combine group $(k - 1)$ and group k and repeat Step (iii) until $\hat{F}(\cdot)$ is a nondecreasing function.

This algorithm produces the non-parametric maximum likelihood estimate (NPMLE)⁶ $\hat{F}(\cdot)$. Since $\hat{F}(\epsilon)$ depends only on the rank but not the magnitude, of $-x_i\beta$, we normalize on the parameters. The conversion adopted in this paper is to set $\beta_1 = 1$. Leurgans (1982) can be used to show that $\hat{F}(t)$ is $\sqrt[3]{n}$ -consistent for any given t .

This algorithm only defines $\hat{F}(\cdot)$ at n points. $\hat{F}(\cdot)$ can be any value besides these n points as long as it is nondecreasing. For convenience, we adopt linear interpolation between any two adjoining points. At the first and the last points, we adopt the following convention: if $\hat{F}(-x_{(1)}\beta)$ is not zero, then we define $\hat{F}(-x_{(1)}\beta - 2) = 0$; if $\hat{F}(-x_{(n)}\beta) \neq 1$, then we define $\hat{F}(-x_{(n)}\beta + 2) = 1$. This is equivalent to connecting an exponential distribution $exp(1)$ to the end points as far as equation (2) is concerned.

⁶It is interesting to note that the NPMLE $\hat{F}(\cdot)$ can also be obtained as an isotonic least squares estimator, i.e.,

$$\min_{F(\cdot)} \sum_1^n (1 - y_{(i)} - F(-x_{(i)}\beta))^2, \quad \text{subject to } F(\cdot) \text{ nondecreasing.}$$

TABLE 1: Comparison among various estimators*, (random design, symmetric error distributions.)

Estimator	Error 1: <i>L</i>			Error 2: <i>U</i>		Error 3: <i>T3</i>	
	VAR	Bias	Rel. Eff.	VAR	Bias	VAR	Bias
n=250							
Logit	.0152	.0020	100%	.0205	.0064	.0273	.0237
MS	.0768	-.0061	20%	.1843	.0656	.1126	.0027
SMS	.0419	.1092	28%	.1203	.1329	.0484	.0797
W-Z	.0193	-.0268	79%	.0200	-.0176	.0170	-.0255
n=500							
Logit	.0076	.0089	100%	.0099	-.0023	.0127	.0080
MS	.0501	.0239	15%	.0802	.0213	.0657	.0224
SMS	.0192	.0826	40%	.0438	.0760	.0426	.0652
W-Z	.0103	-.0178	74%	.0107	-.0101	.0093	-.0169
n=1000							
Logit	.0039	.0003	100%	.0047	-.0014	.0060	.0050
MS	.0261	.0064	15%	.0512	.0196	.0373	.0050
SMS	.0088	.0620	44%	.0180	.0465	.0156	.0377
W-Z	.0045	-.0169	86%	.0047	-.0107	.0043	-.0159
n=2000							
W-Z	.0023	-.0131	87%	.0023	-.0084	.0022	-.0133

* : *L* =logistic, *U* =uniform, *T3* =student's *t* with 3 degrees of freedom. 2 explanatory variables. Explanatory variables and random errors are generated 1000 times.

TABLE 2: Comparison of the rates of convergence among various estimators* as the sample size doubles.

Estimator		Error 1: L	Error 2: U	Error 3: $T3$
n=250 to 500	Asy. Var. Ratio	Var. Ratio	Var. Ratio	Var. Ratio
Logit	0.500	.5013	.4829	.4652
MS	0.630	.6523	.4352	.5835
SMS	0.540	.4575	.3641	.8802
W-Z	0.500**	.5337	.5350	.5471
n=500 to 1000	Asy. Var. Ratio	Var. Ratio	Var. Ratio	Var. Ratio
Logit	0.500	.5118	.4747	.4724
MS	0.630	.5210	.6384	.5677
SMS	0.540	.4570	.4109	.3660
W-Z	0.500**	.4364	.4383	.4624
n=1000 to 2000	Asy. Var. Ratio	Var. Ratio	Var. Ratio	Var. Ratio
W-Z	0.500**	.5111	.4894	.5116

* : L =logistic, U =uniform, $T3$ =student's t with 3 degrees of freedom. 2 explanatory variables, 1000 replications.

** : we expect that the proposed estimator to be \sqrt{n} -consistent.

Table 3: Comparisons among the Klein-Spady (K-S), W-Z and Probit estimators.

Estimator	Bias(β_1)	Var(β_1)	Rel. Eff.
K-S	-0.00042	0.01532	78%
W-Z	0.003765	0.01357	88%
Probit	-0.00013	0.01196	100%

TABLE 4: Comparison among various estimators (Cosslett's designs)

x_1, x_2 are $N(0,1)$, 1000 trials						
Estimator	Error 1: N		Error 2: $M1$		Error 3: $M2$	
n=250	Bias	RMSE	Bias	RMSE	Bias	RMSE
Probit	-0.04	0.29	-0.11	0.49	-0.11	0.50
MS	-0.22	0.76	-0.34	1.16	-0.36	1.27
MRC	-0.05	0.34	-0.11	0.49	-0.11	0.52
SML	-0.08	0.43	-0.02	0.67	-0.20	0.70
SML-1	-0.05	0.31	-0.11	0.48	-0.10	0.47
W-Z	-0.02	0.29	-0.09	0.48	-0.11	0.53
x_1, x_2 are $exp(1) - 1$, 1000 trials.						
Estimator	Error 1: N		Error 2: $M1$		Error 3: $M2$	
n=250	Bias	RMSE	Bias	RMSE	Bias	RMSE
Probit	-0.03	0.35	-0.23	0.72	-0.69	1.24
MS	-0.37	1.29	-0.51	1.87	-0.55	1.64
MRC	-0.05	0.43	-0.13	0.71	-0.27	1.32
SML	-0.10	0.53	-0.23	0.84	-0.29	1.01
SML-1	-0.06	0.39	-0.23	0.73	-0.43	1.38
W-Z	-0.05	0.36	-0.25	0.76	-0.07	0.59
n=1000	Bias	RMSE	Bias	RMSE	Bias	RMSE
Probit	-0.01	0.17	-0.16	0.34	-0.57	0.70
MS	-0.11	0.47	-0.16	0.59	-0.16	0.66
MRC	-0.02	0.19	-0.03	0.26	-0.05	0.34
SML	-0.03	0.25	-0.06	0.35	-0.07	0.40
SML-1	-0.03	0.19	-0.09	0.28	-0.16	0.35
W-Z	-0.02	0.16	-0.09	0.29	-0.01	0.27

TABLE 5: Comparison of the Logit and W-Z estimators* (nonrandom design, 8 explanatory variables.)

Estimator	VAR($\hat{\beta}_0$)	Bias($\hat{\beta}_0$)	VAR($\hat{\beta}_2$)	Bias($\hat{\beta}_2$)	VAR($\hat{\beta}_3$)	Bias($\hat{\beta}_3$)	VAR($\hat{\beta}_4$)
Logit ($n = 1K$)	.1192	-.0291	.0293	.0171	.0566	.0231	.0169
W-Z ($n = 1K$)	.11884	-.0089	.02952	.0120	.05687	.0194	.01689
W-Z ($n = 2K$)	.06188	.0013	.01568	.0059	.02615	.0065	.00694

Bias($\hat{\beta}_4$)	VAR($\hat{\beta}_5$)	Bias($\hat{\beta}_5$)	VAR($\hat{\beta}_6$)	Bias($\hat{\beta}_6$)	VAR($\hat{\beta}_7$)	Bias($\hat{\beta}_7$)	VAR($\hat{\beta}_8$)	Bias($\hat{\beta}_8$)
-.0125	.1282	.0171	.0486	.0258	.0452	.0303	.0058	-.0084
-.0096	.12613	.0080	.04864	.0207	.04773	.0181	.00593	-.0064
-.0022	.05363	.0051	.01897	-.0008	.02234	.0095	.00242	-.0037

* : $\epsilon \sim N(0, 2)$, 8 explanatory variables, 1000 observations. The explanatory variables are generated once. The random errors are generated 1000 times. β_1 is normalized to be 1.

Table 6: Computing times of the W-Z estimator.

sample size	n=500	n=1000
Two explanatory variables	6.9 sec	14.5 sec
Eight explanatory variables	8.9 sec	18.9 sec

TABLE 7: Bootstrap estimate for the variance of the W-Z estimator* (nonrandom design, 8 explanatory variables).

Estimator	VAR($\hat{\beta}_0$)	VAR($\hat{\beta}_2$)	VAR($\hat{\beta}_3$)	VAR($\hat{\beta}_4$)
W-Z	.0909	.0197	.0536	.0123

Estimator	VAR($\hat{\beta}_5$)	VAR($\hat{\beta}_6$)	VAR($\hat{\beta}_7$)	VAR($\hat{\beta}_8$)
W-Z	.1388	.0398	.0395	.0055

* : $\epsilon \sim N(0, 2)$, 8 explanatory variables, 1000 observations, 1000 resamplings.

Figure 1. Histograms of beta in Experiment 1, 1,000 obs, 1,000 replications.

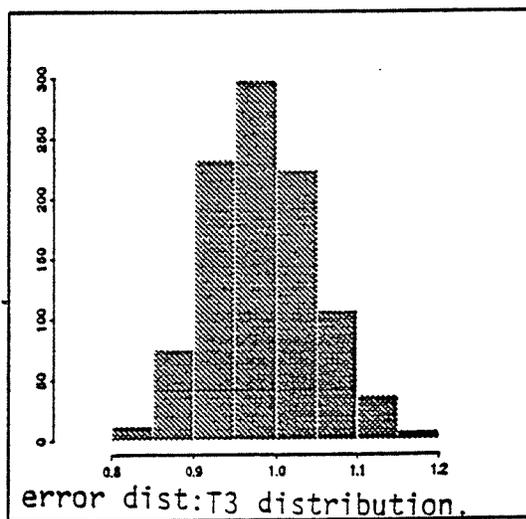
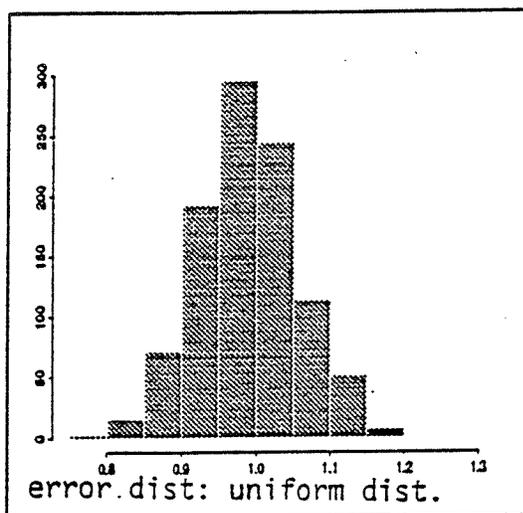
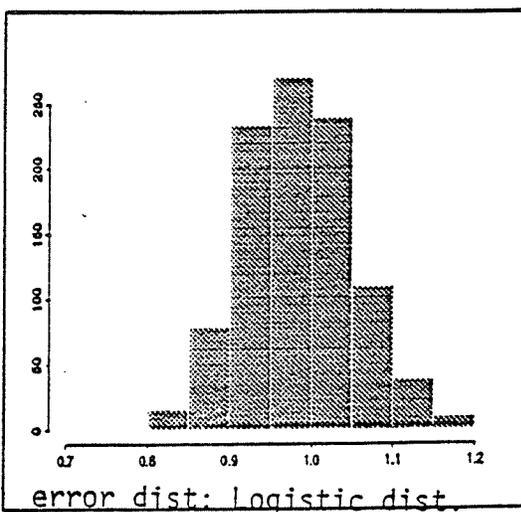
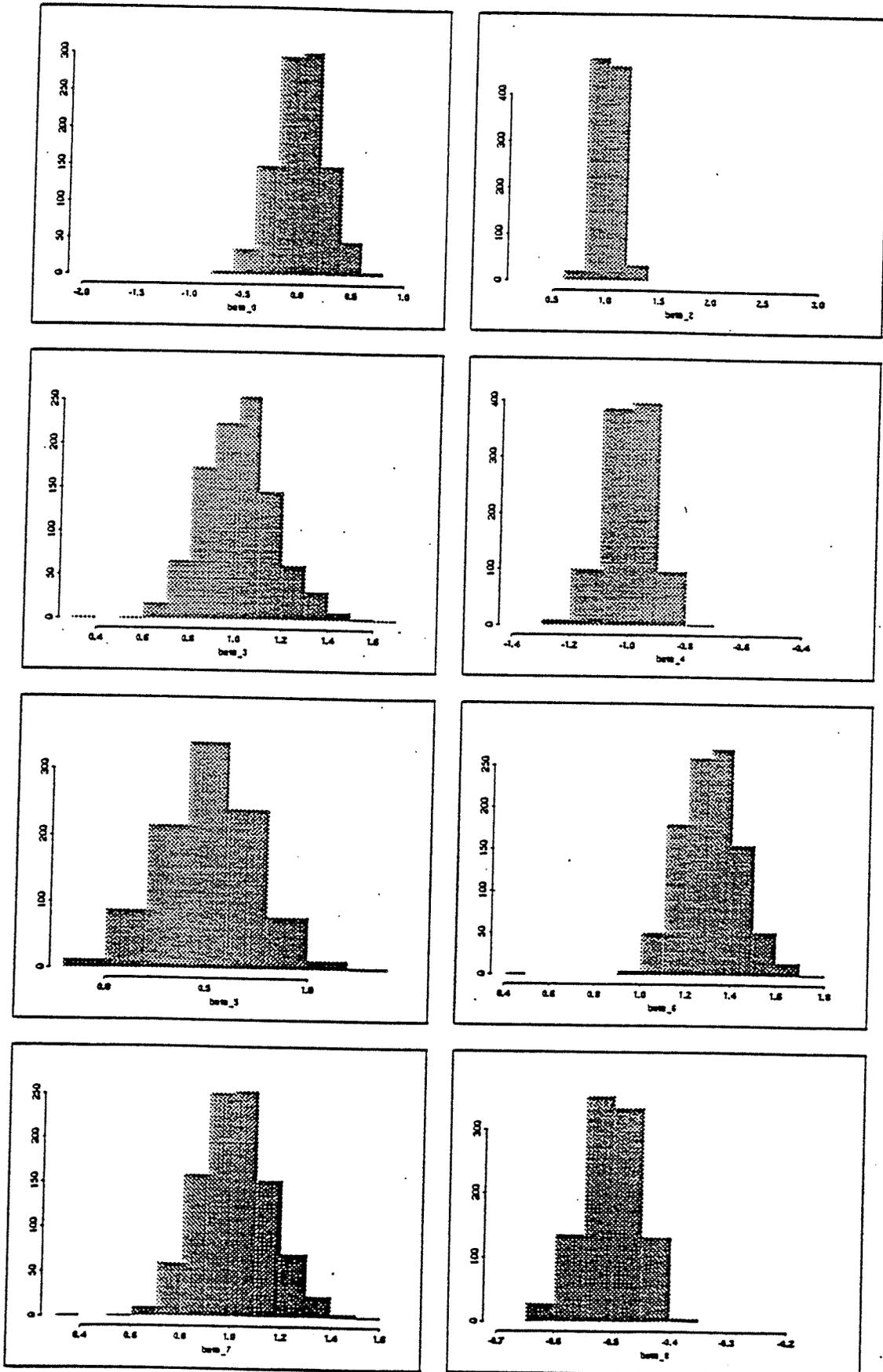


Figure 2. Histograms of betas in Experiment 3, 2,000 obs, 1,000 replications.



Appendix: Review of Previous Works

Early results on semi-parametric estimation of binary choice models were reported by Manski (1975, 1985), Cosslett (1983), Ichimura (1993), Horowitz (1992) and Duan and Li (1991). We will briefly describe their works and provide another interpretation of Manski's maximum score (MS) method in the context of the maximum likelihood principle.

The log-likelihood function for model (1) is

$$LnL(\beta, G) = \sum_{i=1}^n y_i LnG(x_i\beta) + (1 - y_i)Ln[1 - G(x_i\beta)] \quad (5)$$

where $G(x_i\beta) = P(y_i = 1)$, the probability of y_i being one. The relationship between $F(\cdot)$ and $G(\cdot)$ is $G(x) = 1 - F(-x)$. So $G(\cdot)$ is actually the cumulative density function for $-\epsilon$.

In Cosslett (1983), the maximization of (5) is carried out in two steps. First, for any fixed β^* , $LnL(\beta^*, G)$ is maximized with respect to G using the algorithm given by Ayer et al (1955). Denote the solution as $\hat{G}(\cdot|\beta^*)$. The concentrated function $LnL[\beta, \hat{G}(\cdot|\beta^*)]$ is then maximized with respect to β . Cosslett (1983) proves the consistency of his estimator by applying the results of Kiefer and Wolfowitz (1956). As noted in Cosslett (1983), the second step is computationally difficult since $LnL[\beta, \hat{G}(\cdot|\beta^*)]$ is a step function over the parameter space. The standard Newton-Raphson method can not be applied here. Manski's (1975) MS method suffers from the same weakness. Cosslett (1983) adopted the maximization procedure used by Manski (1975): a random set of orthogonal directions is selected in the parameter space and a linear search for the maximum is conducted along each of these directions in turn. This procedure is repeated until the change in the estimated parameters is less than some pre-assigned limit. This produces the trial maximum. However, one cannot be certain that the trial maximum is the actual maximum (Manski and Thompson, 1986.) As the dimension of β increases, the computation becomes extremely intensive. Computational difficulty and uncertainty about the

trial maximum are two major difficulties in applying Cosslett's or Manski's MS method in empirical work.

Manski's (1975) MS method can also be viewed as an application of maximum likelihood estimation method. The sample score function can be written as

$$S_n(\beta) = \sum_{i=1}^n y_i 1(x_i\beta \geq 0) + (1 - y_i)[1 - 1(x_i\beta \geq 0)],$$

where $1(x_i\beta \geq 0)$ is the indicator function. That is, $1(x_i\beta \geq 0) = 1$ if $x_i\beta \geq 0$ and zero otherwise. The MS estimator is defined as the β which maximizes $S_n(\beta)$.

This is actually equation (2) with $\ln G(x_i\beta)$ replaced by an indicator function $1(x_i\beta \geq 0)$ and $\ln[1 - G(x_i\beta)]$ replaced by $[1 - 1(x_i\beta \geq 0)]$. In this context, we can interpret the MS estimator as an improper ML estimator. The consistency of MS estimator is given by Manski (1985). Kim and Pollard (1990) establishes the asymptotic normality of the MS estimator and cubic root-n convergency.

Horowitz (1992) replaces the indicator function $1(x_i\beta \geq 0)$ by a smooth function $K_1(x_i\beta/\sigma_n)$ where $K_1(\cdot)$ is a continuous function satisfying some regularity conditions and $\lim_{n \rightarrow \infty} \sigma_n = 0$. The corresponding estimator is defined as the smoothed maximum score (SMS) estimator. Horowitz (1992) proves asymptotic normality for his SMS estimator. He also proves that the convergence rate is at least $n^{-2/5}$ and can be made arbitrarily close to $n^{-1/2}$, depending on the strength of certain smoothness assumptions.

Matzkin (1992) describes a fully nonparametric maximum likelihood method for estimating both the function $h(\cdot)$ of observable exogenous variables and the cumulative distribution $F(\cdot)$ of the random term. The estimator is shown to be strongly consistent under some regularity conditions. h and F are chosen such that they will minimize the following conditional log-likelihood function.

$$L(y^{(n)}, h, F) = \sum_{i=1}^n \{y_i \log(F(h(x_i))) + (1 - y_i) \log(1 - F(h(x_i)))\}.$$

The algorithm to minimize this function is very similar to the one described in Cosslett (1983). Computationally, it is very difficult.

Klein and Spady (1993) presents a non-parametric version of discriminant analysis of the binary choice model with the distribution of $x\beta$ given y being estimated by a kernel function. More precisely, $G(x_i\beta)$ is replaced by

$$G_n(v) = \frac{P_n g_n(v|y=1)}{P_n g_n(v|y=1) + (1 - P_n) g_n(v|y=0)},$$

where $P_n = n^{-1} \sum_{i=1}^n y_n$. Here $g_n(v|y=1)$ and $g_n(v|y=0)$ are defined as follows:

$$g_n(v|y=1) = (nP_n h_n)^{-1} \sum_{i=1}^n y_n K_2[(v - x_i\beta)/h_n],$$

$$g_n(v|y=0) = [n(1 - P_n)h_n]^{-1} \sum_{i=1}^n (1 - y_n) K_2[(v - x_i\beta)/h_n],$$

where $K_2(\cdot)$ is a kernel function, and h_n satisfy some conditions. Note that $G_n(\cdot)$ is no longer an increasing function of $x_i\beta$, i.e., $G_n(\cdot)$ is not a proper cumulative density function. Klein and Spady (1993) shows that their quasi-maximum-likelihood estimator is \sqrt{n} -consistent and asymptotically normal. There is no obvious data based method for selecting bandwidth h_n . One may use the cross-validation method to determine proper h_n .

Ichimura (1993) provides a semiparametric least square (SLS) estimator and a weighted SLS estimator for single-index models including the binary choice model as a special case. The method is a non-parametric version of the nonlinear least squares estimation. For binary choice models, SLS estimator minimizes

$$J_n(\beta) = \sum_{i=1}^n [y_i - G(x_i\beta)]^2,$$

where $G(x_i\beta)$ is replaced by a kernel estimator.

To adjust for the heteroskedasticity, a weighted SLS estimator is defined to minimize

$$J_n(\beta) = \sum_{i=1}^n W(x_i, \beta) [y_i - G(x_i\beta)]^2,$$

where W is a weighting function which is treated as a known function in Ichimura (1993). Ichimura proves the root- n consistency and asymptotic normality of both the SLS estimator and the WLS estimator. To actually compute SLS estimator, Ichimura uses the grid search method in his paper. This procedure also faces the problems of the multiple maxima and the computational complexity.

Ichimura and Thompson (1993) proposes a maximum likelihood estimation method for binary choice models with random coefficients of unknown distribution. In other word, a nonparametric maximum likelihood estimator of F_0 , the joint distribution of (β, ϵ) is given and shown to be consistent.

Duan and Li (1991) suggest slicing regression, which is very simple and computationally easy. However, this special case of inverse regression is very restrictive. Duan and Li consider the *inverse regression function*: $\xi(y) = E(x|y)$. One crucial requirement (design condition) for slicing regression is that the regressor variable (can be a vector) x is sampled randomly from a nondegenerate elliptically symmetric distribution. Under this condition, the inverse function is easy to estimate since y is a scalar. In fact, $\xi(y)$ is a linear combination of β and $\beta \propto \Sigma^{-1}(\xi(y) - \mu)$, where $\mu = E(x)$, $\Sigma = cov(x)$. It turns out that $\hat{\beta}$ is the principle eigenvector for some estimatable matrix $\hat{\Gamma}$. The computation is very simple and needs no iteration. Duan and Li show that the slicing regression estimator $\hat{\beta}$ is \sqrt{n} -consistent and asymptotically normal but not very efficient. However, this approach depends crucially on the design condition, which is rarely satisfied in practice. They advocate their estimator for use as the initial estimator for further analysis. Another problem associated with the inverse regression method is that this method can only identify the direction of β , not the magnitude of β , even for the multiple choice model.

Remark: a random vector has a elliptically symmetric distribution if its pdf belongs to $\mathcal{F}_0^n(\Sigma) = \{f \in \mathcal{F}^n | f(x) = |\Sigma|^{-1/2} q(x'\Sigma^{-1}x), \quad q \in Q_0\}$ where $Q_0 = \{q \mid q \text{ is a function on } [0, \infty)\}$ and Σ is any given positive definite matrix. One

special case is the normal distribution.