New Wine in Old Bottles: A Sequential Estimation Technique for the LPM

William C. Horrace* and Ronald L. Oaxaca

Revised March 2003

Abstract

The conditions under which ordinary least squares (OLS) is an unbiased and consistent estimator of the linear probability model (LPM) are unlikely to hold in many instances. Yet the LPM still may be the correct model or, perhaps, justified for practical reasons. A sequential least squares (SLS) estimation procedure is introduced that may outperform OLS in terms of finite sample bias and yields a consistent estimator. Monte Carlo simulations reveal that SLS outperforms OLS, probit and logit in terms of mean squared error of the predicted probabilities. An empirical example is provided.

Key Words: Linear Probability Model, Sequential Least Squares, Consistency, Monte Carlo

JEL Code: C25

*Corresponding Author: Center for Policy Research, 426 Eggers Hall, Syracuse University, Syracuse, NY 13244-1020. We gratefully acknowledge valuable comments by Seung Ahn, Badi Baltagi, Gordon Dahl, Dan Houser, Price Fishback, Art Gonzalez, Shawn Kantor, Alan Ker, Paul Ruud and Peter Schmidt. Capable research assistance was provided by Nidhi Thakur.

1. Introduction

The limitations of the Linear Probability Model (LPM) are well known. Estimated probabilities based on OLS are not necessarily bounded on the unit interval and unbiased (OLS) estimation implies that heteroscedasticity is present. Conventional textbook advice points to probit or logit as the standard remedy. These models bound the maximum likelihood estimated probabilities on the unit interval. However, the fact that consistent estimation of the LPM may be difficult does not imply that either probit or logit is the correct specification of the probability model. In some cases it is reasonable to assume that probabilities are generated from bounded linear decision rules. Theoretical rationalizations for the LPM can be found in Rosenthal (1989) and Heckman and Snyder (1997).

Despite the attractiveness of the logit and probit specifications for modeling a binary dependent variable as a function of covariates, OLS on the LPM remains a venerable model in the literature. Recent applications of OLS on the LPM include Klaassen and Magnus (2001), Bettis and Farlie (2001), Lukashin (2000), McGarry (2000), Fairlie and Sundstrom (1999), Lucking-Reiley (2000), and Currie and Gruber (1996). Empirical rationales for the LPM over probit or logit are plentiful. McGarry points to ease of interpretation of marginal effects in the LPM, while Lucking-Reiley cites a perfect correlation problem associated with the probit model for his particular application. Fairlie and Sundstrom prefer LPM because it implies a simple expression for the change in the unemployment rate between two censuses. Bettis and Farlie choose LPM because of an extremely large sample size and other simplifications implied by the linear model. Lukashin uses a linear probability model because it lends itself to a model selection algorithm based on an adaptive gradient criterion. While Currie and Gruber do not give any particular reason why they report LPM results in favor of probit or logit, they do mention that the logit/probit results are similar to the LPM results for their application.

Other rationales for the LPM are suggested by complications arising from the use of probit/logit models in certain contexts. One such occasion is the use of probit/logit models with panel data. Fixed effects and random effects estimation is much more involved in a logit model compared with a linear model. Likewise random effects estimation with a probit model is not as simple as with a linear model, and a fixed effects model cannot be consistently estimated with a probit model. Klaassen and Magnus point to these complications in selecting the LPM over logit or probit in their tennis application. Another rationale is perhaps justified in simultaneous equations/instrumental variable methods. The presence of dummy endogenous regressors is problematic if the DGP is assumed to be probit or logit; these problems were first considered by Heckman (1978). Suffice it to say that, while perhaps less popular than logit and probit, the LPM model still finds its way into the literature for various compelling reasons.

Some well-known LPM theorems are provided in Amemiya (1977), and the quintessential survey on binary dependent variables is Amemiya (1981). Standard econometrics textbooks, such as Greene (2000) and Kmenta (1997), point out LPM modelling complications that can lead to biased and inconsistent OLS estimates. Nevertheless, the literature is not sufficiently clear on the precise conditions under which OLS estimation yields problematic estimators of the parameters of the LPM. The purpose of this paper is: a) to rigorously lay out these conditions; b) to derive the finite-sample and asymptotic biases of OLS when they are present; and c) to provide additional results that shed light on the appropriateness or inappropriateness of OLS estimation of the LPM. Moreover, this paper proposes a consistent sequential estimation strategy that is functionally simpler than probit and logit in the sense that numerical optimization is not required to produce the LPM estimates. Since the estimator is based on a simple bias correction for OLS, the coefficients estimates produced are the true marginal effects and do not need to be transformed to yield readily interpretable results

(as is the case with the non-linear logit and probit specifications). Other potential applications of this estimator are explored later in the paper.

The plan of the paper is as follows. Section 2 provides a few theorems and results from OLS estimation of the LPM that have, heretofore, not been rigorously presented. Section 3 introduces a sequential estimation technique, Sequential Least Squares (SLS), for the LPM that is simple to implement. Section 4 presents a simulation study that compares results from the new sequential estimator to those from OLS and finds that the sequential estimator outperforms OLS in terms of finite sample bias. Section 5 performs a simulation study demonstrating that the sequential estimator outperforms probit and logit when the underlying data generation process is LPM. The metric of performance is the mean squared error of the predicted probabilities of the dependent variable. Section 6 presents an application to the choice of whether or not to purchase health insurance; the empirical results of OLS, SLS, logit and probit are compared. Section 7 summarizes and concludes.

2. LPM Specification and Main Results

A general way in which to specify the Data Generating Process (DGP) for the LPM is as follows. Let y_i be a binary random variable that takes on the values 0 or 1. Let x_i be a continuous random $1 \times k$ vector of explanatory variables on $\Re^k$, β be a $k \times 1$ vector of coefficients, and ε_i be an unobserved random error term. For convenience we will define the following probabilities over the random variable $x_i\beta \in \Re$.

$$\Pr(x_i\beta > 1) = \pi,$$
$$\Pr(x_i\beta \in [0,1]) = \gamma,$$
$$\Pr(x_i\beta < 0) = \rho,$$

where $\pi+\gamma+\rho = 1$. Consider a random sample of data: (y_i, x_i); $i \in N$; $N = \{1, ..., n\}$.

In what follows it will be useful to introduce the following notation. Define the sets

$$\begin{aligned} \kappa_\gamma &= \{i \mid x_i\beta \in [0,1]\}, \\ \kappa_\pi &= \{i \mid x_i\beta > 1\}. \end{aligned} \tag{1}$$

which are the sets of all indices i such that $x_i\beta$ is on the unit interval and of all indices i such that $x_i\beta > 1$, respectively. Notice that $\kappa_\gamma, \kappa_\pi \subseteq N$, $\kappa_\gamma \cap \kappa_\pi = \emptyset$. Equation (1) implies

$$\begin{aligned} &\Pr(i \in \kappa_\pi) = \pi, \\ &\Pr(i \in \kappa_\gamma) = \gamma, \\ &\Pr(i \notin \kappa_\gamma \cup \kappa_\pi) = \rho. \end{aligned} \tag{2}$$

Let the values of y_i be generated according to:

$$\begin{aligned} y_i &= 1 \text{ for } i \in \kappa_\pi, \\ &= \ x_i\beta + \varepsilon_i \text{ for } i \in \kappa_\gamma, \\ &= 0 \text{ otherwise.} \end{aligned} \tag{3}$$

(As convention dictates, we will assume the first element of the vector x_i is always a 1, so that the first element of β is an intercept term.) The conditional probability function for y_i is then

$$\begin{aligned} &\Pr(y_i = 1 \mid x_i,\ i \in \kappa_\pi) = 1, \\ &\Pr(y_i = 1 \mid x_i,\ i \in \kappa_\gamma) = \ x_i\beta\ , \\ &\Pr(y_i = 0 \mid x_i,\ i \in \kappa_\gamma) = 1 - \ x_i\beta, \\ &\Pr(y_i = 1 \mid x_i,\ i \notin \kappa_\gamma \cup \kappa_\pi) = 0. \end{aligned} \tag{4}$$

Therefore, y_i traces the familiar ramp function on $x_i\beta$, which can be thought of as the cumulative distribution function of a continuous uniform random variable on [0,1].

The DGP specification then implies the following error process:

$$\begin{aligned}\varepsilon_i &= 0 \text{ for } i \in \kappa_\pi, \\ &= y_i - x_i\beta \ , \ i \in \kappa_\gamma, \\ &= 0 \text{ for } i \notin \kappa_\gamma \cup \kappa_\pi.\end{aligned}$$

Notice that ε_i is not binary and is realized with the following conditional probabilities

$$\begin{aligned}&\Pr(\varepsilon_i = 0 \mid x_i, \ i \in \kappa_\pi) = 1, \\ &\Pr(\varepsilon_i = 1 - x_i\beta \mid x_i, \ i \in \kappa_\gamma) = x_i\beta \ , \\ &\Pr(\varepsilon_i = - x_i\beta \mid x_i, \ i \in \kappa_\gamma) = 1 - x_i\beta, \\ &\Pr(\varepsilon_i = 0 \mid x_i, \ i \notin \kappa_\gamma \cup \kappa_\pi) = 1.\end{aligned} \tag{5}$$

Estimation of the LPM typically proceeds by OLS on the model:

$$y_i = x_i\beta + u_i, \ i \in N,$$

where it is assumed that u_i is a zero-mean random variable that is independent of the x_i. The resulting estimator will be problematic as proven in the sequel. Notice that OLS specifies an error term u_i, which is different than ε_i :

$$\begin{aligned}u_i &= 1 - x_i\beta \text{ for } i \in \kappa_\pi, \\ &= y_i - x_i\beta \text{ for } i \in \kappa_\gamma, \\ &= -x_i\beta \text{ for } i \notin \kappa_\gamma \cup \kappa_\pi.\end{aligned}$$

The conditional probability function for u_i is

$$\begin{aligned}&\Pr(u_i = 1 - x_i\beta \mid x_i, \ i \in \kappa_\pi) = 1, \\ &\Pr(u_i = 1 - x_i\beta \mid x_i, \ i \in \kappa_\gamma) = x_i\beta \ , \\ &\Pr(u_i = - x_i\beta \mid x_i, \ i \in \kappa_\gamma) = 1 - x_i\beta, \\ &\Pr(u_i = - x_i\beta \mid x_i, \ i \notin \kappa_\gamma \cup \kappa_\pi) = 1.\end{aligned} \tag{6}$$

It is extremely important in what follows to distinguish between u_i, the OLS error, and ε_i, the error of the true DGP, for it is this distinction that induces problems with OLS of the LPM. Figure 1 illustrates this distinction: the first panel is the DGP for the conditional mean function, the second panel is the LPM error (ε_i), and the third panel is the OLS error (u_i). Notice that u_i can assume two different values: $1- x_i\beta$ and $- x_i\beta$, while ε_i can assume three: 0, $1- x_i\beta$ and $- x_i\beta$. The conditional probability functions of equations (4), (5) and (6) imply the following conditional expectations

$$
\begin{aligned}
&E(y_i \mid x_i,\ i \in \kappa_\pi) = 1, \qquad (7)\\
&E(y_i \mid x_i,\ i \in \kappa_\gamma) = \ x_i\beta\ ,\\
&E(y_i \mid x_i,\ i \notin \kappa_\gamma \cup \kappa_\pi) = 0,\\
&E(\varepsilon_i \mid x_i,\ i \in \kappa_\pi) = 0,\\
&E(\varepsilon_i \mid x_i,\ i \in \kappa_\gamma) = 0,\\
&E(\varepsilon_i \mid x_i,\ i \notin \kappa_\gamma \cup \kappa_\pi) = 0,\\
&E(u_i \mid x_i,\ i \in \kappa_\pi) = 1 - x_i\beta,\\
&E(u_i \mid x_i,\ i \in \kappa_\gamma) = \ 0,\\
&E(u_i \mid x_i, i \notin \kappa_\gamma \cup \kappa_\pi) = -x_i\beta.
\end{aligned}
$$

The expectations make clear the obvious difference between u_i and ε_i: u_i only has zero-mean when $i \in \kappa_\gamma$; ε_i always has zero-mean. This is also intuitively obvious in Figure 1.

Theorem 1 *If $\gamma < 1$, Ordinary Least Squares Estimation of the Linear Probability Model is generally biased and inconsistent.*

Proof. The conditional expectation of the usual OLS error is

$$\begin{aligned} E(u_i \mid x_i,\ i \in \kappa_\pi) &= 1 - x_i\beta, \\ E(u_i \mid x_i,\ i \in \kappa_\gamma) &= 0, \\ E(u_i \mid x_i, i \notin \kappa_\gamma \cup \kappa_\pi) &= -x_i\beta. \end{aligned}$$

Therefore, the conditional expectation of the OLS error, u_i, is a function of x_i with probability $(1-\gamma)$. Hence, OLS will be biased and inconsistent, if $\gamma < 1$. ■

The proof formalizes the specification error made when the OLS conditional mean is assumed. This fact has been mentioned by a few authors but has never be stated with any probabilistic rigor. The upshot of the theorem is that only those observations in the set κ_γ have a well-behaved error associated with them, so OLS that includes any observations outside of the set κ_γ will be problematic. In what follows we present a few additional results, derive the finite and asymptotic sample biases, and suggest a bias reduction method for OLS based on a sequential estimation strategy.

Remark 2 *If $\kappa_\gamma \neq N$, OLS estimation is biased and inconsistent . That is, if the sample used to estimate β contains any $i \notin \kappa_\gamma$, then γ is necessarily less than 1, so OLS is problematic.*

Of course the entire problem is due to $\gamma < 1$, so the follow is not surprising.

Remark 3 *If $\gamma = 1$, OLS is unbiased and consistent, because $\pi = \rho = 0$, $E(u_i \mid x_i) = 0$ for all $i \in N$, and the conditional expectation function implied by the DGP is:*

$$E(y_i \mid x_i) = \Pr(y_i = 1 \mid x_i) = x_i\beta,\ i \in N.$$

Therefore the usual OLS results hold under suitable regularity conditions.

Define discrete random variables z_i and w_i where:

$$\begin{aligned} z_i &= 1 \text{ for } i \in \kappa_\gamma, \\ &= 0 \text{ otherwise.} \\ w_i &= 1 \text{ for } i \in \kappa_\pi, \\ &= 0 \text{ otherwise.} \end{aligned}$$

So, $\Pr(z_i = 1) = \gamma$ and $\Pr(w_i = 1) = \pi$. Then an alternative specification of the DGP in equation (3) is:

$$y_i = w_i + z_i x_i \beta + u_i z_i; \quad i \in N \tag{8}$$

This specification is convenient, because it makes explicit the fact that u_i is not the correct error term associated with the DGP, instead $\varepsilon_i = u_i z_i$ is correct. It will also be useful in the sequel. Notice,

$$\begin{aligned} u_i z_i &= 0 \text{ for } i \notin \kappa_\gamma, \\ &= 1 - x_i\beta \text{ for } y_i = 1, \ i \in \kappa_\gamma, \\ &= -x_i\beta \text{ for } y_i = 0, \ i \in \kappa_\gamma. \end{aligned}$$

Moreover, the conditional probability function for $u_i z_i$ is the same as ε_i:

$$\begin{aligned} &\Pr(u_i z_i = 0 \mid x_i, \ i \in \kappa_\pi) = 1, \\ &\Pr(u_i z_i = 1 - x_i\beta \mid x_i, \ i \in \kappa_\gamma) = x_i\beta \ , \\ &\Pr(u_i z_i = -x_i\beta \mid x_i, \ i \in \kappa_\gamma) = 1 - x_i\beta, \\ &\Pr(u_i z_i = 0 \mid x_i, \ i \notin \kappa_\gamma \cup \kappa_\pi) = 1. \end{aligned}$$

It is evident that $E(u_i z_i \mid x_i) = 0$, so the specification in equation (8) has a zero-mean

error, which is independent of x_i. Taking the unconditional mean of equation (8):

$$\begin{aligned} E(y_i) &= \pi + E(z_i x_i)\beta + E(u_i z_i) \\ &= \pi + \gamma E(z_i x_i \mid z_i = 1)\beta + \gamma E(z_i u_i \mid z_i = 1) \\ &= \pi + \gamma E(x_i \mid z_i = 1)\beta + \gamma E(u_i \mid z_i = 1) \\ &= \pi + \gamma \mu_{x\gamma}\beta \end{aligned} \tag{9}$$

where

$$\begin{aligned} \mu_{x\gamma}\beta &= E\left(x_i \mid z_i = 1\right)\beta \\ &= E\left(x_i\beta \mid z_i = 1\right) \\ &= \int_0^1 x_i\beta f\left(x_i\beta \mid z_i = 1\ \right) d\left(x_i\beta\right) \\ &= \frac{1}{\gamma}\int_0^1 x_i\beta f\left(x_i\beta\right) d\left(x_i\beta\right), \end{aligned}$$

and $f(x_i\beta \mid z_i = 1)$ is the bounded conditional probability density and $f\left(x_i\beta\right)$ is the bounded marginal probability density of $x_i\beta$. Since $0 < \mu_{x\gamma}\beta < 1$, and $E\left(y_i\right)$ is a weighted average of 1, $\mu_{x\gamma}\beta$, and 0, it follows that $0 < E\left(y_i\right) < 1$. The mean function of equation (9) will be used in the sequel. Consider the OLS estimator:

$$\widehat{\beta}_n = \left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in N} x_i' y_i.$$

Substituting equation (8):

$$\widehat{\beta}_n = \left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in N} x_i'(w_i + z_i x_i\beta + u_i z_i). \tag{10}$$

The data can be partitioned into those $i \in \kappa_\gamma$, those $i \in \kappa_\pi$ and those that are in neither subset. Taking into consideration the values of z_i and w_i in these three

regimes:

$$\widehat{\beta}_n = \left[\sum_{i\in N} x_i' x_i\right]^{-1} \left[\sum_{i\notin \kappa_\gamma \cup \kappa_\pi} x_i'(0) + \sum_{i\in\kappa_\gamma} x_i'(x_i\beta + u_i) + \sum_{i\in\kappa_\pi} x_i'(1)\right]$$
$$= \left[\sum_{i\in N} x_i' x_i\right]^{-1} \left[\sum_{i\in\kappa_\gamma} x_i' x_i\beta + \sum_{i\in\kappa_\gamma} x_i' u_i + \sum_{i\in\kappa_\pi} x_i'\right]$$

Taking expectations conditional on x_i:

$$E(\widehat{\beta}_n|x_i) = \left[\sum_{i\in N} x_i' x_i\right]^{-1} \left[\sum_{i\in\kappa_\gamma} x_i' x_i\beta + \sum_{i\in\kappa_\gamma} x_i' E(u_i \mid x_i,\ i\in\kappa_\gamma) + \sum_{i\in\kappa_\pi} x_i'\right]$$
$$= \left[\sum_{i\in N} x_i' x_i\right]^{-1} \left[\sum_{i\in\kappa_\gamma} x_i' x_i\beta + \sum_{i\in\kappa_\gamma} x_i'(0) + \sum_{i\in\kappa_\pi} x_i'\right]$$
$$E(\widehat{\beta}_n|x_i) = \left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in\kappa_\gamma} x_i' x_i\beta + \left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in\kappa_\pi} x_i' \neq \beta, \tag{11}$$

which is generally biased because $\gamma < 1$. The bias will persist asymptotically. When $\gamma = 1$, $\kappa_\gamma = N$, the first term on the RHS reduces to β, the second term on the RHS goes to 0, and $\widehat{\beta}_n$ is unbiased.

The inconsistency of $\widehat{\beta}_n$ follows in a similar fashion. Letting C denote the cardinality operator, define $n_\pi = C(\kappa_\pi)$, $n_\gamma = C(\kappa_\gamma)$ and $n_\rho = n - n_\pi - n_\gamma$. Let plim denote the probability limit operator as $n \to \infty$. Assume $\text{plim}\left[n^{-1}\sum_{i\in N} x_i' x_i\right] = Q$ and $\text{plim}\left[n_\gamma^{-1}\sum_{i\in\kappa_\gamma} x_i' x_i\right] = Q_\gamma$ where Q and Q_γ are finite, (non singular) positive definite. Assume $\text{plim}\left[n_\pi^{-1}\sum_{i\in\kappa_\pi} x_i'\right] = \mu_{x\pi}'$, $\text{plim}\left[n^{-1}\sum_{i\in N} x_i'\right] = \mu_x'$ and $\text{plim}\left[n_\gamma^{-1}\sum_{i\in\kappa_\gamma} x_i' u_i\right] = 0$, where $\mu_{x\pi}'$ and μ_x' are finite vectors. Assume $\text{plim}[n^{-1}n_\pi] = \pi$ and $\text{plim}[n_\gamma n^{-1}] = \gamma$. Then it is easy to show that

$$\text{plim}(\widehat{\beta}_n) = Q^{-1}\left(Q_\gamma\beta\gamma + \pi\mu_{x\pi}'\right) \neq \beta.$$

This probability limit seems to imply that even if γ and π were known, $\widehat{\beta}_n$ could not be bias-corrected. Yet, the unconditional mean of y_i in equation (9) seems to imply

that if γ and π were known, an OLS regression of $(y_i - \pi)$ on (γx_i) might produce an unbiased estimate. Define the OLS estimator from such a transformed regression as:

$$\widehat{\beta}_n^* = \left[\sum_{i\in N} \gamma^2 x_i' x_i\right]^{-1} \sum_{i\in N} \gamma x_i' \left(y_i - \pi\right). \tag{12}$$

Theorem 4 *$\widehat{\beta}_n^*$ is biased and inconsistent for β.*

Proof. After some algebra, equation (12) implies

$$\begin{aligned}
\widehat{\beta}_n^* &= \frac{1}{\gamma}\left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in N} x_i' y_i - \frac{1}{\gamma}\left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in N} x_i' \pi \\
&= \frac{1}{\gamma}\hat{\beta}_n - \frac{\pi}{\gamma}\left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in N} x_i'.
\end{aligned}$$

Taking expectations

$$\begin{aligned}
E(\widehat{\beta}_n^*|x_i) &= \frac{1}{\gamma}E(\hat{\beta}_n|x_i) - \frac{\pi}{\gamma}\left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in N} x_i'. \\
&= \frac{1}{\gamma}\left\{\left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in \kappa_\gamma} x_i' x_i \beta + \left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in \kappa_\pi} x_i'\right\} \\
&- \frac{\pi}{\gamma}\left[\sum_{i\in N} x_i' x_i\right]^{-1} \sum_{i\in N} x_i' \\
&\neq \beta.
\end{aligned}$$

Thus, knowledge of π and γ to estimate $E\left(y_i|x_i\right)$ by OLS does not in general lead to an unbiased estimator of β. Moreover it does not lead to consistent estimation by OLS:

$$\begin{aligned}
\text{plim}(\widehat{\beta}_n^*) &= \frac{1}{\gamma}\text{plim}\left(\hat{\beta}_n\right) - \frac{\pi}{\gamma}\left[\text{plim}\left(n^{-1}\sum_{i\in N} x_i' x_i\right)\right]^{-1} \text{plim}\left(n^{-1}\sum_{i\in N} x_i'\right) \\
&= \frac{1}{\gamma}\left[Q^{-1}\left(Q_\gamma \beta\gamma + \mu_{x\pi}'\pi\right)\right] - \frac{\pi}{\gamma}Q^{-1}\mu_x' \\
&= Q^{-1}\left[Q_\gamma \beta + \left(\mu_{x\pi}' - \mu_x'\right)\frac{\pi}{\gamma}\right] \neq \beta.
\end{aligned}$$

■

The bias will persist asymptotically. The problem with the estimators $\widehat{\beta}_n$ and $\widehat{\beta}_n^*$ is not that γ and π are unknown but that κ_γ is unknown, for if we knew κ_γ, we could perform OLS only on those observations therein contained.

Remark 5 *Therefore, sufficient information for unbiased and consistent OLS estimation is knowledge of* κ_γ.

Also notice that if $\kappa_\gamma = N$, then:

$$\sum_{i \in \kappa_\gamma} x_i' x_i = \sum_{i \in N} x_i' x_i, \text{ and } \sum_{i \in \kappa_\pi} x_i' = 0.$$

Therefore, equation (11), reduces to:

$$E(\widehat{\beta}_n | x_i) = \left[\sum_{i \in N} x_i' x_i\right]^{-1} \sum_{i \in N} x_i' x_i \beta + \left[\sum_{i \in N} x_i' x_i\right]^{-1} 0,$$
$$E(\widehat{\beta}_n | x_i) = \beta,$$

unbiased for $\kappa_\gamma = N$. A similar argument can be made to show the consistency of this estimate. Of course if $\gamma = 1$, then $\kappa_\gamma = N$.

Remark 6 *Therefore, without knowledge of* κ_γ *and* κ_π*, a sufficient condition for unbiased OLS estimation when* $\gamma < 1$ *is* $\kappa_\gamma = N$.

$\kappa_\gamma = N$ is a weaker sufficient condition than $\gamma = 1$, but probably unlikely in reasonably large samples. For any given random sample (y_i, x_i); $i \in N$, the $\Pr[\kappa_\gamma = N] = \gamma^n$, so

$$\lim_{n \to \infty} \Pr[\kappa_\gamma \neq N] = \lim_{n \to \infty} (1 - \gamma^n) = 1.$$

Remark 7 *Therefore, without knowledge of* κ_γ *and* κ_π*, if* $\gamma < 1$ *and* $\kappa_\gamma = N$*, then as* $n \to \infty$*,* $\kappa_\gamma \neq N$ *with probability approaching 1, and* $\widehat{\beta}_n$ *is asymptotically biased and inconsistent.*

It should be noted that as the sample size grows, once the first observation $x_i\beta \notin [0,1]$ appears in N then $\kappa_\gamma \neq N$ and finite sample unbiasedness is lost also. Oddly enough the estimator $\widehat{\beta}_n$ could, under the right conditions, be reliable in small samples and unreliable in large samples. If we had knowledge of the sets κ_γ and κ_π, then a consistent estimate of β could be based on the sub-sample:

$$\widehat{\beta}_{\kappa_\gamma} = \left[\sum_{i\in\kappa_\gamma} x_i'x_i\right]^{-1} \sum_{i\in\kappa_\gamma} x_i'y_i, \text{ for } \kappa_\gamma \text{ and } \kappa_\pi \text{ known,}$$

$$\widehat{\beta}_{\kappa_\gamma} = \left[\sum_{i\in\kappa_\gamma} x_i'x_i\right]^{-1} \sum_{i\in\kappa_\gamma} x_i'(w_i + z_ix_i\beta + u_iz_i),$$

$$\widehat{\beta}_{\kappa_\gamma} = \left[\sum_{i\in\kappa_\gamma} x_i'x_i\right]^{-1} \sum_{i\in\kappa_\gamma} x_i'(x_i\beta + u_i),$$

$$E(\widehat{\beta}_{\kappa_\gamma}|x_i) = \beta, \text{ for } \kappa_\gamma \text{ known.}$$

This is tantamount to removing the observations $i \notin \kappa_\gamma$. Then a consistent estimate of γ is $\widehat{\gamma} = C(\kappa_\gamma)/n$, and a consistent estimate of π is $\widehat{\pi} = C(\kappa_\pi)/n$.

3. Sequential Least Squares

Based on the problems associated with OLS on the LPM, it is clear that an alternative estimation approach is warranted. One could certainly envision myriad sophisticated estimators that would be an improvement over OLS: an MLE technique that estimated γ and π as well as β, a non-linear search algorithm that recognizes the constraint $x_i\beta \in [0,1]$, some sort of splines technique that estimates the $x_i\beta = 0$ and $x_i\beta = 1$ break points, etc.. However, our interest is to salvage OLS not to discard it, so we now present a simple OLS correction technique.

If somehow the observations $i \notin \kappa_\gamma$ could be eliminated sequentially, then as the elimination sequence grew: N would decrease to some set of observations that was a

subset of κ_γ, while $\kappa_\pi \to \emptyset$, then $\widehat{\beta}_n$ would converge in some probabilistic sense to β. Therefore, an empirical strategy could involve finding a $\widehat{\beta}_{i\in\kappa_\gamma}$estimate that ensures that the predicted dependent variable is on the unit interval. One specific estimation strategy is to identify the empirical subsets

$$\widehat{\kappa}_\gamma^{(1)} = \{i \mid i \in N \cap x_i\widehat{\beta}_n \in [0,1]\},$$
$$\widehat{\kappa}_{\pi\cup\rho}^{(1)} = \{i \mid i \in N \cap x_i\widehat{\beta}_n \notin [0,1]\},$$

with cardinality $C(\widehat{\kappa}_\gamma^{(1)}) = n_\gamma^{(1)} \leq n$ and $C(\widehat{\kappa}_{\pi\cup\rho}^{(1)}) = n_{\pi\cup\rho}^{(1)} \leq n$. Then a subsample estimate of β is obtained as

$$\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(1)}} = \left[\sum_{i\in\widehat{\kappa}_\gamma^{(1)}} x_i'x_i\right]^{-1} \sum_{i\in\widehat{\kappa}_\gamma^{(1)}} x_i'y_i.$$

Define the subset

$$\widehat{\kappa}_\gamma^{(2)} = \{i \mid i \in \widehat{\kappa}_\gamma^{(1)} \cap x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(1)}} \in [0,1]\},$$
$$\widehat{\kappa}_{\pi\cup\rho}^{(2)} = \{i \mid i \in \widehat{\kappa}_\gamma^{(1)} \cap x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(1)}} \notin [0,1]\},$$

with cardinality $C(\widehat{\kappa}_\gamma^{(2)}) = n_\gamma^{(2)} \leq n_\gamma^{(1)}$ and $C(\widehat{\kappa}_{\pi\cup\rho}^{(2)}) = n_{\pi\cup\rho}^{(2)} \leq n_\gamma^{(1)}$. Then a second subsample estimate is

$$\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(2)}} = \left[\sum_{i\in\widehat{\kappa}_\gamma^{(2)}} x_i'x_i\right]^{-1} \sum_{i\in\widehat{\kappa}_\gamma^{(2)}} x_i'y_i.$$

We can repeat this process in general:

$$\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} = \left[\sum_{i\in\widehat{\kappa}_\gamma^{(j)}} x_i'x_i\right]^{-1} \sum_{i\in\widehat{\kappa}_\gamma^{(j)}} x_i'y_i,\ j = 1, ..., J$$

for

$$\widehat{\kappa}_\gamma^{(j)} = \{i \mid i \in \widehat{\kappa}_\gamma^{(j-1)} \cap x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j-1)}} \in [0,1]\},$$
$$\widehat{\kappa}_{\pi\cup\rho}^{(j)} = \{i \mid i \in \widehat{\kappa}_\gamma^{(j-1)} \cap x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j-1)}} \notin [0,1]\},\ j = 2, ..., J$$

with cardinality $C(\widehat{\kappa}_{\gamma}^{(j)}) = n_{\gamma}^{(j)} \leq n_{\gamma}^{(j-1)}$ and $C(\widehat{\kappa}_{\pi\cup\rho}^{(j)}) = n_{\pi\cup\rho}^{(j)} \leq n_{\gamma}^{(j-1)}$until convergence in the sense that all the observations in the final subsample satisfy $x_i\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(J)}} \in [0,1]$. The final estimate will be

$$\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(J)}} = \left[\sum_{i\in\widehat{\kappa}_{\gamma}^{(J)}} x_i'x_i\right]^{-1} \sum_{i\in\widehat{\kappa}_{\gamma}^{(J)}} x_i'y_i. \tag{13}$$

Call this the Sequential Least Squares (SLS) estimator. The convergence of the sequential estimators imply

$$\frac{n_{\gamma}^{(j)}}{n_{\gamma}^{(j-1)}} \to 1 \text{ and } n_{\pi\cup\rho}^{(j)} \to 0 \tag{14}$$

as $n \to \infty$ and $j \to J$. This condition must hold in order for the trimming to converge in any meaningful way, or else the entire sample would ultimately be discarded.

Theorem 8 *If prediction error $x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta)$ is a continuous random variable on $\Re$ and if $\Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} \in [0,1]\} \to 1$ as $n \to \infty$ and as $j \to J$, then the SLS estimator $\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(J)}}$ is consistent for β.*

Proof. See Appendix. ■

It is not entirely clear when the condition $\Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} \in [0,1]\} \to 1$ as $n \to \infty$ and as $j \to J$ will hold, since the probability in question is a function of the random variable x_i, whose distribution is unknown, in general. However, Horrace and Oaxaca (2001) show that under certain conditions normality of x_i is sufficient to ensure $\Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} \in [0,1]\} \to 1$, so the set of possible distributions of x_i that satisfies the convergence conditions is certainly not empty. In the sequel we perform a simulation study that examines the extent to which the SLS estimator outperforms OLS in terms of sample bias.

Define sets based on the final estimator $\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}}$:

$$\begin{aligned}
\widehat{\kappa}_\gamma^* &= \{i \mid x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} \in [0,1]\},\\
\widehat{\kappa}_\pi^* &= \{i \mid x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} > 1\},\\
\widehat{\kappa}_\rho^* &= \{i \mid x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} < 0\},\\
\widehat{\kappa}_{\pi\cup\rho}^* &= \widehat{\kappa}_\pi^* \cup \widehat{\kappa}_\rho^*\\
&= \{i \mid x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} \notin [0,1]\}.
\end{aligned}$$

Notice the difference between $\widehat{\kappa}_\gamma^*$ and $\widehat{\kappa}_\gamma^{(J)}$, and between $\widehat{\kappa}_{\pi\cup\rho}^*$ and $\widehat{\kappa}_{\pi\cup\rho}^{(J)}$. $\widehat{\kappa}_\gamma^*$, and $\widehat{\kappa}_{\pi\cup\rho}^*$ are based on the entire sample N and $\widehat{\kappa}_\gamma^{(J)}$ and $\widehat{\kappa}_{\pi\cup\rho}^{(J)}$ are based on the subsample $\widehat{\kappa}_\gamma^{(J-1)}$. Then insofar as $\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}}$ is consistent for β, consistent estimates of the probabilities γ, π, and ρ can be obtained from

$$\begin{aligned}
\tilde{\gamma} &= \frac{C(\widehat{\kappa}_\gamma^*)}{n}\\
\tilde{\pi} &= \frac{C(\widehat{\kappa}_\pi^*)}{n}\\
\tilde{\rho} &= 1 - \tilde{\pi} - \tilde{\gamma},
\end{aligned}$$

Note that if $n_\gamma^{(1)} = n$, then no trimming is necessary, $N = \widehat{\kappa}_\gamma^{(J)} = \widehat{\kappa}_\gamma^*$ and $\tilde{\gamma} = 1$. Clearly, this estimator should only be used if the sample size is large, since observations $x_i\beta \in [0,1]$ will be trimmed with positive probability. When the final SLS estimator is used to predict y_i we are assured that $\widetilde{y}_i = x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} \in [0,1]$ for $i \in \widehat{\kappa}_\gamma^*$, however this will not necessarily be the case for all $\widetilde{y}_i$, $i \in N$. As is usually the case, prediction of y_i can be performed as follows:

$$\begin{aligned}
&\widetilde{y}_i = 1 \text{ for } x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} > 1 \text{ for } i \in N\\
&\widetilde{y}_i = x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} \text{ for } x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} \in [0,1] \text{ for } i \in N\\
&\widetilde{y}_i = 0 \text{ for } x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} < 0 \text{ for } i \in N.
\end{aligned}$$

We now present a brief simulation study that demonstrates that the SLS trimming estimator $\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}}$ is generally less biased that the OLS estimator $\widehat{\beta}_n$.

4. Simulation Study: SLS Versus OLS

A simulation study was conducted to assess the performance of the SLS estimator against the OLS estimator in terms of finite sample bias. Initially, we are concerned with understanding the nature of the OLS bias when $\gamma < 1$ and seeing if the SLS estimator is an improvement over OLS. To assess estimator performance for different values of γ, requires selecting γ, π and β then finding an appropriate multivariate distribution for the x_i to generate data such that $\Pr\{x_i\beta \in [0,1]\} = \gamma$ and $\Pr\{x_i\beta > 1\} = \pi$. For x_i with large dimensionality this would be a monumental task, therefore we restrict attention to the bivariate model

$$\begin{aligned} y_i &= 1 \text{ for } i \in \kappa_\pi, \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i \text{ for } i \in \kappa_\gamma, \\ &= 0 \text{ otherwise,} \end{aligned}$$

where β_0, β_1 and x_i are scalars. We also assume that x_i has a normal distribution with mean μ and variance σ. Given these restrictions on the data generation process, it is a relatively simple procedure to select γ, π, β_0 and β_1, and then to calculate μ and σ, such that $\Pr\{\beta_0 + \beta_1 x_i \in [0,1]\} = \gamma$ and $\Pr\{\beta_0 + \beta_1 x_i > 1\} = \pi$. To generate data in this way, notice that for $\beta_1 > 0$

$$\Pr\{\beta_0 + \beta_1 x_i \in [0,1]\} = \Pr\left\{g_i \in [\underline{c}, \overline{c}]\right\},$$

$$\begin{aligned} \underline{c} &= \frac{-\beta_0 - \beta_1\mu}{\beta_1\sigma}, \\ \overline{c} &= \frac{1 - \beta_0 - \beta_1\mu}{\beta_1\sigma}, \end{aligned}$$

where g_i is a standard normal random variate. Given γ, π, β_0 and β_1, the necessary μ and σ can be calculated by solving

$$\Phi^{-1}(1-\gamma-\pi) = \frac{-\beta_0 - \beta_1\mu}{\beta_1\sigma},$$
$$\Phi^{-1}(1-\pi) = \frac{1-\beta_0 - \beta_1\mu}{\beta_1\sigma},$$

where Φ^{-1} is the inverse cumulative distribution function of a standard normal random variate. That is,

$$\mu = -\frac{\beta_0}{\beta_1} - \frac{\Phi^{-1}(1-\gamma-\pi)}{\beta_1\left[\Phi^{-1}(1-\pi) - \Phi^{-1}(1-\gamma-\pi)\right]}$$
$$\sigma = \frac{1}{\beta_1\left[\Phi^{-1}(1-\pi) - \Phi^{-1}(1-\gamma-\pi)\right]}$$

As a practical matter, the sample size, n, must be fairly large to ensure that the SLS procedure doesn't trim the entire sample, (i.e. $\widehat{\kappa}_\gamma^{(J)} = \emptyset$). Since the SLS estimator trims realizations $x_i\beta \in [0,1]$ with positive probability, if there are only a few of these observations in a small sample, then they may all be trimmed during the procedure. Large n is not an unreasonable restriction to impose on the study, since the SLS estimator should only be used in situations where the sample size is fairly large. Therefore we select sample sizes of n = 500, 1000 and 2000. Simulation iterations were arbitrarily set at 100 to calculate the empirical bias of the OLS and SLS estimates of β_0 and β_1. Parameter values were arbitrarily selected as $\gamma = 0.25$ and 0.75, $\pi =$ 0.10, 0.20, β_0 = -0.50, 0.50, and β_1 = 1.00, 2.00. Results are contained in Tables 1 - 4. In all tables the OLS estimators are $\widehat{\beta}_0$ and $\widehat{\beta}_1$, while the SLS estimators are $\tilde{\beta}_0$ and $\tilde{\beta}_1$.

There were 48 simulations runs in total, and in all but 5 cases the SLS estimators had lower magnitude of bias. In those 5 cases where OLS was better the difference in magnitude of bias was at the fourth decimal place. For example in Table 1, the last row shows $Bias(\widehat{\beta}_0) = -0.0010$ and $Bias(\tilde{\beta}_0) = -0.0011$, so the difference in

magnitude of the biases is only 0.0001. In most case where SLS was superior the difference in biases was large (especially when γ was small). For example in the first row of Table 1, we see that $Bias(\widehat{\beta}_0) = 0.2456$ and $Bias(\tilde{\beta}_0) = -0.0151$. This is not an atypical difference. The magnitude of the bias of the SLS is generally decreasing in n, although this is not always the case. However, when the bias does increase as n increases, the increases are small and probably due to sampling variability and not a lack of inconsistency. For example, in Table 3 for the first three rows, $Bias(\tilde{\beta}_1) = 0.0554, 0.0132, 0.0147$ for $n = 500, 1000, 2000$, respectively. Some general observations concerning the OLS estimates are: a) the OLS biases persist as n gets large, b) the OLS biases are larger when γ is smaller (especially for $\widehat{\beta}_1, \gamma$ seems to affect the slope parameter more then the intercept), and c) the bias of the OLS intercept is larger in magnitude when π is larger (i.e. π tends to affect the OLS intercept). It is evident from the simulations that the OLS bias for β_1 is equal to $\beta_1(\gamma - 1)$ which implies $\text{plim}\hat{\beta}_1 = \beta_1\gamma$. In this case knowing γ would lead to a simple consistent estimator of the slope parameter.[1]

The simulations certainly suggest that the SLS estimator generally outperforms the OLS estimator in terms of estimation bias. Figure 2 depicts this bias reduction of SLS over OLS. For the purposes of illustration, this figure contains a single simulation run where $\beta_0 = 0$, $\beta_1 = 1$, $\gamma = 0.75$, and $\pi = 0.10$. The heavy line represents the fitted values for SLS, the medium line are the fitted values for OLS, and the light line is the true LPM data generating process. Clearly, the SLS fitted values reflect a smaller bias than the OLS fitted values.

[1] This is a special case that follows from the assumption of a normal distribution on the $x's$ in a simple regression model. Horrace and Oaxaca (2001) prove the special case and derive the more complicated bias for the OLS estimator for the constant term β_0. An alternative simulation study was performed in which x_i had a uniform distribution. The results, available upon request, show that SLS outperforms OLS in terms of finite sample bias of the estimates.

5. Simulation Study: SLS versus Logit and Probit

A simulation study was conducted to assess the performance of the SLS estimator against Logit and Probit when the underlying data generation process is LPM. Since logit and probit are commonly employed in dependent variable econometric analysis, a comparison seemed necessary even though it is presumed that the DGP is LPM, implying that logit and probit are misspecified models. Such a study presents several problems, but the largest problem stems from the fact that probit and logit are non-linear estimations techniques while SLS and the LPM are linear. Therefore, while bias comparisons of OLS and SLS to the LPM seem natural, bias comparisons of logit and probit to the LPM are not readily forthcoming. Specifically, for logit and probit the estimates of the marginal effect β_1 are functions of the value of x_i, which raises the question of how one should assess the bias of logit or probit on the LPM, which has a constant slope on $x_i\beta \in (0,1)$. Because evaluation of the probit/logit marginal effects at the mean of x_i is so popular among empiricists, we initially sought to evaluate the bias of the marginal effects at the mean. However, logit and probit tended to be highly biased at the mean of x_i. For example, a typical simulation run is depicted in Figure 3. Here, $\beta_0 = 0$, $\beta_1 = 1$, $\gamma = 0.75$, $\pi = 0.10$, and the empirical mean of $x_i = 0.508$. Since $\beta_1 = 1$ and $\beta_0 = 0$, the x-axis in the figure is x_i. It is clear from the figure that near the mean of x_i (about 0.50), probit is upward biased relative to SLS. Therefore, bias at the mean was discarded as a means of evaluating logit and probit on the LPM. However, it is useful to point out that the practice of evaluating the marginal effects at the mean of x_i for logit and probit, may lead to biased results when an LPM DGP is suspected.

An alternative evaluation criterion that could be equally applied to SLS and probit or logit is the mean squared error of the predicted probabilities for SLS and probit or logit. That is, the simulated LPM data implied a known probability that $y_i = 1$, which could be compared to the predicted probabilities of SLS, logit and probit. Predicted

probability errors for the sample could then be squared, summed and averaged to produce an estimate of the mean squared error of the prediction probabilities. For SLS (and OLS which was ultimately included in the comparison), predicted probabilities greater than 1 where converted to 1 and those less than 0 where converted to 0 before constructing the MSE estimates.

To do this, we selected only two sample sizes of $n = 500$ and 1000. Simulation iterations were again set at 100 to calculate the MSE of the predicted probabilities of the OLS, probit, logit and SLS. Again we assumed that x_i was normally distributed. (An alternative simulation study was also performed assuming that x_i followed a uniform distribution; the results, which are not reported, were similar to the present study.) Parameter values were again selected as $\gamma = 0.25$ and 0.75, $\pi = 0.10$, 0.20, β_0 = -0.50, 0.50, and β_1 = 1.00, 2.00 (as in the simulation study of section 4). This implied 32 simulation runs, the results of which are contained in Tables 5, 6, 7 and 8. In all cases probit and logit are superior to OLS, and SLS is superior to probit and logit. The magnitude of the MSE for all models is generally unaffected by the values of β_0 and β_1; for example the first and third rows of Table 5 are very similar (changing β_1 ceteris paribus) as are the first and fifth (changing β_0 ceteris paribus). Not surprisingly, increasing the sample size tends to decrease the MSE for all models. This is always the case for the consistent SLS procedure (and for probit and logit), but is not always true for the inconsistent OLS procedure. Notice that as γ decreases from Table 5 to Table 7, the MSE of OLS tends to significantly increase (by a magnitude of ten-fold), while the MSE of SLS generally does not. This seems to reflect an increased OLS bias associated with smaller γ (remember OLS and SLS are equivalent when $\gamma = 1$). Also notice that as γ decreases from Table 5 to Table 7, the MSE of probit and logit seems to decrease. This is not surprising since logit and probit perform better for extreme values than for median values of $x_i\beta$ when the DGP is LPM, and since smaller γ implies that a greater proportion of the observed $x_i\beta$ will

be extreme. For median values of $x_i\beta$, probit and logit are more highly misspecified, so the larger γ of Table 5 produces more $x_i\beta$ near the median values and, hence, large MSE for probit and logit.[2]

6. Application

An brief empirical example is presented below that illustrates the SLS approach and contrasts its results with OLS LPM, probit, and logit. The data are taken from the NLSY79 and pertain to 1998. The binary event is whether or not an individual is covered by health insurance:

$hins_i = 1$ if the individual is currently covered by health insurance

$= 0$ otherwise.

We define the index function as follows:

$\mathrm{I}_i = \beta_0 + \beta_1 black_i + \beta_2\, other_i + \beta_3 female_i + \beta_4 ntinc_i + \beta_5\,(ntinc)_i^2 + \beta_6 emp_i$, where *black* and *other* are race dummy variables for blacks and other nonwhites, $female$ is a dummy variable for gender, *ntinc* is household net income ($1,000's), and *emp* is a dummy variable for current employment. The LPM specification is given by $hins_i = \mathrm{I}_i + \varepsilon_i$ and the probit and logit specifications are described by $\text{prob}(hins_i = 1) = \text{prob}\,(\mathrm{I}_i + \varepsilon_i \geq 0)$. We examine two variations on the SLS method. The first variation is to use the White Heteroscedastic-Consistent Variance/Covariance matrix to produce robust standard errors for the SLS estimates. Another variation is to employ FGLS with weights $1/\left[(\hat{y}_i)\,(1-\hat{y}_i)\right]$, where $\hat{y}_i$ is the predicted value of $hins_i$ from the SLS estimated equation.

The results are reported in Table 9. The SLS estimator converged on the 10th iteration and trimmed the sample from 6,860 observations to 4,302. Perhaps the most obvious difference between the SLS methods and the other estimators is that

[2]Based on simulations in which x_i was drawn from a uniform distribution, SLS outperformed OLS, logit, and probit in terms of mean squared error of the predicted probabilities. These results are available upon request.

the variable *black* is estimated by the SLS methods to have a positive and statistically significant effect on the probability of having health insurance, whereas the estimated effects of this variable are not statistically significant in the OLS, Probit, and Logit models. The coefficients on the remaining variables retain their signs and statistical significance across the different models. The robust standard errors for the SLS model are very nearly the same as the SLS estimated standard errors. In the case of the SLS/FGLS model, the estimated coefficients are about the same as the SLS estimates, though the estimated standard errors are uniformly smaller than the SLS standard errors and the robust standard errors. As a check we confirmed that the SLS/FGLS model yielded probability predictions that remained bounded in the unit interval.

In Table 10 we report the sample means for the full sample and the trimmed sample. It is clear from a comparison of the two sets of means that violations of the unit interval boundary conditions for the predicted probabilities obtained from OLS estimation of the LPM model are associated with being white, female, higher net family income, and currently employed.

7. Concluding Remarks

Although it is theoretically possible for OLS estimation of the LPM to yield unbiased estimators conditional on the sample, this generally would require fortuitous circumstances. Furthermore, consistency of OLS is shown to be an exceedingly rare occurrence as one would have to accept extraordinary restrictions on the joint distribution of the regressors. Therefore, OLS is frequently a biased estimator and almost always an inconsistent estimator of the LPM. Despite estimation difficulties, the LPM is still frequently used in modeling probabilities. This is partly due to theoretical arguments that justify the linear specification and partly due to the ease of using OLS to estimate the model.

In this paper an alternative estimation strategy has been introduced (SLS) that is

fairly easy to implement and offers the promise of significantly reducing the bias from OLS. The conditions under which SLS is consistent are rigorously derived. Monte Carlo simulations with a two parameter LPM support our conjectures about the bias reducing properties of SLS. These simulations also point to the persistence of OLS bias as the sample size increases, which is to be expected when OLS is not consistent. Monte Carlo simulations also suggest that SLS outperforms logit and probit when the DGP is LPM.

Standard errors have not yet been derived for SLS that incorporate the statistical effect of the sequential trimming procedure. It is clear that SLS is not efficient. The absence of rogue predictions of y_i outside of the unit interval at the outset implies heteroscedastic errors. Conditioning on the final trimmed sample $\widehat{\kappa}_{\gamma}^{(J)}$, one could use a feasible GLS estimator that weights the observations by $1/\sqrt{(\tilde{y}_i)\,(1-\tilde{y}_i)}$, where $\tilde{y}_i = x_i\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(J)}} \in [0,1]$. The estimated standard errors would be obtained in the usual way.

It would be interesting to explore alternative trimming rules for the SLS. For example Seung Ahn (personal communication) has suggested that trimming predicted probabilities outside the interval $[-\omega, 1+\omega]$ for $\omega > 0$ may result in a sequential estimator that has lower MSE for parameter estimates. This is the classical bias/efficiency trade-off. Since SLS with $\omega = 0$, trims "good" observations with positive probability. SLS with $\omega > 0$, may result in a larger final sample size $n_{\gamma}^{(J)}$ (lower variance) at the cost of higher bias of the parameter estimates. Of course this remains to be seen.

Other generalizations of the SLS approach are suggested by complications arising from the use of probit/logit models in certain contexts. One such occasion is the use of probit/logit models with panel data. Fixed effects and random effects estimation is much more involved in a logit model compared with a linear model. Likewise random effects estimation with a probit model is not as simple as with a linear model, and

a fixed effects model cannot be consistently estimated with a probit model. Another example is simultaneous equations/instrumental variable methods. The presence of dummy endogenous regressors is problematic if the DGP is assumed to be probit or logit. Generalization of the SLS approach to these cases has the potential to provide researchers with attractive alternative modeling and estimation strategies.

REFERENCES

Amemiya, Takeshi. "Some Theorems in the Linear Probability Model," *International Economic Review,* 18(3), October 1977, 645-650.

Betts, Julian R. and Robert W. Fairlie. "Explaining Ethnic, Racial, and Immigrant Differences in Private School Attendance," Journal of Urban Economics, February 2001, v50, 26-51.

Currie, Janet and Jonathan Gruber. "Health Insurance Eligibility, Utilization of Medical Care, and Child Health," *Quarterly Journal of Economics,* May 1996, 431-466.

Fairlie, Robert W. and William A. Sundstrom. "The Emergence, Persistence, and Recent Widening of the Racial Unemployment Gap," *Industrial and Labor Relations Review,* 52(2), January 1999, 252-270.

Greene, William H. *Econometric Analysis*, 4th ed. (Upper Saddle River, NJ: Prentice-Hall, 2000).

Heckman, James J. "Dummy Endogenous Variables in a Simultaneous Equation System," Econometrica v46, n4 (July 1978): 931-59.

Heckman, James J. and James M. Snyder, Jr. "Linear Probability Models of the Demand for Attributes With an Empirical Application to Estimating the Preferences of Legislators," *Rand Journal of Economics,* 28(0), 1997, S142-S189.

Horrace, William C. and Ronald L. Oaxaca. "Distributional Assumptions and OLS Bias in the LPM," mimeo, May 2001.

Klaassen, Franc J. G. M. and Magnus, Jan R. "Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model," Journal of the American Statistical Association v96, n454 (June 2001): 500-509

Kmenta, Jan. *Elements of Econometrics*, 2nd ed. (New York: Macmillan, 1986).

Lucking-Reiley, David. "Field Experiments on the Effects of Reserve Prices in Auctions: More Magic on the Internet," mimeo, December 2000.

Lukashin, Youri Pavlovich. "Econometric Analysis of Managers' Judgements on the Determinants of the Financial Situation in Russia," *Economics of Planning,* 33, 2000, 85-101.

McGarry, Kathleen. "Testing Parental Altruism: Implications of a Dynamic Model," NBER Working Paper 7593, March 2000.

Rosenthal, R.W. "A Bounded-Rationality Approach to the Study of Noncooperative Games," 18, 1989, 273-292.

Table 1. $\gamma = 0.75$, $\pi = 0.10$

β_0	β_1	n	$Bias(\widehat{\beta}_0)$	$Bias(\tilde{\beta}_0)$	$Bias(\widehat{\beta}_1)$	$Bias(\tilde{\beta}_1)$
-0.50	1.00	500	0.2456	-0.0151	-0.2470	0.0139
		1000	0.2486	-0.0089	-0.2470	0.0112
		2000	0.2489	-0.0082	-0.2501	0.0072
-0.50	2.00	500	0.2523	-0.0072	-0.5022	0.0204
		1000	0.2485	-0.0085	-0.4983	0.0158
		2000	0.2499	-0.0041	-0.4997	0.0082
0.50	1.00	500	0.0000	-0.0005	-0.2476	0.0072
		1000	0.0003	0.0014	-0.2516	0.0032
		2000	0.0003	0.0001	-0.2504	-0.0001
0.50	2.00	500	-0.0018	-0.0024	-0.4956	0.0257
		1000	0.0009	0.0010	-0.4969	0.0079
		2000	-0.0010	-0.0011	-0.5090	-0.0124

Table 2. $\gamma = 0.75$, $\pi = 0.20$

β_0	β_1	n	$Bias(\widehat{\beta}_0)$	$Bias(\tilde{\beta}_0)$	$Bias(\widehat{\beta}_1)$	$Bias(\tilde{\beta}_1)$
-0.50	1.00	500	0.2576	-0110	-0.2487	0.0173
		1000	0.2502	-0121	-0.2480	0.0096
		2000	0.2522	-0018	-0.2497	-0.0002
-0.50	2.00	500	0.2626	-0066	-0.5103	0.0208
		1000	0.2523	-0014	-0.4981	0.0015
		2000	0.2516	-0036	-0.4963	0.0068
0.50	1.00	500	0.0030	-0012	-0.2443	0.0160
		1000	0.0041	-0001	-0.2487	0.0092
		2000	0.0038	-0002	-0.2473	0.0110
0.50	2.00	500	0.0044	0.0003	-0.4947	0.0356
		1000	0.0022	-0.0029	-0.4901	0.0237
		2000	0.0063	0.0023	-0.5042	-0.0048

Table 3. $\gamma = 0.25$, $\pi = 0.10$

β_0	β_1	n	$Bias(\widehat{\beta}_0)$	$Bias(\tilde{\beta}_0)$	$Bias(\widehat{\beta}_1)$	$Bias(\tilde{\beta}_1)$
-0.50	1.00	500	0.6916	-0.0533	-0.7498	0.0554
		1000	0.6920	-0.0140	-0.7498	0.0132
		2000	0.6934	-0.0122	-0.7498	0.0147
-0.50	2.00	500	0.6930	-0.0456	-1.4993	0.0889
		1000	0.6930	-0.0095	-1.5006	0.0192
		2000	0.6937	-0.0058	-1.4990	0.0184
0.50	1.00	500	-0.0610	0.0093	-0.7515	0.0391
		1000	-0.0574	0.0014	-0.7501	0.0100
		2000	-0.0579	0.0014	-0.7502	0.0108
0.50	2.00	500	-0.0535	0.0071	-1.4970	0.0838
		1000	-0.0569	0.0022	-1.4985	0.0339
		2000	-0.0591	0.0023	-1.5014	0.0096

Table 4. $\gamma = 0.25$, $\pi = 0.20$

β_0	β_1	n	$Bias(\widehat{\beta}_0)$	$Bias(\tilde{\beta}_0)$	$Bias(\widehat{\beta}_1)$	$Bias(\tilde{\beta}_1)$
-0.50	1.00	500	0.7371	-0.0301	-0.7488	0.0327
		1000	0.7354	-0.0116	-0.7493	0.0084
		2000	0.7382	-0.0155	-0.7498	0.0185
-0.50	2.00	500	0.7344	-0.0430	-1.5009	0.0775
		1000	0.7368	-0.0221	-1.5000	0.0409
		2000	0.7362	-0.0035	-1.4985	0.0071
0.50	1.00	500	-0.0135	-0.0032	-0.7491	0.0270
		1000	-0.0119	0.0050	-0.7497	0.0421
		2000	-0.0131	-0.0002	-0.7501	0.0146
0.50	2.00	500	-0.0121	0.0035	-1.4988	0.0861
		1000	-0.0111	0.0024	-1.5001	0.0343
		2000	-0.0135	0.0007	-1.4989	0.0391

Table 5. $\gamma = 0.75$, $\pi = 0.10$

β_0	β_1	n	$MSE(\widehat{y}_{OLS})$	$MSE(\widehat{y}_{probit})$	$MSE(\widehat{y}_{Logit})$	$MSE(\widehat{y}_{SLS})$
-0.50	1.00	500	0.00407	0.00150	0.00193	0.00061
		1000	0.00418	0.00131	0.00172	0.00034
-0.50	2.00	500	0.00432	0.00137	0.00176	0.00053
		1000	0.00410	0.00118	0.00158	0.00024
0.50	1.00	500	0.00419	0.00172	0.00217	0.00085
		1000	0.00424	0.00121	0.00161	0.00030
0.50	2.00	500	0.00427	0.00153	0.00195	0.00063
		1000	0.00417	0.00128	0.00169	0.00033

Table 6. $\gamma = 0.75$, $\pi = 0.20$

β_0	β_1	n	$MSE(\widehat{y}_{OLS})$	$MSE(\widehat{y}_{probit})$	$MSE(\widehat{y}_{Logit})$	$MSE(\widehat{y}_{SLS})$
-0.50	1.00	500	0.00417	0.00152	0.00194	0.00063
		1000	0.00406	0.00124	0.00165	0.00030
-0.50	2.00	500	0.00419	0.00152	0.00196	0.00063
		1000	0.00408	0.00131	0.00172	0.00035
0.50	1.00	500	0.00417	0.00141	0.00182	0.00051
		1000	0.00397	0.00130	0.00171	0.00032
0.50	2.00	500	0.00417	0.00156	0.00199	0.00060
		1000	0.00419	0.00118	0.00157	0.00029

Table 7. $\gamma = 0.25$, $\pi = 0.10$

β_0	β_1	n	$MSE(\widehat{y}_{OLS})$	$MSE(\widehat{y}_{probit})$	$MSE(\widehat{y}_{Logit})$	$MSE(\widehat{y}_{SLS})$
-0.50	1.00	500	0.03653	0.00095	0.00124	0.00061
		1000	0.03623	0.00068	0.00095	0.00026
-0.50	2.00	500	0.03637	0.00085	0.00113	0.00049
		1000	0.03630	0.00074	0.00103	0.00030
0.50	1.00	500	0.03636	0.00097	0.00126	0.00069
		1000	0.03623	0.00068	0.00095	0.00026
0.50	2.00	500	0.03628	0.00081	0.00108	0.00046
		1000	0.03614	0.00076	0.00105	0.00030

Table 8. $\gamma = 0.25$, $\pi = 0.20$

β_0	β_1	n	$MSE(\widehat{y}_{OLS})$	$MSE(\widehat{y}_{probit})$	$MSE(\widehat{y}_{Logt})$	$MSE(\widehat{y}_{SLS})$
-0.50	1.00	500	0.04090	0.00090	0.00118	0.00057
		1000	0.04080	0.00072	0.00099	0.00028
-0.50	2.00	500	0.04122	0.00092	0.00120	0.00049
		1000	0.04073	0.00076	0.00103	0.00031
0.50	1.00	500	0.04076	0.00094	0.00123	0.00059
		1000	0.04098	0.00068	0.00095	0.00026
0.50	2.00	500	0.04117	0.00098	0.00128	0.00067
		1000	0.04108	0.00069	0.00097	0.00030

Table 9. Application

Variable	OLS Coeff.	SLS Coeff.	SLS/FGLS Coeff.	Probit Coeff.	Logit Coeff.
constant	0.4671	0.3127	0.3177	-0.4002	-0.8335
(*s.e.*)	(0.0145)	(0.0204, 0.0237)	(0.0215)	(0.0615)	(0.1072)
black	-0.0025	0.0243	0.0146	-0.0091	0.0215
(*s.e.*)	(0.0097)	(0.0133, 0.0138)	(0.0074)	(0.0436)	(0.0784)
other	-0.0444	-0.0493	-0.0646	-0.1805	-0.2937
(*s.e.*)	(0.0181)	(0.0239, 0.0253)	(0.0224)	(0.0792)	(0.1406)
female	0.0636	0.0946	0.1076	0.2738	0.4833
(*s.e.*)	(0.0085)	(0.0122, 0.0123)	(0.0075)	(0.0398)	(0.0721)
ntinc	0.0057	0.0111	0.0102	0.0255	0.0487
(*s.e.*)	(0.0002)	(0.0005, 0.0005)	(0.0004)	(0.0011)	(0.0021)
$ntinc^2$	-1.72E-05	-3.32E-05	-3.05E-05	-7.55E-05	-1.44E-04
(*s.e.*)	(8.61E-07)	(1.59E-06, 1.61E-06)	(1.24E-06)	(4.24E-06)	(8.23E-06)
emp	0.1367	0.1427	0.1644	0.4816	0.7968
(*s.e.*)	(0.0113)	(0.0150, 0.0162)	(0.0118)	(0.0484)	(0.0855)
Obs.	6,860	4,302	4,302	6,860	6,860
R^2	0.136	0.152	0.180	0.216	0.222

s.e. = standard error. For SLS, second standard error is "robust standard error".

Probit and logit marginal effects evaluated at the sample mean of each variable.

For probit and logit reported R^2 is pseudo-R^2.

Table 10. Sample Means

Variable	**Full Sample**	**Trimmed Sample**	**% Change**
hins	0.8329	0.7678	-7.82
black	0.2774	0.3173	14.38
other	0.0583	0.0686	17.67
$female$	0.5135	0.4633	-9.78
ntinc	54.8350	35.1407	-35.92
$ntinc^2$	5,266.0	2,942.2	-44.13
emp	0.8245	0.7915	-4.00
Observations	6,860	4,302	-37.29

APPENDIX 1

Proof of Theorem 8:

Let

$$\widetilde{y}_i = x_i \tilde{\beta}_{i \in \widehat{\kappa}_\gamma^{(j)}}.$$

Since $y_i = x_i\beta + \varepsilon_i$,

$$\widetilde{y}_i = y_i + x_i(\tilde{\beta}_{i \in \widehat{\kappa}_\gamma^{(j)}} - \beta) - \varepsilon_i,$$

so the predicted y_i is a function of the true y_i and a prediction error. We would like to investigate how the magnitude and direction of the prediction error affects the trimming process and vice versa. Therefore, consider the following cases

Case 1. Suppose some $x_i\beta > 1$. This implies $y_i = 1$ and $\varepsilon_i = 0$, therefore

$$\widetilde{y}_i = 1 + x_i(\tilde{\beta}_{i \in \widehat{\kappa}_\gamma^{(j)}} - \beta).$$

That is the predicted value will equal 1 plus the prediction error and the $x_i\beta > 1$ will be trimmed if the prediction error is such that:

$$x_i(\tilde{\beta}_{i \in \widehat{\kappa}_\gamma^{(j)}} - \beta) > 0, \quad \text{or}$$
$$x_i(\tilde{\beta}_{i \in \widehat{\kappa}_\gamma^{(j)}} - \beta) < -1.$$

Case 2. Suppose $x_i\beta < 0$. This implies $y_i = 0$ and $\varepsilon_i = 0$, therefore

$$\widetilde{y}_i = x_i(\tilde{\beta}_{i \in \widehat{\kappa}_\gamma^{(j)}} - \beta),$$

and $x_i\beta < 0$ will be trimmed if either

$$x_i(\tilde{\beta}_{i \in \widehat{\kappa}_\gamma^{(j)}} - \beta) > 1, \quad \text{or}$$
$$x_i(\tilde{\beta}_{i \in \widehat{\kappa}_\gamma^{(j)}} - \beta) < 0.$$

Case 3. Suppose $x_i\beta \in [0,1]$. This implies $y_i = x_i\beta$, therefore

$$\widetilde{y}_i = x_i\beta + x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta),$$

and $x_i\beta \in [0,1]$ will be trimmed if either

$$x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) > 1 - x_i\beta, \quad \text{or}$$
$$x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) < -x_i\beta.$$

These cases imply the following relationship between the size of the prediction error and the trimming of different classes of observations $x_i\beta$:

If $x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) < -1$, trim $x_i\beta < 0$ or $x_i\beta \in [0,1]$ or $x_i\beta > 1$.

If $x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [-1, -x_i\beta]$, trim $x_i\beta < 0$ or $x_i\beta \in [0,1]$.

If $x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [-x_i\beta, 0]$, trim $x_i\beta < 0$ only.

If $x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [0, 1 - x_i\beta]$, trim $x_i\beta > 1$ only.

If $x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [1 - x_i\beta, 1]$, trim $x_i\beta \in [0,1]$ or $x_i\beta > 1$.

If $x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) > 1$, trim $x_i\beta < 0$ or $x_i\beta \in [0,1]$ or $x_i\beta > 1$.

Notice that when the prediction error is small in magnitude (close to zero) then with probability zero an observation $x_i\beta \in [0,1]$ will be trimmed. When the prediction error is large then with positive probability an observation $x_i\beta \in [0,1]$ will be trimmed. Notice that for all values of the prediction error some $x_i\beta \notin [0,1]$ will be trimmed with positive probability. Define the following probabilities (all conditional on the subsample $i \in \widehat{\kappa}_\gamma^{(j)}$):

$$\Pr\{\text{Trim } x_i\beta > 1 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) < -1\} = \widehat{\pi}_{j1}$$
$$\Pr\{\text{Trim } x_i\beta \in [0,1] \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) < -1\} = \widehat{\gamma}_{j1}$$
$$\Pr\{\text{Trim } x_i\beta < 0 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) < -1\} = \widehat{\rho}_{j1}$$

$$\Pr\{\text{Trim } x_i\beta > 1 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [-1, -x_i\beta]\} = \widehat{\pi}_{j2} = 0$$

$$\Pr\{\text{Trim } x_i\beta \in [0,1] \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [-1, -x_i\beta]\} = \widehat{\gamma}_{j2}$$

$$\Pr\{\text{Trim } x_i\beta < 0 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [-1, -x_i\beta]\} = \widehat{\rho}_{j2}$$

$$\Pr\{\text{Trim } x_i\beta > 1 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [-x_i\beta, 0]\} = \widehat{\pi}_{j3} = 0$$

$$\Pr\{\text{Trim } x_i\beta \in [0,1] \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [-x_i\beta, 0]\} = \widehat{\gamma}_{j3} = 0$$

$$\Pr\{\text{Trim } x_i\beta < 0 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [-x_i\beta, 0]\}] = \widehat{\rho}_{j3} = 1$$

$$\Pr\{\text{Trim } x_i\beta > 1 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [0, 1 - x_i\beta]\} = \widehat{\pi}_{j4} = 1$$

$$\Pr\{\text{Trim } x_i\beta \in [0,1] \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [0, 1 - x_i\beta]\} = \widehat{\gamma}_{j4} = 0$$

$$\Pr\{\text{Trim } x_i\beta < 0 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [0, 1 - x_i\beta]\} = \widehat{\rho}_{j4} = 0$$

$$\Pr\{\text{Trim } x_i\beta > 1 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [1 - x_i\beta, 1]\} = \widehat{\pi}_{j5}$$

$$\Pr\{\text{Trim } x_i\beta \in [0,1] \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [1 - x_i\beta, 1]\} = \widehat{\gamma}_{j5}$$

$$\Pr\{\text{Trim } x_i\beta < 0 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) \in [1 - x_i\beta, 1]\} = \widehat{\rho}_{j5} = 0$$

$$\Pr\{\text{Trim } x_i\beta > 1 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) > 1\} = \widehat{\pi}_{j6}$$

$$\Pr\{\text{Trim } x_i\beta \in [0,1] \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) > 1\} = \widehat{\gamma}_{j6}$$

$$\Pr\{\text{Trim } x_i\beta < 0 \mid x_i(\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} - \beta) > 1\} = \widehat{\rho}_{j6}$$

Also, define conditional on the subsample $i \in \widehat{\kappa}_{\gamma}^{(j)}$

$$\begin{aligned}
&\Pr\{x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta) < -1\} = \delta_{j1} \\
&\Pr\{x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta)\} \in [-1, -x_i\beta]\} = \delta_{j2} \\
&\Pr\{x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta)\} \in [-x_i\beta, 0]\} = \delta_{j3} \\
&\Pr\{x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta)\} \in [0, 1 - x_i\beta]\} = \delta_{j4} \\
&\Pr\{x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta) \in [1 - x_i\beta, 1]\} = \delta_{j5} \\
&\Pr\{x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta) > 1\} = \delta_{j6}.
\end{aligned}$$

Then conditional on the subsample $i \in \widehat{\kappa}_{\gamma}^{(j)}$:

$$\begin{aligned}
&\Pr\{\text{Trim } x_i\beta \in [0,1]\} = \widehat{\gamma}_{j1}\delta_{j1} + \widehat{\gamma}_{j2}\delta_{j2} + \widehat{\gamma}_{j5}\delta_{j5} + \widehat{\gamma}_{j6}\delta_{j6} \\
&\Pr\{\text{Trim } x_i\beta \notin [0,1]\} = \widehat{\rho}_{j1}\delta_{j1} + \widehat{\rho}_{j2}\delta_{j2} + \delta_{j3} + \widehat{\rho}_{j6}\delta_{j6} + \widehat{\pi}_{j1}\delta_{j1} + \delta_{j4} + \widehat{\pi}_{j5}\delta_{j5} + \widehat{\pi}_{j6}\delta_{j6}
\end{aligned}$$

but $\widehat{\rho}_{jt} = 1 - \widehat{\gamma}_{jt} - \widehat{\pi}_{jt}$, $t = 1, 2, ...6$. Hence:

$$\begin{aligned}
&\Pr\{\text{Trim } x_i\beta \in [0,1]\} = \widehat{\gamma}_{j1}\delta_{j1} + \widehat{\gamma}_{j2}\delta_{j2} + \widehat{\gamma}_{j5}\delta_{j5} + \widehat{\gamma}_{j6}\delta_{j6} \\
&\Pr\{\text{Trim } x_i\beta \notin [0,1]\} = (1 - \widehat{\gamma}_{j1})\delta_{j1} + (1 - \widehat{\gamma}_{j2} - \widehat{\pi}_{j2})\delta_{j2} + \delta_{j3} + \delta_{j4} + \widehat{\pi}_{j5}\delta_{j5} + (1 - \widehat{\gamma}_{j6})\delta_{j6}
\end{aligned}$$

If $\delta_{j3} + \delta_{j4} \to 1$, as $n \to \infty$, $j \to J$ then $\Pr\{\text{Trim } x_i\beta \notin [0,1]\} \to 1$, implying $\Pr\{\text{Trim } x_i\beta \in [0,1]\} \to 0$. That is , the probability of making a trimming mistake approaches zero. But

$$\begin{aligned}
\delta_{j3} + \delta_{j4} &= \Pr\{x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta) \in [-x_i\beta, 0] \text{ or } [0, 1 - x_i\beta]\} \\
&= \Pr\{x_i(\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} - \beta) \in [-x_i\beta, 1 - x_i\beta]\} \\
&= \Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} \in [0,1]\}
\end{aligned}$$

Therefore, a sufficient condition for convergence is

$$\Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_{\gamma}^{(j)}} \in [0,1]\} \to 1.$$

Insofar as $\frac{n_\gamma^{(j)}}{n_\gamma^{(j-1)}}$ is an estimate for $\Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} \in [0,1]\}$ the requirement that $\Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} \in [0,1]\} \to 1$ should ensure convergence $\frac{n_\gamma^{(j)}}{n_\gamma^{(j-1)}} \to 1$ in the sample, but this is not guaranteed. If $\Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} \in [0,1]\} \to 1$ then $\Pr\{\text{Trim } x_i\beta \in [0,1]\} \to 0$. This is equivalent to saying that

$$\text{if } \Pr\{x_i\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(j)}} \in [0,1]\} \to 1, \text{ then } \Pr\{x_i\beta \in [0,1] \ \forall i \in \widehat{\kappa}_\gamma^{(J)}\} \to 1$$

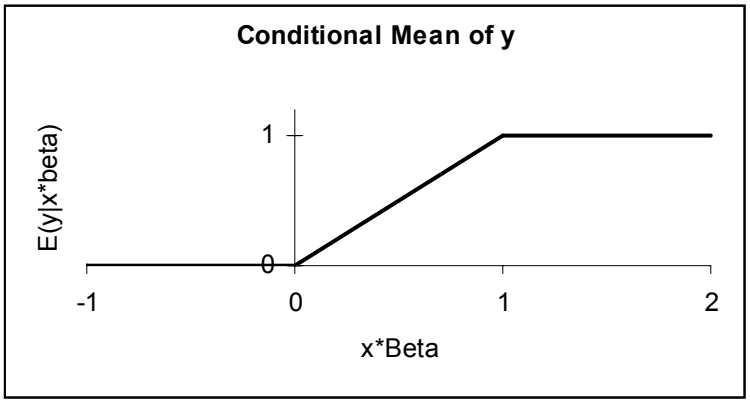
in the limit as $n \to \infty$, $j \to J$. Therefore, in the probability limit $\widehat{\kappa}_\gamma^{(j)}$ will consist only of $x_i\beta \in [0,1]$. This is not to say that $\widehat{\kappa}_\gamma^{(j)} \underset{p}{\to} \kappa_\gamma$. Indeed,

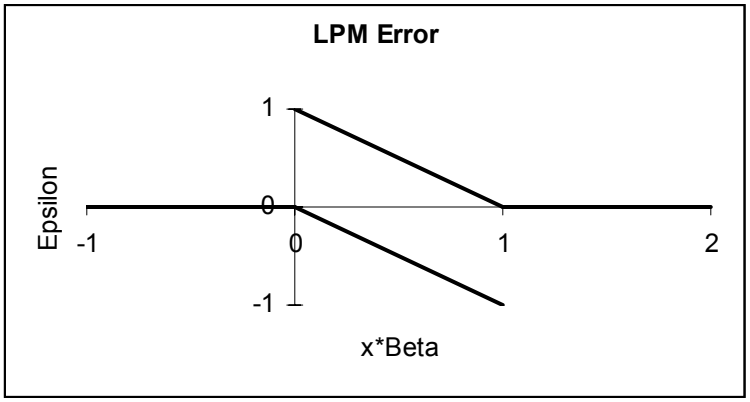
$$\widehat{\kappa}_\gamma^{(j)} \underset{p}{\to} \Lambda_n = \{i = 1, 2, ... \mid x_i\beta \in [0,1] \ \forall i\},$$

$\Lambda_n \subseteq \kappa_\gamma$. Then it is easy to show that

$$\tilde{\beta}_{i\in\widehat{\kappa}_\gamma^{(J)}} \underset{p}{\to} \tilde{\beta}_{i\in\Lambda_n} \underset{p}{\to} \beta.$$

Figure 1. Comparison of OLS and LPM Errors.





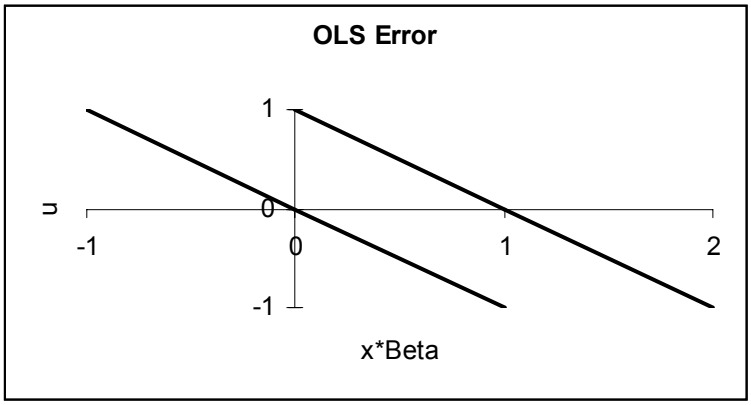


Figure 2. Comparison of SLS and OLS on the LPM

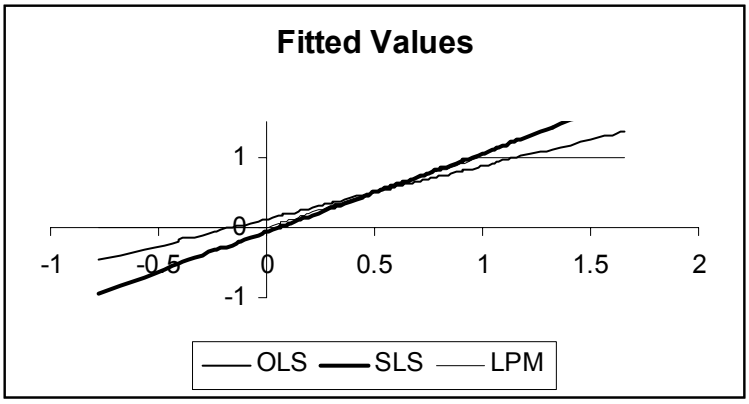


Figure 3. Comparison of SLS and Probit on the LPM.