

Distributional Assumptions and OLS Bias in the LPM

William C. Horrace & Ronald L. Oaxaca

Department of Economics

University of Arizona

2001

Introduction:

Horrace and Oaxaca (2001) detail a Sequential Least Squares (SLS) estimator that out-performs OLS, probit and logit in terms of mean-squared error of the predicted probabilities when the data generation process is the linear probability model (LPM). In their Theorem 8, they provide conditions under which SLS is consistent. These conditions hinge critically on the distribution of the explanatory variable in the DGP. This paper derives a few consistency results when the DGP is a simple linear model and the dependent variable has a normal distribution to demonstrate that the consistency assumptions of the SLS are met under the a normality assumption. The notation follows that of Horrace and Oaxaca (2001).

Result 1:

Consider a univariate random sample: $x_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$, and a linear transformation: $\beta_0 + x_i\beta_1$. Let $\gamma = \Pr\{\beta_0 + x_i\beta_1 \in [0, 1]\}$ and $\pi = \Pr\{\beta_0 + x_i\beta_1 > 1\}$. Consider simple OLS on the LPM. Then equation (11) becomes:

$$E(\hat{\beta}_n|x_i) = \begin{bmatrix} n & \sum_{i \in N} x_i \\ \sum_{i \in N} x_i & \sum_{i \in N} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \beta_0 n_\gamma + \beta_1 \sum_{i \in \kappa_\gamma} x_i + n_\pi \\ \beta_0 \sum_{i \in \kappa_\gamma} x_i + \beta_1 \sum_{i \in \kappa_\gamma} x_i^2 + \sum_{i \in \kappa_\pi} x_i \end{bmatrix}$$

This is a 2x1 matrix. The row one element is the conditional expectation of LPM intercept and the row two element is the conditional expectation of the LPM slope. Denote the OLS slope estimate as $\hat{\beta}_{1n}$. Then the conditional expectation of the slope estimate can be written as (after inverting the moment matrix above):

$$E(\hat{\beta}_{1n}|x_i) = \frac{-\left(\beta_0 n_\gamma + \beta_1 \sum_{i \in \kappa_\gamma} x_i + n_\pi\right) \sum_{i \in N} x_i + n \left(\beta_0 \sum_{i \in \kappa_\gamma} x_i + \beta_1 \sum_{i \in \kappa_\gamma} x_i^2 + \sum_{i \in \kappa_\pi} x_i\right)}{n \sum_{i \in N} x_i^2 - \left(\sum_{i \in N} x_i\right)^2}$$

After some algebra, taking the limit over x as $n \rightarrow \infty$ and appealing to the law of large numbers:

$$\lim E(\hat{\beta}_{1n}|x_i) = \frac{\gamma\beta_0}{\sigma^2} [E(x_\gamma) - \mu] + \frac{\gamma\beta_1}{\sigma^2} \{V(x_\gamma) + E(x_\gamma)[E(x_\gamma) - \mu]\} + \frac{\pi}{\sigma^2} [E(x_\pi) - \mu]$$

where $V(x_\gamma)$ and $E(x_\gamma)$ are the variance and mean of the distribution of the x in set κ_γ , respectively. The $E(x_\pi)$ is the mean of the distribution of the x in set κ_π . For the normal distribution these moments have closed form solutions:

$$V(x_\gamma) = \sigma^2 \left\{ 1 - \left[\frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\gamma} \right]^2 + \frac{\frac{a-\mu}{\sigma} \phi\left(\frac{a-\mu}{\sigma}\right) - \frac{b-\mu}{\sigma} \phi\left(\frac{b-\mu}{\sigma}\right)}{\gamma} \right\}$$

$$E(x_\gamma) = \mu + \sigma \frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\gamma}$$

$$E(x_\pi) = \mu + \sigma \frac{\phi\left(\frac{b-\mu}{\sigma}\right)}{\pi}$$

Where ϕ is the p.d.f. of the standard normal distribution, $a = -\beta_0/\beta_1$ and $b = (1 - \beta_0)/\beta_1$. Substitution of these moments into the above equation we get (after some tedious algebra):

$$\lim E(\widehat{\beta}_{1n}|x_i) = \gamma\beta_1.$$

Using similar (tedious) techniques it can also be shown that

$$\begin{aligned} \lim E(\widehat{\beta}_{0n}|x_i) &= \frac{\gamma\beta_0}{\sigma^2} [\sigma^2 + \mu^2 - \mu E(x_\gamma)] + \frac{\gamma\beta_1}{\sigma^2} \{(\sigma^2 + \mu^2)E(x_\gamma) - [\text{Var}(x_\gamma) + E(x_\gamma)]\mu\} \\ &\quad + \frac{\pi}{\sigma^2} [(\sigma^2 + \mu^2) - \mu E(x_\pi)] \\ \lim E(\widehat{\beta}_{0n}|x_i) &= \pi + \gamma\beta_0 + \beta_1 \left[\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right) \right]. \end{aligned}$$

Notice that when $\beta_0 + \mu\beta_1 = \frac{1}{2}$, implying $\phi\left(\frac{a-\mu}{\sigma}\right) = \phi\left(\frac{b-\mu}{\sigma}\right)$ and $\lim E(\widehat{\beta}_{0n}|x_i) = \pi + \gamma\beta_0$.

Result 2:

Suppose β_{0n}^j , β_{1n}^j , β_{0n}^{j+1} , and β_{1n}^{j+1} are any two successive SLS estimates of the intercept and slope parameters from the simple regression where (as before) $x_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$. Define $a_j = -\beta_{0n}^j/\beta_{1n}^j$ and $b_j = (1 - \beta_{0n}^j)/\beta_{1n}^j$. and let $\gamma_j = \Pr\{\beta_{0n}^j + x_i\beta_{1n}^j \in [0, 1]\}$. The sample for the j th iteration of the SLS procedure is a random sample from a truncation of x_i . Call this new random variable x_i^j , a truncation of x_i below a_j and above b_j . Then:

$$\begin{aligned} V(x^j) &= \sigma^2 \left\{ 1 - \left[\frac{\phi\left(\frac{a_j-\mu}{\sigma}\right) - \phi\left(\frac{b_j-\mu}{\sigma}\right)}{\gamma_j} \right]^2 + \frac{\frac{a_j-\mu}{\sigma} \phi\left(\frac{a_j-\mu}{\sigma}\right) - \frac{b_j-\mu}{\sigma} \phi\left(\frac{b_j-\mu}{\sigma}\right)}{\gamma_j} \right\} = \sigma_j^2 \\ E(x^j) &= \mu + \sigma \frac{\phi\left(\frac{a_j-\mu}{\sigma}\right) - \phi\left(\frac{b_j-\mu}{\sigma}\right)}{\gamma_j} = \mu_j \end{aligned}$$

Then using limiting arguments

$$\begin{aligned} \lim E(\beta_{1n}^{j+1}|x_i, x_i^j, a_j, b_j) &= \frac{\gamma\beta_0}{\sigma_j^2} [E(x_\gamma) - \mu_j] + \frac{\gamma\beta_1}{\sigma_j^2} \{V(x_\gamma) + E(x_\gamma)[E(x_\gamma) - \mu_j]\} \\ &\quad + \frac{\pi}{\sigma_j^2} [E(x_\pi) - \mu_j]. \end{aligned}$$

Using substitutions and tedious algebra similar to those used in Result 1 it can be shown that:

$$\lim E(\beta_{1n}^{j+1}|x_i^j, a_j, b_j) = \frac{\gamma\beta_1\sigma^2}{\sigma_j^2} - \left[\phi\left(\frac{a_j-\mu}{\sigma}\right) - \phi\left(\frac{b_j-\mu}{\sigma}\right) \right]$$

$$\times \left\{ \frac{(\pi + \gamma\beta_0 + \gamma\beta_1\mu)\sigma}{\gamma_j\sigma_j^2} + \frac{\gamma\beta_1\sigma^2}{\gamma_j\sigma_j^2} \left[\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right) \right] \right\}$$

If the next SLS estimate is β_{1n}^{j+2} , then x^{j+1} is a truncation of x below a_{j+1} and above b_{j+1} with similarly defined mean and variance: μ_{j+1} and σ_{j+1}^2 , respectively and $\gamma_{j+1} = \Pr\{\beta_{0n}^{j+1} + x_i\beta_{1n}^{j+1} \in [0, 1]\}$. Therefore,

$$\begin{aligned} \lim E(\beta_{1n}^{j+2} | x_i^{j+1}, a_{j+1}, b_{j+1}) &= \frac{\gamma\beta_1\sigma^2}{\sigma_{j+1}^2} - \left[\phi\left(\frac{a_{j+1}-\mu}{\sigma}\right) - \phi\left(\frac{b_{j+1}-\mu}{\sigma}\right) \right] \\ &\times \left\{ \frac{(\pi + \gamma\beta_0 + \gamma\beta_1\mu)\sigma}{\gamma_{j+1}\sigma_{j+1}^2} + \frac{\gamma\beta_1\sigma^2}{\gamma_{j+1}\sigma_{j+1}^2} \left[\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right) \right] \right\} \end{aligned}$$

Result 3:

Now let's restrict the model so that $\beta_0 + \mu\beta_1 = \frac{1}{2}$, implying $\phi\left(\frac{a-\mu}{\sigma}\right) = \phi\left(\frac{b-\mu}{\sigma}\right)$. In this case, the SLS estimators will always pass through the mean of the sample, which in the limit will always occur at $\beta_0 + \mu\beta_1 = \frac{1}{2}$. This being the case, the SLS regression lines will rotate about the point $\left(\frac{1}{2}, \frac{1}{2}\right)$ in the Cartesian space of $\beta_0 + x_i\beta_1$ and $\beta_{0n}^j + x_i\beta_{1n}^j$. Also, $E(x_\gamma) = \mu$ in the limit. This symmetry causes the trimming to also be symmetric about $\beta_0 + \mu\beta_1 = \frac{1}{2}$, so that $\phi\left(\frac{a_j-\mu}{\sigma}\right) = \phi\left(\frac{b_j-\mu}{\sigma}\right)$, in the limit as well. Then the SLS estimators reduce to:

$$\lim E(\beta_{1n}^{j+1} | x_i^j, a_j, b_j) = \frac{\gamma\beta_1\sigma^2}{\sigma_j^2} \quad \text{and} \quad \lim E(\beta_{1n}^{j+2} | x_i^{j+1}, a_{j+1}, b_{j+1}) = \frac{\gamma\beta_1\sigma^2}{\sigma_{j+1}^2}.$$

Clearly, $\sigma^2 \geq \sigma_j^2 \geq \sigma_{j+1}^2$, so as j increases, the $\frac{\sigma^2}{\sigma_{j+1}^2} \geq \frac{\sigma^2}{\sigma_j^2} \geq 1$ offsets the $\gamma < 1$ in the expressions $\frac{\gamma\beta_1\sigma^2}{\sigma_{j+1}^2} \geq \frac{\gamma\beta_1\sigma^2}{\sigma_j^2}$. Consequently the SLS slope estimators will always have decreasing asymptotic bias in the limit as j and n get large. Therefore we have found a case where the SLS estimator is asymptotically unbiased.

Now consider the OLS estimator in this case ($j = 1$):

$$\lim E(\hat{\beta}_{0n} + x_i\hat{\beta}_{1n} | x_i) = \pi + \gamma(\beta_0 + \beta_1)$$

In this case if we knew γ then we would know π , because the restriction $\beta_0 + \mu\beta_1 = \frac{1}{2}$ implies that $\pi = 1 - \gamma - \pi$ or $\pi = (1 - \gamma)/(2\pi)$. Then we could simply correct the bias of the OLS estimate.

References:

Horrace, W.C. and R.L. Oaxaca (2001) New wine in old bottles: a sequential estimation technique for the LPM. Unpublished Manuscript, Department of Economics, University of Arizona.