The 'lobit' Model

Miles D. Williams

June 11, 2019

Censored Regression

It is well known that in the context of a censored outcome variable, the least squares estimator for the linear association between a set of explanatory variables and said outcome is biased. To counteract this problem, James Tobin devised what would come to be affectionately called the "Tobit" model. This model blends the components of a probit and Gaussian model (OLS under restrictive assumptions) to fit a linear predictor, not to the observed outcome, but to an unobserved (*latent*) outcome. In the decades since Tobit was developed, many new generalized forms and adaptations of the original model have been created; though, each abides by similar principles, chief of which is a reliance on the probit link function to model the "censored" component of the outcome variable in question. With relative ease, I replace the probit link function with the logit link function. Much as logit and probit models are viewed as viable alternatives for fitting binary data, so can this new logistic Tobit, or *lobit* model be viewed as an alternative or fitting censored data.

The Likelihood Function

The likelihood function for a basic binary outcome model is given as

$$L = \prod_i [\Pr(y_i|X)^{y_i} \Pr(1-y_i|X)^{1-y_i}] : y \in \{0,1\} \tag{1}$$

where for a probit model,

$$\Pr(y_i|X) \equiv \Phi\left(\frac{X'\beta}{\sigma}\right),$$

where $\Phi(\cdot)$ denotes the cumulative density function where β and σ a parameters to be estimated. Meanwhile, for a logit model,

$$\Pr(y_i|X) \equiv \frac{e^{X'\beta}}{1+e^{X'\beta}}.$$

The likelihood function for a Gaussian (normal) model is given as

$$L = \prod_i \sigma^{-1}\varphi\left(\frac{y_i - X'\beta}{\sigma}\right) \tag{2}$$

where $\varphi(\cdot)$ is the probability density function and β and σ are again parameters to be estimated.

The Tobit likelihood function is simply a combination of the likelihood functions for a binary outcome with a probit link and for a normal model. For the classic case where y is a censored outcome with a lower bound of 0 and upper bound of ∞, that is:

$$L = \prod_i \left[1 - \Phi\left(\frac{X'\beta}{\sigma}\right)\right]^{1-D_i} \left[\sigma^{-1}\varphi\left(\frac{y_i - X'\beta}{\sigma}\right)\right]^{D_i} : y_i \geq 0 \tag{3}$$

where D_i is a dummy that takes the value 1 when $y_i > 0$, 0 otherwise.

Adopting a logit link function in lieu of a probit link function is straightforward, simply requiring specifying the likelihood function as

$$L = \prod_i \left[1 - \frac{e^{X'\beta}}{1+e^{X'\beta}}\right]^{1-D_i} \left[\sigma^{-1}\varphi\left(\frac{y_i - X'\beta}{\sigma}\right)\right]^{D_i} : y_i \geq 0 \tag{4}$$

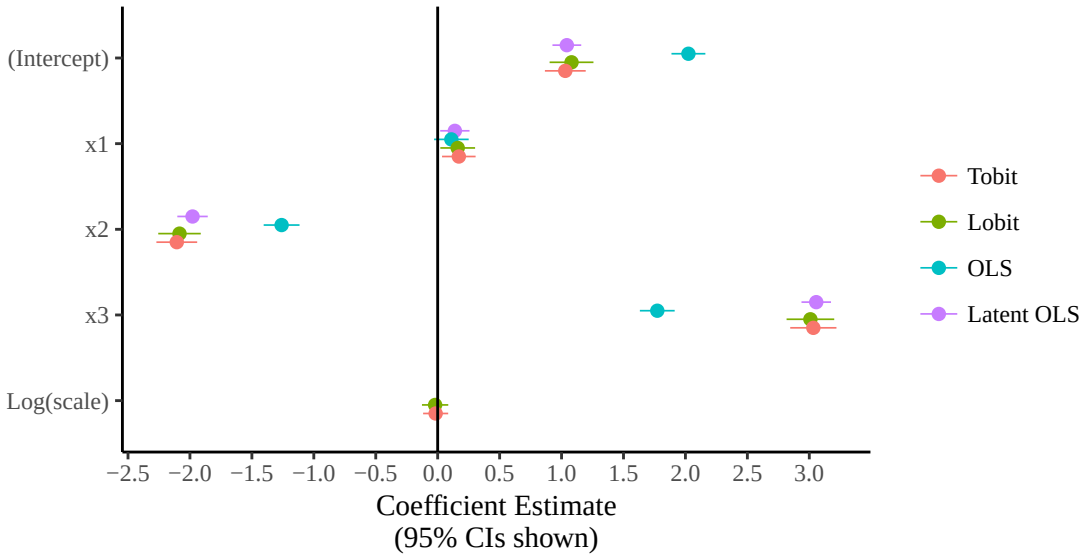

Figure 1: Coefficient plot of Tobit, lobit, OLS, and latent model estimates. 95 percent confidence intervals shown.

A Simple Comparison

I conduct a simple simulation to demonstrate the functionality of the lobit model vis-à-vis Tobit. I begin by generating a $N \times K$ matrix of random covariates, $\mathbf{X}$, which has $N = 300$ rows, denoting 300 observations, and $K = 4$ columns, denoting one constant (the first column) and 3 random covariates with $\mu = 0$ and $\sigma = 1$.

I then generate values for a *latent* outcome variable, y^*, via the following data generating process:

$$y_i^* = \mathbf{X}\boldsymbol{\beta} + \epsilon_i : \epsilon_i \sim N(0, \sigma) \tag{5}$$

where I specify $\boldsymbol{\beta}$ as a $K \times 1$ matrix of linear predictor parameters. I specify β parameters as 1, 0.3, -2, and 3. Subscript i denotes the i^{th} observation.

With the d.g.p. for the latent outcome defined, I specify the *observed* outcome y_i as:

$$y_i = \begin{cases} y_i^* & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases}.$$

This results in a censored outcome with a disproportionate number of observations with $y = 0$.

With simulated censored outcome in hand, we can compare how well Tobit and lobit allow us to correctly generate estimates of $\boldsymbol{\beta}$ used to generate the *latent* outcome as given in equation 5. To do this I simply estimate an OLS model where y_i^* is regressed on $\mathbf{X}$ and compare the model summary to Tobit and lobit estimates where y_i is regressed on $\mathbf{X}$.

Results are shown in Figures 1 and 2. Figure 1 shows coefficient estimates for parameters of interest. 95 percent confidence intervals are shown. Plain to the eye is the close similarity between latent model estimates and those obtained via Tobit and lobit. For reference, I have included results from an OLS model where y_i is regressed on $\mathbf{X}$. This model clearly demonstrates the bias inherent to least squares estimation of the linear association between covariates and a censored outcome.

The matter that most concerns us is the relative performance of lobit and Tobit. The results from the simulation reveal little difference in the parameter estimates provided by each estimator. Comparing linear predictions with y_i^* reveals similar consistency. Figure 2 shows the loess line for y_i^* over fitted values given by

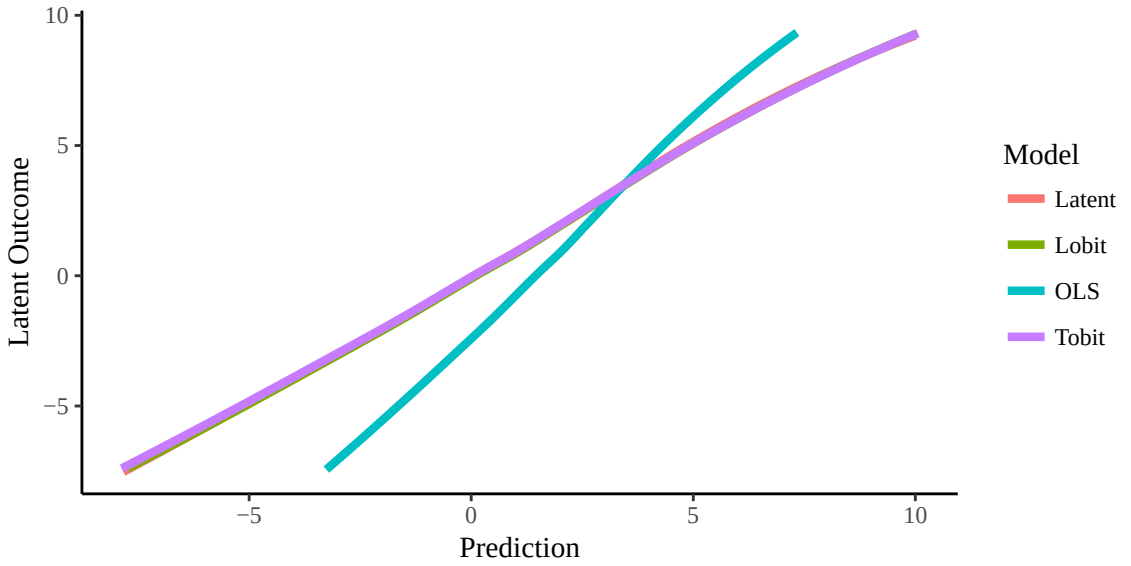

Figure 2: Predicting the *latent* outcome with *observed* data. The relative performance of Tobit, lobit, and OLS estimates.

Tobit, lobit, OLS, and latent models. The loess curve nearly completely overlaps for Tobit, lobit, and latent models. Compare this to OLS, which clearly predicts the latent outcome poorly relative to either Tobit or lobit.

Summary

The lobit model, at least in this preliminary trial, performs quite well, comparable to the more classic Tobit model. I suspect that preference for either estimator will follow logic similar to the perennial logit vs. probit debate (both work well, so go with the one you prefer). For logit adherents, the draw centers on the ease of interpretation of linear coefficients. Probit coefficients do not have a direct interpretation, whereas logit estimates may be interpreted as the linear effect of a covariate on the logged odds of an outcome being 1. Similarly, while Tobit and lobit estimates may be rightly interpreted as the linear effect of a covariate on a latent outcome, they also may be interpreted as the linear effect of a covariate on uncensored values of the outcome *weighted* by the probability of the outcome being greater than zero (at least in the classic case where the outcome is restricted to values greater than or equal to zero). In the case of lobit, the probability component of the likelihood function models the linear effect of a predictor on the logged odds of a censored outcome taking a value greater than zero. This adds more sense than nonsense when a researcher draws substantive conclusions about the relationship between covariates and a censored outcome.