(Re)estimating marginal changes after "truncreg" and "tobit" in Stata

Drichoutis, Andreas

Department of Economics, University of Ioannina

23 March 2011

(Re)estimating marginal changes after "truncreg" and "tobit" in Stata[1]

Andreas C. Drichoutis

Dept. of Economics, University of Ioannina, Greece

1. Introduction

In censored and truncated regression models, coefficient estimates represent the effect of an independent variable on a latent variable. This effect is generally not useful, because the latent variable is unobserved. Thus we are mostly interested in estimating marginal changes which represent the effect of an independent variable on the conditional mean function. A typo in Stata's help file has contributed to the incorrect estimation of marginal changes in such cases. This has significant implications for empirical practice, and results from some published studies may need to be reconsidered.

Popular software programs, such as Stata and Limdep, include automatic routines that ease the computational burden of estimating marginal changes and corresponding standard errors. Unfortunately, a typo that appears in the help files of some of Stata's commands could potentially lead some users in choosing the wrong prediction option when estimating marginal changes after the associated commands (i.e., the `truncreg` and `tobit` commands). I should note that this is not a problem with the code but a problem with the help files. Stata has already acknowledged this error in personal communication. The same problem is pervasive in other commands of the censored regression model like `xttobit` (the random effects variant of tobit) and `ivtobit` (tobit model with continuous endogenous regressors). Stata has corrected the help files for `ivtobit` and published an erratum for the reference manuals on their website while an erratum was still pending at time of writing for the `truncreg`, `tobit` and `xttobit` commands. However, this typo has been fixed in the final free update of Stata 11 (as of September 1, 2011) and does not appear in Stata 12.

[1] I would like to thank, without implicating, William Greene for providing a preview version of Limdep ver. 10 and for sorting out several of my queries. All errors and omissions are my own.

The purpose of this note is to alert researchers about the potential flaws of past empirical studies and to avoid similar flaws in future empirical research. Relying on a single program for doing empirical economic analysis is potentially dangerous. It is unsettling to think of how many papers may have misreported the real effects of their models, with significant implications for empirical practice and policy making. I first discuss the formulas for estimating marginal changes in censored and truncated regression models and then pinpoint the error in Stata's help files. Then I provide an empirical illustration in which I compare Stata's output with Limdep's output and my own calculations in Stata.

2. The censored regression model

In the censored regression model, the latent underlying regression is:

$$Y_i^* = \mathbf{b}'\mathbf{X}_i + u_i \ , \qquad (1)$$

while the observed dependent variable is:

$Y_i = L_i$ if $Y_i^* \le L_i$ (lower tail censoring),

$Y_i = U_i$ if $Y_i^* \ge U_i$ (upper tail censoring), and

$Y_i = Y_i^* = \mathbf{b}'\mathbf{X}_i + u_i$ if $L_i < Y_i^* < U_i$.

For the latent variable, the conditional mean function is simply $E\left(Y_i^* \mid \mathbf{X}_i\right) = \mathbf{b}'\mathbf{X}_i$. For an observation randomly drawn from the population, which may or may not be censored, the conditional mean function is (Greene, 2003, 2007):

$$E\left(Y_i \mid \mathbf{X}_i\right) = L_i\Phi\left(\alpha_i\right) + U_i\left(1-\Phi\left(\gamma_i\right)\right) + \left(\Phi\left(\gamma_i\right)-\Phi\left(\alpha_i\right)\right)\left(\mathbf{b}'\mathbf{X}_i + \sigma\frac{\phi\left(\alpha_i\right)-\phi\left(\gamma_i\right)}{\Phi\left(\gamma_i\right)-\Phi\left(\alpha_i\right)}\right) , \qquad (2)$$

where $\alpha_i = \dfrac{L_i - \mathbf{b}'\mathbf{X}_i}{\sigma}$ and $\gamma_i = \dfrac{U_i - \mathbf{b}'\mathbf{X}_i}{\sigma}$. The most familiar censored regression model is the Tobit model, which arises by setting $U_i = +\infty$ and $L_i = 0$ in (2).

The marginal effect of a continuous variable X_j is given by differentiating (2) with respect to it:

$$ME_{X_j} = \partial E\left(Y_i \mid \mathbf{X}_i\right)/\partial X_{ji} = b_j\left(\Phi\left(\gamma_i\right) - \Phi\left(\alpha_i\right)\right) = b_j \times \text{Prob}\left(\alpha_i < Y_i^* < \gamma_i\right) \ . \qquad (3)$$

If X_j is a dummy, then marginal effects are only approximately correct, and discrete changes should be estimated. From (2) we get:

$$DC_{X_j} = E\left(Y_i \mid \bar{\mathbf{X}}, X_j = 1\right) - E\left(Y_i \mid \bar{\mathbf{X}}, X_j = 0\right). \qquad (4)$$

3. The truncated regression model

The truncated regression model can be seen as a special case of the censored regression model where only data from the third group are observed. The latent regression is still given by (1), and the observed dependent variable is:

$$Y_i = Y_i^* = \mathbf{b}'\mathbf{X}_i + u_i \qquad \text{if} \qquad L_i < Y_i^* < U_i \ .$$

The conditional mean function in this case is:

$$E\left(Y_i \mid \mathbf{X}_i, L_i < Y_i < U_i\right) = \mathbf{b}'\mathbf{X}_i + \sigma\frac{\phi\left(\alpha_i\right) - \phi\left(\gamma_i\right)}{\Phi\left(\gamma_i\right) - \Phi\left(\alpha_i\right)} \ , \qquad (5)$$

where $\alpha_i = \dfrac{L_i - \mathbf{b}'\mathbf{X}_i}{\sigma}$ and $\gamma_i = \dfrac{U_i - \mathbf{b}'\mathbf{X}_i}{\sigma}$.

If X_j is continuous, then marginal effects can be derived by differentiating (5):

$$ME_{X_j} = b_j\left[1 + \left(\frac{\alpha_i\phi\left(\alpha_i\right) - \gamma_i\phi\left(\gamma_i\right)}{\Phi\left(\gamma_i\right) - \Phi\left(\alpha_i\right)}\right) - \left(\frac{\phi\left(\alpha_i\right) - \phi\left(\gamma_i\right)}{\Phi\left(\gamma_i\right) - \Phi\left(\alpha_i\right)}\right)^2\right] = b_j h\left(b, \sigma\right) \ , \qquad (6)$$

where $h\left(b, \sigma\right)$ is the scale factor.

If X_j is a dummy, then discrete changes should be estimated as:

$$DC_{X_j} = E\left(Y_i \mid \bar{\mathbf{X}}, X_j = 1, L_i < Y_i < U_i\right) - E\left(Y_i \mid \bar{\mathbf{X}}, X_j = 0, L_i < Y_i < U_i\right) \ . \qquad (7)$$

4. Estimating marginal changes in Stata

To estimate marginal changes in Stata, one needs to use the `mfx` module (the `mfx` module has been superseded by `margins` in Stata 11, although the error in the help file remained at time of writing). The help file of Stata indicates that the specified options in `mfx, predict(`*`options`*`)` can be, among others:

`e(a,b)` which calculates $E\left(xb+u \mid a<xb+u<b\right)$, the expected value of $y \mid x$ conditional on $y \mid x$ being in the interval *(a,b)*, meaning that $y \mid x$ is censored,

or

`ystar(a,b)` which calculates $E\left(y^*\right)$, where $y^*=a$ if $xb+u \le a$, $y^*=b$ if $xb+u \ge b$, and $y^*=xb+u$ otherwise, meaning that y^* is truncated.

However, if one tries to use `e(a,b)` after a censored regression model or `ystar(a,b)` after a truncated regression model, one gets incorrect results. Stata's technical support has acknowledged in personal communication that "the prediction option for `'ystar(a,b)'` corresponds to the case when y^* is censored (not truncated), whereas the prediction option for `e(a,b)` corresponds to the case when $y \mid x$ is truncated (not censored)."

To illustrate these errors with an empirical application, I use one of Stata's example datasets and similar specifications that appear in the help files for `tobit` and `truncreg`. The dataset is a 1978 automobile dataset which contains information on different car models, their selling price, fuel efficiency (miles per gallon), weight, gear ratio as well as the origin of the car (foreign vs. domestic) etc.

I regress miles per gallon (*mpg*) on a continuous variable (*weight*) and a dummy variable indicating origin of the car (*foreign*). This specification is similar to the one that appears in the example of the viewer window of Stata if one types: `help tobit postestimation##predict`. The only addition is the *foreign* variable. The lower limit is 17 and the upper limit is 24. I used the `nlcom` module in Stata to calculate standard errors for marginal changes of own calculations.

It is obvious from Table 1 that, in the censored regression model, Limdep's output, the `ystar(a,b)` option, and my own calculations coincide. According to Stata's official help file, the `e(a,b)` option should have been used, but it provides a very different set of results. Note that Limdep results were obtained from the preview version of Limdep ver. 10. The effect of the dummy variable *foreign* in Limdep, is slightly different than the corrected output of Stata and my own calculations in Stata. This is because the routine that requests estimation of marginal changes at the means, uses the scaled coefficient, i.e., it treats the dummy as a continuous variable. In essence, the program uses formula (3) instead of (4) and formula (6) instead of (7) which, however, produces only approximately correct results.

For the truncated regression model I use the same dataset and a specification similar to the one that appears in the example of the viewer window of Stata if one types: `help truncreg postestimation##predict.` I regress price of cars on miles per gallon (*mpg*) and origin of the car (*foreign*). The lower limit is 4,000 and the upper limit is 9,500. Similarly, in the truncated regression model, the `e(a,b)` option, Limdep's output[2], and own calculations coincide. However, according to Stata's help file, the `ystar(a,b)` option rather than the `e(a,b)` option should have been used[3]. The former evidently produces a different set of results.

5. Conclusions

Estimating marginal changes in Stata after a censored or a truncated regression model has been a black box for many researchers. Unfortunately, a simple typo in the help file of Stata may have caused researchers to use the wrong formulas for calculating marginal changes. This could have had significant impact on policy recommendations if Stata's users ignored the mathematics (which appear correct) but not the labels in the help file. If this is the case, many studies may

[2] Users of Limdep ver. 9 may also notice that standard errors in the case of the truncated regression are different than the ones reported in Table 1 that were obtained with ver. 10. This is due to a simplification of an extremely complicated derivative that was used by the program in version 9. Limdep ver. 10 now uses the full correct derivative and standard errors will coincide with output of Stata, as evident in Table 1.

[3] In the case of a truncated regression the results obtained with the wrong option i.e., the `ystar(a,b)` option, are meaningless. In contrast, in a censored regression model the results obtained with the wrong option, i.e., the `e(a,b)` option, correspond to the marginal effects on the expected value of the dependent variable conditional on being uncensored.

need to publish errata of their empirical findings after re-estimating marginal changes for truncated or censored regression models.

6. References

Greene, W.H. 2003. *Econometric Analysis (5th ed.)*. Prentice Hall: New Jersey.

Greene, W.H. 2007. *Limdep version 9.0, Econometric modeling guide (Vol. 1).* Econometric Software Inc.: Plainview, New York.

Table 1. Estimation results using Stata and Limdep

		Stata's output				**Limdep's output**		**Own calculations in Stata**	
		e(a,b)		***ystar(a,b)***					
	Variables	Coef.	Std. Error	Coef.	Std. Error	Coef.	Std. Error	Coef.	Std. Error
Censored regression (lower limit=17, upper limit=24, Dependent variable is *mpg*)	*weight*	-0.328	0.059	-0.582	0.066	-0.582	0.066	-0.582	0.066
	foreign	-1.541	0.492	-2.706	0.780	-2.846	0.914	-2.706	0.780
Truncated regression (lower limit=4000, upper limit=9500, Dependent variable is *price*)	*mpg*	-78.901	31.836	-89.264	60.430	-78.905	31.836	-78.901	31.836
	foreign	1119.988	439.570	1541.508	676.680	876.649	283.876	1119.978	439.570