

# Censored Data

Nathaniel Beck  
Department of Politics  
NYU  
New York, NY 10012  
nathaniel.beck@nyu.edu  
[http://www.nyu.edu/gsas/dept/politics/faculty/beck/beck\\_home.html](http://www.nyu.edu/gsas/dept/politics/faculty/beck/beck_home.html)

Jan. 2004

## Intro - Censoring and Truncation

### Truncation

Only some observations get into the sample

Example - NJ Negative Income Tax Experiment - Only those with incomes under \$10,000 in the experiment.

Only a problem if truncation is on the dependent variable

### Censoring

All individuals (or a sample) are observed, but only some observations can be made on the dependent variables

Examples

Wages of Woman (not PC!) - some don't work (for wages!) and so dependent variable is not observed

Spending on appliances - dependent variable not observed for those who don't buy

Selection bias - grades are not observed for those not admitted to college

Whether someone released on bail returns is not granted bail

## Truncation

Suppose we observe data where

$$y_i = x_i' \beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \quad (1)$$

but that the sample is truncated (from below) at  $a$ . Thus, CONDITIONAL ON  $i$  BEING IN THE SAMPLE,  $\epsilon$  is no longer normal but is distributed like a truncated normal.  $y_i > a$  implies that  $\epsilon_i > a - x_i' \beta$ . By the definition of conditional densities, a truncated density has

$$f(y|y > c) = \frac{f(y)}{1 - F(c)} \quad (2)$$

Thus the density of observation  $i$  in our truncated sample is

$$\frac{\frac{1}{\sigma} \phi((y_i - x_i' \beta) / \sigma)}{1 - \Phi((a - x_i' \beta) / \sigma)} \quad (3)$$

Assuming that  $a$  is a known constant we can thus do maximum likelihood fairly easily.

## Tobit

Censoring is more interesting

Suppose that, for Ms.  $i$ , that  $y$  is observed hours per week in the labor market. Let  $y^*$  represent the number of hours she desires to work. Then

$$y = 0 \text{ if } y^* \leq 0 \quad (4)$$

$$= y^* \text{ if } y^* > 0 \quad (5)$$

Further, suppose

$$y_i^* = x_i' \beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \quad (6)$$

What values should go in the likelihood.

For the censored values, it is just  $P(\text{being censored})$ , that is,

$$P(y_i^* \leq 0) = P(x_i' \beta \leq -\epsilon_i) = 1 - \Phi\left(\frac{x_i' \beta}{\sigma}\right) \quad (7)$$

while for the non-censored values it is just a normal density,  $\phi(y_i; x_i'\beta, \sigma^2)$ .

$$L = \prod_{y_i > 0} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}} \prod_{y_i \leq 0} 1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right) \quad (8)$$

which can be maximized fairly easily

## Issues

If the censoring point is not 0 but must be estimated, use

$$\min(y_i | y_i > 0) \quad (9)$$

This will exceed the true censoring point, but is better than using 0.

Why not use OLS on the uncensored obs? Consider the draws of  $\epsilon$  for  $i$ 's with low  $x_i'\beta$ .

They must, on average, be positive. But for those with high  $x_i'\beta$  can have any  $\epsilon$ . The  $x$  and  $\epsilon$  not independent, so OLS not good.

## Selection bias

Selection bias models have two equations. One is for whether an individual is selected or not ( $s_i = 0$  or  $1$ ) and an outcome equation ( $y_i$ , which is scored continuously, though it could easily be made into a quantal variable) which gives the outcome (wages, grades) for those who are selected.

Why not just estimate outcome model for those who are selected. Suppose the error terms in both equations are related (they share a common omitted variable). In terms of college admissions/outcomes this might be some measures of ability that are not in the analyst's record. Then you are more likely to get selected with a positive 'error' and since errors over the two equations are correlated the usual OLS assumptions about the error process for the outcome equation are incorrect. (OLS is fine if the error terms are independent.)

## Model setup

Selection

$$s_i^* = w_i' \gamma + \mu_i \quad (10)$$

$$s_i = 1 \text{ if } s_i^* > 0 \quad (11)$$

$$s_i = 0 \text{ if } s_i^* \leq 0 \quad (12)$$

Outcome

$$y_i = x_i \beta + \epsilon_i, \text{ Observed if } s_i = 1 \quad (13)$$

Stochastic Model

$$(s_i^*, y_i) \sim N(w_i' \gamma, x_i' \beta, 1, \sigma_\epsilon^2, \rho) \quad (14)$$

$$(\mu_i, \epsilon_i) \sim N(0, 0, 1, \sigma_\epsilon^2, \rho) \quad (15)$$

(Why is the variance of  $\mu = 1$  and the censoring point 0. For the same reason that we can't estimate cutoff and variance of a probit!)

## Moments of BVN

Let y and z be bivariate normal with correlation  $\rho$ .

$$E(y|z > a) = \mu_y + \rho \sigma_y \lambda(\alpha_z) \quad (16)$$

$$Var(y|z > a) = \sigma_y^2 (1 - \rho^2 \delta(\alpha_z)) \quad (17)$$

where

$$\alpha_z = \frac{a - \mu_z}{\sigma_z} \quad (18)$$

$$\lambda(\alpha_z) = \frac{\phi(\alpha_z)}{1 - \Phi(\alpha_z)} \text{ 'Inverse Mills Ratio' } \quad (19)$$

$$\delta(\alpha_z) = \lambda(\alpha_z)(\lambda(\alpha_z) - \alpha_z) \quad (20)$$

This gives an easy (non-ml) way to estimate these models, due to Heckman.

1. First do probit to estimate  $\hat{\gamma}$ .
2. Use this to calculate  $\hat{\lambda}_i = \frac{\phi(w_i' \hat{\gamma})}{\Phi(w_i' \hat{\gamma})}$ .
3. Estimate  $\beta$  in outcome equation by running a regression of  $y$  on  $x$  and  $\hat{\lambda}$ .

## Likelihoods

### Censored Observations

For a censored individual all we know is that  $s_i^* \leq 0$  so its contribution to the likelihood is

$$L_i^c = P(s_i^* \leq 0) \quad (21)$$

$$= P(w_i' \gamma + \mu_i \leq 0) \quad (22)$$

$$= P(\mu_i \leq -w_i' \gamma) \quad (23)$$

$$= \Phi(-w_i' \gamma) \quad (24)$$

### Uncensored Observations

For an individual whose outcome is observed the contribution would be (IF WE KNEW  $s_0^*$ !)

$$L_i^o = \phi(y_i, s_0^*) \quad (25)$$

But all we know is that  $s_i^* > 0$  so we have to integrate the above over all positive values of  $s_i^*$ .

$$L_i^o = \int_0^{\infty} \phi_{BVN}(y_i, s_i^*) ds_i^* \quad (26)$$

This can be simplified, since any joint density is the product of one conditional and one marginal ( $\phi_{BVN}(y_i, s_i^*) = \phi(s_i^* | y_i) \phi(y_i)$ ).

$$L_i^o = \phi(y_i) \int_0^{\infty} f(s_i^* | y_i) ds_i^* \quad (27)$$

$$= \phi(y_i; x_i' \beta, \sigma_\epsilon^2) (1 - \Phi(0; \theta_i, \sigma_i^2)) \quad (28)$$

where the parameters of  $\Phi$  come from the equation of the previous slide (on moments of conditional normal) - the notation gets messy! (but is conceptually easy)

$$L_i^O = \frac{1}{\sigma_\epsilon} \phi\left(\frac{y_i - x_i'\beta}{\sigma_\epsilon}\right) \quad (29)$$

$$\Phi\left(\frac{w_i'\gamma + \frac{\rho}{\sigma_\epsilon}(y_i - x_i'\beta)}{\sqrt{1 - \frac{\rho^2}{\sigma_\epsilon^2}}}\right) \quad (30)$$

Thus the log likelihood of the sample is

$$\ln L = \sum_{i=1}^n s_i \ln L_i^O + (1 - s_i) \ln L_i^C \quad (31)$$

## Double Probit

Suppose that the outcome equation is also a probit

$$y_i^* = x_i'\beta + \epsilon_i \quad (32)$$

$$y_i = 1 \text{ if } y_i^* > 0 \quad (33)$$

where we now (as in probit) set  $\sigma_\epsilon^2 = 1$ .

The likelihood for a censored observation is as above (Equation 24).

To simplify notation, let

$$\eta_i = x_i'\beta \quad (34)$$

$$\zeta_i = z_i'\gamma \quad (35)$$

For uncensored observations the contribution to the likelihood is

$$L_i^O = \frac{1}{2\pi\sqrt{1-\rho^2}} \quad (36)$$

$$\int_{-\infty}^{\eta_i} \int_{-\infty}^{\zeta_i} e^{-\left[\frac{1}{2(1-\rho^2)}(\eta_i^2 - \rho\eta_i\zeta_i + \zeta_i^2)\right]} d\eta_i d\zeta_i \quad (37)$$

with the log likelihood being then formed as for the previous case (Equation 31)

This will take the computer a while to estimate!