

ARE213

Econometrics

Spring 2006 UC Berkeley Department of Agricultural and Resource Economics

MAXIMUM LIKELIHOOD ESTIMATION V:

TRUNCATION, CENSORING, AND CORNER SOLUTIONS (W 16.4-16.8, T)

I. INTRODUCTION

Here we look at a set of complications with the standard linear model where part of the information is missing. Suppose we have a normal linear model

$$Y_i^* = X_i' \beta + \varepsilon_i,$$

with

$$\varepsilon_i | X_i \sim \mathcal{N}(0, \sigma^2).$$

If we observe (Y_i, X_i) for a random sample, we can estimate β by least squares,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}^*),$$

which is optimal (minimum variance unbiased estimator, best linear unbiased estimator, maximum likelihood estimator, etcetera).

Here we want to look at three complications. First, the truncated regression model. Suppose we do not have a random sample from the population, but a random sample conditional on $Y_i^* \geq 0$. (More generally, we can have a random sample conditional on $Y_i^* \in \mathcal{Y} \subset \mathbb{R}$, but the main ideas are illustrated just as well in the simple case. One generalization, known as stratified sampling, is concerned with the case where \mathbb{R} is partitioned in J strata, and we have a J random samples, one from each of the strata, with the sampling probabilities for each of the strata potentially different from their population shares. See for example Imbens and Lancaster (1995).) The second is censoring. In that case we have a random sample from

the population, but we only observe Y_i^* if Y_i^* is positive. If Y_i^* is positive we only observe X_i . The difference with truncated samples is (a) we know whether Y_i^* is negative, and (b) we always observe X_i . The third case is that of what Wooldridge calls corner solutions. This is often not distinguished from censoring. We observe the same data as in censoring, but here we are interested not in the distribution of Y_i^* , but in the distribution of $Y_i = \max(Y_i^*, 0)$. What is the difference? An example of censoring is topcoding in social security earnings data sets: we only observe earnings up to the social security maximum and otherwise observe the maximum. In that case we are obviously interested in the actual earnings and its relation to covariates, not the observed minimum of actual earnings and the social security maximum. An example of a corner solution is hours worked. These are non-negative, and to take account of that we may wish to model a latent variable Y_i^* as linear in covariates, with the observed Y_i equal to the maximum of Y_i^* and zero. We remained interested though in the distribution of the observed variable, actual hours worked, not in the distribution of the latent variable.

II. TRUNCATION(W 17.3)

Conditional on being observed (and we have to condition on that since we do not know how many observations were not observed), that is conditional on $Y_i^* \geq 0$, the distribution of Y_i^* given X_i is

$$f_{Y^*|X, Y^* \geq 0}(y; x) = \frac{f_{Y^*|X}(y|x)}{1 - F_{Y^*|X}(0|x)}.$$

With normality this means the distribution of the observed dependent variable Y is

$$f_{Y|X}(y|x) = \frac{(1/\sigma)\phi((y - x'\beta)/\sigma)}{1 - \Phi(-x'\beta/\sigma)},$$

where

$$\phi(a) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}a^2\right),$$

and

$$\Phi(a) = \int_{-\infty}^a \phi(z) dz.$$

So, the log likelihood function is

$$L(\beta, \sigma^2) = \sum_{i=1}^N \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_i - x_i'\beta)^2 - \ln(1 - \Phi(-x_i'\beta/\sigma)) \right).$$

To estimate the parameters we typically use maximum likelihood. This requires nonlinear optimization due to the last term in the log likelihood function.

What happens if we ignored the truncation and just did least squares? First note that if $Z \sim \mathcal{N}(0, 1)$, we have

$$\mathbb{E}[Z|Z > c] = \frac{\phi(c)}{1 - \Phi(c)}.$$

The ratio $\lambda(c) = \phi(c)/\Phi(c)$ is known as the inverse of the Mill's ratio. Now consider the conditional expectation Y given X in the sample:

$$\begin{aligned} \mathbb{E}[Y|X, Y > 0] &= X'\beta + \mathbb{E}[\varepsilon|\varepsilon > -X'\beta, X] \\ &= X'\beta + \sigma \cdot \mathbb{E}\left[\frac{\varepsilon}{\sigma} \mid \frac{\varepsilon}{\sigma} > -\frac{X'\beta}{\sigma}, X\right] = X'\beta + \sigma \cdot \frac{\phi(X'\beta/\sigma)}{\Phi(X'\beta/\sigma)} = X'\beta + \sigma \cdot \lambda(X'\beta/\sigma). \end{aligned}$$

Hence we overestimate the conditional expectation. It is more difficult to assess the effect on the slope coefficients.

The likelihood functions are not always well behaved for such models, and it can be difficult to find the maximum likelihood estimates.

III. CENSORING (W 16.1, 16.4)

A leading example of censoring arises in duration models where the dependent variable is the duration of some event, e.g., the duration of a spell of unemployment, or the survival

time after surgery. Often there are some durations that are not completed when the study is finished: they are censored at that point in time.

Here we first look at the case where all censoring is at zero. Let D_i be an indicator for censoring:

$$D_i = \begin{cases} 1 & \text{if not censored } (Y_i^* \geq 0), \\ 0 & \text{if censored } (Y_i^* < 0). \end{cases} .$$

Hence we observe a random sample of $(X_i, Y_i^* \cdot D_i, D_i)$. Then the contribution to the likelihood function for the i th observation is

$$\mathcal{L}_i = f(y_i|x_i; \beta, \sigma^2)^{d_i} F(0|x_i; \beta, \sigma^2)^{1-d_i},$$

which given normality for the conditional distribution of Y_i given X_i reduces to

$$\mathcal{L}_i = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right) \right)^{d_i} \Phi\left(-\frac{x_i'\beta}{\sigma}\right)^{1-d_i} .$$

The log likelihood function is then

$$L(\beta, \sigma^2) = \sum_{i=1}^N d_i \cdot \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \right) + (1 - d_i) \cdot \ln \Phi\left(-\frac{x_i'\beta}{\sigma}\right) .$$

Note that the data here differ from those in the standard probit case in that we observe the value of Y^* if $Y^* > 0$. In the probit case we only observed whether Y^* was positive or negative, leading to the log likelihood function

$$L(\beta, \sigma^2) = \sum_{i=1}^N d_i \cdot \Phi\left(\frac{x_i'\beta}{\sigma}\right) + (1 - d_i) \cdot \Phi\left(-\frac{x_i'\beta}{\sigma}\right) .$$

In that case we cannot estimate σ and we typically normalize by setting $\sigma^2 = 1$.

For the censored case let us again look at the conditional expectation of the outcome:

$$\mathbb{E}[Y|X] = E[Y|X, Y \geq 0] \cdot \Pr(Y \geq 0|X) + 0 \cdot \Pr(Y < 0|X) .$$

The first expectation is the same as in the truncated regression:

$$\mathbb{E}[Y|X, Y > 0] = X'\beta + \sigma \cdot \frac{\phi(X'\beta/\sigma)}{\Phi(X'\beta/\sigma)}.$$

The probability is

$$\Pr(Y \geq 0|X) = \Pr(\varepsilon > -X'\beta) = 1 - \Phi(-X'\beta/\sigma) = \Phi(X'\beta/\sigma).$$

Hence

$$\mathbb{E}[Y|X] = \left(X'\beta + \sigma \cdot \frac{\phi(X'\beta/\sigma)}{\Phi(X'\beta/\sigma)} \right) \cdot \Phi(X'\beta/\sigma). \quad (1)$$

Since this is not equal to $X'\beta$, OLS is in general inconsistent.

Now there is an alternative to maximum likelihood estimation. We can first estimate a probit model for the indicator that $Y > 0$. This gives us an estimate of β/σ . Then we can calculate the inverse of the Mills ratio

$$\hat{\lambda}_i = \frac{\phi(X'\hat{\beta}/\hat{\sigma})}{1 - \Phi(-X'\hat{\beta}/\hat{\sigma})}.$$

Note that

$$\mathbb{E}[Y|X, Y > 0, \lambda] = X'\beta + \sigma \cdot \lambda(X'\beta/\sigma).$$

We therefore regress, using only the positive Y observations, Y_i on X_i and $\hat{\lambda}_i$. This type of estimator was proposed by Heckman (1979).

IV. CORNER SOLUTIONS (W 16.2)

If we are interested in the distribution of Y given X , the parameter estimates themselves are of little value. In that case we want to convert them to derivatives of the conditional expectation with respect to the covariates, the same way we did this for the binary response models. Wooldridge shows that the derivative of the inverse of the Mills' ratio is

$$\frac{\partial \lambda}{\partial c}(c) = -\lambda(c) \cdot (c + \lambda(c)).$$

This can be used to show that

$$\frac{\partial \mathbb{E}}{\partial x}[Y|Y > 0, X = x] = \beta \cdot (1 - \lambda(x'\beta/\sigma) \cdot (x'\beta/\sigma + \lambda(x'\beta/\sigma))).$$

Since

$$\frac{\partial \Pr}{\partial x}(Y > 0|X = x) \frac{\partial 1 - \Phi(-X'\beta/\sigma)}{\partial x} = (\beta/\sigma) \cdot \phi(x'\beta/\sigma).$$

we can use the chain rule to get the derivative of the conditional expectation in (1), and it follows that

$$\frac{\partial \mathbb{E}}{\partial x}[Y|X = x] = \Phi(x'\beta/\sigma) \cdot \beta.$$

So, we can evaluate this at a particular value of x , or average over the sample distribution of the covariates. In both cases the delta method can be used to estimate the standard error. Alternatively, one can simply use the bootstrap. Generate a number of bootstrap samples. For each bootstrap sample calculate the maximum likelihood estimates, calculate the object of interest (the derivative at a particular value of x or averaged over the original sample distribution), and calculate its sample variance. That is obviously much simpler than using the delta method.

V. GENERALIZATIONS

Other sampling schemes are also possible. We already mentioned stratified sampling, where the range of the dependent variable is partitioned, and we have random sampling conditional on the stratum but not between the strata. If the dependent variable is discrete this type of sampling is referred to as choice-based sampling or response-based sampling. See Cosslett (1981) and Imbens (1992). This type of sampling strategy is very convenient if some of the choices are rare. Rather than risk getting one or two individuals in a sample of size 100 with a particular choice, a more informative sample would be obtained by sampling fifty individuals with each choice.

Another case that arises often in duration models is length-biased sampling. Suppose some individuals experience unemployment spells. One may sample such individuals by going to the unemployment registry and sampling individuals as they come in to register. Alternatively one may sample from the stock of unemployed. This means individuals with longer spells are more likely to be sampled. In fact, the probability of being sampled is proportional to the length of the spell. Such sampling schemes require careful calculation of the likelihood functions.

VI. AN APPLICATION

To see how these things work out in practice, I took the NLS data we have looked at before. There are 935 observations on earnings and education. The normal distributed based likelihood function is

$$\mathcal{L}(\beta, \sigma) = \prod_{i=1}^N \frac{1}{\sqrt{\sigma^2}} \phi\left(\frac{y_i - x'_i \beta}{\sigma}\right).$$

The first column of results in Table 1 gives the ols results for a regression of log weekly wages on a constant and years of education. The returns to education are estimated to be 6.67%. This gives the standard to compare the subsequent results to. With complete data this is what we would have gotten, and what we hope to get by using the corrections for censoring and truncation.

Then I (artificially) censored all weekly wages at \$600. This led to 12.5% censored observations. The log likelihood function is now

$$\mathcal{L}(\beta, \sigma) = \prod_{i=1}^N \left(\frac{1}{\sqrt{\sigma^2}} \phi\left(\frac{y_i - x'_i \beta}{\sigma}\right) \right)^{1-d_i} \cdot \left(1 - \Phi\left(\frac{600 - x'_i \beta}{\sigma}\right) \right)^{d_i}.$$

Ignoring the censoring and still doing ols leads to the results in the second column. The third column gives the maximum likelihood results for this case. The ml results are clearly much closer to the original complete data results.

Next, I completely discarded all observations with weekly wages in excess of \$600 to

generate a truncated sample. This changes the likelihood function to

$$\mathcal{L}(\beta, \sigma) = \prod_{i=1}^N \frac{\frac{1}{\sqrt{\sigma^2}} \phi\left(\frac{y_i - x_i' \beta}{\sigma}\right)}{\Phi\left(\frac{600 - x_i' \beta}{\sigma}\right)}.$$

Again I estimated the coefficients both by ols and ml. You can see that the simple ols results are very far from the ml results and from the complete data result. Despite the fact that only 12.5% of the data are truncated, this leads to severe biases in the ols estimates. The ml results are almost identical to the complete data ols results.

Table 1: REGRESSION ESTIMATES, NLS DATA (12.5% CENSORING/TRUNCATION)

variable	full sample	censored sample		truncated sample	
	ols/ml	ols	ml	ols	ml
intercept	5.0455 (0.0849)	5.1964 (0.0751)	5.0597 (0.0828)	5.3626 (0.0806)	5.1398 (0.1458)
education	0.0667 (0.0062)	0.0531 (0.0055)	0.0654 (0.0060)	0.0363 (0.0060)	0.0667 (0.0114)

REFERENCES

COSSLETT, S. R., (1981), "Maximum Likelihood Estimation for Choice-based Samples", *Econometrica*, vol. 49, 1289–1316,

HECKMAN, J, (1979), "Sample Selection as a Specification Error," *Econometrica*, Vol 47, No. 1, 153-162.

IMBENS, G. W., (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-based Sampling", *Econometrica*, vol. 60.

IMBENS, G., AND T. LANCASTER, "Efficient Estimation and Stratified Sampling", with T. Lancaster, *Journal of Econometrics*, Vol 74, No 2, 289–318.