ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

LaLonde Paper Detour:

Truncation Models and Heckman Selection Models

LaLonde (1986) presents a regression of the form $Y_i = \delta D_i + X_i\beta + rH_i + u_i$ that he says comes from a Heckman Selection Model. How is this equation derived and where does it come from?

1 Truncation Model

Consider first a data truncation model of the form:

$$y_i^* = x_i\beta + u_i$$

Assume that you observe a unit i if and only if $y_i^* > 0$. Otherwise, the data are *missing* – the distribution of y_i is truncated. Will a regression of y_i on x_i yield an unbiased estimate of β? In general, no. To see this, suppose that β is positive. In that case, if x is large, then you are likely to observe y_i^* even when u_i is small. But if x is small, you will only observe y_i^* when u_i is large. So the observations with small x will tend to have larger u, and the observations with large x will tend to have smaller u. This will cause your estimate of β to be attenuated towards zero when you run the regression of y on x. Truncation models of this form produce attenuation bias.

Can we correct for it? Yes, *if we make strict distributional assumptions* (big if there). Suppose that $u \sim N(0,1)$.[1] What is $E[y_i^*|x_i,\ y_i^* > 0]$? This is equivalent to asking, what is

[1] We can alternatively assume that it has some constant variance σ^2. The form of the derivation is unchanged, but you just have to normalize by dividing through by σ and then keep track of the σ's as you go.

$E[u_i|x_i,\ x_i\beta + u_i > 0]$? First, we need to compute the conditional density of u_i given that $x_i\beta + u_i > 0$.

$$P(u_i < u|x_i, -x_i\beta < u_i) = \frac{P(u_i < u \ \& - x_i\beta < u_i)}{P(-x_i\beta < u_i)} = \frac{P(-x_i\beta < u_i < u)}{P(-x_i\beta < u_i)}$$

$$P(u_i < u|x_i, -x_i\beta < u_i) = \begin{cases} \frac{\Phi(u)-\Phi(-x_i\beta)}{\Phi(x_i\beta)} & \text{if } -x_i\beta < u \\ 0 & \text{otherwise} \end{cases}$$

Thus the conditional density of u_i given that the observation is observed (i.e., $x_i\beta+u_i > 0$) is the derivative of the function above with respect to u:

$$f(u|x_i, -x_i\beta < u_i) = \begin{cases} \frac{\phi(u)}{\Phi(x_i\beta)} & \text{if } -x_i\beta < u \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

So what is $E[u_i|x_i,\ x_i\beta + u_i > 0]$?

$$E[u_i|x_i,\ x_i\beta + u_i > 0] = \int_{-x_\beta}^{\infty} u\frac{\phi(u)}{\Phi(x_i\beta)}du \ = \ \frac{1}{\Phi(x_i\beta)}\int_{-x_\beta}^{\infty} u\frac{1}{\sqrt{2\pi}}\exp(\frac{-u^2}{2})du =$$

$$\frac{1}{\Phi(x_i\beta)}[\frac{-1}{\sqrt{2\pi}}\exp(\frac{-u^2}{2})|_{-x_i\beta}^{\infty}] \ = \ 0 - \frac{1}{\Phi(x_i\beta)}[\frac{-1}{\sqrt{2\pi}}\exp(\frac{-(x_i\beta)^2}{2})] \ = \ \frac{\phi(x_i\beta)}{\Phi(x_i\beta)} \tag{2}$$

So $E[u_i|x_i,\ x_i\beta + u_i > 0] = \frac{\phi(x_i\beta)}{\Phi(x_i\beta)}$. We therefore know that in our *observed* data,

$$E[y_i|x_i] = x_i\beta + \frac{\phi(x_i\beta)}{\Phi(x_i\beta)}$$

How do we estimate this given that it is a nonlinear function? Two possibilities are nonlinear least squares (NLLS) and maximum likelihood estimation (MLE). For NLLS, observe that we can minimize the sum of squared residuals for the function above just like we minimize the sum of squared residuals when deriving OLS:

$$\text{Min wrt } \beta \quad \sum_{i=1}^{N}(y_i - x_i\beta + -\frac{\phi(x_i\beta)}{\Phi(x_i\beta)})^2$$

Of course, unlike with OLS, in general we cannot solve this problem analytically, so you will have to use a computer to numerically solve for the value of β that minimizes that function.

Alternatively, we could use MLE. Note that $P(Y_i < y_i|\text{observed}) = P(x_i\beta + u_i < y_i|\text{observed}) = P(u_i < y_i - x_i\beta + |\text{observed})$. From equation (1) we know the likelihood function for an individual observation is $\frac{\phi(y_i - x_i\beta)}{\Phi(x_i\beta)}$, so the likelihood for all the data is:

$$\prod_{i=1}^{N} \frac{\phi(y_i - x_i\beta)}{\Phi(x_i\beta)}$$

Maximizing this function with respect to β, or more practically, maximizing the log of it, will produce the MLE for β. Again, you will have to do this via computer rather than analytically.

2 Heckman Selection Model

We just reviewed the truncation model, which is about missing data, where the missingness is nonrandom. The Heckman Selection Model that LaLonde presents is also about missing data; in this case it is the counterfactual potential outcomes, $Y_i(1)$ and $Y_i(0)$ that are missing. Consider a model similar to LaLonde's of the form:

$$y_i(0) = x_i\beta + u_i$$

$$y_i(1) = \delta + x_i\beta + u_i$$

Suppose that the following rule for selecting into the treatment holds, where z_i contains all the variables in x_i and possibly some extra regressors:

We observe $y_i(0)$ if $d_i = z_i\gamma + \eta_i < 0$ (i.e., $D_i = 0$)

We observe $y_i(1)$ if $d_i = z_i\gamma + \eta_i > 0$ (i.e., $D_i = 1$)

Further suppose that both u_i and η_i are distributed $N(0,1)$ with some nonzero covariance, and that $(u_i, \eta_i) \perp z_i$. Note that we have a missing data problem – we want to know $\delta = y_i(1) - y_i(0)$, but $y_i(1)$ is missing when $D_i = 0$ and $y_i(0)$ is missing when $D_i = 1$. So what is $E[u_i|z_i, D_i = 1]$? First, note that since u_i and η_i are joint normal, $E[u_i|\eta_i, z_i] = E[u_i|\eta_i] = \alpha\eta_i$ (the first equality is because $u_i \perp z_i$, the second equality is due to the Regression-CEF Theorem from the first lecture). Therefore:

$$
\begin{aligned}
E[u_i|z_i, D_i = d] &= E[E[u_i|\eta_i, z_i, D_i = d]|z_i, D_i = d] \\
&= E[E[u_i|\eta_i, z_i]|z_i, D_i = d] \\
&= E[\alpha\eta_i|z_i, D_i = d] = \alpha E[\eta_i|z_i, D_i = d]
\end{aligned}
$$

The first step is just iterated expectations. In the second step, we can drop D_i from the inner expectation because it is a deterministic function of z_i and η_i (fixing z_i and η_i completely determines D_i). In the third step, we substitute in using the equality in the paragraph above.

Therefore, $E[u_i|z_i, D_i = 1] = \alpha E[\eta_i|z_i, D_i = 1] = \alpha E[\eta_i|z_i,\ z_i\gamma + \eta_i > 0]$. Applying our result from equation (2), we know that:

$$E[\alpha\eta_i|z_i,\ z_i\gamma + \eta_i > 0] = \alpha\frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)}$$

Thus $E[y_i|z_i, D_i = 1] = \delta + x_i\beta + \alpha\frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)}$. A similar set of derivations shows that $E[y_i|z_i, D_i = 0] = x_i\beta - \alpha\frac{\phi(z_i\gamma)}{1-\Phi(z_i\gamma)}$

$E[y_i|z_i, D_i]$ is thus:

$$E[y_i|z_i, D_i] = \delta D_i + x_i\beta + \alpha[D_i\frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)} - (1 - D_i)\frac{\phi(z_i\gamma)}{1 - \Phi(z_i\gamma)}]$$

This is virtually identical to what LaLonde shows on p. 615, though for some reason the denominators on his bias correction terms are switched around.[2] Regardless, we can write $rH_i = \alpha[D_i \frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)} - (1 - D_i)\frac{\phi(z_i\gamma)}{1-\Phi(z_i\gamma)}]$ and estimate the following equation:

$$y_i = \delta D_i + x_i\beta + rH_i + v_i$$

To estimate H_i, we can run a probit of D_i on z_i – this will give us the fitted values $z_i\hat{\gamma}$ that we need to construct $\hat{H}_i$. Then we can regress y_i on D_i, X_i, and $\hat{H}_i$. Hence it is a two-step procedure.

Is the procedure really that useful? Probably not. On the one hand, if z_i contains more elements than x_i, then you can simply run 2SLS, with the elements in z_i that are excluded from x_i as your instruments – the only thing the Heckman model is really getting you is a nonlinear first stage. The problem is that if the distributional assumptions or the functional assumptions are wrong, you can still get an inconsistent estimate even if you have a good instrument. On the other hand, if $z_i = x_i$, then you are identified purely off getting the functional form of $E[y_i|x_i]$ correct, and modeling the choice equation correctly, and making the correct distributional assumptions about the residuals, which is a very tenuous form of identification. Nevertheless, the model has a long history in econometrics (and it is central to the work for which Heckman was awarded the Nobel), so you should at least be aware of it.

[2] I have run Stata simulations to confirm that my derivation gives consistent estimates under the stated assumptions, while LaLonde's does not.