Appendix: Statistical Models of Attrition

An alternative to the nonparametric approach discussed in Chapter 7 is to posit a parametric model of the attrition process and the potential outcomes. Parametric models make assumptions about the functions linking cause and effect and the distributions from which unobserved causes are drawn. The most widely used parametric model, first proposed by Heckman (1979), involves two regression equations. The first equation offers a model of the outcome:

$$Y_i^* = \beta_0 + \beta_1 z_i + u_i. \tag{A7.1}$$

This "outcome equation" may also be expressed in terms of potential outcomes using the same form as equation (4.7).

The second equation offers a model of the process that determines whether the outcome is observed or missing. This "selection equation" predicts each subject's propensity to render observed outcomes:

$$r_i^* = \gamma_0 + \gamma_1 z_i + \gamma_1 Q_i + e_i. \tag{A7.2}$$

where Q_i is a variable (or collection of variables) that predict attrition but are unrelated to u_i.[1] This exclusion restriction is similar to the assumption we encountered in Chapter 5, when we discussed an "instrumental variable" that predicted whether a subject received treatment but had no causal effect on outcomes. The strongest case for excludability of Q_i occurs when the intensity of effort to obtain outcome data is randomly allocated. DiNardo and McCrary (2010), for example, discuss an experiment in which researchers randomly varied the amount of effort they devoted to obtaining outcomes.

The two equations work together in the following way. Let $r_i = 1$ if $r_i^* > 0$; otherwise $r_i = 0$. In other words, we observe outcomes for subjects whose propensity to be observed is above a certain threshold. In the case of the voucher lottery example, we would observe students whose potential test scores were above a certain cutoff level. Thus,

$$Y_i = \beta_0 + \beta_1 z_i + u_i \text{ if } r_i = 1\text{, otherwise } Y_i \text{ is missing.} \tag{A7.3}$$

Because we observe outcomes for some observations and not others, a regression of Y_i on z_i may generate biased estimates. Equation (A7.4) shows that the bias in the estimated treatment effect stems from the relationship between u_i and e_i. The expected value of Y_i can be written as

$$E[Y_i|Z_i, R_i = 1] = \beta_0 + \beta_1 z_i + E[u_i|e_i > -\gamma_0 - \gamma_1 Z_i - \gamma_1 Q_i]. \tag{A7.4}$$

---

[1] Although this approach can produce estimates even when there is no excluded variable in the selection equation, it is rarely applied without an excluded variable because in the absence of an instrument the results are entirely driven by the specific distributional and modeling assumptions.

If u_i and e_i are independent, there is no bias. However, the omitted factors predicting attrition (u_i) are typically correlated with omitted factors predicting outcomes (e_i).

In order to eliminate this bias, the researcher imposes assumptions that allow the third term in equation (A7.4) to be measured and included as a control variable in a regression. The critical assumptions are that u_i and e_i are bivariate normal (each is normally distributed, but they are correlated), with standard deviations normalized to 1, and *s*. It is important to appreciate how strong these assumptions are: the errors are not only assumed to be bivariate normal, they are also assumed to be homoskedastic.

Under these assumptions,

$$E[Y_i|Z_i, R_i = 1] = \beta_0 + \beta_1 z_i + s\,\frac{\varphi(\hat{r}_i)}{\Phi(\hat{r}_i)}, \qquad \text{(A7.5)}$$

where φ is the normal pdf and Φ is the normal cdf, and $\hat{r}_i$ is the predicted probability of $r_i = 1$ based on a probit estimation of the selection model. In order to estimate the average treatment effect, one regresses Y_i on z_i and a "correction term," $\frac{\varphi(\hat{r}_i)}{\Phi(\hat{r}_i)}$. Although in general, controlling for a covariate is an inadequate remedy for systematic attrition, under the assumptions of this model, adding the correction term to the regression model produces unbiased estimates of β_1. Another interesting property of this model is that regressing Y_i on z_i alone yields unbiased estimates so long as Z_i is independent of the correction term, a condition that is satisfied when the expected rate of attrition is the same for the treatment and control group.

The selection model above rests on strong assumptions. Homoskedasticity, for example, presupposes that the variance of u_i is the same for all subjects, an assumption that does not follow from random allocation of subjects to treatment. A different set of assumptions leads to a different estimation approach, known as Tobit. Recall from equation (7.10) that applying the difference-in-means estimator to observed outcomes is, in expectation, equivalent to estimating the treatment effect among those who would be observed if assigned to the control group, plus a term that represents the selection effect. Suppose we assume that the treatment effect is positive for all subjects. Under this assumption, the subjects who would be observed if assigned to the control group will also be observed if assigned to treatment, since $Y_i(1) > c$ if $Y_i(0) > c$. As for the term that represents the selection effect, if the treatment effect is positive, $E[Y_i(1)|(R_i(1) = 1) - Y_i(1)|(R_i(0) = 1)]$ will be positive. The intuition here is that the set of subjects for whom $r_i(0) = 1$ is a more select set with higher potential outcomes than the set of subjects for whom $r_i(1) = 1$. When $r_i(0) = 1$, $Y_i(0) > c$. When $r_i(1) = 1$, $Y_i(1) > c$, but this is an easier hurdle because $Y_i(1)$ is greater than $Y_i(0)$.

Based on the assumption of positive treatment effects, Angrist et al. (2006) propose a parametric model of the effects of vouchers on test scores:

$$Y_i^* = \beta_0 + \beta_1 z_i + u_i, \qquad \text{(A7.6)}$$

where $Y_i = Y_i^*$ if $Y_i^* > c$; otherwise Y_i is missing. This model is similar to the selection model above except that now missingness is a function of latent outcomes, not covariates. If the outcomes were not truncated, the parameters of equation (A7.6) could be estimated using regression. Given truncation, regression is biased. Angrist et al. assume that the u_i are drawn independently from a normal distribution and estimate this regression model for different values of the cutoff parameter, c. They term this approach *artificial censoring*, because for a small fraction of subjects, contrary to the model, observed values fall below the proposed censoring value c, in which case the researchers treat these subjects as though they were missing.

Table A7.1 shows the results of tobit estimation based on the assumption that outcomes are missing whenever the score the subject would have received is less than 32, which is the 1st percentile score among the observed scores. In contrast to the missing at random assumption, this censoring value suggests that the missing subjects would have done very poorly on the exam. The details of estimating these models are as follows: To prepare the data for the command used to estimate the tobit moded, all outcome values less than or equal to 32 are set to 32. The Stata command option `ll` in the command line below indicates that the smallest value in the data will be used as the left censoring value.

The tobit estimates imply that attrition severely distorts simple regression results. If the model is correct, the estimated effect of vouchers on reading scores is 3.3, rather than 0.7. Repeating this exercise for a range of censoring values indicates the sensitivity of the estimates to different assumptions about truncation.

```
. tobit readcens1 vouch0 age sex_name, ll

Tobit regression                                  Number of obs   =       3541
                                                  LR chi2(3)      =     888.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -6143.5532                       Pseudo R2       =     0.0675

------------------------------------------------------------------------------
   readcens1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      vouch0 |   3.289796   .7048559     4.67   0.000     1.907831    4.671761
         age |  -9.029631   .3660284   -24.67   0.000    -9.747279   -8.311982
    sex_name |  -1.673547   .6865377    -2.44   0.015    -3.019596   -.3274971
       _cons |   137.9839   4.393478    31.41   0.000     129.3699    146.5979
-------------+----------------------------------------------------------------
      /sigma |   16.29016    .385755                      15.53384    17.04649
------------------------------------------------------------------------------
  Obs. summary:       2334  left-censored observations at readcens1<=32
                      1207     uncensored observations
                         0 right-censored observations
```