TABLE 24.4 Estimated Earnings Equations

	Migration	Migrant Earnings	Nonmigrant Earnings
Constant	−1.509	9.041	8.593
SE	−0.708 (−5.72)	−4.104 (−9.54)	−4.161 (−57.71)
ΔEMP	−1.488 (−2.60)	—	—
ΔPCI	1.455 (3.14)	—	—
Age	−0.008 (−5.29)	—	—
Race	−0.065 (−1.17)	—	—
Sex	−0.082 (−2.14)	—	—
ΔSIC	0.948 (24.15)	−0.790 (−2.24)	−0.927 (−9.35)
λ	—	0.212 (0.50)	0.863 (2.84)

the net benefit are factors that also affect the income received in either place. An analysis of income in a sample of migrants must account for the incidental truncation of the mover's income on a positive net benefit. Likewise, the income of the stayer is incidentally truncated on a nonpositive net benefit. The model implies an income after moving for all observations, but we observe it only for those who actually do move. Nakosteen and Zimmer (1980) applied the selectivity model to a sample of 9,223 individuals with data for 2 years (1971 and 1973) sampled from the Social Security Administration's Continuous Work History Sample. Over the period, 1,078 individuals migrated and the remaining 8,145 did not. The independent variables in the migration equation were as follows:

SE = self-employment dummy variable; 1 if yes,
ΔEMP = rate of growth of state employment,
ΔPCI = growth of state per capita income,
$\mathbf{x}$ = age, race (nonwhite = 1), sex (female = 1),
ΔSIC = 1 if individual changes industry.

The earnings equations included ΔSIC and SE. The authors reported the results given in Table 24.4. The figures in parentheses are asymptotic t ratios.

19.6.1 ~~24.5.4~~ REGRESSION ANALYSIS OF TREATMENT EFFECTS

AU: Term "treatment effects" was KT on msp 19-68. Here also?

The basic model of selectivity outlined earlier has been extended in an impressive variety of directions.[33] An interesting application that has found wide use is the measurement of **treatment effects** and program effectiveness.[34]

An earnings equation that accounts for the value of a college education is

$$\text{earnings}_i = \mathbf{x}_i'\beta + \delta C_i + \varepsilon_i,$$

where C_i is a dummy variable indicating whether or not the individual attended college. The same format has been used in any number of other analyses of programs, experiments, and treatments. The question is: Does δ measure the value of a college education (assuming that the rest of the regression model is correctly specified)? The answer is

~~[27] For a survey, see Maddala (1983).~~

[34] This is one of the fundamental applications of this body of techniques, and is also the setting for the most longstanding and contentious debate on the subject. A *Journal of Business and Economic Statistics* symposium [Angrist (2001)] raised many of the important questions on whether and how it is possible to measure treatment effects.

19-72

no if the typical individual who chooses to go to college would have relatively high earnings whether or not he or she went to college. The problem is one of self-selection. If our observation is correct, then least squares estimates of δ will actually overestimate the treatment effect. The same observation applies to estimates of the treatment effects in other settings in which the individuals themselves decide whether or not they will receive the treatment.

To put this in a more familiar context, suppose that we model program participation (e.g., whether or not the individual goes to college) as

$$C_i^* = \mathbf{w}_i'\gamma + u_i,$$
$$C_i = 1 \quad \text{if } C_i^* > 0, 0 \text{ otherwise.}$$

We also suppose that, consistent with our previous conjecture, u_i and ε_i are correlated. Coupled with our earnings equation, we find that

$$E[y_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i] = \mathbf{x}_i'\beta + \delta + E[\varepsilon_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i]$$
$$= \mathbf{x}_i'\beta + \delta + \rho\sigma_\varepsilon\lambda(-\mathbf{w}_i'\gamma) \qquad \text{19-34 } \cancel{(24\text{-}22)}$$

19-24 once again. [See (24-21).] Evidently, a viable strategy for estimating this model is to use the two-step estimator discussed earlier. The net result will be a different estimate of δ that will account for the self-selected nature of program participation. For nonparticipants, the counterpart to (24-22) is 19-34

$$E[y_i \mid C_i = 0, \mathbf{x}_i, \mathbf{z}_i] = \mathbf{x}_i'\beta + \rho\sigma_\varepsilon\left[\frac{-\phi(\mathbf{w}_i'\gamma)}{1-\Phi(\mathbf{w}_i'\gamma)}\right]. \qquad \text{19-35 } \cancel{(24\text{-}23)}$$

The difference in expected earnings between participants and nonparticipants is, then,

$$E[y_i \mid C_i = 1, \mathbf{x}_i, \mathbf{z}_i] - E[y_i \mid C_i = 0, \mathbf{x}_i, \mathbf{z}_i] = \delta + \rho\sigma_\varepsilon\left[\frac{\phi_i}{\Phi_i(1-\Phi_i)}\right]. \qquad \text{19-36 } \cancel{(24\text{-}24)}$$

If the selectivity correction λ_i is omitted from the least squares regression, then this difference is what is estimated by the least squares coefficient on the treatment dummy variable. But because (by assumption) all terms are positive, we see that least squares overestimates the treatment effect. Note, finally, that simply estimating separate equations for participants and nonparticipants does not solve the problem. In fact, doing so would be equivalent to estimating the two regressions of Example 24.9 by least squares, which, as we have seen, would lead to inconsistent estimates of both sets of parameters. 19-12

A insert next page

There are many variations of this model in the empirical literature. They have been applied to the analysis of education,[35] the Head Start program,[36] and a host of other settings.[37] This strand of literature is particularly important because the use of dummy variable models to analyze treatment effects and program participation has a long history in empirical economics. This analysis has called into question the interpretation of a number of received studies.

35 ~~29~~ Willis and Rosen (1979).

36 ~~30~~ Goldberger (1972), 2008

37 ~~31~~ A useful summary of the issues is Barnow, Cain, and Goldberger (1981). ~~See also Maddala (1983) for a long list of applications. A related application is the switching regression model. See, for example, Quandt (1982, 1988).~~ See, also, Imbens and Wooldridge (2009).

(A)

Insert on msp 19-72 where indicated

KT

To describe the problem created by **selection on the unobservables**, we will drop the independence assumptions. The model with endogenous participation and different outcome equations would be

$$
\begin{aligned}
C_i^* &= \mathbf{w}_i'\boldsymbol{\gamma} + u_i,\ C_i = 1 \text{ if } C_i^* > 0 \text{ and } 0 \text{ otherwise,}\\
y_{i0} &= \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_{i0},\\
y_{i1} &= \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{i1}.
\end{aligned}
$$

It is useful to combine the second and third equations in

$$y_{ij} = C_i(\mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{i1}) + (1 - C_i)(\mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_{i0}),\ j = 0,1.$$

We assume joint normality for the three disturbances;

$$\begin{pmatrix} u_i \\ \varepsilon_{i0} \\ \varepsilon_{i1} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0\theta_0 & \rho_1\theta_1 \\ \rho_0\theta_0 & \theta_0^2 & \theta_{01} \\ \rho_1\theta_1 & \theta_{01} & \theta_1^2 \end{pmatrix}\right].$$

The variance in the participation equation is normalized to one for a binary outcome, as described earlier (Section 17.2). Endogeneity of the participation is implied by the nonzero values of the correlations ρ_0 and ρ_1. The familiar problem of the missing counterfactual appears here in our inability to estimate θ_{01}. The data will never contain information on both states simultaneously, so it will be impossible to estimate a covariance of y_{i0} and y_{i1} (conditioned on $\mathbf{x}_i$ or otherwise). Thus, the parameter θ_{01} is not identified (estimable) – we normalize it to zero. The parameters of this model after the two normalizations can be estimated by two step least squares as suggested in Section 19.XX, or by full information maximum likelihood. The average treatment effect on the treated would be

AQ: Provide missing Section number

$$\text{ATET} = E[y_{i1}|C_i=1,\mathbf{x}_i,\mathbf{w}_i] - E[y_{i0}|C_i=1,\mathbf{x}_i,\mathbf{w}_i] = \mathbf{x}_i'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\rho_1\theta_1 - \rho_0\theta_0)\frac{\phi(\mathbf{w}_i'\boldsymbol{\gamma})}{\Phi(\mathbf{w}_i'\boldsymbol{\gamma})}.$$

minus minus

[See (19-34).] If the treatment assignment is completely random, then $\rho_1 = \rho_0 = 0$, and we are left with the first term. But, of course, it is the nonrandomness of the treatment assignment that brought us to this point. Finally, if the two coefficient vectors differ only in their constant terms, $\beta_{0,0}$ and $\beta_{1,0}$, then we are left with the same δ that appears in (19-36) – the ATET would be $\beta_{0,1} + C_i(\beta_{1,0} - \beta_{0,0})$.

minus

end of insert

19.6.1.a ~~24.5.5~~ THE NORMALITY ASSUMPTION

Some research has cast some skepticism on the selection model based on the normal distribution. [See Goldberger (1983) for an early salvo in this literature.] Among the findings are that the parameter estimates are surprisingly sensitive to the distributional assumption that underlies the model. Of course, this fact in itself does not invalidate the normality assumption, but it does call its generality into question. On the other hand, the received evidence is convincing that sample selection, in the abstract, raises serious problems, distributional questions aside. The literature—for example, Duncan (1986b), Manski (1989, 1990), and Heckman (1990)—has suggested some promising approaches based on robust and nonparametric estimators. These approaches obviously have the virtue of greater generality. Unfortunately, the cost is that they generally are quite limited in the breadth of the models they can accommodate. That is, one might gain the robustness of a nonparametric estimator at the cost of being unable to make use of the rich set of accompanying variables usually present in the panels to which selectivity models are often applied. For example, the nonparametric bounds approach of Manski (1990) is defined for two regressors. Other methods [e.g., Duncan (1986b)] allow more elaborate specifications.

~~37~~ 38

Recent research includes specific attempts to move away from the normality assumption.[~~37~~ 38] An example is Martins (2001), building on Newey (1991), which takes the core specification as given in (24-20) as the platform, but constructs an alternative to the assumption of bivariate normality. Martins's specification modifies the Heckman model by employing an equation of the form

19-22

$$E[y_i \mid z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \mu(\mathbf{w}_i'\boldsymbol{\gamma})$$

delete comma

where the latter, "selectivity correction" is not the inverse Mills ratio, but some other result from a different model. The correction term is estimated using the Klein and Spady model discussed in Section 23.6.1. This is labeled a "semiparametric" approach. Whether the conditional mean in the selected sample should even remain a linear index function remains to be settled. Not surprisingly, Martins's results, based on two-step least squares differ only slightly from the conventional results based on normality. This approach is arguably only a fairly small step away from the tight parameterization of the Heckman model. Other non- and semiparametric specifications, e.g., Honorè and Kyriazidou (1997, 2000) represent more substantial departures from the normal model, but are much less operational.[~~38~~ 39] The upshot is that the issue remains unsettled. For better or worse, the empirical literature on the subject continues to be dominated by Heckman's original model built around the joint normal distribution.

grave

for example

AU: OK to spell out "e.g."?

FN 38

19.6.1.b ~~24.5.6~~ ESTIMATING THE EFFECT OF TREATMENT ON THE TREATED

Consider a regression approach to analyzing treatment effects in a two-period setting,

$$y_{it} = \theta_t + \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma C_i + u_i + \varepsilon_{it}, \quad t = 0, 1,$$

[~~37~~ 38] Again, Angrist (2001) is an important contribution to this literature.

[~~38~~ 39] This particular work considers selection in a "panel" (mainly two periods). But, the panel data setting for sample selection models is more involved than a cross-section analysis. In a panel data set, the "selection" is likely to be a decision at the beginning of Period 1 to be in the data set for all subsequent periods. As such, something more intricate than the model we have considered here is called for.

19-75

where C_i is the treatment dummy variable and u_i is the unobserved individual effect. The setting is the pre- and post-treatment analysis of the sort considered in this section, where we examine the impact of a job training program on post training earnings. Because there are two periods, a natural approach to the analysis is to examine the changes,

$$\Delta y_i = (\theta_1 - \theta_0) + \gamma \Delta C_i + (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \Delta \varepsilon_{it}$$

where $\Delta C_i = 1$ for the treated and 0 for the nontreated individuals, and the first differences eliminate the unobserved individual effects. In the absence of controls (regressors, $\mathbf{x}_{it}$), or assuming that the controls are unchanged, the estimator of the effect of the treatment will be

$$\hat{\gamma} = \overline{\Delta y} \,|\, (\Delta C_i = 1) - \overline{\Delta y} \,|\, (C_i = 0),$$

which is the KT **difference in differences** estimator. This simplifies the problem considerably, but has several shortcomings. Most important, by using the simple differences, we have lost our ability to discern what induced the change, whether it was the program or something else, presumably in $\mathbf{x}_{it}$.

Even without the normality assumption, the preceding regression approach is more tightly structured than many are comfortable with. A considerable amount of research has focused on what assumptions are needed to reach that model and whether they are likely to be appropriate in a given setting.[39 40] The overall objective of the analysis of the preceding two sections is to evaluate the effect of a treatment, C_i, on the individual treated. The implicit counterfactual is an observation on what the "response" (dependent variable) of the treated individual would have been had they not been treated. But, of course, an individual will be in one state or the other, not both. Denote by y_0 the random variable that is the outcome variable in the absence of the treatment and by y_1 the outcome when the treatment has taken place. The **average treatment effect**, averaged over the entire population is

FN 39

AU: These 2 terms have already been KTS (msp 69). Here also?

$$ATE = E[y_1 - y_0].$$

This is the impact of the treatment on an individual drawn at random from the entire population. However, the desired quantity is not necessarily the ATE, but the **average treatment effect on the treated**, which would be

$$ATE \,|\, T = E[y_1 - y_0 \,|\, C = 1].$$

The difficulty of measuring this is, once again, the counterfactual, $E[y_0 \,|\, C = 1]$. Whether these two measures will be the same is at the center of the much of the discussion on this subject. If treatment is completely randomly assigned, then $E[y_j \,|\, C = 1] = E[y_j \,|\, C = 0] = E[y_j \,|\, C = j],\ j = 0, 1$. This means that with completely random treatment assignment

$$ATE = E[y_1 \,|\, C = 1] - E[y_0 \,|\, C = 0].$$

To put this in our example, if college attendance were completely randomly distributed throughout the population, then the impact of college attendance on income (neglecting other covariates at this point), could be measured simply by averaging the incomes of

39 40 A sampling of the more important parts of the literature on this issue includes Heckman (1992, 1997), Imbens and Angrist (1994), Manski (1996), and Wooldridge (2002a, Chapter 18).

19-76

college attendees and subtracting the average income of nonattendees. The preceding theory might work for the treatment "having brown eyes," but it is unlikely to work for college attendance. Not only is the college attendance treatment not randomly distributed, but the treatment "assignment" is surely related to expectations about y_1 versus y_0, and, at a minimum, y_0 itself. (College is expensive.) More generally, the researcher faces the difficulty in calculating treatment effects that assignment to the treatment might not be exogenous.

19-34 19-36

The **control function** approach that we used in (~~24-22~~)–(~~24-24~~) is used to account for the endogeneity of the treatment assignment in the regression context. The very specific assumptions of the bivariate normal distribution of the unobservables somewhat simplifies the estimation, because they make explicit what control function (λ_i) is appropriate to use in the regression. As Wooldridge (2002a, p. 622) points out, however, the binary variable in the treatment effects regression represents simply an endogenous variable in a linear equation, amenable to **instrumental variable estimation** (assuming suitable instruments are available). Barnow, Cain, and Goldberger (1981) proposed a two-stage least squares estimator, with instrumental variable equal to the predicted probability from the probit treatment assignment model. This is slightly less **parametric** than 19-36 (~~22-24~~) because, in principle, its validity does not rely on joint normality of the disturbances. [Wooldridge (2002a, pp. 621–633) discusses the underlying assumptions.]

AU: Term "parametric" is not in chap. list

If the treatment assignment is "completely ignorable," then, as noted, estimation of the treatment effects is greatly simplified. Suppose, as well, that there are observable variables that influence both the outcome and the treatment assignment. Suppose it is possible to obtain pairs of individuals matched by a common $\mathbf{x}_i$, one with $C_i = 0$, the other with $C_i = 1$. If done with a sufficient number of pairs so as to average over the population of $\mathbf{x}_i$'s, then a **matching estimator,** the average value of $(y_i \mid C_i = 1) - (y_i \mid C_i = 0)$ would estimate $E[y_1 - y_0]$, which is what we seek. Of course, it is optimistic to hope to find a large sample of such matched pairs, both because the sample overall is finite and because there may be many regressors, and the "cells" in the distribution of $\mathbf{x}_i$ are likely to be thinly populated. This will be worse when the regressors are continuous, for example, with a "family income" variable. Rosenbaum and Rubin (1983) 41 and others[35] suggested, instead, matching on the **propensity score,** $F(\mathbf{x}_i) = \text{Prob}(C_i = 1 \mid \mathbf{x}_i)$. Individuals with similar propensity scores are paired and the average treatment effect is then estimated by the differences in outcomes. Various strategies are suggested by the authors for obtaining the necessary subsamples and for verifying the conditions under which the procedures will be valid. [See, e.g., Becker and Ichino (2002) and Greene (2007c).]

Example 24.10 Treatment Effects on Earnings

LaLonde (1986) analyzed the results of a labor market experiment, The National Supported Work Demonstration, in which a group of disadvantaged workers lacking basic job skills were given work experience and counseling in a sheltered environment. Qualified applicants were assigned to training positions randomly. The treatment group received the benefits of the program. Those in the control group "were left to fend for themselves." [The demonstration was run in numerous cities in the mid-1970s. See LaLonde (1986, pp. 605–609) for institutional

[35] Other important references in this literature are Becker and Ichino (1999), Dehejia and Wahba (1999), LaLonde (1986), Heckman, Ichimura, and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998), Heckman, LaLonde, and Smith (1999), Heckman, Tobias, and Vytlacil (2003), and Heckman and Vytlacil (2000).

19-77

19.6.2 Propensity Score Matching

If the treatment assignment is "completely ignorable," then, as noted, estimation of the treatment effects is greatly simplified. Suppose, as well, that there are observable variables that influence both the outcome and the treatment assignment. Suppose it is possible to obtain pairs of individuals matched by a common $\mathbf{x}_i$, one with $C_i = 0$, the other with $C_i = 1$. If done with a sufficient number of pairs so as to average over the population of $\mathbf{x}_i$'s, then a **matching estimator,** the average value of $(y_i \mid C_i = 1) - (y_i \mid C_i = 0)$ would estimate $E[y_1 - y_0]$, which is what we seek. Of course, it is optimistic to hope to find a large sample of such matched pairs, both because the sample overall is finite and because there may be many regressors, and the "cells" in the distribution of $\mathbf{x}_i$ are likely to be thinly populated. This will be worse when the regressors are continuous, for example, with a "family income" variable. Rosenbaum and Rubin (1983) and others[41] suggested, instead, matching on the **propensity score,** $F(\mathbf{x}_i) = \text{Prob}(C_i = 1|\mathbf{x}_i)$. Individuals with similar propensity scores are paired and the average treatment effect is then estimated by the differences in outcomes. Various strategies are suggested by the authors for obtaining the necessary subsamples and for verifying the conditions under which the procedures will be valid. [See, e.g., Becker and Ichino (2002) and Greene (2007c).]

Example 19.15 Treatment Effects on Earnings

LaLonde (1986) analyzed the results of a labor market experiment, The National Supported Work Demonstration, in which a group of disadvantaged workers lacking basic job skills were given work experience and counseling in a sheltered environment. Qualified applicants were assigned to training positions randomly. The treatment group received the benefits of the program. Those in the control group "were left to fend for themselves." [The demonstration was run in numerous cities in the mid-1970s. See LaLonde (1986, pp. 605–609) for institutional

[41] Other important references in this literature are Becker and Ichino (1999), Dehejia and Wahba (1999), LaLonde (1986), Heckman, Ichimura, and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998), Heckman, LaLonde, and Smith (1999), Heckman, Tobias, and Vytlacil (2003), and Heckman and Vytlacil (2000).

19-78

details on the NSW experiments.] The training period was 1976–1977; the outcome of interest for the sample examined here was post-training 1978 earnings. LaLonde reports a large variety of estimates of the treatment effect, for different subgroups and using different estimation methods. Nonparametric estimates for the group in our sample are roughly \$900 for the income increment in the post-training year. (See LaLonde, p. 609.) Similar results are reported from a two-step regression-based estimator similar to ~~(24-22)~~ 19-34 to ~~(24-24)~~ 19-36. (See LaLonde's footnote to Table 6, p. 616.)

LaLonde's data are fairly well traveled, having been used in replications and extensions in, for examples ~~e.g.,~~ Dehejia and Wahba (1999), Becker and Ichino (2002), and Greene (2007b, c). We have reestimated the matching estimates reported in Becker and Ichino. The data in the file used there (and here) contain 2,490 control observations and 185 treatment observations on the following variables:

t = treatment dummy variable,
age = age in years,
educ = education in years,
marr = dummy variable for married,
black = dummy variable for black,
hisp = dummy variable for Hispanic,
nodegree = dummy for no degree (not used),
*re*74 = real earnings in 1974,
*re*75 = real earnings in 1975,
*re*78 = real earnings in 1978.

Transformed variables added to the equation are

*age*2 = *age* squared,
*educ*2 = *educ* squared,
*re*742 = *re*74 squared,
*re*752 = *re*75 squared,
*blacku*74 = *black* times 1(*re*74 = 0).

3 We also scaled all earnings variables by 10,000 before beginning the analysis. (See Appendix URL
19.X Table F24.2) The data are downloaded from the website http://www.nber.org/%7Erdehejia/nswdata.html. The two specific subsamples are in http://www.nber.org/%7Erdehejia//psid_controls.txt and http://www.nber.org/%7Erdehejia/nswre74_treated.txt.) (We note that URL Becker and Ichino report they were unable to replicate Dehejia and Wahba's results, although they could come reasonably close. We, in turn, were not able to replicate either set of results, though we, likewise, obtained quite similar results.)

The analysis proceeded as follows: A logit model in which the included variables were a *constant*, *age*, *age*^{2}, *education*, *education*^{2}, *marr*, *black*, *hisp*, *re74*, *re75*, *re742*, *re752*, and *black74* was computed for the treatment assignment. The fitted probabilities are used for the propensity scores. By means of an iterative search, the range of propensity scores was partitioned into 8 regions within which, by a simple *F* test, the mean scores of the treatments and controls were not statistically different. The partitioning is shown in Table ~~24.5.~~ 19.10 The 1,347 observations are all the treated observations and the 1,162 control observations are those whose propensity scores fell within the range of the scores for the treated observations.

eight/

TB 19.10

Within each interval, each treated observation is paired with a small number of the nearest control observations. We found the average difference between treated observation and control to equal \$1,574.35. Becker and Ichino reported \$1,537.94.

As an experiment, we refit the propensity score equation using a probit model, retaining the fitted probabilities. We then used the two-step estimator described earlier to fit ~~(24-22)~~

19-34

19-79

19.10

TABLE ~~24.5~~ Empirical Distribution of Propensity Scores

Percent	*Lower*	*Upper*
0–5	0.000591	0.000783
5–10	0.000787	0.001061
10–15	0.001065	0.001377
15–20	0.001378	0.001748
20–25	0.001760	0.002321
25–30	0.002340	0.002956
30–35	0.002974	0.004057
35–40	0.004059	0.005272
40–45	0.005278	0.007486
45–50	0.007557	0.010451
50–55	0.010563	0.014643
55–60	0.014686	0.022462
60–65	0.022621	0.035060
65–70	0.035075	0.051415
70–75	0.051415	0.076188
75–80	0.076376	0.134189
80–85	0.134238	0.320638
85–90	0.321233	0.616002
90–95	0.624407	0.949418
95–100	0.949418	0.974835

Sample size = 1,347
Average score = 0.137238
Std. Dev score = 0.274079

	Lower	Upper	# obs
1	0.000591	0.098016	1041
2	0.098016	0.195440	63
3	0.195440	0.390289	65
4	0.390289	0.585138	36
5	0.585138	0.779986	32
6	0.779986	0.877411	17
7	0.877411	0.926123	7
8	0.926123	0.974835	86

19-35

and (~~24-23~~) using the entire sample. The estimates of δ, ρ, and σ were $(-1.01437, 0.35519, 1.38426)$. Using the results from the probit model, we averaged the result in (~~24-24~~) for the entire sample, obtaining an estimated treatment effect of \$1,476.30. 19-36

24.5.7 SAMPLE SELECTION IN NONLINEAR MODELS

The preceding analysis has focused on an extension of the linear regression (or the estimation of simple averages of the data). The method of analysis changes in nonlinear models. To begin, it is not necessarily obvious what the impact of the sample selection is on the response variable, or how it can be accommodated in a model. Consider the model analyzed by Boyes, Hoffman, and Lowe (1989):

$$y_{i1} = 1 \text{ if individual } i \text{ defaults on a loan, 0 otherwise,}$$
$$y_{i2} = 1 \text{ if the individual is granted a loan, 0 otherwise.}$$

Wynand and van Praag (1981) also used this framework to analyze consumer insurance purchases in the first application of the selection methodology in a nonlinear model. Greene (1992) applied the same model to y_1 = default on credit card loans, in which y_{i2} denotes whether an application for the card was accepted or not. [Mohanty (2002) also used this model to analyze teen employment in California.] For a given individual, y_1 is not observed unless $y_{i2} = 1$. Following the lead of the linear regression case in Section 24.5.3, a natural approach might seem to be to fit the second (selection) equation using a univariate probit model, compute the inverse Mills ratio, λ_i, and add it to the first equation as an additional "control" variable to accommodate the selection effect. [This is the approach used by Wynand and van Praag (1981).] The problems with this control function approach are, first, it is unclear what in the model is being "controlled" and, second, assuming the first model is correct, the appropriate model conditioned

19-80

19.6.3 Regression Discontinuity

There are many situations in which there is no possibility of randomized assignment of treatments. Examples include student outcomes and policy interventions in schools. Angrist and Lavy (1999), for example, studied the effect of class sizes on test scores. Van der Klaauw studied financial aid offers that were tied to SAT scores and grade point averages. In these cases, the natural experiment approach advocated by Angrist and Pischke (2009) is an appealing way to proceed, when it is feasible. The **regression discontinuity design** presents an alternative strategy. The conditions under which the approach can be effective are: (1) The outcome, y, is a continuous variable; (2) the outcome varies smoothly with an assignment variable, A, (3) treatment is "sharply" assigned based on the value of A, specifically, $C = 1(A > A^*)$ where A^* is a fixed threshold or cutoff value. [A "**fuzzy design** is based on $\text{Prob}(C = 1|A) = F(A)$. The identification problems with fuzzy design are much more complicated than with sharp design. Readers are referred to Van der Klaauw (2002) for further discussion of fuzzy design.] We assume, then, that

when/ and/

AU: Term "fuzzy design" not in chap. list

$$y = f(A,C) + \varepsilon.$$

Suppose, for example, the outcome variable is a test score, and that an administrative treatment such as a special education program is funded based on the poverty rates of certain communities. The ideal conditions for a regression discontinuity design based on these assumptions is shown in Figure 19.8. The logic of the calculation is that the points near the threshold value, which have "essentially" the same stimulus value, constitute a nearly random sample of observations which are segmented by the treatment.

FIG 19.8

The method requires that $E[\varepsilon|A,C] = E[\varepsilon|A]$ – the assignment variable is exogenous to the experiment. The result in Figure 19.8 is consistent with

$$y = f(A) + \alpha C + \varepsilon,$$

where α will be the treatment effect to be estimated. The specification of $f(A)$ can be problematic. Assuming a linear function when something more general will bias the estimate of α. For this reason, nonparametric methods, such as the LOWESS regression (see Section 12.3.5) might be attractive. This is likely to enable the analyst to make fuller use of the observations that are moer distant from the cutoff point. [See Van der Klaaus (2002).] Identification of the treatment effect begins with the assumption that $f(A)$ is continuous at A^*, so that

$$\lim_{A\uparrow A^*} f(A) = \lim_{A\downarrow A^*} f(A) = f(A^*).$$

then

$$\lim_{A\downarrow A^*} E[y\,|\,A] - \lim_{A\uparrow A^*} E[y\,|\,A] = f(A^*) + \alpha + \lim_{A\downarrow A^*} E[\varepsilon\,|\,A] - f(A^*) - \lim_{A\uparrow A^*} E[\varepsilon\,|\,A]$$
$$= \alpha.$$

With this in place, the treatment effect can be estimated by the difference of the average outcomes for those individuals "close" to the threshold value, A^*. Details on regression discontinuity design are provided by Trochim (1984,2000) and Van der Klaauw (2002).

19-81

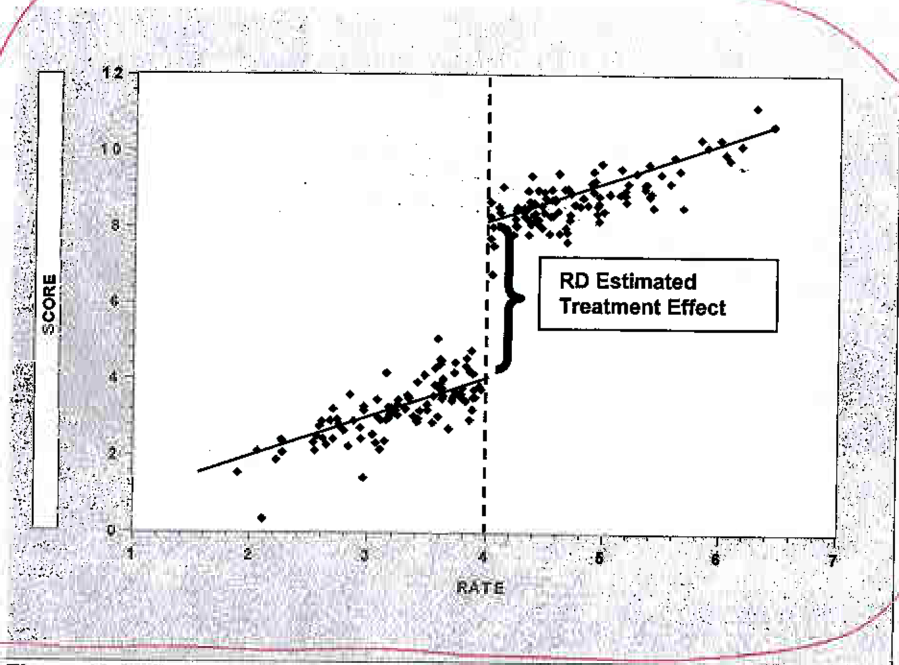

Figure 19.8 Regression Discontinuity

19-82

19.7 SUMMARY AND CONCLUSIONS

This chapter has examined settings in which, in principle, the linear regression model of Chapter 2 would apply, but the data generating mechanism produces a nonlinear form: truncation, censoring, and sample selection or endogenous sampling. For each case, we develop the basic theory of the effect, then use the results in a major area of research in econometrics.

and

In the truncated regression model, the range of the dependent variable is restricted substantively. Certainly all economic data are restricted in this way—aggregate income data cannot be negative, for example. But when data are truncated so that plausible values of the dependent variable are precluded, for example, when zero values for expenditure are discarded, the data that remain are analyzed with models that explicitly account for the truncation. The stochastic frontier model is based on a composite disturbance in which one part follows the assumptions of the familiar regression model while the second component is built on a platform of the truncated regression.

When data are censored, values of the dependent variable that could in principle be observed are masked. Ranges of values of the true variable being studied are observed as a single value. The basic problem this presents for model building is that in such a case, we observe variation of the independent variables without the corresponding variation in the dependent variable that might be expected. Consistent estimation, and useful interpretation of estimation results are based on maximum likelihood or some other technique that explicitly accounts for the censoring mechanism. The most common case of censoring in observed data arises in the context of duration analysis, or survival functions (which borrows a term from medical statistics where this style of model building originated). It is useful to think of duration, or survival data, as the measurement of time between transitions or changes of state. We examined three modeling approaches that correspond to the description in Chapter 12, nonparametric (survival tables), semiparametric (the proportional hazard models), and parametric (various forms such as the Weibull model).

Finally, the issue of sample selection arises when the observed data are not drawn randomly from the population of interest. Failure to account for this nonrandom sampling produces a model that describes only the nonrandom subsample, not the larger population. In each case, we examined the model specification and estimation techniques which are appropriate for these variations of the regression model. Maximum likelihood is usually the method of choice, but for the third case, a two-step estimator has become more common. The leading contemporary application of selection methods and endogenous sampling is in the measure of treatment effects. We considered three approaches to analysis of treatment efficets, regression methods, propensity score matching and regression discontinuity.

(effects)

19–83

AU: Terms with blue checks were not bold KTs in text

Key Terms and Concepts

- Accelerated failure time model
- Attrition
- Average treatment effect
- Average treatment effect on the treated
- Attenuation
- Censored regression model
- Censored variable
- Censoring
- Conditional moment test
- Control function
- Corner solution model
- Data envelopment analysis
- Degree of truncation
- Delta method
- Difference in differences
- Duration dependence
- Duration model
- Endogeneity
- Exponential model
- Generalized residual
- Hazard function
- Hazard rate
- Heterogeneity
- Heteroscedasticity
- Hurdle model
- Incidental truncation
- Instrumetal variable estimation
- Integrated hazard function
- Intensity equation
- Inverse probability weighted estimator
- Inverse Mills ratio
- Lagrange multiplier test
- Marginal effects
- Matching estimator
- Mean independence assumption
- Missing counterfactual
- Negative duration dependence
- Olsen's reparameterization
- Parametric model
- Partial likelihood
- Participation equation
- Positive duration dependence
- Product limit estimator
- Propensity score
- Proportional hazard
- Regression discontinuity design
- Risk set
- Rubin causal model
- Sample selection
- Selection on observables
- Selection on unobservables
- Semiparametric estimator
- Semiparametric model
- Specification error
- Stochastic frontier model
- Survival function
- Time-varying covariate
- Tobit model
- Treatment effect
- Truncated bivariate normal distribution

19-84

- Truncated distribution
- Truncated mean
- Truncated normal distribution
- Truncated random variable
- Truncated standard normal distribution
- Truncated variance
- Truncation
- Two-part model
- Two-step estimation
- Type II tobit model
- Weibull survival model

1,9 - 85

24.6 SUMMARY AND CONCLUSIONS

This chapter has examined three settings in which, in principle, the linear regression model of Chapter 2 would apply, but the data generating mechanism produces a nonlinear form. In the truncated regression model, the range of the dependent variable is restricted substantively. Certainly all economic data are restricted in this way—aggregate income data cannot be negative, for example. But when data are truncated so that plausible values of the dependent variable are precluded, for example, when zero values for expenditure are discarded, the data that remain are analyzed with models that explicitly account for the truncation. When data are censored, values of the dependent variable that could in principle be observed are masked. Ranges of values of the true variable being studied are observed as a single value. The basic problem this presents for model building is that in such a case, we observe variation of the independent variables without the corresponding variation in the dependent variable that might be expected. Finally, the issue of sample selection arises when the observed data are not drawn randomly from the population of interest. Failure to account for this nonrandom sampling produces a model that describes only the nonrandom subsample, not the larger population. In each case, we examined the model specification and estimation techniques which are appropriate for these variations of the regression model. Maximum likelihood is usually the method of choice, but for the third case, a two-step estimator has become more common.

Key Terms and Concepts

- Attrition
- Average treatment effect
- Average treatment effect on the treated
- Attenuation
- Censored regression model
- Censored variable
- Censoring
- Conditional moment test
- Control function
- Corner solution
- Degree of truncation
- Delta method
- Difference in differences
- Generalized residual
- Hazard function
- Heterogeneity
- Heteroscedasticity
- Incidental truncation
- Instrumetal variable estimation
- Inverse probability weighted estimator
- Inverse Mills ratio
- Lagrange multiplier test
- Marginal effects
- Matching estimator
- Mean independence assumption
- Olsen's reparameterization
- Parametric model
- Propensity score
- Sample selection
- Semiparametric estimator
- Specification error
- Tobit model
- Treatment effect
- Truncated bivariate normal distribution
- Truncated distribution
- Truncated mean
- Truncated normal distribution
- Truncated random variable
- Truncated standard normal distribution
- Truncated variance
- Two-step estimation

Exercises

1. The following 20 observations are drawn from a censored normal distribution:

3.8396	7.2040	0.00000	0.00000	4.4132	8.0230
5.7971	7.0828	0.00000	0.80260	13.0670	4.3211
0.00000	8.6801	5.4571	0.00000	8.1021	0.00000
1.2526	5.6016				

19-86

The applicable model is

$$y_i^* = \mu + \varepsilon_i,$$
$$y_i = y_i^* \quad \text{if } \mu + \varepsilon_i > 0, 0 \text{ otherwise},$$
$$\varepsilon_i \sim N[0, \sigma^2].$$

Exercises 1 through 4 in this section are based on the preceding information. The OLS estimator of μ in the context of this tobit model is simply the sample mean. Compute the mean of all 20 observations. Would you expect this estimator to over- or underestimate μ? If we consider only the nonzero observations, then the truncated regression model applies. The sample mean of the nonlimit observations is the least squares estimator in this context. Compute it and then comment on whether this sample mean should be an overestimate or an underestimate of the true mean.

2. We now consider the tobit model that applies to the full data set.
   a. Formulate the log-likelihood for this very simple tobit model.
   b. Reformulate the log-likelihood in terms of $\theta = 1/\sigma$ and $\gamma = \mu/\sigma$. Then derive the necessary conditions for maximizing the log-likelihood with respect to θ and γ.
   c. Discuss how you would obtain the values of θ and γ to solve the problem in part b.
   d. Compute the maximum likelihood estimates of μ and σ.
3. Using only the nonlimit observations, repeat Exercise 2 in the context of the truncated regression model. Estimate μ and σ by using the method of moments estimator outlined in Example ~~24.2~~ 19.2. Compare your results with those in the previous exercises.
4. Continuing to use the data in Exercise 1, consider once again only the nonzero observations. Suppose that the sampling mechanism is as follows: y^* and another normally distributed random variable z have population correlation 0.7. The two variables, y^* and z, are sampled jointly. When z is greater than zero, y is reported. When z is less than zero, both z and y are discarded. Exactly 35 draws were required to obtain the preceding sample. Estimate μ and σ. (Hint: Use Theorem ~~24.5~~ 19.5.)
5. Derive the ~~marginal~~ partial effects for the tobit model with heteroscedasticity that is described in Section ~~24.3.4.a.~~ 19.3.5.a.
6. Prove that the Hessian for the tobit model in (~~24-14~~ 19-14) is negative definite after Olsen's transformation is applied to the parameters.

Applications

1. ~~In Section 25.5.2,~~ We will examine Ray Fair's famous analysis of a *Psychology Today* survey on extramarital affairs. Although the dependent variable used in that study was a count, Fair used the tobit model as the platform for his study. Our analysis ~~in Section 25.5.2 will examine~~ the study, using a Poisson model for counts instead. Fair's original study also included but did not analyze a second data set that was a similar survey conducted by *Redbook* magazine. The data are reproduced in

in Example 18.9

examined

19-87

Applications

1. We examined Ray Fair's famous analysis (*Journal of Political Economy*, 1978) of a *Psychology Today* survey on extramarital affairs in Example 18.9 using a Poisson regression model. Although the dependent variable used in that study was a count, Fair (1978) used the tobit model as the platform for his study. You can reproduce the tobit estimates in Fair's paper easily with any software package that contains a tobit estimator most do. The data appear in Appendix Table F18.1. Reproduce Fair's least squares and tobit estimates. Compute the partial effects for the model and interpret all results.

2. Fair's original study also included but did not analyze a second data set that was a similar 17.2 survey conducted by *Redbook* magazine. The data are reproduced in Appendix Table F~~19~~.2. (Our thanks to Ray Fair for providing these data.) This sample contains observations on 6,366 women and the following variables:

> *id* = an identification number,
> *C* = constant, value = 1,
> *yrb* = a constructed measure of time spent in extramarital affairs,
> *v1* = a rating of the marriage, coded 1 to 4,
> *v2* = age, in years, aggregated,
> *v3* = number of years married,
> *v4* = number of children, top coded at 5,
> *v5* = religiosity, 1 to 4, 1 = not, 4 = very,
> *v6* = education, coded 9, 12, 14, 16, 17, 20,
> *v7* = occupation,
> *v8* = husband's occupation.

Three other variables were not used. Details on the variables in the model are given in Fair's (1978) *Journal of Political Economy* paper. Using these data, conduct a parallel study to the *Psychology Today* study that was done in Fair (1978). Are the results consistent? Report all results, including partial effects and relevant diagnostic statistics.

3. Continuing the analysis of the previous application, note that these data conform precisely to the description of "corner solutions" in Section 19.3.4. The dependent variable is not censored in the fashion usually assumed for a tobit model. To investigate whether the dependent variable is determined by a two-part decision process (yes/no and, if yes, how much), specify and estimate a two-equation model in which the first equation analyzes the binary decision *A* = 1 if *yrb* > 0 and 0 otherwise and the second equation analyzes *yrb* | *yrb* > 0. What is the appropriate model? What do you find? Report all results. (Note, if you analyze the second dependent variable using the truncated regression, you should remove some extreme observations from your sample. The truncated regression estimator refuses to converge with the full data set, but works nicely for the example if you omit observations with *yrb* > 5.)

19-88

END 19

AU: Check style of application title, OK bold?

4. **StochasticFrontier Model.** Section 10.5.1 presents estimates of a Cobb-Douglas cost function using Nerlove's 1955 data on the U.S. electric power industry. Christensen and Greene's 1976 update of this study used 1970 data for this industry. The Christensen and Greene data are given in Appendix Table F4.3.These data have provided a standard test data set for estimating different forms of production and cost functions, including the stochastic frontier model discussed in Section 19.2.4. It has been suggested that one explanation for the apparent finding of economies of scale in these data is that the smaller firms were inefficient for other reasons. The stochastic frontier might allow one to disentangle these effects. Use these data to fit a frontier cost function which includes a quadratic term in log output in addition to the linear term and the factor prices. Then examine the estimated Jondrow et al. residuals to see if they do indeed vary negatively with output, as suggested. (This will require either some programming on your part or specialized software. The stochastic frontier model is provided as an option in Stata, TSP and LIMDEP. Or, the likelihood function can be programmed fairly easily for RATS, MatLab or GAUSS. Note, for a cost frontier as opposed to a production frontier, it is necessary to reverse the sign on the argument in the Φ function that appears in the log-likelihood.)