this example, but not in all such examples, alternative consistent estimators for a subset of the regression parameters are available.

4.8. Instrumental Variables

A major complication that is emphasized in microeconometrics is the possibility of inconsistent parameter estimation caused by endogenous regressors. Then regression estimates measure only the magnitude of association, rather than the magnitude and direction of causation, both of which are needed for policy analysis.

The instrumental variables estimator provides a way to nonetheless obtain consistent parameter estimates. This method, widely used in econometrics and rarely used elsewhere, is conceptually difficult and easily misused.

We provide a lengthy expository treatment that defines an instrumental variable and explains how the instrumental variables method works in a simple setting.

4.8.1. Inconsistency of OLS

Consider the scalar regression model with dependent variable y and single regressor x. The goal of regression analysis is to estimate the conditional mean function $\mathrm{E}[y|x]$. A linear conditional mean model, without intercept for notational convenience, specifies

$$\mathrm{E}[y|x] = \beta x. \tag{4.42}$$

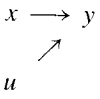
This model without intercept subsumes the model with intercept if dependent and regressor variables are deviations from their respective means. Interest lies in obtaining a consistent estimate of β as this gives the change in the conditional mean given an *exogenous* change in x. For example, interest may lie in the effect in earnings caused by an increase in schooling attributed to exogenous reasons, such as an increase in the minimum age at which students leave school, that are not a choice of the individual.

The OLS regression model specifies

$$y = \beta x + u, \tag{4.43}$$

where u is an error term. Regression of y on x yields OLS estimate $\widehat{\beta}$ of β.

Standard regression results make the assumption that the regressors are uncorrelated with the errors in the model (4.43). Then the only effect of x on y is a direct effect via the term βx. We have the following path analysis diagram:



where there is no association between x and u. So x and u are independent causes of y.

However, in some situations there may be an association between regressors and errors. For example, consider regression of log-earnings (y) on years of schooling (x). The error term u embodies all factors other than schooling that determine earnings,

such as ability. Suppose a person has a high level of u, as a result of high (unobserved) ability. This increases earnings, since $y = \beta x + u$, but it may also lead to higher levels of x, since schooling is likely to be higher for those with high ability. A more appropriate path diagram is then the following:

$$\begin{array}{ccc} x & \longrightarrow & y \\ \uparrow & \nearrow & \\ u & & \end{array}$$

where now there is an association between x and u.

What are the consequences of this correlation between x and u? Now higher levels of x have two effects on y. From (4.43) there is both a direct effect via βx and an indirect effect via u affecting x, which in turn affects y. The goal of regression is to estimate only the first effect, yielding an estimate of β. The OLS estimate will instead combine these two effects, giving $\widehat{\beta} > \beta$ in this example where both effects are positive. Using calculus, we have $y = \beta x + u(x)$ with total derivative

$$\frac{dy}{dx} = \beta + \frac{du}{dx}. \tag{4.44}$$

The data give information on dy/dx, so OLS estimates the total effect $\beta + du/dx$ rather than β alone. The OLS estimator is therefore biased and inconsistent for β, unless there is no association between x and u.

A more formal treatment of the linear regression model with K regressors leads to the same conclusion. From Section 4.7.1 a necessary condition for consistency of OLS is that plim $N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}$. Consistency requires that the regressors are asymptotically uncorrelated with the errors. From (4.37) the magnitude of the inconsistency of OLS is $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$, the OLS coefficient from regression of u on $\mathbf{x}$. This is just the OLS estimate of $du/d\mathbf{x}$, confirming the intuitive result in (4.44).

4.8.2. Instrumental Variable

The inconsistency of OLS is due to endogeneity of x, meaning that changes in x are associated not only with changes in y but also changes in the error u. What is needed is a method to generate only exogenous variation in x. An obvious way is through a randomized experiment, but for most economics applications such experiments are too expensive or even infeasible.

Definition of an Instrument

A crude experimental or treatment approach is still possible using observational data, provided there exists an **instrument** z that has the property that changes in z are associated with changes in x but do not lead to change in y (aside from the indirect route via x). This leads to the following path diagram:

$$\begin{array}{ccccc} z & \longrightarrow & x & \longrightarrow & y \\ & & \uparrow & \nearrow & \\ & & u & & \end{array}$$

which introduces a variable z that is causally associated with x but not u. It is still the case that z and y will be correlated, but the only source of such correlation is the indirect path of z being correlated with x, which in turn determines y. The more direct path of z being a regressor in the model for y is ruled out.

More formally, a variable z is called an **instrument** or **instrumental variable** for the regressor x in the scalar regression model $y = \beta x + u$ if (1) z is uncorrelated with the error u and (2) z is correlated with the regressor x.

The first assumption excludes the instrument z from being a regressor in the model for y, since if instead y depended on both x and z, and y is regressed on x alone, then z is being absorbed into the error so that z will then be correlated with the error. The second assumption requires that there is some association between the instrument and the variable being instrumented.

Examples of an Instrument

In many microeconometric applications it is difficult to find legitimate instruments. Here we provide two examples.

Suppose we want to estimate the response of market demand to exogenous changes in market price. Quantity demanded clearly depends on price, but prices are not exogenously given since they are determined in part by market demand. A suitable instrument for price is a variable that is correlated with price but does not directly affect quantity demanded. An obvious candidate is a variable that affects market supply, since this also affect prices, but is not a direct determinant of demand. An example is a measure of favorable growing conditions if an agricultural product is being modeled. The choice of instrument here is uncontroversial, provided favorable growing conditions do not directly affect demand, and is helped greatly by the formal economic model of supply and demand.

Next suppose we want to estimate the returns to exogenous changes in schooling. Most observational data sets lack measures of individual ability, so regression of earnings on schooling has error that includes unobserved ability and hence is correlated with the regressor schooling. We need an instrument z that is correlated with schooling, uncorrelated with ability, and more generally uncorrelated with the error term, which means that it cannot directly determine earnings.

One popular candidate for z is proximity to a college or university (Card, 1995). This clearly satisfies condition 2 because, for example, people whose home is a long way from a community college or state university are less likely to attend college. It most likely satisfies 1, though since it can be argued that people who live a long way from a college are more likely to be in low-wage labor markets one needs to estimate a multiple regression for y that includes additional regressors such as indicators for nonmetropolitan area.

A second candidate for the instrument is month of birth (Angrist and Krueger, 1991). This clearly satisfies condition 1 as there is no reason to believe that month of birth has a direct effect on earnings if the regression includes age in years. Surprisingly condition 2 may also be satisfied, as birth month determines age of first entry

into school in the USA, which in turn may affect years of schooling since laws often specify a minimum school-leaving age. Bound, Jaeger, and Baker (1995) provide a critique of this instrument.

The consequences of choosing poor instruments are considered in detail in Section 4.9.

4.8.3. Instrumental Variables Estimator

For regression with scalar regressor x and scalar instrument z, the **instrumental variables (IV) estimator** is defined as

$$\widehat{\beta}_{\text{IV}} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}, \tag{4.45}$$

where, in the scalar regressor case $\mathbf{z}$, $\mathbf{x}$ and $\mathbf{y}$ are $N \times 1$ vectors. This estimator provides a consistent estimator for the slope coefficient β in the linear model $y = \beta x + u$ if z is correlated with x and uncorrelated with the error term.

There are several ways to derive (4.45). We provide an intuitive derivation, one that differs from derivations usually presented such as that in Section 6.2.5.

Return to the earnings–schooling example. Suppose a one-unit change in the instrument z is associated with 0.2 more years of schooling and with a \$500 increase in annual earnings. This increase in earnings is a consequence of the indirect effect that increase in z led to increase in schooling, which in turn increases income. Then it follows that 0.2 years additional schooling is associated with a \$500 increase in earnings, so that a one-year increase in schooling is associated with a $\$500/0.2 = \$2{,}500$ increase in earnings. The causal estimate of β is therefore 2,500. In mathematical notation we have estimated the changes dx/dz and dy/dz and calculated the causal estimator as

$$\beta_{\text{IV}} = \frac{dy/dz}{dx/dz}. \tag{4.46}$$

This approach to identification of the causal parameter β is given in Heckman (2000, p. 58); see also the example in Section 2.4.2.

All that remains is consistent estimation of dy/dz and dx/dz. The obvious way to estimate dy/dz is by OLS regression of y on z with slope estimate $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$. Similarly, estimate dx/dz by OLS regression of x on z with slope estimate $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}$. Then

$$\widehat{\beta}_{\text{IV}} = \frac{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}}{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}. \tag{4.47}$$

4.8.4. Wald Estimator

A leading simple example of IV is one where the instrument z is a **binary instrument**. Denote the subsample averages of y and x by $\bar{y}_1$ and $\bar{x}_1$, respectively, when $z = 1$ and by $\bar{y}_0$ and $\bar{x}_0$, respectively, when $z = 0$. Then $\Delta y/\Delta z = (\bar{y}_1 - \bar{y}_0)$ and

$\Delta x/\Delta z = (\bar{x}_1 - \bar{x}_0)$, and (4.46) yields

$$\widehat{\beta}_{\text{Wald}} = \frac{(\bar{y}_1 - \bar{y}_0)}{(\bar{x}_1 - \bar{x}_0)}. \tag{4.48}$$

This estimator is called the **Wald estimator**, after Wald (1940), or the **grouping estimator**.

The Wald estimator can also be obtained from the formula (4.45). For the no-intercept model variables are measured in deviations from means, so $\mathbf{z}'\mathbf{y} = \sum_i (z_i - \bar{z})(y_i - \bar{y})$. For binary z this yields $\mathbf{z}'\mathbf{y} = N_1(\bar{y}_1 - \bar{y}) = N_1 N_0(\bar{y}_1 - \bar{y}_0)/N$, where N_0 and N_1 are the number of observations for which $z = 0$ and $z = 1$. This result uses $\bar{y}_1 - \bar{y} = (N_0\bar{y}_1 + N_1\bar{y}_1)/N - (N_0\bar{y}_0 + N_1\bar{y}_1)/N = N_0(\bar{y}_1 - \bar{y}_0)/N$. Similarly, $\mathbf{z}'\mathbf{x} = N_1 N_0(\bar{x}_1 - \bar{x}_0)/N$. Combining these results, we have that (4.45) yields (4.48).

For the earnings–schooling example it is being assumed that we can define two groups where group membership does not directly determine earnings, though it does affect level of schooling and hence indirectly affects earnings. Then the IV estimate is the difference in average earnings across the two groups divided by the difference in average schooling across the two groups.

4.8.5. Sample Covariance and Correlation Analysis

The IV estimator can also be interpreted in terms of covariances or correlations.

For sample covariances we have directly from (4.45) that

$$\widehat{\beta}_{\text{IV}} = \frac{\text{Cov}[z, y]}{\text{Cov}[z, x]}, \tag{4.49}$$

where here Cov[ ] is being used to denote sample covariance.

For sample correlations, note that the OLS estimator for the model (4.43) can be written as $\widehat{\beta}_{\text{OLS}} = r_{xy}\sqrt{\mathbf{y}'\mathbf{y}}/\sqrt{\mathbf{x}'\mathbf{x}}$, where $r_{xy} = \mathbf{x}'\mathbf{y}/\sqrt{(\mathbf{x}'\mathbf{x})(\mathbf{y}'\mathbf{y})}$ is the **sample correlation** between x and y. This leads to the interpretation of the OLS estimator as implying that a one standard deviation change in x is associated with an r_{xy} standard deviation change in y. The problem is that the correlation r_{xy} is contaminated by correlation between x and u. An alternative approach is to measure the correlation between x and y indirectly by the correlation between z and y divided by the correlation between z and x. Then

$$\widehat{\beta}_{\text{IV}} = \frac{r_{zy}}{r_{zx}} \frac{\sqrt{\mathbf{y}'\mathbf{y}}}{\sqrt{\mathbf{x}'\mathbf{x}}}, \tag{4.50}$$

which can be shown to equal $\widehat{\beta}_{\text{IV}}$ in (4.45).

4.8.6. IV Estimation for Multiple Regression

Now consider the multiple regression model with typical observation

$$y = \mathbf{x}'\boldsymbol{\beta} + u,$$

with K regressor variables, so that $\mathbf{x}$ and $\boldsymbol{\beta}$ are $K \times 1$ vectors.

Instruments

Assume the existence of an $r \times 1$ vector of **instruments z**, with $r \geq K$, satisfying the following:

1. **z** is uncorrelated with the error u.
2. **z** is correlated with the regressor vector **x**.
3. **z** is strongly correlated, rather than weakly correlated, with the regressor vector **x**.

The first two properties are necessary for consistency and were presented earlier in the scalar case. The third property, defined in Section 4.9.1, is a strengthening of the second to ensure good finite-sample performance of the IV estimator.

In the multiple regression case **z** and **x** may share some common components. Some components of **x**, called **exogenous regressors**, may be uncorrelated with u. These components are clearly suitable instruments as they satisfy conditions 1 and 2. Other components of **x**, called **endogenous regressors**, may be correlated with u. These components lead to inconsistency of OLS and are also clearly unsuitable instruments as they do not satisfy condition 1. Partition **x** into $\mathbf{x} = [\mathbf{x}_1' \ \mathbf{x}_2']'$, where $\mathbf{x}_1$ contains endogenous regressors and $\mathbf{x}_2$ contains exogenous regressors. Then a valid instrument is $\mathbf{z} = [\mathbf{z}_1' \ \mathbf{x}_2']'$, where $\mathbf{x}_2$ can be an instrument for itself, but we need to find at least as many instruments $\mathbf{z}_1$ as there are endogenous variables $\mathbf{x}_1$.

Identification

Identification in a simultaneous equations model was presented in Section 2.5. Here we have a single equation. The **order condition** requires that the number of instruments must at least equal the number of independent endogenous components, so that $r \geq K$. The model is said to be **just-identified** if $r = K$ and **overidentified** if $r > K$.

In many multiple regression applications there is only one endogenous regressor. For example, the earnings on schooling regression will include many other regressors such as age, geographic location, and family background. Interest lies in the coefficient on schooling, but this is an endogenous variable most likely correlated with the error because ability is unobserved. Possible candidates for the necessary single instrument for schooling have already been given in Section 4.8.2.

If an instrument fails the first condition the instrument is an **invalid instrument**. If an instrument fails the second condition the instrument is an **irrelevant instrument**, and the model may be **unidentified** if too few instruments are relevant. The third condition fails when very low correlation exists between the instrument and the endogenous variable being instrumented. The model is said to be **weakly identified** and the instrument is called a **weak instrument**.

Instrumental Variables Estimator

When the model is just-identified, so that $r = K$, the **instrumental variables estimator** is the obvious matrix generalization of (4.45)

$$\widehat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}, \tag{4.51}$$

where $\mathbf{Z}$ is an $N \times K$ matrix with ith row $\mathbf{z}_i'$. Substituting the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ for $\mathbf{y}$ in (4.51) yields

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{IV}} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}] \\ &= \boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u} \\ &= \boldsymbol{\beta} + (N^{-1}\mathbf{Z}'\mathbf{X})^{-1}N^{-1}\mathbf{Z}'\mathbf{u}.\end{aligned}$$

It follows immediately that the IV estimator is consistent if

$$\text{plim}\, N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$$

and

$$\text{plim}\, N^{-1}\mathbf{Z}'\mathbf{X} \neq \mathbf{0}.$$

These are essentially conditions 1 and 2 that $\mathbf{z}$ is uncorrelated with $\mathbf{u}$ and correlated with $\mathbf{x}$. To ensure that the inverse of $N^{-1}\mathbf{Z}'\mathbf{X}$ exists it is assumed that $\mathbf{Z}'\mathbf{X}$ is of full rank K, a stronger assumption than the order condition that $r = K$.

With heteroskedastic errors the IV estimator is asymptotically normal with mean $\boldsymbol{\beta}$ and variance matrix consistently estimated by

$$\widehat{\text{V}}[\widehat{\boldsymbol{\beta}}_{\text{IV}}] = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\mathbf{Z}(\mathbf{Z}'\mathbf{X})^{-1}, \tag{4.52}$$

where $\widehat{\boldsymbol{\Omega}} = \text{Diag}[\widehat{u}_i^2]$. This result is obtained in a manner similar to that for OLS given in Section 4.4.4.

The IV estimator, although consistent, leads to a loss of efficiency that can be very large in practice. Intuitively IV will not work well if the instrument $\mathbf{z}$ has low correlation with the regressor $\mathbf{x}$ (see Section 4.9.3).

4.8.7. Two-Stage Least Squares

The IV estimator in (4.51) requires that the number of instruments equals the number of regressors. For overidentified models the IV estimator can be used, by discarding some of the instruments so that the model is just-identified. However, an asymptotic efficiency loss can occur when discarding these instruments.

Instead, a common procedure is to use the **two-stage least-squares (2SLS) estimator**

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = \left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\right], \tag{4.53}$$

presented and motivated in Section 6.4.

The 2SLS estimator is an IV estimator. In a just-identified model it simplifies to the IV estimator given in (4.51) with instruments $\mathbf{Z}$. In an overidentified model the 2SLS estimator equals the IV estimator given in (4.51) if the instruments are $\widehat{\mathbf{X}}$, where $\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ is the predicted value of $\mathbf{x}$ from OLS regression of $\mathbf{x}$ on $\mathbf{z}$.

The 2SLS estimator gets its name from the result that it can be obtained by two consecutive OLS regressions: OLS regression of $\mathbf{x}$ on $\mathbf{z}$ to get $\widehat{\mathbf{x}}$ followed by OLS of y on $\widehat{\mathbf{x}}$, which gives $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$. This interpretation does not necessarily generalize to nonlinear regressions; see Section 6.5.6.

The 2SLS estimator is often expressed more compactly as

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left[\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}\right]^{-1}\left[\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y}\right], \tag{4.54}$$

where

$$\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$$

is an idempotent **projection matrix** that satisfies $\mathbf{P}_{\mathbf{Z}} = \mathbf{P}_{\mathbf{Z}}'$, $\mathbf{P}_{\mathbf{Z}}\mathbf{P}_{\mathbf{Z}}' = \mathbf{P}_{\mathbf{Z}}$, and $\mathbf{P}_{\mathbf{Z}}\mathbf{Z} = \mathbf{Z}$. The 2SLS estimator can be shown to be asymptotically normal distributed with estimated asymptotic variance

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}_{2SLS}] = N\left[\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}\right]^{-1}\left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\widehat{\mathbf{S}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]\left[\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}\right]^{-1}, \tag{4.55}$$

where in the usual case of heteroskedastic errors $\widehat{\mathbf{S}} = N^{-1}\sum_i \widehat{u}_i^2\mathbf{z}_i\mathbf{z}_i'$ and $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{2SLS}$. A commonly used small-sample adjustment is to divide by $N - K$ rather than N in the formula for $\widehat{\mathbf{S}}$.

In the special case that errors are homoskedastic, simplification occurs and $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}_{2SLS}] = s^2[\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}]^{-1}$. This latter result is given in many introductory treatments, but the more general formula (4.55) is preferred as the modern approach is to treat errors as potentially heteroskedastic.

For overidentified models with heteroskedastic errors an estimator that White (1982) calls the **two-stage instrumental variables estimator** is more efficient than 2SLS. Moreover, some commonly used model specification tests require estimation by this estimator rather than 2SLS. For details see Section 6.4.2.

4.8.8. IV Example

As an example of IV estimation, consider estimation of the slope coefficient of x for the dgp

$$y = 0 + 0.5x + u,$$
$$x = 0 + z + v,$$

where $z \sim \mathcal{N}[2, 1]$ and (u, v) are joint normal with means 0, variances 1, and correlation 0.8.

OLS of y on x yields inconsistent estimates as x is correlated with u since by construction x is correlated with v, which in turn is correlated with u. IV estimation yields consistent estimates. The variable z is a valid instrument since by construction is uncorrelated with u but is correlated with x. Transformations of z, such as z^3, are also valid instruments.

Various estimates and associated standard errors from a generated data sample of size 10,000 are given in Table 4.4. We focus on the slope coefficient.

The OLS estimator is inconsistent, with slope coefficient estimate of 0.902 being more than 50 standard errors from the true value of 0.5. The remaining estimates are consistent and are all within two standard errors of 0.5.

There are several ways to compute the IV estimator. The slope coefficient from OLS regression of y on z is 0.5168 and from OLS regression of x on z it is 1.0124,

Table 4.4. *Instrumental Variables Example*[a]

	OLS	IV	2SLS	IV (z^3)
Constant	−0.804	−0.017	−0.017	−0.014
	(0.014)	(0.022)	(0.032)	(0.025)
x	0.902	0.510	0.510	0.509
	(0.006)	(0.010)	(0.014)	(0.012)
R^2	0.709	0.576	0.576	0.574

[a] Generated data for a sample size of 10,000. OLS is inconsistent and other estimators are consistent. Robust standard errors are reported though they are unnecessary here as errors are homoskedastic. The 2SLS standard errors are incorrect. The data-generating process is given in the text.

yielding an IV estimate of $0.5168/1.0124 = 0.510$ using (4.47). In practice one instead directly computes the IV estimator using (4.45) or (4.51), with z used as the instrument for x and standard errors computed using (4.52). The 2SLS estimator (see (4.54)) can be computed by OLS regression of y on $\widehat{x}$, where $\widehat{x}$ is the prediction from OLS regression of x on z. The 2SLS estimates exactly equal the IV estimates in this just-identified model, though the standard errors from this OLS regression of y on $\widehat{x}$ are incorrect as will be explained in Section 6.4.5.

The final column uses z^3 rather than z as the instrument for x. This alternative IV estimator is consistent, since z^3 is uncorrelated with u and correlated with x. However, it is less efficient for this particular dgp, and the standard error of the slope coefficient rises from 0.010 to 0.012.

There is an efficiency loss in IV estimation compared to OLS estimation, see (4.61) for a general result for the case of single regressor and single instrument. Here $r^2_{x,z} = 0.510$, not given in Table 4.4, is high so the loss is not great and the standard error of the slope coefficient increases somewhat from 0.006 to 0.010. In practice the efficiency loss can be much greater than this.

4.9. Instrumental Variables in Practice

Important practical issues include determining whether IV methods are necessary and, if necessary, determining whether the instruments are valid. The relevant specification tests are presented in Section 8.4. Unfortunately, the validity of tests are limited. They require the assumption that in a just-identified model the instruments are valid and test only overidentifying restrictions.

Although IV estimators are consistent given valid instruments, as detailed in the following, IV estimators can be much less efficient than the OLS estimator and can have a finite-sample distribution that for usual finite-sample sizes differs greatly from the asymptotic distribution. These problems are greatly magnified if instruments are weakly correlated with the variables being instrumented. One way that weak instruments can arise is if there are many more instruments than needed. This is simply dealt with by dropping some of the instruments (see also Donald and Newey, 2001). A

more fundamental problem arises when even with the minimal number of instruments one or more of the instruments is weak.

This section focuses on the problem of weak instruments.

4.9.1. Weak Instruments

There is no single definition of a weak instrument. Many authors use the following signals of a **weak instrument**, presented here for progressively more complex models.

- Scalar regressor x and scalar instrument z: A weak instrument is one for which $r^2_{x,z}$ is small.
- Scalar regressor x and vector of instruments $\mathbf{z}$: The instruments are weak if the R^2 from regression of x on $\mathbf{z}$, denoted $R^2_{x,\mathbf{z}}$, is small or if the F-statistic for test of overall fit in this regression is small.
- Multiple regressors $\mathbf{x}$ with only one endogenous: A weak instrument is one for which the partial R^2 is low or the partial F-statistic is small, where these partial statistics are defined toward the end of Section 4.9.1.
- Multiple regressors $\mathbf{x}$ with several endogenous: There are several measures.

R^2 Measures

Consider a single equation

$$y = \beta_1 x_1 + \mathbf{x}_2'\beta_2 + u, \tag{4.56}$$

where just one regressor x_1 is endogenous and the remaining regressors in the vector $\mathbf{x}_2$ are exogenous. Assume that the instrument vector $\mathbf{z}$ includes the exogenous instruments $\mathbf{x}_2$, as well as least one other instrument.

One possible R^2 measure is the usual R^2 from regression of x_1 on $\mathbf{z}$. However, this could be high only because x_1 is highly correlated with $\mathbf{x}_2$ whereas intuitively we really need x_1 to be highly correlated with the instrument(s) other than $\mathbf{x}_2$.

Bound, Jaeger, and Baker (1995) therefore proposed use of a **partial $\mathbf{R}^2$**, denoted R^2_p, that purges the effect of $\mathbf{x}_2$. R^2_p is obtained as R^2 from the regression

$$(x_1 - \widetilde{x}_1) = (\mathbf{z} - \widetilde{\mathbf{z}})'\gamma + v, \tag{4.57}$$

where $\widetilde{x}_1$ and $\widetilde{\mathbf{z}}$ are the fitted values from regressions of x_1 on $\mathbf{x}_2$ and $\mathbf{z}$ on $\mathbf{x}_2$. In the just-identified case $\mathbf{z} - \widetilde{\mathbf{z}}$ will reduce to $z_1 - \widetilde{z}_1$, where z_1 is the single instrument other than $\mathbf{x}_2$ and $\widetilde{z}_1$ is the fitted value from regression of z_1 on $\mathbf{x}_2$.

It is not unusual for R^2_p to be much lower than $R^2_{x_1,\mathbf{z}}$. The formula for R^2_p simplifies to r^2_{xz} when there is only one regressor and it is endogenous. It further simplifies to $\text{Cor}[x, z]$ when there is only one instrument.

When there is more than one endogenous variable, analysis is less straightforward as a number of generalizations of R^2_p have been proposed.

Consider a single equation with more than one endogenous variable model and focus on estimation of the coefficient of the first endogenous variable. Then in (4.56)

x_1 is endogenous and additionally some of the variables in $\mathbf{x}_2$ are also endogenous. Several alternative measures replace the right-hand side of (4.57) with a residual that controls for the presence of other endogenous regressors. Shea (1997) proposed a partial R^2, say R_p^{*2}, that is computed as the squared sample correlation between $(x_1 - \widetilde{x}_1)$ and $(\widehat{x}_1 - \widetilde{\widehat{x}}_1)$. Here $(x_1 - \widetilde{x}_1)$ is again the residual from regression of x_1 on $\mathbf{x}_2$, whereas $(\widehat{x}_1 - \widetilde{\widehat{x}}_1)$ is the residual from regression of $\widehat{x}_1$ (the fitted value from regression of x_1 on $\mathbf{z}$) on $\widehat{\mathbf{x}}_2$ (the fitted value from regression of $\mathbf{x}_2$ on $\mathbf{z}$). Poskitt and Skeels (2002) proposed an alternative partial R^2, which, like Shea's R_p^{*2}, simplifies to R_p^2 when there is only one endogenous regressor. Hall, Rudebusch, and Wilcox (1996) instead proposed use of canonical correlations.

These measures for the coefficient for the first endogenous variable can be repeated for the other endogenous variables. Poskitt and Skeels (2002) additionally consider an R^2 measure that applies jointly to instrumentation of all the endogenous variables.

The problems of inconsistency of estimators and loss of precision are magnified as the partial R^2 measures fall, as detailed in Sections 4.9.2 and 4.9.3. See especially (4.60) and (4.62).

Partial F-Statistics

For poor finite-sample performance, considered in Section 4.9.4, it is common to use a related measure, the F-statistic for whether coefficients are zero in regression of the endogenous regressor on instruments.

For a single regressor that is endogenous we use the usual overall F-statistic, for a test of $\boldsymbol{\pi} = \mathbf{0}$ in the regression $x = \mathbf{z}'\boldsymbol{\pi} + v$ of the endogenous regressor on the instruments. This F-statistic is a function of $R^2_{x,\mathbf{z}}$.

More commonly, some exogenous regressors also appear in the model, and in model (4.56) with single endogenous regressor $\mathbf{x}_1$ we use the F-statistic for a test of $\boldsymbol{\pi}_1 = \mathbf{0}$ in the regression

$$x = \mathbf{z}_1'\boldsymbol{\pi}_1 + \mathbf{x}_2'\boldsymbol{\pi}_2 + v, \tag{4.58}$$

where $\mathbf{z}_1$ are the instruments other than the exogenous regressors and $\mathbf{x}_2$ are the exogenous regressors. This is the first-stage regression in the two-stage least-squares interpretation of IV.

This statistic is used as a signal of potential finite-sample bias in the IV estimator. In Section 4.9.4 we explain results of Staiger and Stock (1997) that suggest a value less than 10 is problematic and a value of 5 or less is a sign of extreme finite-sample bias and we consider extension to more than one endogenous regressor.

4.9.2. Inconsistency of IV Estimators

The essential condition for consistency of IV is condition 1 in Section 4.8.6, that the instrument should be uncorrelated with the error term. No test is possible in the just-identified case. In the overidentified case a test of the overidentifying assumptions is possible (see Section 6.4.3). Rejection then could be due to either instrument

endogeneity or model failure. Thus condition 1 is difficult to test directly and determining whether an instrument is exogenous is usually a subjective decision, albeit one often guided by economic theory.

It is always possible to create an exogenous instrument through **functional form restrictions**. For example, suppose there are two regressors so that $y = \beta_1 x_1 + \beta_2 x_2 + u$, with x_1 uncorrelated with u and x_2 correlated with u. Note that throughout this section all variables are assumed to be measured in departures from means, so that without loss of generality the intercept term can be omitted. Then OLS is inconsistent, as x_2 is endogenous. A seemingly good instrument for x_2 is x_1^2, since x_1^2 is likely to be uncorrelated with u because x_1 is uncorrelated with u. However, the validity of this instrument requires the functional form restriction on the conditional mean that x_1 only enters the model linearly and not quadratically. In practice one should view a linear model as only an approximation, and obtaining instruments in such an artificial way can be easily criticized.

A better way to create a valid instrument is through alternative **exclusion restrictions** that do not rely so heavily on choice of functional form. Some practical examples have been given in Section 4.8.2.

Structural models such as the classical linear simultaneous equations model (see Sections 2.4 and 6.10.6) make such exclusion restrictions very explicit. Even then the restrictions can often be criticized for being too ad hoc, unless compelling economic theory supports the restrictions.

For panel data applications it may be reasonable to assume that only current data may belong in the equation of interest – an exclusion restriction permitting past data to be used as instruments under the assumption that errors are serially uncorrelated (see Section 22.2.4). Similarly, in models of decision making under uncertainty (see Section 6.2.7), lagged variables can be used as instruments as they are part of the information set.

There is no formal test of instrument exogeneity that does not additionally test whether the regression equation is correctly specified. Instrument exogeneity inevitably relies on a priori information, such as that from economic or statistical theory. The evaluation by Bound et al. (1995, pp. 446–447) of the validity of the instruments used by Angrist and Krueger (1991) provides an insightful example of the subtleties involved in determining instrument exogeneity.

It is especially important that an instrument be exogenous if an instrument is weak, because with weak instruments even very mild endogeneity of the instrument can lead to IV parameter estimates that are much more inconsistent than the already inconsistent OLS parameter estimates.

For simplicity consider linear regression with one regressor and one instrument; hence $y = \beta x + u$. Then performing some algebra, left as an exercise, yields

$$\frac{\text{plim}\,\widehat{\beta}_{\text{IV}} - \beta}{\text{plim}\,\widehat{\beta}_{\text{OLS}} - \beta} = \frac{\text{Cor}[z,u]}{\text{Cor}[x,u]} \times \frac{1}{\text{Cor}[z,x]}. \tag{4.59}$$

Thus with an invalid instrument and low correlation between the instrument and the regressor, the IV estimator can be even more inconsistent than OLS. For example, suppose the correlation between z and x is 0.1, which is not unusual for cross-section

data. Then IV becomes more inconsistent than OLS as soon as the correlation coefficient between z and u exceeds a mere 0.1 times the correlation coefficient between x and u.

Result (4.59) can be extended to the model (4.56) with one endogenous regressor and several exogenous regressors, iid errors, and instruments that include all the exogenous regressors. Then

$$\frac{\text{plim}\,\widehat{\beta}_{1,\text{2SLS}} - \beta_1}{\text{plim}\,\widehat{\beta}_{1,\text{OLS}} - \beta_1} = \frac{\text{Cor}[\widehat{x}, u]}{\text{Cor}[x, u]} \times \frac{1}{R_p^2}, \tag{4.60}$$

where R_p^2 is defined after (4.56). For extension to more than one endogenous regressor see Shea (1997).

These results, emphasized by Bound et al. (1995), have profound implications for the use of IV. If instruments are weak then even mild instrument endogeneity can lead to IV being even more inconsistent than OLS. Perhaps because the conclusion is so negative, the literature has neglected this aspect of weak instruments. A notable recent exception is Hahn and Hausman (2003a).

Most of the literature assumes that condition 1 is satisfied, so that IV is consistent, and focuses on other complications attributable to weak instruments.

4.9.3. Low Precision

Although IV estimation can lead to consistent estimation when OLS is inconsistent, it also leads to a loss in precision. Intuitively, from Section 4.8.2 the instrument z is a treatment that leads to exogenous movement in x but does so with considerable noise.

The loss in precision increases, and standard errors increase, with weaker instruments. This is easily seen in the simplest case of a single endogenous regressor and single instrument with iid errors. Then the asymptotic variance is

$$\begin{aligned} \text{V}[\widehat{\beta}_{\text{IV}}] &= \sigma^2(\mathbf{x}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{z}(\mathbf{z}'\mathbf{x})^{-1} \\ &= [\sigma^2/\mathbf{x}'\mathbf{x}]/[(\mathbf{z}'\mathbf{x})^2/(\mathbf{z}'\mathbf{z})(\mathbf{x}'\mathbf{x})] \\ &= \text{V}[\widehat{\beta}_{\text{OLS}}]/r_{xz}^2. \end{aligned} \tag{4.61}$$

For example, if the squared sample correlation coefficient between z and x equals 0.1, then IV standard errors are 10 times those of OLS. Moreover, the IV estimator has larger variance than the OLS estimator unless $\text{Cor}[z, x] = 1$.

Result (4.61) can be extended to the model (4.56) with one endogenous regressor and several exogenous regressors, iid errors, and instruments that include all the exogenous regressors. Then

$$\text{se}[\widehat{\beta}_{1,\text{2SLS}}] = \text{se}[\widehat{\beta}_{1,\text{OLS}}]/R_p, \tag{4.62}$$

where $\text{se}[\cdot]$ denotes asymptotic standard error and R_p^2 is defined after (4.56). For extension to more than one endogenous regressor this R_p^2 is replaced by the R_p^{*2} proposed by Shea (1997). This provided the motivation for Shea's test statistic.

The poor precision is concentrated on the coefficients for endogenous variables. For exogenous variables the standard errors for 2SLS coefficient estimates are similar to

those for OLS. Intuitively, exogenous variables are being instrumented by themselves, so they have a very strong instrument.

For the coefficients of an endogenous regressor it is a low *partial* R^2, rather than R^2, that leads to a loss of estimator precision. This explains why 2SLS standard errors can be much higher than OLS standard errors despite the high raw correlation between the endogenous variable and the instruments. Going the other way, 2SLS standard errors for coefficients of endogenous variables that are much larger than OLS standard errors provide a clear signal that instruments are weak.

Statistics used to detect low precision of IV caused by weak instruments are called measures of **instrument relevance**. To some extent they are unnecessary as the problem is easily detected if IV standard errors are much larger than OLS standard errors.

4.9.4. Finite-Sample Bias

This section summarizes a relatively challenging and as yet unfinished literature on "weak instruments" that focuses on the practical problem that even in "large" samples asymptotic theory can provide a poor approximation to the distribution of the IV estimator. In particular the IV estimator is biased in finite samples even if asymptotically consistent. The bias can be especially pronounced when instruments are weak.

This bias of IV, which is toward the inconsistent OLS estimator, can be remarkably large, as demonstrated in a simple Monte Carlo experiment by Nelson and Startz (1990), and by a real data application involving several hundred thousand observations but very weak instruments by Bound et al. (1995). Moreover, the standard errors can also be very biased, as also demonstrated by Nelson and Startz (1990).

The theoretical literature entails quite specialized and advanced econometric theory, as it is actually difficult to obtain the sample mean of the IV estimator. To see this, consider adapting to the IV estimator the usual proof of unbiasedness of the OLS estimator given in Section 4.4.8. For $\widehat{\boldsymbol{\beta}}_{\text{IV}}$ defined in (4.51) for the just-identified case this yields

$$\begin{aligned} \mathrm{E}[\widehat{\boldsymbol{\beta}}_{\text{IV}}] &= \boldsymbol{\beta} + \mathrm{E}_{\mathbf{Z},\mathbf{X},\mathbf{u}}[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u}] \\ &= \boldsymbol{\beta} + \mathrm{E}_{\mathbf{Z},\mathbf{X}}\left[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}' \times [\mathrm{E}[\mathbf{u}|\mathbf{Z},\mathbf{X}]\right], \end{aligned}$$

where the unconditional expectation with respect to all stochastic variables, $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{u}$, is obtained by first taking expectation with respect to $\mathbf{u}$ conditional on $\mathbf{Z}$ and $\mathbf{X}$, using the law of Iterated Expectations (see Section A.8.). An obvious sufficient condition for the IV estimator to have mean $\boldsymbol{\beta}$ is that $\mathrm{E}[\mathbf{u}|\mathbf{Z},\mathbf{X}] = \mathbf{0}$. This assumption is too strong, however, because it implies $\mathrm{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$, in which case there would be no need to instrument in the first place. So there is no simple way to obtain $\mathrm{E}[\widehat{\boldsymbol{\beta}}_{\text{IV}}]$. A similar problem does not arise in establishing consistency. Then $\widehat{\boldsymbol{\beta}}_{\text{IV}} = \boldsymbol{\beta} + \left(N^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1} N^{-1}\mathbf{Z}'\mathbf{u}$, where the term $N^{-1}\mathbf{Z}'\mathbf{u}$ can be considered in isolation of $\mathbf{X}$ and the assumption $\mathrm{E}[\mathbf{u}|\mathbf{Z}] = \mathbf{0}$ leads to plim $N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$.

Therefore we need to use alternative methods to obtain the mean of the IV estimator. Here we merely summarize key results.

Initial research made the strong assumption of joint normality of variables and homoskedastic errors. Then the IV estimator has a Wishart distribution (defined in Chapter 13). Surprisingly, the mean of the IV estimator does not even exist in the just-identified case, a signal that there may be finite-sample problems. The mean does exist if there is at least one overidentifying restriction, and the variance exists if there are at least two overidentifying restrictions. Even when the mean exists the IV estimator is biased, with bias in the direction of OLS. With more overidentifying restrictions the bias increases, eventually equaling the bias of the OLS estimator. A detailed discussion is given in Davidson and MacKinnon (1993, pp. 221–224). Approximations based on power-series expansions have also been used.

What determines the size of the finite-sample bias? For regression with a single regressor x that is endogenous and is related to the instruments $\mathbf{z}$ by the reduced form model $x = \mathbf{z}\boldsymbol{\pi} + v$, the **concentration parameter** τ^2 is defined as $\tau^2 = \boldsymbol{\pi}'\mathbf{Z}\mathbf{Z}'\boldsymbol{\pi}/\sigma_v^2$. The bias of IV can be shown to be an increasing function of τ^2. The quantity τ^2/K, where K is the number of instruments, is the population analogue of the F-statistic for a test of whether $\boldsymbol{\pi} = \mathbf{0}$. The statistic $F - 1$, where F is the actual F-statistic in the first-stage reduced form model, can be shown to be an approximately unbiased estimate of τ^2/K. This leads to tests for finite-sample bias being based on the F-statistic given in Section 4.9.2.

Staiger and Stock (1997) obtained results under weaker distributional assumptions. In particular, normality is no longer needed. Their approach uses weak instrument asymptotics that find the limit distribution of IV estimators for a sequence of models with τ^2/K held constant as $N \to \infty$. In a simple model $1/F$ provides an approximate estimate of the finite-sample bias of the IV estimator relative to OLS. More generally, the extent of the bias for given F varies with the number of endogenous regressors and the number of instruments. Simulations show that to ensure that the maximal bias in IV is no more than 10% that of OLS we need $F > 10$. This threshold is widely cited but falls to around 6.5, for example, if one is comfortable with bias in IV of 20% of that for OLS. So a less strict rule of thumb is $F > 5$. Shea (1997) demonstrated that low partial R^2 is also associated with finite-sample bias but there is no similar rule of thumb for use of partial R^2 as a diagnostic for finite-sample bias.

For models with more than one endogenous regressor, separate F-statistics can be computed for each endogenous regressor. For a joint statistic Stock, Wright and Yogo (2002) propose using the minimum eigenvalue of a matrix analogue of the first-stage test F-statistic. Stock and Yogo (2003) present relevant critical values for this eigenvalue as the desired degree of bias, the number of endogenous variables, and the number of overidentifying restrictions vary. These tables include the single endogenous regressor as a special case and presume at least two overidentifying restrictions, so they do not apply to just-identified models.

Finite-sample bias problems arise not only for the IV estimate but also for IV standard errors and test statistics. Stock et al. (2002) present a similar approach to Wald tests whereby a test of $\beta = \beta_0$ at a nominal level of 5% is to have actual size of, say, no more than 15%. Stock and Yogo (2003) also present detailed tables taking this size distortion approach that include just-identified models.

4.9.5. Responses to Weak Instruments

What can the practitioner do in the face of weak instruments?

As already noted one approach is to limit the number of instruments used. This can be done by dropping instruments or by combining instruments.

If finite-sample bias is a concern then alternative estimators may have better small-sample properties than 2SLS. A number of alternatives, many variants of IV, are presented in Section 6.4.4.

Despite the emphasis on finite-sample bias the other problems created by weak instruments may be of greater importance in applications. It is possible with a large enough sample for the first-stage reduced form F-statistic to be large enough that finite-sample bias is not a problem. Meanwhile, the partial R^2 may be very small, leading to fragility to even slight correlation between the model error and instrument. This is difficult to test for and to overcome.

There also can be great loss in estimator precision, as detailed in Sections 4.9.3 and 4.9.4. In such cases either larger samples are needed or alternative approaches to estimating causal marginal effects must be used. These methods are summarized in Section 2.8 and presented elsewhere in this book.

4.9.6. IV Application

Kling (2001) analyzed in detail the use of college proximity as an instrument for schooling. Here we use the same data from the NLS young men's cohort on 3,010 males aged 24 to 34 years old in 1976 as used to produce Table 1 of Kling (2001) and originally used by Card (1995). The model estimated is

$$\ln w_i = \alpha + \beta_1 s_i + \beta_2 e_i + \beta_3 e_i^2 + \mathbf{x}_{2i}'\gamma + u_i,$$

where s denotes years of schooling, e denotes years of work experience, e^2 denotes experience squared, and $\mathbf{x}_2$ is a vector of 26 control variables that are mainly geographic indicators and measure of parental education.

The schooling variable is considered endogenous, owing to lack of data on ability. Additionally, the two work experience variables are endogenous, since work experience is calculated as age minus years of schooling minus six, as is common in this literature, and schooling is endogenous. At least three instruments are needed.

Here exactly three instruments are used, so the model is just-identified. The first instrument is *col4*, an indicator for whether a four-year college is nearby. This instrument has already been discussed in Section 4.8.2. The other two instruments are age and age squared. These are highly correlated with experience and experience squared, yet it is believed they can be omitted from the model for log-wage since it is work experience that matters. The remaining regressor vector $\mathbf{x}_2$ is used as an instrument for itself.

Although age is clearly exogenous, some unobservables such as social skills may be correlated with both age and wage. Then the use of age and age squared as instruments can be questioned. This illustrates the general point that there can be disagreement on assumptions of instrument validity.

Table 4.5. *Returns to Schooling: Instrumental Variables Estimates*[a]

	OLS	IV
Schooling (s)	0.073	0.132
	(0.004)	(0.049)
R^2	0.304	0.207
Shea's partial R^2	–	0.006
First-stage F-statistic for s	–	8.07

[a] Sample of 3,010 young males. Dependent variable is log hourly wage. Coefficient and standard error for schooling given; estimates for experience, experience squared, 26 control variables, and an intercept are not reported. For the three endogenous regressors – schooling (s), experience (e), and experience squared (e^2) – the three instruments are an indicator for whether a four-year college (col) is nearby, age, and age squared. The partial R^2 and first-stage F-statistic are weak instruments diagnostics explained in the test.

Results are given in Table 4.5. The OLS estimate of β_1 is 0.073, so that wages rise by 7.6% ($= 100 \times (e^{.073} - 1)$) on average with each extra year of schooling. This estimate is an inconsistent estimate of β_1 given omitted ability. The IV estimate, or equivalently the 2SLS estimate since the model is just-identified, is 0.132. An extra year of schooling is estimated to lead to a 14.1% ($= 100 \times (e^{.132} - 1)$) increase in wage.

The IV estimator is much less efficient than OLS. A formal test does not reject homoskedasticity and we follow Kling (2001) and use the usual standard errors, which are very close to the heteroskedastic-robust standard errors. The standard error of $\widehat{\beta}_{1,\text{OLS}}$ is 0.004 whereas that for $\widehat{\beta}_{1,\text{IV}}$ is 0.049, over 10 times larger. The standard errors for the other two endogenous regressors are about 4 times larger and the standard errors for the exogenous regressors are about 1.2 times larger. The R^2 falls from 0.304 to 0.207.

R^2 measures confirm that the instruments are not very relevant for schooling. A simple test is to note that the regression (4.58) of schooling on all of the instruments yields $R^2 = 0.297$, which only falls a little to $R^2 = 0.291$ if the three additional instruments are dropped. More formally, Shea's partial R^2 here equals $0.0064 = 0.08^2$, which from (4.62) predicts that the standard error of $\widehat{\beta}_{1,\text{IV}}$ will be inflated by a multiple $12.5 = 1/0.08$, very close to the inflation observed here. This reduces the t-statistic on schooling from 19.64 to 2.68. In many applications such a reduction would lead to statistical insignificance. In addition, from Section 4.9.2 even slight correlation between the instrument $col4_i$ and the error term u_i will lead to inconsistency of IV.

To see whether finite-sample bias may also be a problem we run the regression (4.58) of schooling on all of the instruments. Testing the joint significance of the three additional instruments yields an F-statistic of 8.07, suggesting that the bias of IV may be 10 or 20% that of OLS. A similar regression for the other two endogenous variables yields much higher F-statistics since, for example, age is a good additional instrument

for experience. Given that there are three endogenous regressors it is actually better to use the method of Stock et al. (2002) discussed in Section 4.9.4, though here the problem is restricted to schooling since for experience and experience squared, respectively, Shea's partial R^2 equals 0.0876 and 0.0138, whereas the first-stage F-statistics are 1,772 and 1,542.

If additional instruments are available then the model becomes overidentified and standard procedure is to additionally perform a test of overidentifying restrictions (see Section 8.4.4).

4.10. Practical Considerations

The estimation procedures in this chapter are implemented in all standard econometrics packages for cross-section data, except that not all packages implement quantile regression. Most provide robust standard errors as an option rather than the default.

The most difficult estimator to apply can be the instrumental variables estimator, as in many potential applications it can be difficult to obtain instruments that are uncorrelated with the error yet reasonably correlated with the regressor or regressors being instrumented. Such instruments can be obtained through specification of a complete structural model, such as a simultaneous equations system. Current applied research emphasizes alternative approaches such as natural experiments.

4.11. Bibliographic Notes

The results in this chapter are presented in many first-year graduate texts, such as those by Davidson and MacKinnon (2004), Greene (2003), Hayashi (2000), Johnston and diNardo (1997), Mittelhammer, Judge, and Miller (2000), and Ruud (2000). We have emphasized regression with stochastic regressors, robust standard errors, quantile regression, endogeneity, and instrumental variables.

4.2 Manski (1991) has a nice discussion of regression in a general setting that includes discussion of the loss functions given in Section 4.2.

4.3 The returns to schooling example is well studied. Angrist and Krueger (1999) and Card (1999) provide recent surveys.

4.4 For a history of least squares see Stigler (1986). The method was introduced by Legendre in 1805. Gauss in 1810 applied least squares to the linear model with normally distributed error and proposed the elimination method for computation, and in later work he proposed the theorem now called the Gauss–Markov theorem. Galton introduced the concept of regression, meaning mean-reversion in the context of inheritance of family traits, in 1887. For an early "modern" treatment with application to pauperism and welfare availability see Yule (1897). Statistical inference based on least-squares estimates of the linear regression model was developed most notably by Fisher. The heteroskedastic-consistent estimate of the variance matrix of the OLS estimator, due to White (1980a) building on earlier work by Eicker (1963), has had a profound impact on statistical inference in microeconometrics and has been extended to many settings.

4.6 Boscovich in 1757 proposed a least absolute deviations estimator that predates least squares; see Stigler (1986). A review of quantile regression, introduced by Koenker and

recalling that $\mathbf{G}_N(\boldsymbol{\theta}) = \partial \mathbf{g}_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$, where $\boldsymbol{\theta}^+$ is a point between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}$. Substituting (6.30) back into (6.29) yields

$$\mathbf{G}_N(\widehat{\boldsymbol{\theta}})'\mathbf{W}_N\left[\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) + \mathbf{G}_N(\boldsymbol{\theta}^+)\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right] = \mathbf{0}.$$

Solving for $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ yields

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\left[\mathbf{G}_N(\widehat{\boldsymbol{\theta}})'\mathbf{W}_N\mathbf{G}_N(\boldsymbol{\theta}^+)\right]^{-1}\mathbf{G}_N(\widehat{\boldsymbol{\theta}})'\mathbf{W}_N\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0). \tag{6.31}$$

Equation (6.31) is the key result for obtaining the limit distribution of the GMM estimator. We obtain the probability limits of each of the first five terms using $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, given consistency, in which case $\boldsymbol{\theta}^+ \xrightarrow{p} \boldsymbol{\theta}_0$. The last term on the right-hand side of (6.31) has a limit normal distribution by assumption (v). Thus

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} -(\mathbf{G}_0'\mathbf{W}_0\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{W}_0 \times \mathcal{N}[0, \mathbf{S}_0],$$

where $\mathbf{G}_0$, $\mathbf{W}_0$, and $\mathbf{S}_0$ have been defined in Proposition 6.1. Applying the limit normal product rule (Theorem A.17) yields (6.11).

This derivation treats the GMM first-order conditions as being q linear combinations of the r sample moments $\mathbf{g}_N(\widehat{\boldsymbol{\theta}})$, since $\mathbf{G}_N(\widehat{\boldsymbol{\theta}})'\mathbf{W}_N$ is a $q \times r$ matrix. The MM estimator is the special case $q = r$, since then $\mathbf{G}_N(\widehat{\boldsymbol{\theta}})'\mathbf{W}_N$ is a full-rank square matrix, so $\mathbf{G}_N(\widehat{\boldsymbol{\theta}})'\mathbf{W}_N\mathbf{g}_N(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ implies that $\mathbf{g}_N(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$.

To derive the distribution of the OIR test statistic in (6.26), begin with a first-order Taylor series expansion of $\sqrt{N}\mathbf{g}_N(\widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ to obtain

$$\begin{aligned}\sqrt{N}\mathbf{g}_N(\widehat{\boldsymbol{\theta}}_{\text{OGMM}}) &= \sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) + \mathbf{G}_N(\boldsymbol{\theta}^+)\sqrt{N}(\widehat{\boldsymbol{\theta}}_{\text{OGMM}} - \boldsymbol{\theta}_0)\\ &= \sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) - \mathbf{G}_0(\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{S}_0^{-1}\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) + o_p(1)\\ &= [\mathbf{I} - \mathbf{M}_0\mathbf{S}_0^{-1}]\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) + o_p(1),\end{aligned}$$

where the second equality uses (6.31) with $\mathbf{W}_N$ consistent for $\mathbf{S}_0^{-1}$, $\mathbf{M}_0 = \mathbf{G}_0(\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0'$, and $o_p(1)$ is defined in Definition A.22. It follows that

$$\begin{aligned}\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\widehat{\boldsymbol{\theta}}_{\text{OGMM}}) &= \mathbf{S}_0^{-1/2}[\mathbf{I} - \mathbf{M}_0\mathbf{S}_0^{-1}]\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) + o_p(1)\\ &= [\mathbf{I} - \mathbf{S}_0^{-1/2}\mathbf{M}_0\mathbf{S}_0^{-1/2}]\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) + o_p(1).\end{aligned} \tag{6.32}$$

Now $[\mathbf{I} - \mathbf{S}_0^{-1/2}\mathbf{M}_0\mathbf{S}_0^{-1/2}] = [\mathbf{I} - \mathbf{S}_0^{-1/2}\mathbf{G}_0(\mathbf{G}_0'\mathbf{S}_0^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{S}_0^{-1/2}]$ is an idempotent matrix of rank $(r - q)$, and $\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[0, \mathbf{I}]$ given $\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[0, \mathbf{S}_0]$. From standard results for quadratic forms of normal variables it follows that the inner product

$$\tau_N = (\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\widehat{\boldsymbol{\theta}}_{\text{OGMM}}))'(\mathbf{S}_0^{-1/2}\sqrt{N}\mathbf{g}_N(\widehat{\boldsymbol{\theta}}_{\text{OGMM}}))$$

converges to the $\chi^2(r - q)$ distribution.

6.4. Linear Instrumental Variables

Correlation of regressors with the error term leads to inconsistency of least-squares methods. Examples of such failure include omitted variables, simultaneity,

measurement error in the regressors, and sample selection bias. Instrumental variables methods provide a general approach that can handle any of these problems, provided suitable instruments exist.

Instrumental variables methods fall naturally into the GMM framework as a surplus of instruments leads to an excess of moment conditions that can be used for estimation. Many IV results are most easily obtained using the GMM framework.

Linear IV is important enough to appear in many places in this book. An introduction was given in Sections 4.8 and 4.9. This section presents single-equation linear IV as a particular application of GMM. For completeness the section also presents the earlier literature on a special case, the two-stage least-squares estimator. Systems linear IV estimation is summarized in Section 6.9.5. Tests of endogeneity and tests of overidentifying restrictions for linear models are detailed in Section 8.4. Chapter 22 presents linear IV estimation with panel data.

6.4.1. Linear GMM with Instruments

Consider the linear regression model

$$y_i = \mathbf{x}_i'\beta + u_i, \tag{6.33}$$

where each component of $\mathbf{x}$ is viewed as being an **exogenous regressor** if it is uncorrelated with the error in model (6.33) or an **endogenous regressor** if it is correlated. If all regressors are exogenous then LS estimators can be used, but if any components of $\mathbf{x}$ are endogenous then LS estimators are inconsistent for β.

From Section 4.8, consistent estimates can be obtained by IV estimation. The key assumption is the existence of an $r \times 1$ vector of **instruments z** that satisfies

$$\mathrm{E}[u_i|\mathbf{z}_i] = \mathbf{0}. \tag{6.34}$$

Exogenous regressors can be instrumented by themselves. As there must be at least as many instruments as regressors, the challenge is to find additional instruments that at least equal the number of endogenous variables in the model. Some examples of such instruments have been given in Section 4.8.2.

Linear GMM Estimator

From Section 6.2.5, the conditional moment restriction (6.34) and model (6.33) imply the unconditional moment restriction

$$\mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\beta)] = \mathbf{0}, \tag{6.35}$$

where for notational simplicity the following analysis uses β rather than the more formal β_0 to denote the true parameter value. A quadratic form in the corresponding sample moments leads to the GMM objective function $Q_N(\beta)$ given in (6.4).

In matrix notation define $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ as usual and let $\mathbf{Z}$ denote the $N \times r$ matrix of instruments with ith row $\mathbf{z}_i'$. Then $\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\beta) = \mathbf{Z}'\mathbf{u}$ and (6.4) becomes

$$Q_N(\beta) = \left[\frac{1}{N}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{Z}\right]\mathbf{W}_N\left[\frac{1}{N}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)\right], \tag{6.36}$$

where $\mathbf{W}_N$ is an $r \times r$ full-rank symmetric weighting matrix with leading examples given at the end of this section. The first-order conditions

$$\frac{\partial Q_N(\beta)}{\partial \beta} = -2\left[\frac{1}{N}\mathbf{X}'\mathbf{Z}\right]\mathbf{W}_N\left[\frac{1}{N}\mathbf{Z}'(\mathbf{y}-\mathbf{X}\beta)\right] = \mathbf{0}$$

can actually be solved for β in this special case of GMM, leading to the **GMM estimator in the linear IV model**

$$\widehat{\beta}_{\text{GMM}} = \left[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y}, \tag{6.37}$$

where the divisions by N have canceled out.

Distribution of Linear GMM Estimator

The general results of Section 6.3 can be used to derive the asymptotic distribution. Alternatively, since an explicit solution for $\widehat{\beta}_{\text{GMM}}$ exists the analysis for OLS given in Section 4.4. can be adapted. Substituting $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ into (6.37) yields

$$\widehat{\beta}_{\text{GMM}} = \beta + \left[\left(N^{-1}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_N\left(N^{-1}\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left(N^{-1}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_N\left(N^{-1}\mathbf{Z}'\mathbf{u}\right). \tag{6.38}$$

From the last term, consistency of the GMM estimator essentially requires that plim $N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$. Under pure random sampling this requires that (6.35) holds, whereas under other common sampling schemes (see Section 24.3) the stronger assumption (6.34) is needed.

Additionally, the **rank condition** for **identification** of β that plim $N^{-1}\mathbf{Z}'\mathbf{X}$ is of rank K ensures that the inverse in the right-hand side exists, provided $\mathbf{W}_N$ is of full rank. A weaker **order condition** is that $r \geq K$.

The limit distribution is based on the expression for $\sqrt{N}(\widehat{\beta}_{\text{GMM}} - \beta)$ obtained by simple manipulation of (6.38). This yields an asymptotic normal distribution for $\widehat{\beta}_{\text{GMM}}$ with mean β and estimated asymptotic variance

$$\widehat{\text{V}}[\widehat{\beta}_{\text{GMM}}] = N\left[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}\right]^{-1}\left[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\widehat{\mathbf{S}}\mathbf{W}_N\mathbf{Z}'\mathbf{X}\right]\left[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}\right]^{-1}, \tag{6.39}$$

where $\widehat{\mathbf{S}}$ is a consistent estimate of

$$\mathbf{S} = \lim \frac{1}{N}\sum_{i=1}^{N} \text{E}\left[u_i^2\mathbf{z}_i\mathbf{z}_i'\right],$$

given the usual cross-section assumption of independence over i. The essential additional assumption needed for (6.39) is that $N^{-1/2}\mathbf{Z}'\mathbf{u} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{S}]$. Result (6.39) also follows from Proposition 6.1 with $\mathbf{h}(\cdot) = \mathbf{z}(\mathbf{y} - \mathbf{x}'\beta)$ and hence $\partial\mathbf{h}/\partial\beta' = -\mathbf{z}\mathbf{x}'$.

For cross-section data with heteroskedastic errors, $\mathbf{S}$ is consistently estimated by

$$\widehat{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{z}_i\mathbf{z}_i' = \mathbf{Z}'\mathbf{D}\mathbf{Z}/\mathbf{N}, \tag{6.40}$$

where $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\beta}_{\text{GMM}}$ is the GMM residual and $\mathbf{D}$ is an $N \times N$ diagonal matrix with entries $\widehat{u}_i^2$. A commonly used small-sample adjustment is to divide by $N - K$

Table 6.2. *GMM Estimators in Linear IV Model and Their Asymptotic Variance*[a]

Estimator	Definition and Asymptotic Variance
GMM	$\widehat{\boldsymbol{\beta}}_{\text{GMM}} = [\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y}$
(general $\mathbf{W}_N$)	$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = N[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\widehat{\mathbf{S}}\mathbf{W}_N\mathbf{Z}'\mathbf{X}][\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1}$
Optimal GMM	$\widehat{\boldsymbol{\beta}}_{\text{OGMM}} = [\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{y}$
($\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$)	$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = N[\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$
2SLS	$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$
($\mathbf{W}_N = [N^{-1}\mathbf{Z}'\mathbf{Z}]^{-1}$)	$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = N[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\widehat{\mathbf{S}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]$ $\times [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$
	$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = s^2[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$ if homoskedastic errors
IV	$\widehat{\boldsymbol{\beta}}_{\text{IV}} = [\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'\mathbf{y}$
(just-identified)	$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = N(\mathbf{Z}'\mathbf{X})^{-1}\widehat{\mathbf{S}}(\mathbf{X}'\mathbf{Z})^{-1}$

[a] Equations are based on a linear regression model with dependent variable $\mathbf{y}$, regressors $\mathbf{X}$, and instruments $\mathbf{Z}$. $\widehat{\mathbf{S}}$ is defined in (6.40) and s^2 is defined after (6.41). All variance matrix estimates assume errors that are independent across observations and heteroskedastic, aside from the simplification for homoskedastic errors given for the 2SLS estimator. Optimal GMM uses the optimal weighting matrix.

rather than N in the formula for $\widehat{\mathbf{S}}$. In the more restrictive case of homoskedastic errors, $\mathrm{E}[u_i^2|\mathbf{z}_i] = \sigma^2$ and so $\mathbf{S} = \lim N^{-1}\sum_i \sigma^2\mathrm{E}[\mathbf{z}_i\mathbf{z}_i']$, leading to estimate

$$\widehat{\mathbf{S}} = s^2\mathbf{Z}'\mathbf{Z}/N, \tag{6.41}$$

where $s^2 = (N-K)^{-1}\sum_{i=1}^N \widehat{u}_i^2$ is consistent for σ^2. These results mimic similar results for OLS presented in Section 4.4.5.

6.4.2. Different Linear GMM Estimators

Implementation of the results of Section 6.4.1 requires specification of the weighting matrix $\mathbf{W}_N$. For just-identified models all choices of $\mathbf{W}_N$ lead to the same estimator. For overidentified models there are two common choices of $\mathbf{W}_N$, given in the following.

Table 6.2 summarizes these estimators and gives the appropriate specialization of the estimated variance matrix formula given in (6.39), assuming independent heteroskedastic errors.

Instrumental Variables Estimator

In the just-identified case $r = K$ and $\mathbf{X}'\mathbf{Z}$ is a square matrix that is invertible. Then $[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}]^{-1} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{W}_N^{-1}(\mathbf{X}'\mathbf{Z})^{-1}$ and (6.37) simplifies to the **instrumental variables** estimator

$$\widehat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}, \tag{6.42}$$

introduced in Section 4.8.6. For just-identified models the GMM estimator for any choice of $\mathbf{W}_N$ equals the IV estimator.

The simple IV estimator can also be used in overidentified models, by discarding some of the instruments so that the model is just-identified, but this results in an efficiency loss compared to using all the instruments.

Optimal-Weighted GMM

From Section 6.3.5, for overidentified models the most efficient GMM estimator, meaning GMM with optimal choice of weighting matrix, sets $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$ in (6.37).

The **optimal GMM estimator or two-step GMM estimator** in the linear IV model is

$$\widehat{\boldsymbol{\beta}}_{\text{OGMM}} = \left[(\mathbf{X}'\mathbf{Z})\widehat{\mathbf{S}}^{-1}(\mathbf{Z}'\mathbf{X})\right]^{-1}(\mathbf{X}'\mathbf{Z})\widehat{\mathbf{S}}^{-1}(\mathbf{Z}'\mathbf{y}). \tag{6.43}$$

For heteroskedastic errors, $\widehat{\mathbf{S}}$ is computed using (6.40) based on a consistent first-step estimate $\widehat{\boldsymbol{\beta}}$ such as the 2SLS estimator defined in (6.44). White (1982) called this estimator a **two-stage IV estimator**, since both steps entail IV estimation.

The estimated asymptotic variance matrix for optimal GMM given in Table 6.2 is of relatively simple form as (6.39) simplifies when $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$. In computing the estimated variance one can use $\widehat{\mathbf{S}}$ as presented in Table 6.2, but it is more common to instead use an estimator $\widetilde{\mathbf{S}}$, say, that is also computed using (6.40) but evaluates the residual at the optimal GMM estimator rather than the first-step estimate used to form $\widehat{\mathbf{S}}$ in (6.43).

Two-Stage Least Squares

If errors are homoskedastic rather than heteroskedastic, $\widehat{\mathbf{S}}^{-1} = [s^2N^{-1}\mathbf{Z}'\mathbf{Z}]^{-1}$ from (6.41). Then $\mathbf{W}_N = (N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$ in (6.37), leading to the **two-stage least-squares estimator**, introduced in Section 4.8.7, that can be expressed compactly as

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = \left[\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X}\right]^{-1}\left[\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y}\right], \tag{6.44}$$

where $\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}'$. The basis of the term two-stage least-squares is presented in the next section. The 2SLS estimator is also called the **generalized instrumental variables (GIV) estimator** as it generalizes the IV estimator to the overidentified case of more instruments than regressors. It is also called the **one-step GMM** because (6.44) can be calculated in one step, whereas optimal GMM requires two steps.

The 2SLS estimator is asymptotically normal distributed with estimated asymptotic variance given in Table 6.2. The general form should be used if one wishes to guard against heteroskedastic errors whereas the simpler form, presented in many introductory textbooks, is consistent only if errors are indeed homoskedastic.

Optimal GMM versus 2SLS

Both the optimal GMM and the 2SLS estimator lead to efficiency gains in overidentified models. Optimal GMM has the advantage of being more efficient than 2SLS, if errors are heteroskedastic, though the efficiency gain need not be great. Some of the GMM testing procedures given in Section 7.5 and Chapter 8 assume estimation

using the optimal weighting matrix. Optimal GMM has the disadvantage of requiring additional computation compared to 2SLS. Moreover, as discussed in Section 6.3.5, asymptotic theory may provide a poor small-sample approximation to the distribution of the optimal GMM estimator.

In cross-section applications it is common to use the less efficient 2SLS, though with inference based on heteroskedastic robust standard errors.

Even More Efficient GMM Estimation

The estimator $\widehat{\boldsymbol{\beta}}_{\text{OGMM}}$ is the most efficient estimator based on the unconditional moment condition $\mathrm{E}[\mathbf{z}_i u_i] = \mathbf{0}$, where $u_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$. However, this is not the best moment condition to use if the starting point is the conditional moment condition $\mathrm{E}[u_i|\mathbf{z}_i] = \mathbf{0}$ and errors are heteroskedastic, meaning $\mathrm{V}[u_i|\mathbf{z}_i]$ varies with $\mathbf{z}_i$.

Applying the general results of Section 6.3.7, we can write the optimal moment condition for GMM estimation based on $\mathrm{E}[u_i|\mathbf{z}_i] = \mathbf{0}$ as

$$\mathrm{E}\left[\mathrm{E}\left[\mathbf{x}_i|\mathbf{z}_i\right] u_i / \mathrm{V}\left[u_i|\mathbf{z}_i\right]\right] = \mathbf{0}. \tag{6.45}$$

As with the LS regression example in Section 6.3.7, one should divide by the error variance $\mathrm{V}[u|\mathbf{z}]$. Implementation is more difficult than in the LS case, however, as a model for $\mathrm{E}[\mathbf{x}|\mathbf{z}]$ needs to be specified in addition to one for $\mathrm{V}[u|\mathbf{z}]$. This may be possible with additional structure. In particular, for a linear simultaneous equations system $\mathrm{E}[\mathbf{x}_i|\mathbf{z}_i]$ is linear in $\mathbf{z}$ so that estimation is based on $\mathrm{E}[\mathbf{x}_i u_i / \mathrm{V}[u_i|\mathbf{z}_i]] = 0$.

For linear models the GMM estimator is usually based on the simpler condition $\mathrm{E}[\mathbf{z}_i u_i] = \mathbf{0}$. Given this condition, the optimal GMM estimator defined in (6.43) is the most efficient GMM estimator.

6.4.3. Alternative Derivations of Two-Stage Least Squares

The 2SLS estimator, the standard IV estimator for overidentified models, was derived in Section 6.4.2 as a GMM estimator.

Here we present three other derivations of the 2SLS estimator. One of these derivations, due to Theil, provided the original motivation for 2SLS, which predates GMM. Theil's interpretation is emphasized in introductory treatments. However, it does not generalize to nonlinear models, whereas the GMM interpretation does.

We consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{6.46}$$

with $\mathrm{E}[\mathbf{u}|\mathbf{Z}] = \mathbf{0}$ and additionally $\mathrm{V}[\mathbf{u}|\mathbf{Z}] = \sigma^2\mathbf{I}$.

GLS in a Transformed Model

Premultiplication of (6.46) by the instruments $\mathbf{Z}'$ yields the transformed model

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}'\mathbf{u}. \tag{6.47}$$

This transformed model is often used as motivation for the IV estimator when $r = K$, since ignoring $\mathbf{Z}'\mathbf{u}$ since $N^{-1}\mathbf{Z}'\mathbf{u} \rightarrow \mathbf{0}$ and solving yields $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$.

Here instead we consider the overidentified case. Conditional on $\mathbf{Z}$ the error $\mathbf{Z}'\mathbf{u}$ has mean zero and variance $\sigma^2\mathbf{Z}'\mathbf{Z}$ given the assumptions after (6.46). The efficient GLS estimator of $\boldsymbol{\beta}$ in model (6.46) is then

$$\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{Z}(\sigma^2\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Z}(\sigma^2\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}, \tag{6.48}$$

which equals the 2SLS estimator in (6.44) since the multipliers σ^2 cancel out. More generally, note that if the transformed model (6.47) is instead estimated by WLS with weighting matrix $\mathbf{W}_N$ then the more general estimator (6.37) is obtained.

Theil's Interpretation

Theil (1953) proposed estimation by OLS regression of the original model (6.46), except that the regressors $\mathbf{X}$ are replaced by a prediction $\widehat{\mathbf{X}}$ that is asymptotically uncorrelated with the error term.

Suppose that in the **reduced form model** the regressors $\mathbf{X}$ are a linear combination of the instruments plus some error, so that

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Pi} + \mathbf{v}, \tag{6.49}$$

where $\boldsymbol{\Pi}$ is a $K \times r$ matrix. Multivariate OLS regression of $\mathbf{X}$ on $\mathbf{Z}$ yields estimator $\widehat{\boldsymbol{\Pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ and OLS predictions $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Pi}}$ or

$$\widehat{\mathbf{X}} = \mathbf{P}_{\mathbf{Z}}\mathbf{X},$$

where $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. OLS regression of $\mathbf{y}$ on $\widehat{\mathbf{X}}$ rather than $\mathbf{y}$ on $\mathbf{X}$ yields estimator

$$\widehat{\boldsymbol{\beta}}_{\text{Theil}} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y}. \tag{6.50}$$

Theil's interpretation permits computation by two OLS regressions, with the first-stage OLS giving $\widehat{\mathbf{X}}$ and the second-stage OLS giving $\widehat{\boldsymbol{\beta}}$, leading to the term **two-stage least-squares estimator**.

To establish consistency of this estimator reexpress the linear model (6.46) as

$$\mathbf{y} = \widehat{\mathbf{X}}\boldsymbol{\beta} + (\mathbf{X}-\widehat{\mathbf{X}})\boldsymbol{\beta} + \mathbf{u}.$$

The second-stage OLS regression of $\mathbf{y}$ on $\widehat{\mathbf{X}}$ yields a consistent estimator of $\boldsymbol{\beta}$ if the regressor $\widehat{\mathbf{X}}$ is asymptotically uncorrelated with the composite error term $(\mathbf{X}-\widehat{\mathbf{X}})\boldsymbol{\beta} + \mathbf{u}$. If $\widehat{\mathbf{X}}$ were any proxy variable there is no reason for this to hold; however, here $\widehat{\mathbf{X}}$ is uncorrelated with $(\mathbf{X}-\widehat{\mathbf{X}})$ as an OLS prediction is orthogonal to the OLS residual. Thus plim $N^{-1}\widehat{\mathbf{X}}'(\mathbf{X}-\widehat{\mathbf{X}})\boldsymbol{\beta} = \mathbf{0}$. Also,

$$N^{-1}\widehat{\mathbf{X}}'\mathbf{u} = N^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u} = N^{-1}\mathbf{X}'\mathbf{Z}(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}N^{-1}\mathbf{Z}'\mathbf{u}.$$

Then $\widehat{\mathbf{X}}$ is asymptotically uncorrelated with $\mathbf{u}$ provided $\mathbf{Z}$ is a valid instrument so that plim $N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$. This consistency result for $\widehat{\boldsymbol{\beta}}_{\text{Theil}}$ depends heavily on the linearity of the model and does not generalize to nonlinear models.

Theil's estimator in (6.50) equals the 2SLS estimator defined earlier in (6.44). We have

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{Theil}} &= (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{P}_{\mathbf{Z}}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y},\end{aligned}$$

the 2SLS estimator, using $\mathbf{P}_{\mathbf{Z}}'\mathbf{P}_{\mathbf{Z}} = \mathbf{P}_{\mathbf{Z}}$ in the final equality.

Care is needed in implementing 2SLS using Theil's method. The second-stage OLS will give the wrong standard errors, even if errors are homoskedastic, as it will estimate σ^2 using the second-stage OLS regression residuals ($\mathbf{y} - \widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}}$) rather than the actual residuals ($\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$). In practice one may also make adjustment for heteroskedastic errors. It is much easier to use a program that offers 2SLS as an option and directly computes (6.44) and the associated variance matrix given in Table 6.2.

The 2SLS interpretation does not always carry over to nonlinear models, as detailed in Section 6.5.4. The GMM interpretation does, and for this reason it is emphasized here more than Theil's original derivation of linear 2SLS.

Theil actually considered a model where only some of the regressors $\mathbf{X}$ are endogenous and the remaining are exogenous. The preceding analysis still applies, provided all the exogenous components of $\mathbf{X}$ are included in the instruments $\mathbf{Z}$. Then the first-stage OLS regression of the exogenous regressors on the instruments fits perfectly and the predictions of the exogenous regressors equal their actual values. So in practice at the first-stage just the endogenous variables are regressed on the instruments, and the second-stage regression is of $\mathbf{y}$ on the exogenous regressors and the first-stage predictions of the endogenous regressors.

Basmann's Interpretation

Basmann (1957) proposed using as instruments the OLS reduced form predictions $\widehat{\mathbf{X}} = \mathbf{P}_{\mathbf{Z}}\mathbf{X}$ for the simple IV estimator in the just-identified case, since there are then exactly as many instruments $\widehat{\mathbf{X}}$ as regressors $\mathbf{X}$. This yields

$$\widehat{\boldsymbol{\beta}}_{\text{Basmann}} = (\widehat{\mathbf{X}}'\mathbf{X})^{-1}\widehat{\mathbf{X}}'\mathbf{y}. \tag{6.51}$$

This is consistent since plim $N^{-1}\widehat{\mathbf{X}}'\mathbf{u} = \mathbf{0}$, as already shown for Theil's estimator.

The estimator (6.51) actually equals the 2SLS estimator defined in (6.44), since $\widehat{\mathbf{X}}' = \mathbf{X}'\mathbf{P}_{\mathbf{Z}}$.

This IV approach will lead to correct standard errors and can be extended to nonlinear settings.

6.4.4. Alternatives to Standard IV Estimators

The IV-based optimal GMM and 2SLS estimators presented in Section 6.4.2 are the standard estimators used when regressors are endogenous. Chernozhukov and Hansen (2005) present an IV estimator for quantile regression.

Here we briefly discuss leading alternative estimators that have received renewed interest given the poor finite-sample properties of 2SLS with weak instruments detailed in Section 4.9. We focus on single-equation linear models. At this stage there is no method that is relatively efficient yet has small bias in small samples.

Limited-Information Maximum Likelihood

The limited-information maximum likelihood (LIML) **estimator** is obtained by joint ML estimation of the single equation (6.46) plus the reduced form for the endogenous regressors in the right-hand side of (6.46) assuming homoskedastic normal errors. For details see Greene (2003, p. 402) or Davidson and MacKinnon (1993, pp. 644–651). More generally the k class of estimators (see, for example, Greene, 2003, p. 403) includes LIML, 2SLS, and OLS.

The LIML estimator due to Anderson and Rubin (1949) predates the 2SLS estimator. Unlike 2SLS, the LIML estimator is invariant to the normalization used in a simultaneous equations system. Moreover, LIML and 2SLS are asymptotically equivalent given homoskedastic errors. Yet LIML is rarely used as it is more difficult to implement and harder to explain than 2SLS. Bekker (1994) presents small-sample results for LIML and a generalization of LIML. See also Hahn and Hausman (2002).

Split-Sample IV

Begin with Basmann's interpretation of 2SLS as an IV estimator given in (6.51). Substituting for $\mathbf{y}$ from (6.46) yields

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\widehat{\mathbf{X}}'\mathbf{X})^{-1}\widehat{\mathbf{X}}'\mathbf{u}.$$

By assumption plim $N^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{0}$ so plim $N^{-1}\widehat{\mathbf{X}}'\mathbf{u} = \mathbf{0}$ and $\widehat{\boldsymbol{\beta}}$ is consistent. However, correlation between $\mathbf{X}$ and $\mathbf{u}$, the reason for IV estimation, means that $\widehat{\mathbf{X}} = \mathbf{P}_{\mathbf{Z}}\mathbf{X}$ is correlated with $\mathbf{u}$. Thus $\mathrm{E}[\widehat{\mathbf{X}}'\mathbf{u}] \neq \mathbf{0}$, which leads to bias in the IV estimator. This bias arises from using $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Pi}}$ rather than $\widehat{\mathbf{X}} = \mathbf{Z}\boldsymbol{\Pi}$ as the instrument.

An alternative is to instead use as instrument predictions $\widetilde{\mathbf{X}}$, which have the property that $\mathrm{E}[\widetilde{\mathbf{X}}'\mathbf{u}] = \mathbf{0}$ in addition to plim $N^{-1}\widehat{\mathbf{X}}'\mathbf{u} = \mathbf{0}$, and use estimator

$$\widetilde{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}'\mathbf{X})^{-1}\widetilde{\mathbf{X}}'\mathbf{y}.$$

Since $\mathrm{E}[\widetilde{\mathbf{X}}'\mathbf{u}] = \mathbf{0}$ does not imply $\mathrm{E}[(\widetilde{\mathbf{X}}'\mathbf{X})^{-1}\widetilde{\mathbf{X}}'\mathbf{u}] = \mathbf{0}$, this estimator will still be biased, but the bias may be reduced.

Angrist and Krueger (1995) proposed obtaining such instruments by splitting the sample into two subsamples $(\mathbf{y}_1, \mathbf{X}_1, \mathbf{Z}_1)$ and $(\mathbf{y}_2, \mathbf{X}_2, \mathbf{Z}_2)$. The first sample is used to obtain estimate $\widehat{\boldsymbol{\Pi}}_1$ from regression of $\mathbf{X}_1$ on $\mathbf{Z}_1$. The second sample is used to obtain the IV estimator where the instrument $\widetilde{\mathbf{X}}_2 = \mathbf{Z}_2\widehat{\boldsymbol{\Pi}}_1$ uses $\widehat{\boldsymbol{\Pi}}_1$ obtained from the separate first sample. Angrist and Krueger (1995) define the **unbiased split-sample IV estimator** as

$$\widetilde{\boldsymbol{\beta}}_{\text{USSIV}} = (\widetilde{\mathbf{X}}_2'\mathbf{X}_2)^{-1}\widetilde{\mathbf{X}}_2'\mathbf{y}_2.$$

The **split-sample IV estimator** $\widetilde{\boldsymbol{\beta}}_{\text{SSIV}} = (\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2)^{-1}\widetilde{\mathbf{X}}_2'\mathbf{y}_2$ is a variant based on Theil's interpretation of 2SLS. These estimators have finite-sample bias toward zero, unlike 2SLS, which is biased toward OLS. However, considerable efficiency loss occurs because only half the sample is used at the final stage.

Jackknife IV

A more efficient variant of this estimator implements a similar procedure but generates instruments observation by observation.

Let the subscript $(-i)$ denote the leave-one-out operation that drops the ith observation. Then for the ith observation we obtain estimate $\widehat{\boldsymbol{\Pi}}_i$ from regression of $\mathbf{X}_{(-i)}$ on $\mathbf{Z}_{(-i)}$ and use as instrument $\widetilde{\mathbf{x}}_i' = \mathbf{z}_i'\widehat{\boldsymbol{\Pi}}_i$. Repeating N times gives an instrument vector denoted $\widetilde{\mathbf{X}}_{(-i)}$ with ith row $\widetilde{\mathbf{x}}_i'$. This leads to the **jackknife IV estimator**

$$\widetilde{\boldsymbol{\beta}}_{\text{JIV}} = (\widetilde{\mathbf{X}}_{(-i)}'\mathbf{X})^{-1}\widetilde{\mathbf{X}}_{(-i)}'\mathbf{y}_2.$$

This estimator was originally proposed by Phillips and Hale (1977). Angrist, Imbens and Krueger (1999) and Blomquist and Dahlberg (1999) called it a jackknife estimator since the jackknife (see Section 11.5.5) is a leave-one-out method for bias reduction. The computational burden of obtaining the N jackknife predicted values $\widetilde{\mathbf{x}}_i'$ is modest by use of the recursive formula given in Section 11.5.5. The Monte Carlo evidence given in the two recent papers is mixed, however, indicating a potential for bias reduction but also an increase in the variance. So the jackknife version may not be better than the conventional version in terms of mean-square error. The earlier paper by Phillips and Hale (1977) presents analytical results that the finite-sample bias of the JIV estimator is smaller than that of 2SLS only for appreciably overidentified models with $r > 2(K+1)$. See also Hahn, Hausman and Kuersteiner (2001).

Independently Weighted 2SLS

A related method to split-sample IV is the independently weighted GMM estimator of Altonji and Segal (1996) given in Section 6.3.5. Splitting the sample into G groups and specializing to linear IV yields the **independently weighted IV estimator**

$$\widehat{\boldsymbol{\beta}}_{\text{IWIV}} = \frac{1}{G}\sum_{g=1}^{G}\left[\mathbf{X}_g'\mathbf{Z}_g\widehat{\mathbf{S}}_{(-g)}^{-1}\mathbf{Z}_g'\mathbf{X}_g\right]^{-1}\mathbf{X}_g'\mathbf{Z}_g\widehat{\mathbf{S}}_{(-g)}^{-1}\mathbf{Z}_g'\mathbf{y}_g,$$

where $\widehat{\mathbf{S}}_{(-g)}$ is computed using $\widehat{\mathbf{S}}$ defined in (6.40) except that observations from the gth group are excluded. In a panel application Ziliak (1997) found that the independently weighted IV estimator performed much better than the unbiased split-sample IV estimator.

6.5. Nonlinear Instrumental Variables

Nonlinear IV methods, notably nonlinear 2SLS proposed by Amemiya (1974), permit consistent estimates of nonlinear regression models in situations where the NLS

CHAPTER 16

Tobit and Selection Models

16.1. Introduction

In this chapter we consider two closely related topics: regression when the dependent variable of interest is **incompletely observed** and regression when the dependent variable is completely observed but is observed in a **selected sample** that is not representative of the population. This includes limited dependent variable models, latent variable models, generalized Tobit models, and selection models.

All these models share the common feature that even in the simplest case of population conditional mean linear in regressors, OLS regression leads to inconsistent parameter estimates because the sample is not representative of the population. Alternative estimation procedures, most relying on strong distributional assumptions, are necessary to ensure consistent parameter estimation.

Leading causes of incompletely observed data are truncation and censoring. For **truncated data** some observations on both the dependent variable and regressors are lost. For example, income may be the dependent variable and only low-income people are included in the sample. For **censored data** information on the dependent variable is lost, but not data on the regressors. For example, people of all income levels may be included in the sample, but for confidentiality reasons the income of high-income people may be top-coded and reported only as exceeding, say, $100,000 per year. Truncation entails greater information loss than does censoring. A leading example of truncation and censoring is the **Tobit model**, named after Tobin (1958), who considered linear regression under normality. Similar issues arise for truncation and censoring in other models introduced in later chapters, most notably for censored duration data presented in Chapter 17. More generally, truncation and censoring are examples of missing data problems that are studied in Chapter 27.

The first-generation estimation methods require strong distributional assumptions. Even seemingly minor departures from assumptions, such as heteroskedastic errors when homoskedastic errors are assumed, can lead to inconsistent parameter estimates. For this reason the models presented in this chapter provide a leading econometrics application of semiparametric regression methods. Semiparametric methods for simple

forms of censoring and truncation such as top-coding have been successfully applied. However, for more general models with selection on unobservables there is to date no widely accepted procedure.

Section 16.2 presents general theory for censored and truncated nonlinear regression models, with specialization to the Tobit model given in Section 16.3. An alternative model for censored data, the two-part model, is introduced in Section 16.4. The sample selection model is presented in Section 16.5. An application to health expenditures in Section 16.6 contrasts the two-part and sample selection models. The Roy model for unobserved counterfactuals is presented in Section 16.7. Section 16.8 considers fully structural models obtained by utility maximization with corner solutions or by extension of simultaneous equation models to selected samples. Semiparametric estimation is presented in Section 16.9.

16.2. Censored and Truncated Models

We present general methods for estimation of fully parametric models when data are censored or truncated. These methods can be applied to models presented in later chapters such as count and duration models. The leading example, the Tobit model for censoring or truncation in linear models, is introduced in Section 16.2.1 and given separate treatment in Section 16.3.

16.2.1. Censoring and Truncation Example

Let y^* denote a variable that is incompletely observed. For truncation from below, y^* is only observed if y^* exceeds a threshold. For simplicity, let that threshold be zero. Then we observe $y = y^*$ if $y^* > 0$. Since negative values do not appear in the sample, the truncated mean exceeds the mean of y^*. For censoring from below at zero, y^* is not completely observed when $y^* \leq 0$, but it is known that $y^* < 0$ and for simplicity y is then set to 0. Since negative values are scaled up to zero, the censored mean also exceeds the mean of y^*. Clearly, sample means in truncated or censored samples cannot be used without adjustment to estimate the original population mean.

This chapter studies similar issues for regression models. With luck, truncation and censoring might lead only to a shift up or down in the intercept, leaving slope coefficients unchanged; however, this is not the case. For example, if $\mathrm{E}[y^*|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ in the original model then truncation or censoring leads to $\mathrm{E}[y|\mathbf{x}]$ being nonlinear in $\mathbf{x}$ and $\boldsymbol{\beta}$ so that OLS gives inconsistent estimates of $\boldsymbol{\beta}$ and hence inconsistent estimates of marginal effects.

As an illustration we consider the following labor supply example with simulated data. The relationship between desired annual hours worked, y^*, and hourly wage, w, is specified to be of linear-log form with data-generation process

$$\begin{aligned} y^* &= -2500 + 1000 \ln w + \varepsilon, \\ \varepsilon &\sim \mathcal{N}[0, 1000^2], \\ \ln w &\sim \mathcal{N}[2.75, 0.60^2]. \end{aligned} \tag{16.1}$$

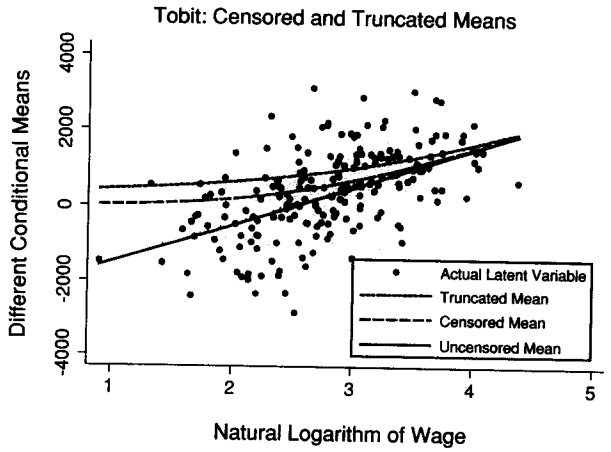


Figure 16.1: Tobit regression of hours on log wage: uncensored conditional mean (bottom), censored conditional mean (middle), and truncated conditional mean (top) for censoring/truncation from below at zero hours. Data are generated from a classical linear regression model.

This is a Tobit model, studied in detail in Section 16.3. The model implies that the wage elasticity is $1000/y^*$, which equals, for example, 0.5 for full-time work (2,000 hours). For each 1% increase in wage, annual hours increase by 10 hours.

Figure 16.1 presents a scatter plot of y^* and $\ln w$ for a generated sample of 200 observations. The unconditional mean for y^*, which is $-2500 + 1000 \ln w$, is given by the lowest curve, which is a straight line.

With censoring at zero, negative values of y^* are set to zero because people with negative desired hours of work choose not to work. For this particular sample this is the case for about 35% of the observations. This pushes up the mean for low wages, since the many negative values of the y^* are shifted up to zero. It has little impact for high wages, since then few observations on y^* are zero. The middle curve in Figure 16.1 gives the resulting censored mean, using the formula given later in (16.23).

With truncation at zero the 35% of the population with negative values of y^* are dropped altogether. This increases the mean above the censored mean, since zero values are no longer included in the data used to form the mean. The upper curve in Figure 16.1 gives the resulting truncated mean, using the formula given later in (16.23).

It is clear that censored and truncated conditional means are nonlinear in $\mathbf{x}$ even if the underlying population mean is linear. OLS estimation using truncated or censored data will lead to inconsistent estimation of the slope parameter, since by visual inspection of Figure 16.1 a linear approximation to the nonlinear truncated and censored means will have flatter slope than that for the original untruncated mean. Analysis should instead be based on the formulas for the censored or truncated conditional mean. Unfortunately these are based on strong distributional assumptions, as we will see.

16.2.2. Censoring and Truncation Mechanisms

As is customary for regression analysis, we let y denote the observed value of the dependent variable. The departure from usual analysis is that y is the incompletely observed value of a **latent dependent variable** y^*, where the observation rule is

$$y = g(y^*),$$

for some specified function $g(\cdot)$. Leading examples of $g(\cdot)$ immediately follow.

Censoring

With **censoring** we always observe the regressors $\mathbf{x}$, completely observe y^* for a subset of the possible values of y^*, and incompletely observe y for the remaining possible values of y^*. If censoring is **from below** (or from the left), we observe

$$y = \begin{cases} y^* & \text{if } y^* > L \\ L & \text{if } y^* \leq L. \end{cases} \tag{16.2}$$

For example, all consumers may be sampled with some having positive durable goods expenditures ($y^* > 0$) and others having zero expenditures ($y^* \leq 0$). If censoring is **from above** (or from the right) we observe

$$y = \begin{cases} y^* & \text{if } y^* < U \\ U & \text{if } y^* \geq U. \end{cases} \tag{16.3}$$

For example, annual income data may be top-coded at $U = \$100{,}000$. This form of censoring is called type 1 censoring in the duration literature (see Section 17.4.1).

The incompletely observed observations on y^* are set to L or U for simplicity. More generally, we require that for incompletely observed observations y^* is known to be missing (i.e., we observe that y^* lies outside the relevant bound) and regressors $\mathbf{x}$ continue to be completely observed.

Truncation

Truncation entails additional information loss as all data on observations at the bound are lost. With truncation from below we observe only

$$y = y^* \quad \text{if } y^* > L. \tag{16.4}$$

For example, only consumers who purchased durable goods may be sampled ($L = 0$). With truncation from above we observe only

$$y = y^* \quad \text{if } y^* < U. \tag{16.5}$$

For example, only low-income individuals may be sampled.

Interval Data

Interval data are data recorded in intervals. Survey data are often collected in this way to aid recall and to provide some greater anonymity in responses to more personal

questions. For example, income may be reported in intervals of $10,000 and then top-coded at $100,000. Such data are censored at multiple points, with the observed data y being the particular interval in which the unobserved y^* lies.

16.2.3. Censored and Truncated MLE

Censoring and truncation are easily dealt with if the researcher applies a fully parametric approach. This may be the case with interval data or top-coded data where, for example, it may be reasonable to assume a log-normal distribution for earnings or a negative binomial model for number of doctor visits.

If the conditional distribution of y^* given regressors $\mathbf{x}$ is specified, then the parameters of this distribution can be consistently and efficiently estimated by ML estimation based on the conditional distribution of the censored or truncated y. Specifically, let $f^*(y^*|\mathbf{x})$ and $F^*(y^*|\mathbf{x})$ denote the conditional probability density function (or probability mass function) and cumulative distribution function of the latent variable y^*. Then one can always obtain $f(y|\mathbf{x})$ and $F(y|\mathbf{x})$, the corresponding conditional pdf and cdf of the observed dependent variable y, since $y = g(y^*)$ is a transformation of y^*.

The limitation of the parametric approach is its reliance on strong distributional assumptions. For example, for the linear regression model under normality the MLE remains consistent even if the errors are nonnormal, but the censored MLE becomes inconsistent if the errors are nonnormal (see Section 16.3.2). More flexible models and semiparametric methods are presented in later sections.

Censored MLE

Censoring and truncation change both the conditional mean and the conditional density. We begin with the density.

Consider ML estimation given censoring from below. For $y > L$ the density of y is the same as that for y^*, so $f(y|\mathbf{x}) = f^*(y|\mathbf{x})$. For $y = L$, the lower bound, the density is discrete with mass equal to the probability of observing $y^* \leq L$, or $F^*(L|\mathbf{x})$. Thus for censoring from below

$$f(y|\mathbf{x}) = \begin{cases} f^*(y|\mathbf{x}) & \text{if } y > L, \\ F^*(L|\mathbf{x}) & \text{if } y = L. \end{cases}$$

As mentioned after (16.3), setting $y = L$ when $y^* \leq L$ is not necessary. Even if no value of y is observed when $y^* \leq L$ the density is still $F^*(L|\mathbf{x})$.

The density is a hybrid of the pdf and cdf of y^*. Similar to analysis for binary outcome models, it is notationally convenient to introduce an indicator variable

$$d = \begin{cases} 1 & \text{if } y > L, \\ 0 & \text{if } y = L. \end{cases} \tag{16.6}$$

Then the conditional density given censoring from below can be written as

$$f(y|\mathbf{x}) = f^*(y|\mathbf{x})^d F^*(L|\mathbf{x})^{1-d}. \tag{16.7}$$

For a sample of N independent observations, the censored MLE maximizes

$$\ln L_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left\{ d_i \ln f^*(y_i|\mathbf{x}_i, \boldsymbol{\theta}) + (1 - d_i) \ln F^*(L_i|\mathbf{x}_i, \boldsymbol{\theta}) \right\}, \tag{16.8}$$

where $\boldsymbol{\theta}$ are the parameters of the distribution of y^*. For generality the censoring lower bound L_i is permitted to vary across individuals, though usually $L_i = L$. The censored MLE is consistent and asymptotically normal, provided the original density of the uncensored variable $f^*(y^*|\mathbf{x}, \boldsymbol{\theta})$ is correctly specified.

When censoring is instead from above, the log-likelihood is similar to (16.8), except now $d = 1$ if $y < U$ and $d = 0$ otherwise, and $F^*(L|\mathbf{x}, \boldsymbol{\theta})$ is replaced by $1 - F^*(U|\mathbf{x}, \boldsymbol{\theta})$. A leading example is right-censored duration data (see Section 17.4).

Truncated MLE

For truncation from below at L, and suppressing dependence on $\mathbf{x}$, the conditional density of the observed y is

$$\begin{aligned} f(y) &= f^*(y|y > L) \\ &= f^*(y)/\Pr[y|y > L] \\ &= f^*(y)/[1 - F^*(L)]. \end{aligned}$$

The truncated MLE therefore maximizes

$$\ln L_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left\{ \ln f^*(y_i|\mathbf{x}_i, \boldsymbol{\theta}) - \ln[1 - F^*(L_i|\mathbf{x}_i, \boldsymbol{\theta})] \right\}. \tag{16.9}$$

If instead truncation is from above, the log-likelihood is (16.9), except that $1 - F^*(L|\mathbf{x}, \boldsymbol{\theta})$ is replaced by $F^*(U|\mathbf{x}, \boldsymbol{\theta})$.

Ignoring censoring or truncation leads to inconsistency. For example, if truncation is ignored the MLE maximizes $\sum_i \ln f^*(y_i|\mathbf{x}_i, \boldsymbol{\theta})$, which is the wrong likelihood function as it drops the second term in (16.9). Consistency of the censored and truncated MLE requires correct specification of $f(\cdot)$, which in turn requires correct specification of the latent variable density $f^*(\cdot)$. Even if $f^*(\cdot)$ is an LEF density (see Section 5.7.3), the density, and not just the mean, must be correctly specified if censoring or truncation are present.

Interval Data MLE

Suppose the latent variable y^* is only observed to lie in the $(J + 1)$ mutually exclusive intervals $(-\infty, a_1], (a_1, a_2], \ldots, (a_J, \infty)$, where $a_1, a_2, \ldots, a_J$ are known. Then since

$$\begin{aligned} \Pr[a_j < y^* \leq a_{j+1}] &= \Pr[y^* \leq a_{j+1}] - \Pr[y^* \leq a_j] \\ &= F^*(a_{j+1}) - F^*(a_j), \end{aligned}$$

the interval data MLE maximizes

$$\ln L_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=0}^{J} d_{ij} \ln \left[F^*(a_{j+1}|\mathbf{x}_i, \boldsymbol{\theta}) - F^*(a_j|\mathbf{x}_i, \boldsymbol{\theta}) \right], \tag{16.10}$$

where the d_{ij}, $j = 0, \ldots, J$, are binary indicators equal to one if $y_{ij} \in (a_j, a_{j+1}]$ and zero otherwise. This is similar to an ordered probit or logit model (see Section 15.9.1), except here the interval boundaries $a_1, \ldots, a_J$ are known.

16.2.4. Poisson Censored and Truncated MLE Example

Assume that y^* is Poisson distributed, so that $f^*(y) = e^{-\mu}\mu^y/y!$ and $\ln f^*(y) = -\mu + y \ln \mu - \ln y!$, with mean $\mu = \exp(\mathbf{x}'\boldsymbol{\beta})$.

Suppose the number of visits to a health clinic is modeled, but data are only available for people who visited the health clinic. Then the data are truncated from below at zero and we only observe $y = y^*$ if $y^* > 0$. Then $F^*(0) = \Pr[y^* \leq 0] = \Pr[y^* = 0] = e^{-\mu}$, and from (16.9) the truncated MLE for $\boldsymbol{\beta}$ maximizes

$$\ln L_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ -\exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i \mathbf{x}_i'\boldsymbol{\beta} - \ln y_i! - \ln[1 - \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))] \right\}.$$

Suppose instead that data are censored from above at 10 because of top-coding, so that we observe $y = y^*$ if $y^* < 10$ and that $y = 10$ if $y^* \geq 10$. Then $\Pr[y^* \geq 10] = 1 - \Pr[y^* < 10] = 1 - \sum_{k=0}^{9} f^*(k)$. From (16.8) the censored MLE for $\boldsymbol{\beta}$ maximizes

$$\ln L_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ d_i \left[-\exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i \mathbf{x}_i'\boldsymbol{\beta} - \ln y_i! \right] + (1 - d_i) \ln \left[\sum_{k=0}^{9} e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})} (\exp(\mathbf{x}_i'\boldsymbol{\beta}))^k / k! \right] \right\}.$$

In both cases the resulting first-order conditions are considerably more complicated than those for the Poisson MLE without truncation or censoring. Also, in both cases ignoring the truncation or censoring and maximizing the original density leads to inconsistent parameter estimates.

16.2.5. Censored and Truncated Conditional Means

Censoring and truncation change the conditional mean.

For example, consider the Poisson truncated from below at zero. The truncated density is $f^*(y)/[1 - F^*(0)]$, $y = 1, 2, \ldots$, so the truncated mean is $\sum_{k=1}^{\infty} k f^*(k)/[1 - F^*(0)] = \sum_{k=0}^{\infty} k f^*(k)/[1 - F^*(0)] = \mu/(1 - e^{-\mu})$. Thus

$$\mathrm{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})/[1 - \exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))],$$

rather than $\exp(\mathbf{x}'\boldsymbol{\beta})$ if there were no truncation.

This expression for $\mathrm{E}[y|\mathbf{x}]$ can be used for NLS estimation. There is little advantage to NLS rather than ML estimation, however, as given truncation the NLS estimator relies on distributional assumptions that are essentially as strong as those needed for consistency of the more efficient ML estimator.

16.3. Tobit Model

Truncation and censoring arise most often in econometrics in the linear regression model with normally distributed error, when only positive outcomes are completely observed. This model is called the Tobit model after Tobin (1958), who applied it to individual expenditures on consumer durable goods. The model in practice is usually too restrictive. It is nonetheless presented in some detail, as it provides the basis for more general models presented in subsequent sections of this chapter.

16.3.1. Tobit Model

The censored normal regression model, or **Tobit model**, is one with censoring from below at zero where the latent variable is linear in regressors with additive error that is normally distributed and homoskedastic. Thus

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \tag{16.11}$$

where the error term

$$\varepsilon \sim \mathcal{N}[0, \sigma^2] \tag{16.12}$$

has variance σ^2 constant across observations. This implies that the latent variable $y^* \sim \mathcal{N}[\mathbf{x}'\boldsymbol{\beta}, \sigma^2]$. The observed y is defined by (16.2) with $L = 0$, so

$$y = \begin{cases} y^* & \text{if } y^* > 0, \\ - & \text{if } y^* \leq 0, \end{cases} \tag{16.13}$$

where – means that y is observed to be missing. No particular value of y is necessarily observed when $y^* \leq 0$, though in some settings such as durable goods expenditures we observe $y = 0$.

Equations (16.11) – (16.13) define the prototypical Tobit model analyzed by Tobin (1958). More generally, Tobit models begin with (16.11) and (16.12) for the latent variable but can have other censoring mechanisms including censoring from above, censoring from both below and above (the **two-limit Tobit model**), and interval-censored data. The results in this section are restricted to the censoring mechanism given in (16.13). The models of later sections are sometimes called generalized Tobit models.

The normalization $L = 0$ is not only natural in many settings, but some such normalization is necessary for a linear model with intercept and constant threshold parameter L. Then we observe y if $y^* > L$, or equivalently if $\beta_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon > L$ or $(\beta_1 - L) + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon > 0$. Thus only the difference $(\beta_1 - L)$ is identified. More generally, the latent model $y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ with variable censoring threshold $L = \mathbf{x}'\boldsymbol{\gamma}$ is observationally equivalent to the latent model $y^* = \mathbf{x}'(\boldsymbol{\beta} - \boldsymbol{\gamma}) + \varepsilon$ with fixed threshold $L = 0$. These results are a consequence of censoring arising in a linear model with additive error and do not carry over to nonlinear models, such as the preceding Poisson example.

Applying the general expression (16.7) for the censored density, here $f^*(y)$ is the $\mathcal{N}[\mathbf{x}'\boldsymbol{\beta}, \sigma^2]$ density and

$$\begin{aligned} F^*(0) &= \Pr[y^* \leq 0] \\ &= \Pr[\mathbf{x}'\boldsymbol{\beta} + \varepsilon \leq 0] \\ &= \Phi(-\mathbf{x}'\boldsymbol{\beta}/\sigma) \\ &= 1 - \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma), \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cdf and the last equality uses symmetry of the standard normal distribution. Thus the censored density can be expressed as

$$f(y) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \mathbf{x}'\boldsymbol{\beta})^2\right\}\right]^d \left[1 - \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\right]^{1-d}, \tag{16.14}$$

where the binary indicator d is defined in (16.6) with $L = 0$.

The Tobit MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\sigma}^2)'$ maximizes the censored log-likelihood function (16.8). Given (16.14) this becomes

$$\begin{aligned} \ln L_N(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{N} \Bigg\{ & d_i \left(-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right) \\ & + (1 - d_i)\ln\left(1 - \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right)\Bigg\}, \end{aligned} \tag{16.15}$$

a mixture of discrete and continuous densities. The first-order conditions are

$$\begin{aligned} \frac{\partial \ln L_N}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N} \frac{1}{\sigma^2}\left(d_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) - (1 - d_i)\frac{\sigma\phi_i}{(1 - \Phi_i)}\right)\mathbf{x}_i = \mathbf{0} \\ \frac{\partial \ln L_N}{\partial \sigma^2} &= \sum_{i=1}^{N} \left\{ d_i \left(-\frac{1}{2\sigma^2} + \frac{\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2}{2\sigma^4}\right) + (1 - d_i)\frac{\phi_i \mathbf{x}_i'\boldsymbol{\beta}}{(1 - \Phi_i)}\frac{1}{2\sigma^3}\right\} = 0, \end{aligned} \tag{16.16}$$

using $\partial\Phi(z)/\partial z = \phi(z)$ where $\phi(\cdot)$ is the standard normal pdf, and with the definitions $\phi_i = \phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)$ and $\Phi_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)$. As usual $\widehat{\boldsymbol{\theta}}$ is consistent if the density is correctly specified, that is, if the dgp is (16.11) and (16.12) and the censoring mechanism is (16.13). The MLE is asymptotic normal distributed with variance matrix given in, for example, Maddala (1983, p. 155) and Amemiya (1985, p. 373).

Tobin (1958) proposed ML estimation of the Tobit model and asserted that the usual ML theory applied. Amemiya (1973) provided a formal proof that the usual theory did apply, despite the mixed discrete–continuous nature of the censored density. The appendix of this classic paper of Amemiya details the asymptotic theory for extremum estimators presented in Section 5.3.

If data are truncated, rather than censored, from below at zero then the Tobit MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\sigma}^2)'$ maximizes the truncated normal log-likelihood function

$$\ln L_N(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{N} \left\{ -\frac{1}{2}\ln\sigma^2 - \frac{1}{2}\ln 2\pi - \frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 - \ln\Phi\left(\mathbf{x}_i'\boldsymbol{\beta}/\sigma\right) \right\}, \tag{16.17}$$

obtained using (16.9) for y^* distributed as in (16.11) and (16.12).

16.3.2. Inconsistency of the Tobit MLE

A very major weakness of the Tobit MLE is its heavy reliance on distributional assumptions. If the error ε is either heteroskedastic or nonnormal the MLE is inconsistent.

This can be seen from the ML first-order conditions (16.16), which are a quite complicated function of variables including d_i, y_i, ϕ_i, and Φ_i. The first equation in (16.16) satisfies $\mathrm{E}[\partial \ln L_N/\partial\boldsymbol{\beta}] = \mathbf{0}$, a necessary condition for consistency (see Section 5.3.7), if

$$\begin{aligned} \mathrm{E}[d_i] &= \Phi_i, \\ \mathrm{E}[d_i y_i] &= \Phi_i \mathbf{x}_i'\boldsymbol{\beta} + \sigma\phi_i. \end{aligned}$$

These moment conditions can be shown to hold if the dgp is (16.11) and (16.12) and the censoring mechanism is (16.13). However, they are unlikely to hold under any other specification of the dgp, as they rely heavily on both normality and homoskedasticity. For example, with *heteroskedastic errors* the estimator is inconsistent, since then $\mathrm{E}[d_i] = \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma_i) \neq \Phi_i$ unless $\sigma_i^2 = \sigma^2$.

Consistent estimation with heteroskedastic normal errors is possible by specifying a model for heteroskedasticity, say $\sigma_i^2 = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$. For censoring from below at zero the log-likelihood $\ln L_N(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is that given in (16.15) with σ^2 replaced by $\exp(\mathbf{z}_i'\boldsymbol{\gamma})$. Consistency then requires normal errors and correct specification of the functional form of the heteroskedasticity.

Clearly, with censoring or truncation, distributional assumptions become important even for distributions somewhat robust to misspecification in the uncensored or untruncated case. Specification tests for the Tobit model are discussed in Section 16.3.7. In many censored data applications the Tobit model is not appropriate. More general models presented in subsequent sections of this chapter are instead used.

16.3.3. Censored and Truncated Means in Linear Regression

Censoring and truncation in the linear regression model (16.11) lead to observed dependent variable y that has distribution with conditional mean other than $\mathbf{x}'\boldsymbol{\beta}$, conditional variance other than σ^2 even if ε is homoskedastic, and distribution that is nonnormal even if ε is normally distributed. We present general results for linear regression in this section before specializing to normally distributed errors in Sections 16.3.4–

16.3.7. The results provide additional insights regarding the consequences of truncation and censoring and form the basis for non-ML estimation methods presented in later sections.

We begin with the truncated mean. The effects of truncation are intuitively predictable. Left-truncation excludes small values, so the mean should increase, whereas with right-truncation the mean should decrease. Since truncation reduces the range of variation, the variance should decrease.

For *left-truncation* at zero we only observe y if $y^* > 0$. If we suppress dependence of expectations on $\mathbf{x}$ for notational simplicity, the left-truncated mean becomes

$$\begin{aligned} \mathrm{E}[y] &= \mathrm{E}[y^* | y^* > 0] \\ &= \mathrm{E}\left[\mathbf{x}'\boldsymbol{\beta} + \varepsilon | \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0\right] \\ &= \mathrm{E}\left[\mathbf{x}'\boldsymbol{\beta} | \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0\right] + \mathrm{E}\left[\varepsilon | \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0\right] \\ &= \mathbf{x}'\boldsymbol{\beta} + \mathrm{E}\left[\varepsilon | \varepsilon > -\mathbf{x}'\boldsymbol{\beta}\right], \end{aligned} \tag{16.18}$$

where the second equality uses (16.11), and the last equality assumes ε is independent of $\mathbf{x}$. As expected the truncated mean exceeds $\mathbf{x}'\boldsymbol{\beta}$, since $\mathrm{E}[\varepsilon | \varepsilon > c]$ for any constant c will exceed $\mathrm{E}[\varepsilon]$.

For data *left-censored* at zero suppose we observe $y = 0$, rather than merely that $y^* \leq 0$. The censored mean is obtained by first conditioning the observable y on the binary indicator d defined in (16.6) with $L = 0$ and then unconditioning. Suppressing dependence on $\mathbf{x}$ for notational simplicity again, we have the left-censored mean

$$\begin{aligned} \mathrm{E}[y] &= \mathrm{E}_d[\mathrm{E}_{y|d}[y|d]] \\ &= \Pr[d = 0] \times \mathrm{E}[y|d = 0] + \Pr[d = 1] \times \mathrm{E}[y|d = 1] \\ &= 0 \times \Pr[y^* \leq 0] + \Pr[y^* > 0] \times \mathrm{E}[y^* | y^* > 0] \\ &= \Pr[y^* > 0] \times \mathrm{E}[y^* | y^* > 0], \end{aligned} \tag{16.19}$$

where $\Pr[y^* > 0] = 1 - \Pr[y^* \leq 0] = \Pr[\varepsilon > -\mathbf{x}'\boldsymbol{\beta}]$ is one minus the censoring probability and $\mathrm{E}[y^* | y^* > 0]$ is the truncated mean already derived in (16.18).

In summary, for the linear regression model with censoring or truncation from below at zero, the conditional means are given by

$$\begin{aligned} &\text{latent variable:} && \mathrm{E}[y^* | \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} \\ &\text{left-truncated (at 0):} && \mathrm{E}[y | \mathbf{x}, y > 0] = \mathbf{x}'\boldsymbol{\beta} + \mathrm{E}\left[\varepsilon | \varepsilon > -\mathbf{x}'\boldsymbol{\beta}\right], \\ &\text{left-censored (at 0):} && \mathrm{E}[y | \mathbf{x}] = \Pr[\varepsilon > -\mathbf{x}'\boldsymbol{\beta}] \left\{\mathbf{x}'\boldsymbol{\beta} + \mathrm{E}\left[\varepsilon | \varepsilon > -\mathbf{x}'\boldsymbol{\beta}\right]\right\}. \end{aligned} \tag{16.20}$$

It is clear that even though the original conditional mean is linear, censoring or truncation leads to conditional means that are nonlinear so that OLS estimates will be inconsistent.

One possible approach to take is a parametric one of assuming a distribution for ε. This leads to expressions for $\mathrm{E}\left[\varepsilon | \varepsilon > -\mathbf{x}'\boldsymbol{\beta}\right]$ and $\Pr[\varepsilon > -\mathbf{x}'\boldsymbol{\beta}]$ and hence the truncated or censored conditional mean. We do this in the next section for normally distributed errors.

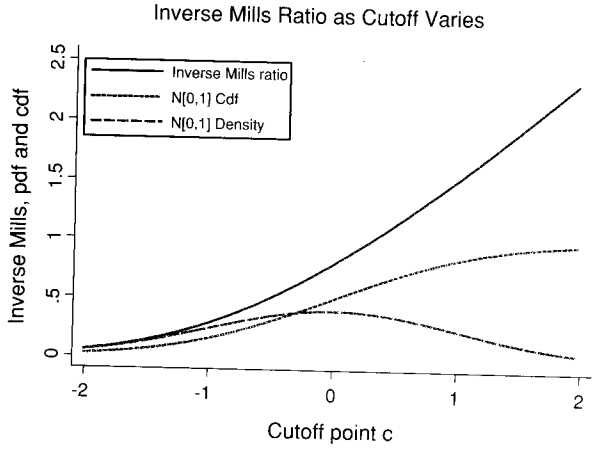


Figure 16.2: Inverse Mills ratio for the standard normal distribution as the censoring or cutoff point c increases. Standard normal cdf and density also plotted.

A second approach seeks to avoid or minimize such parametric assumptions. We consider this in a later section, but note here that regardless of the distribution for ε the truncated mean is a single-index model with correction term decreasing in $\mathbf{x}'\boldsymbol{\beta}$ since $\mathrm{E}[\varepsilon|\varepsilon > -\mathbf{x}'\boldsymbol{\beta}]$ is a monotonically decreasing function in $\mathbf{x}'\boldsymbol{\beta}$.

16.3.4. Censored and Truncated Means in the Tobit Model

For the Tobit model the regression error ε is normal and we use the following result, derived in Section 16.10.1.

Proposition 16.1 (Truncated Moments of the Standard Normal): *Suppose* $z \sim \mathcal{N}[0, 1]$. *Then the left-truncated moments of* z are

(i) $\mathrm{E}[z|z > c] = \phi(c)/[1 - \Phi(c)]$, *and* $\mathrm{E}[z|z > -c] = \phi(c)/\Phi(c)$,

(ii) $\mathrm{E}[z^2|z > c] = 1 + c\phi(c)/[1 - \Phi(c)]$, *and*

(iii) $\mathrm{V}[z|z > c] = 1 + c\phi(c)/[1 - \Phi(c)] - \phi(c)^2/[1 - \Phi(c)]^2$

Result (i) of Proposition 16.1 is shown in Figure 16.2. We consider truncation of $z \sim \mathcal{N}[0, 1]$ from below at c, where c ranges from -2 to 2. The lowest curve is the standard normal density $\phi(c)$ evaluated at c. The middle curve is the standard normal cdf $\Phi(c)$ evaluated at c and gives the probability of truncation when truncation is at c. This probability is approximately 0.023 at $c = -2$ and 0.977 at $c = 2$. The upper curve gives the truncated mean $\mathrm{E}[z|z > c] = \phi(c)/[1 - \Phi(c)]$. As expected this is close to $\mathrm{E}[z] = 0$ for $c = -2$, since then there is little truncation, and $\mathrm{E}[z|z > c] > c$. What is not expected a priori is that $\phi(c)/[1 - \Phi(c)]$ is approximately linear, especially for $c > 0$. Moments when truncation is from above can be obtained using, for example, $\mathrm{E}[z|z < c] = -\mathrm{E}[-z| - z > -c] = -\phi(c)/\Phi(c)$.

Appling this result to (16.18), the error term has truncated mean

$$\begin{aligned} \mathrm{E}\left[\varepsilon|\varepsilon > -\mathbf{x}'\boldsymbol{\beta}\right] &= \sigma \mathrm{E}\left[\tfrac{\varepsilon}{\sigma}|\tfrac{\varepsilon}{\sigma} > \tfrac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right] \\ &= \sigma\phi(-\tfrac{\mathbf{x}'\boldsymbol{\beta}}{\sigma})/[1-\Phi(-\tfrac{\mathbf{x}'\boldsymbol{\beta}}{\sigma})] \\ &= \sigma\phi(\tfrac{\mathbf{x}'\boldsymbol{\beta}}{\sigma})/[\Phi(\tfrac{\mathbf{x}'\boldsymbol{\beta}}{\sigma})] \\ &= \sigma\lambda(\tfrac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}), \end{aligned} \tag{16.21}$$

where the second line uses Proposition 16.1, the third line uses symmetry about zero of $\phi(z)$, and we define

$$\lambda(z) = \frac{\phi(z)}{\Phi(z)}. \tag{16.22}$$

We follow the definition and terminology of Amemiya (1985) and many others in defining $\lambda(\cdot)$ as in (16.22) and calling it the **inverse Mills ratio**. From Johnson and Kotz (1970, p. 278), Mills actually tabulated the ratio $(1-\Phi(z))/\phi(z)$ whose inverse $\phi(z)/[1-\Phi(z)] = \phi(z)/\Phi(-z)$ is the hazard function of the normal distribution. Some authors therefore instead write (16.21) as $\mathrm{E}\left[\varepsilon|\varepsilon > -\mathbf{x}'\boldsymbol{\beta}\right] = \sigma\lambda^*(-\mathbf{x}'\boldsymbol{\beta}/\sigma)$, where $\lambda^*(z) = \phi(z)/\Phi(-z)$ is referred to as the inverse Mills ratio.

Also, $\Pr[\varepsilon > -\mathbf{x}'\boldsymbol{\beta}] = \Pr[-\varepsilon < \mathbf{x}'\boldsymbol{\beta}] = \Pr[-\varepsilon/\sigma < \mathbf{x}'\boldsymbol{\beta}/\sigma] = \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)$. Then the conditional means in (16.20) specialize to

$$\begin{aligned} &\text{latent variable:} && \mathrm{E}[y^*|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}, \\ &\text{left-truncated (at 0):} && \mathrm{E}[y|\mathbf{x},\, y>0] = \mathbf{x}'\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma), \\ &\text{left-censored (at 0):} && \mathrm{E}[y|\mathbf{x}] = \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)\mathbf{x}'\boldsymbol{\beta} + \sigma\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma). \end{aligned} \tag{16.23}$$

The variance is similarly obtained (see Exercise 16.1). Defining $w = \mathbf{x}'\boldsymbol{\beta}/\sigma$, we have

$$\begin{aligned} &\text{latent variable:} && \mathrm{V}[y^*|\mathbf{x}] = \sigma^2, \\ &\text{left-truncated (at 0):} && \mathrm{V}[y|\mathbf{x},\, y>0] = \sigma^2\left[1 - w\lambda(w) - \lambda(w)^2\right], \\ &\text{left-censored (at 0):} && \mathrm{V}[y|\mathbf{x}] = \sigma^2\Phi(w)\left\{w^2 + w\lambda(w) + 1 - \Phi(w)[w+\lambda(w)]\right\}^2. \end{aligned} \tag{16.24}$$

Clearly truncation and censoring induce heteroskedasticity, and for truncation $\mathrm{V}[y|\mathbf{x}] < \sigma^2$ so that truncation reduces variability, as expected.

These results assume normal errors. Maddala (1983, p. 369) gives results similar to Proposition 16.1 for the log-normal, logistic, uniform, Laplace, exponential, and gamma distributions.

16.3.5. Marginal Effects in the Tobit Model

The marginal effect is the effect on the conditional mean of the dependent variable of changes in the regressors. This effect varies according to whether interest lies in the latent variable mean $\mathbf{x}'\boldsymbol{\beta}$ or the truncated or censored means given in (16.23).

Differentiating each with respect to $\mathbf{x}$ yields

$$\begin{aligned} &\text{latent variable:} && \partial \mathrm{E}[y^*|\mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\beta}, \\ &\text{left-truncated (at 0):} && \partial \mathrm{E}[y, y > 0|\mathbf{x}]/\partial \mathbf{x} = \{1 - w\lambda(w) - \lambda(w)^2\}\boldsymbol{\beta}, \\ &\text{left-censored (at 0):} && \partial \mathrm{E}[y|\mathbf{x}]/\partial \mathbf{x} = \Phi(w)\boldsymbol{\beta}, \end{aligned} \tag{16.25}$$

where $w = \mathbf{x}'\boldsymbol{\beta}/\sigma$ and we use $\partial\Phi(z)/\partial z = \phi(z)$ and $\partial\phi(z)/\partial z = -z\phi(z)$. The simple expression for the censored mean is obtained after some manipulation. It can be decomposed into two effects, one for $y = 0$ and one for $y > 0$ (see McDonald and Moffitt, 1980).

In some cases truncation or censoring is just an artifact of data collection, so the truncated and censored means are of no intrinsic interest and we are interested in $\partial \mathrm{E}[y^*|\mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\beta}$. For example, with top-coded earnings data we are clearly interested in measuring the effect of schooling on mean earnings rather than earnings of those not top-coded.

In other cases truncation or censoring has behavioral implications. In a model for hours worked, for example, the three marginal effects in (16.25) correspond to the effect of a change in a regressor on, respectively, (1) desired hours of work, (2) actual hours of work for workers, and (3) actual hours of work for workers and nonworkers. For (1) we clearly need an estimate of $\boldsymbol{\beta}$, but for (2) and (3) OLS slope coefficients, although inconsistent for $\boldsymbol{\beta}$, may actually provide a reasonable crude estimate of the marginal effect since the truncated and censored means are still fairly linear in $\mathbf{x}$.

16.3.6. Alternative Estimators for the Tobit Model

In addition to the MLE, consistent estimation is possible by NLS based on the correct expression for the truncated or censored mean. We consider the NLS estimator and other least-squares estimators.

NLS Estimator

The results in (16.23) can be used to permit consistent estimation of the Tobit model parameters by NLS. For example, with truncated data we minimize

$$S_N(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \sigma\lambda(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)\right)^2$$

with respect to both $\boldsymbol{\beta}$ and σ^2, but then perform inference controlling for the heteroskedasticity given in (16.24). A similar estimator can be obtained for censored data.

This estimator is not used in practice. Consistency requires correct specification of the truncated mean, which from (16.21) requires both normality and homoskedasticity of the errors. One might as well estimate by ML since this relies on assumptions just as strong and is fully efficient. Moreover, in practice the NLS estimator can be imprecise. From Figure 16.2 it is clear that $\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma)$ is approximately linear in $\mathbf{x}'\boldsymbol{\beta}/\sigma$, leading to near collinearity because $\mathbf{x}$ is also a regressor. In Section 16.5 we consider models that permit correction terms similar to $\sigma\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma)$ in (16.23) that have the advantage of depending in part on regressors other than those in $\mathbf{x}$.

Heckman Two-Step Estimator

From (16.23) the truncated (at zero) mean is

$$\mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}'\boldsymbol{\beta}/\sigma). \tag{16.26}$$

Rather than use NLS, this can be estimated in the following two-step procedure if censored data are available. First, for the full sample do probit regression of d on $\mathbf{x}$, where the binary variable d equals one if $y > 0$ is observed, to give consistent estimate $\widehat{\boldsymbol{\alpha}}$, where $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma$. Second, for the truncated sample do OLS regression of y on $\mathbf{x}$ and $\lambda(\mathbf{x}'\widehat{\boldsymbol{\alpha}})$ to give consistent estimates of $\boldsymbol{\beta}$ and σ.

This estimation procedure, due to Heckman (1976, 1979), is presented in Section 16.5.4 where it is applied to the more general sample selection model. Section 16.10.2 derives the standard error of $\widehat{\boldsymbol{\beta}}$ that accounts for the regressor $\lambda(\mathbf{x}'\widehat{\boldsymbol{\alpha}})$ depending on estimated parameters and for heteroskedasticity induced by truncation.

OLS Estimation of the Tobit Model

The OLS estimates using censored or truncated data are inconsistent for $\boldsymbol{\beta}$. This is because the censored and truncated means given in (16.23) are not equal to $\mathbf{x}'\boldsymbol{\beta}$, violating the essential condition for consistency of OLS.

For censored data, OLS provides a linear approximation to the nonlinear censored regression curve. It is clear from Figure 16.1 and (16.25) that this line is flatter than the regression line for uncensored data, which has slope equal to the true slope parameter. Goldberger (1981) showed analytically that if y and $\mathbf{x}$ are joint normally distributed and there is censoring from below at zero, then the OLS slope parameters converge to p times the true slope parameter, where p is the fraction of the sample with positive values of y. These conditions are restrictive but were relaxed somewhat by Ruud (1986). In practice this proportionality result provides a good empirical approximation to the inconsistency of OLS if a Tobit model is instead appropriate.

Similarly, with truncation the regression line is flatter than the untruncated regression line. Goldberger (1981) obtained an analytical result similar to that for the censored case. If y and $\mathbf{x}$ are joint normally distributed and there is censoring from below at zero, then the OLS slope parameters converge to a multiple of the true slope parameter. The multiple, the expression for which is quite lengthy, lies between zero and one, and the shrinkage is the same for all slope coefficients. Truncated OLS therefore understates the absolute magnitude of the true slope parameters.

16.3.7. Specification Tests for the Tobit Model

Given the fragility of the Tobit model it is good practice to test for distributional misspecification. There are four broad strategies.

The first approach is to nest the Tobit model within a richer parametric model and apply a Wald, LR, or LM test. Since the null hypothesis model, the Tobit model, is most easily estimated it is natural to use LM tests. This is particularly straightforward for testing against heteroskedasticity of the form $\sigma_i^2 = \exp(\mathbf{x}_i'\boldsymbol{\alpha})$ in the censored

regression model. Using the OPG form of the LM test (see Section 7.3.5) we compute N times the uncentered R^2 from auxiliary regression of 1 on $\tilde{\mathbf{s}}_{1i}$ and $\tilde{\mathbf{s}}_{2i}$, where $f_i = f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is the density given in (16.14) with σ replaced by $\exp(\mathbf{x}'\boldsymbol{\alpha})$, the expressions for $\mathbf{s}_{1i} = \partial \ln f_i/\partial\boldsymbol{\beta}$ and $\mathbf{s}_{2i} = \partial \ln f_i/\partial\boldsymbol{\alpha}$ are obtained by minor adaptation of the expressions in (16.16), and tilde denotes evaluation at the censored Tobit MLE with all components of $\boldsymbol{\alpha}$ except that for the intercept equal to zero. A similar approach for testing the assumption of normally distributed errors is more difficult as there is no standard generalization of the normal.

A second approach is to use conditional moment tests (see Section 8.2) that do not require specification of an alternative hypothesis model. In particular, the first-order conditions (16.16) for the censored Tobit MLE suggest conditional moment tests based on the generalized residual

$$e_i = d_i \frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma^2} - (1 - d_i)\frac{\phi_i}{\sigma(1 - \Phi_i)}.$$

If the Tobit model is correctly specified then $\mathrm{E}[e_i|\mathbf{x}_i] = 0$ since the regularity conditions imply that $\mathrm{E}[\partial \ln f(y_i)/\partial\boldsymbol{\beta}] = 0$. Then we can implement an m-test of H_0 : $\mathrm{E}[e\mathbf{z}] = \mathbf{0}$ against H_a : $\mathrm{E}[e\mathbf{z}] \neq \mathbf{0}$ using $N^{-1}\sum_{i=1}^{N} \widehat{e}_i\mathbf{z}_i$, where $\widehat{e}_i = e_i$ evaluated at the Tobit MLE $(\widehat{\boldsymbol{\beta}},\widehat{\sigma}^2)$. From Section 8.2.2 this test can be implemented by computing N times the uncentered R^2 from auxiliary regression of 1 on $\widehat{e}_i\mathbf{z}_i$, $\widehat{\mathbf{s}}_{1i}$, and $\widehat{s}_{2i}$, where $f_i = f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)$ is the density given in (16.14) and $\mathbf{s}_{1i} = \partial \ln f_i/\partial\boldsymbol{\beta}$ and $s_{2i} = \partial \ln f_i/\partial\sigma^2$ given in (16.16) are evaluated at $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$. The variables $\mathbf{z}_i$ may be variables other than $\mathbf{x}_i$, in which case the test can be interpreted as a test of omitted regressors, or powers of the components of $\mathbf{x}_i$. Conditional moment tests based on higher order moments have also been developed. For details see Chesher and Irish (1987) and Pagan and Vella (1989).

A third approach is to adapt some of the diagnostic and testing methods developed for right-censored duration data (see Chapter 19) to left-censored normally distributed data.

A final approach contrasts the Tobit MLE $\widehat{\boldsymbol{\beta}}$ with alternative estimates of $\boldsymbol{\beta}$, notably the semiparametric estimates presented in Section 16.9, that are consistent under weaker distributional assumptions.

For further details see Pagan and Vella (1989), who present theory with some application, and Melenberg and Van Soest (1996), who provide a more complete application. Both papers consider specification tests for the richer sample selection model (see Section 16.5) in addition to those for the Tobit model.

16.4. Two-Part Model

The preceding models for censored data restrict the censoring mechanism to be from the same model as that generating the outcome variable. More generally, the censoring mechanism and outcome may be modeled using separate processes. For example, in explaining individual annual hospital expenses one process may determine hospitalization and a second process may explain consequent hospital expenses. The case for

postulating two separate mechanisms is strong if there is compelling reason to believe that certain realized values occur with too large or too small a frequency than is consistent with a simpler model. For example, one might observe many more zeros than is consistent with, for example, the Poisson distribution. A two-part model that permits the zeros and non-zeros to be generated by different densities adds flexibility. Indeed it is a specific type of mixture model.

There are two approaches to such generalization. The two-part model, given in this section, specifies a model for the censoring mechanism and a model for the outcome *conditional* on the outcome being observed. The sample selection model, presented in the subsequent section, instead specifies a joint distribution for the censoring mechanism and outcome, and then finds the implied distribution conditional on the outcome observed. These approaches are contrasted in Section 16.5.7.

16.4.1. Two-Part Model

Let an individual with fully observed outcome be called a **participant** in the activity being studied. Define a binary indicator variable $d = 1$ for participants and $d = 0$ for nonparticipants. Suppose that $y > 0$ is observed for participants and $y = 0$ is observed for nonparticipants. For nonparticipants we observe only $\Pr[d = 0]$. For participants the *conditional density* of y given $y > 0$ is specified to be $f(y|d = 1)$, for some choice of density $f(\cdot)$. The **two-part model** for y is then given by

$$f(y|\mathbf{x}) = \begin{cases} \Pr[d = 0|\mathbf{x}] & \text{if } y = 0, \\ \Pr[d = 1|\mathbf{x}] f(y|d = 1, \mathbf{x}) & \text{if } y > 0. \end{cases} \tag{16.27}$$

This model was presented in detail by Cragg (1971) as a generalization of the Tobit model, which can be presented as a special case of (16.27). An obvious model for the participation decision d is a probit or logit model. A latent variable formulation is that $d = 1$ if $I = \mathbf{x}'\beta + \varepsilon$ exceeds zero, and the model is then viewed as a **hurdle model** since crossing a hurdle or threshold leads to participation. To ensure positive values for the participants, the density $f(y|d = 1, \mathbf{x})$ should be that for a positive-valued random variable, such as the log-normal, or an appropriate density such as the normal truncated from below at zero.

For simplicity the same regressors usually appear in both parts of the model, but this can be relaxed and should be if there are obvious exclusion restrictions. Maximum likelihood estimation is straightforward as it separates into estimation of a discrete choice model using all observations and estimation of the parameters of the density $f(y|d = 1, \mathbf{x})$ using only observations with $y > 0$.

16.4.2. Two-Part Model Examples

Duan et al. (1983) present a leading application of this model to forecasting medical expenses using data from the Rand Health Insurance Experiment. They specified a probit model for whether or not any medical expenses were incurred during the year, so $\Pr[d = 1|\mathbf{x}] = \Phi\left(\mathbf{x}_1'\beta_1\right)$, and a log-normal model for medical expenses given that some expenses were incurred, so $\ln y|d = 1, \mathbf{x} \sim \mathcal{N}[\mathbf{x}_2'\beta_2, \sigma_2^2]$. Then expected

medical expenses over the entire population are given by

$$\mathrm{E}[y|\mathbf{x}] = \Phi\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right)\exp[\sigma_2^2/2 + \mathbf{x}_2'\boldsymbol{\beta}_2], \tag{16.28}$$

where the second term uses the result that if $\ln y \sim \mathcal{N}[\mu, \sigma^2]$ then $\mathrm{E}[y] = \exp(\mu + \sigma^2/2)$. Mullahy (1998) considers such retransformation in further detail.

Two-part models are especially popular for modeling count data. For example, in modeling the number of doctor visits there is one model to determine whether or not a patient visits a physician at all and a second model to determine the consequent number of visits for those with at least one visit. Then $\Pr[d = 1]$ is specified to be the probability that a Poisson or negative binomial variable exceeds zero, whereas the density $f(y|d = 1)$ is specified to be a Poisson or negative binomial density truncated from below at zero. This model, due to Mullahy (1986), is called a hurdle model in the count literature and is detailed in Section 20.4.5.

For continuous data two-part models are used for expenditure models with excess zeros (Cragg's original motivation). An alternative, a sample selection model, is presented next.

16.5. Sample Selection Models

Sample selection can arise in many setttings and so there are many sample selection models. This section begins with a general discussion of sample selection before focusing on a leading example, the **bivariate sample selection model** studied by Heckman (1979). Another leading example, the **Roy model**, is treated separately in Section 16.7.

16.5.1. Sample Selection Models

Observational studies are rarely based on pure random samples. Most often exogenous sampling is used (see Section 3.2.4) and the usual estimators can be applied. If instead a sample, intentionally or unintentionally, is based in part on values taken by a dependent variable, parameter estimates may be inconsistent unless corrective measures are taken. Such samples can be broadly defined as **selected samples**.

There are many **selection models**, since there are many ways that a selected sample may be generated. Indeed it is very easy to be unaware that a selected sample is being used. For example, consider interpretation of average scores over time on an achievement test such as the Scholastic Aptitude Test, when test taking is voluntary. A decline over time may be due to real deterioration in student knowledge. However, it may just reflect the selection effect that relatively more students have been taking the test over time and the new test takers are the relatively weaker students.

Selection may be due to **self-selection**, with the outcome of interest determined in part by individual choice of whether or not to participate in the activity of interest. It can also result from **sample selection**, with those who participate in the activity of interest deliberately oversampled – an extreme case being sampling only participants. In either case, similar issues arise and selection models are usually called sample selection models.

This chapter presents only three of the many selection models in the literature. The simplest model is the Tobit model already presented in Section 16.3. A prototypical commonly used model that we call the bivariate sample selection model is presented in the remainder of this section. This model generalizes the Tobit model by introducing a censoring latent variable that differs from the latent variable generating the outcome of interest. Another popular model called the Roy model is presented in Section 16.7. This model considers an outcome that takes one of two values depending on the value taken by a censoring random variable. These models correspond to, respectively, the Tobit model types 1, 2, and 5 in the terminology of Amemiya (1985, p. 384).

Consistent estimation in the presence of sample selection on unobservables relies on relatively strong distributional assumptions, even in the case of semiparametric estimation. Experimental data studies provide an attractive alternative as selection problems can then be avoided by random assignment. However, experiments can be difficult to implement in economics applications for cost and ethical reasons. The treatment effects approach, detailed in Chapter 25, seeks to apply the experimental approach to observational data.

16.5.2. A Bivariate Sample Selection Model (Type 2 Tobit)

Let y_2^* denote the outcome of interest. In the standard truncated Tobit model this outcome is observed if $y_2^* > 0$. A more general model introduces a different latent variable, y_1^*, and the outcome y_2^* is observed if $y_1^* > 0$. For example, y_1^* determines whether or not to work and y_2^* determines how much to work, and $y_1^* \neq y_2^*$ since there are fixed costs to work such as commuting costs that are more important in determining participation than hours of work once working.

The **bivariate sample selection model** comprises a **participation equation** that

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0, \\ 0 & \text{if } y_1^* \leq 0 \end{cases} \tag{16.29}$$

and a resultant **outcome equation** that

$$y_2 = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ - & \text{if } y_1^* \leq 0. \end{cases} \tag{16.30}$$

This model specifies that y_2 is observed when $y_1^* > 0$, whereas y_2 need not take on any meaningful value when $y_1^* \leq 0$. The standard model specifies a linear model with additive errors for the latent variables, so

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \end{aligned} \tag{16.31}$$

with problems arising in estimating $\boldsymbol{\beta}_2$ if ε_1 and ε_2 are correlated. The Tobit model is clearly the special case where $y_1^* = y_2^*$.

There is no generally accepted name for this model. Heckman (1979) used it to illustrate estimation given sample selection. The model is equivalent to a **Tobit model with stochastic threshold** (Nelson, 1977). Suppose we observe y_2^* if $y_2^* > L^*$, where y_2^* is defined as in (16.31) and the threshold is $L^* = \mathbf{z}'\boldsymbol{\gamma} + v$ rather than $L^* = 0$ in

Section 16.3. Then, equivalently, we observe y_2^* if $y_1^* > 0$, where $y_1^* = y_2^* - L^* = (\mathbf{x}_2'\boldsymbol{\beta}_2 - \mathbf{z}'\boldsymbol{\gamma}) + (\varepsilon_2 - v) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1$ and where $\mathbf{x}_1$ denotes the union of $\mathbf{x}_2$ and $\mathbf{z}$, and $\boldsymbol{\beta}_1$ and ε_1 are defined in an obvious manner. Amemiya (1985, p. 384) calls the model a **type 2 Tobit model**. Wooldridge (2002, p. 506) calls the model one with a **probit selection equation**. Others call this model the generalized Tobit model or the sample selection model, though there are many such models.

Estimation by ML is straightforward given the additional assumption that the correlated errors are joint normally distributed and homoskedastic, with

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right]. \tag{16.32}$$

As for the probit model in Section 14.4.1, the normalization $\sigma_1^2 = 1$ is used since only the sign of y_1^* is observed.

Given (16.29) and (16.30), for $y_1^* > 0$ we observe y_2^*, with probability equal to the probability that $y_1^* > 0$ times the conditional probability of y_2^* given that $y_1^* > 0$. Thus for positive y_2 the density of the observables is $f^*(y_2^*|y_1^* > 0) \times \Pr[y_1^* > 0]$. For $y_1^* \leq 0$ all that is observed is that this event has occurred, and the density is the probability of this event occurring. The bivariate sample selection model therefore has likelihood function

$$\mathrm{L} = \prod_{i=1}^{n} \left\{\Pr[y_{1i}^* \leq 0]\right\}^{1-y_{1i}} \left\{f(y_{2i} \mid y_{1i}^* > 0) \times \Pr[y_{1i}{}^* > 0]\right\}^{y_{1i}}, \tag{16.33}$$

where the first term is the discrete contribution when $y_{1i}^* \leq 0$, since then $y_{1i} = 0$, and the second term is the continuous contribution when $y_{1i}^* > 0$. This likelihood function is applicable to quite general models, not just linear models with joint normal errors.

Specializing to linear models with joint normal errors gives a bivariate density $f^*(y_1^*, y_2^*)$ that is normal, leading to a conditional density in the second term that is univariate normal and easily handled. Amemiya (1985, pp. 385–387) provides details, including the exact form of the likelihood function.

The classic early application of this model was to labor supply, where y_1^* is the unobserved desire or propensity to work, whereas y_2 is actual hours worked. The model is also conceptually more appealing for labor supply than the Tobit model in Section 14.2.1 which required the artifice of "desired" hours of work. This prototypical application does have the complication that data on a key regressor, the offered wage, is missing for those individuals who do not work. This complication is handled by adding an equation for the offered wage and substituting this in, though the model is then strictly speaking not just a bivariate sample selection model. See Mroz (1987) for an excellent application to labor supply.

16.5.3. Conditional Means in the Bivariate Sample Selection Model

In this section we obtain the conditional truncated mean in the bivariate sample selection model. It differs from $\mathbf{x}_2'\boldsymbol{\beta}_2$, so that OLS regression of y_2 on $\mathbf{x}_2$ leads to inconsistent parameter estimates. Nonetheless, the expression for the conditional mean can be

used to motivate an alternative estimation procedure given in the subsequent section that relies on weaker distributional assumptions than those of the MLE.

We consider the truncated mean in the sample selectivity model where only positive values of y_2 are used. In general this is

$$\begin{aligned} \mathrm{E}[y_2|\mathbf{x}, y_1^* > 0] &= \mathrm{E}[\mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 | \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 > 0] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \mathrm{E}[\varepsilon_2 | \varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1], \end{aligned} \tag{16.34}$$

where $\mathbf{x}$ denotes the union of $\mathbf{x}_1$ and $\mathbf{x}_2$. If the errors ε_1 and ε_2 are independent then the last term simplifies to $\mathrm{E}[\varepsilon_2] = 0$, and OLS regression of y_2 on $\mathbf{x}_2$ will give a consistent estimate of $\boldsymbol{\beta}_2$. However, any correlation between the two errors means that the truncated mean is no longer $\mathbf{x}_2'\boldsymbol{\beta}_2$ and we need to account for selection.

To obtain $\mathrm{E}[\varepsilon_2|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]$ when ε_1 and ε_2 are correlated, Heckman (1979) noted that if the errors $(\varepsilon_1, \varepsilon_2)$ in (16.31) are joint normal as in (16.32) then Equation (16.36) in the following implies that

$$\varepsilon_2 = \sigma_{12}\varepsilon_1 + \xi, \tag{16.35}$$

where the random variable ξ is independent of ε_1. To obtain this result, note that in general the joint normal distribution

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right],$$

implies the conditional normal distribution

$$\mathbf{z}_2|\mathbf{z}_1 \sim \mathcal{N}\left[\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right],$$

a result that implies that

$$\mathbf{z}_2 = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\xi}, \tag{16.36}$$

where $\boldsymbol{\xi} \sim \mathcal{N}[\mathbf{0}, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}]$ is independent of $\mathbf{z}_1$. For the joint density given in (16.32) we have scalars and $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = 1$, so (16.36) specializes to (16.35).

By using (16.35), the truncated mean (16.34) becomes

$$\begin{aligned} \mathrm{E}[y_2|\mathbf{x}, y_1^* > 0] &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \mathrm{E}\left[(\sigma_{12}\varepsilon_1 + \xi)\,|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1\right] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\mathrm{E}[\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1], \end{aligned}$$

where we use independence of ξ and ε_1. The selection term is similar to that in the simpler Tobit model and again using the expression for $\mathrm{E}[z|z > -c]$ in Proposition 16.1 we obtain

$$\mathrm{E}[y_2|\mathbf{x}, y_1^* > 0] = \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right), \tag{16.37}$$

where $\lambda(z) = \phi(z)/\Phi(z)$ and we have used $\sigma_1^2 = 1$. Similarly, Proposition 16.1(iii) yields the truncated variance

$$\mathrm{V}[y_2|\mathbf{x}, y_1^* > 0] = \sigma_2^2 - \sigma_{12}^2\lambda\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right)(\mathbf{x}_1'\boldsymbol{\beta}_1 + \lambda\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right)). \tag{16.38}$$

The preceding analysis specifies no value for y_2 when $y_1^* \leq 0$. In some applications y_2 may equal zero when $y_1^* < 0$. Then it is meaningful to consider the censored mean.

Conditioning the observable y_2 on the unobservables y_1^* and y_2^* and then unconditioning yields

$$\begin{aligned} \mathrm{E}[y_2|\mathbf{x}] &= \mathrm{E}_{y_1^*}[\mathrm{E}[y_2|\mathbf{x}, y_1^*]] \\ &= \Pr[y_1^* \leq 0|\mathbf{x}] \times 0 + \Pr[y_1^* > 0|\mathbf{x}] \times \mathrm{E}[y_2^*|\mathbf{x}, y_2^* > 0] \\ &= 0 + \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)\left\{\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right)\right\} \\ &= \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\phi\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right), \end{aligned} \tag{16.39}$$

where the third line uses (16.37) and the last line uses $\lambda(z) = \phi(z)/\Phi(z)$. The censored variance can be shown to be heteroskedastic.

16.5.4. Heckman Two-Step Estimator

An important result is that OLS regression of y_2 on $\mathbf{x}_2$ alone using just the observed positive values of y_2 leads to inconsistent estimation of $\boldsymbol{\beta}$ unless the errors are uncorrelated so that $\sigma_{12} = 0$. This is clear from the truncated mean formula (16.37), which additionally includes the "regressor" $\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)$.

Heckman's two-step procedure, sometimes called the **Heckit estimator**, augments the OLS regression by an estimate of the omitted regressor $\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)$. Thus using positive values of y_2 estimate by OLS the model

$$y_{2i} = \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \sigma_{12}\lambda(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1) + v_i, \tag{16.40}$$

where v is an error term, $\widehat{\boldsymbol{\beta}}_1$ is obtained by first-step probit regression of y_1 on $\mathbf{x}_1$ since $\Pr[y_1^* > 0] = \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)$, and $\lambda(\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1) = \phi(\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1)/\Phi(\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1)$ is the estimated inverse Mills ratio. This regression does not directly provide an estimate of σ_2^2, but the truncated variance formula (16.38) leads to estimate $\widehat{\sigma}_2^2 = N^{-1}\sum_i[\widehat{v}_i^2 + \widehat{\sigma}_{12}^2\widehat{\lambda}_i(\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1 + \widehat{\lambda}_i)]$, where $\widehat{v}_i$ is the OLS residual from (16.40) and $\widehat{\lambda}_i = \lambda(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)$. The correlation between the two errors in (16.32) can then be estimated by $\widehat{\rho} = \widehat{\sigma}_{12}/\widehat{\sigma}_2$.

A test of whether or not $\sigma_{12} = 0$ or $\rho = 0$ is a test of whether or not the errors are correlated and sample selection correction is needed. One such test is a Wald test based on $\widehat{\sigma}_{12}$, the estimated coefficient of the inverse Mills ratio.

It is important to note that both the usual OLS standard errors and heteroskedasticity-robust standard errors reported from the regression (16.40) are incorrect. Correct formulas for the standard errors take account of two complications in the second-stage regression. First, even if $\boldsymbol{\beta}_1$ were known, the error in (16.40) is heteroskedastic from (16.38). Second, in fact $\boldsymbol{\beta}_1$ is replaced by an estimate, a complication studied in Section 6.6 and analyzed in Section 16.10.2 for the simpler Tobit model. Formulas for the correct standard errors are given in Heckman (1979); see also Greene (1981). Section 16.10.2 derives these formulas for the simpler Tobit model. Implementation is not simple so it is best to use a package that automatically handles this complication or to use the bootstrap.

The resulting estimator of $\boldsymbol{\beta}_2$ is consistent. Despite an efficiency loss compared to the MLE under joint normality of the errors that can be quite large, the estimator is very popular for the following reasons: (1) It is simple to implement; (2) the approach is applicable to a range of selection models including those given in Section 16.7; (3) the estimator requires distributional assumptions weaker than joint normality of ε_1

and ε_2; and (4) these distributional assumptions can be weakened even further to permit semiparametric estimation as in Section 16.9.

The **key assumption** needed is (16.35), essentially that

$$\varepsilon_2 = \delta\varepsilon_1 + \xi, \tag{16.41}$$

where ξ is independent of ε_1. This seems to be a quite sensible model. In the case of expenditures on a durable good, say, this says that the error in the expenditure equation is a multiple of the error in the purchase decision equation, plus some noise that is independent of the purchase decision; essentially a linear regression model for the errors. Given assumption (16.41) the conditional mean (16.34) becomes

$$\mathrm{E}[y_2|y_1^* > 0] = \mathbf{x}_2'\boldsymbol{\beta}_2 + \delta\mathrm{E}[\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]. \tag{16.42}$$

If ε_1 is standard normal distributed this leads to (16.37), the basis for the OLS regression (16.40).

More generally, Heckman's two-step method can be applied to (16.42) with distributions for ε_1 other than normal; see, for example, Olsen (1980). One can also use semiparametric methods that do not impose a functional form for $\mathrm{E}[\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]$ (see Section 16.9).

16.5.5. Identification Considerations

The bivariate sample selection model with normal errors is theoretically identified without any restriction on the regressors. In particular, exactly the same regressors can appear in the equations for y_1^* and y_2^*.

The model with normally distributed errors is close to unidentified, however, if exactly the same regressors are used. If $\mathbf{x}_1 = \mathbf{x}_2$ then $\mathrm{E}[y_2|y_1^* > 0] \simeq \mathbf{x}_2'\boldsymbol{\beta}_2 + a + b\mathbf{x}_2'\boldsymbol{\beta}_1$, using (16.37) and the observation from Section 16.3.2 that the inverse Mills ratio term $\lambda(\cdot)$ is approximately linear over a wide range of its argument. This leads to obvious multicollinearity problems, discussed in many articles including those by Nawata (1993), Nawata and Nagase (1996), and Leung and Yu (1996). Multicollinearity can be detected using the condition number given in Section 10.4.2, where from (16.40) the regressors are $\mathbf{x}_2$ and $\lambda(\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1)$. The problem is less severe the greater the variation in $\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1$ across observations, that is, the better a probit model can discriminate between participants and nonparticipants.

Semiparametric variants of the Heckman two-step method (see Section 16.9.3) do require an exclusion restriction. So **identification** in the bivariate sample selection model with normal errors is being achieved by functional form assumptions.

For practical purposes therefore, estimation of the bivariate sample selection model may require that at least one regressor in the participation equation (y_1^*) be excluded from the outcome equation (y_2^*). For example, fixed costs of working unrelated to hours worked will affect the decision to work but not hours worked. This can be a great limitation as in many applications, such as that in Section 16.6, it can be very difficult to make defensible exclusion restrictions.

16.5.6. Marginal Effects

The marginal effects in the bivariate sample selection model vary according to whether we consider the latent variable mean or the truncated mean given in (16.37) or the censored mean (if it is appropriate).

It is convenient to define $\mathbf{x}$ to be the vector formed by union of $\mathbf{x}_1$ and $\mathbf{x}_2$ and rewrite $\mathbf{x}_1'\boldsymbol{\beta}_1$ as $\mathbf{x}'\boldsymbol{\gamma}_1$ and $\mathbf{x}_2'\boldsymbol{\beta}_2$ as $\mathbf{x}'\boldsymbol{\gamma}_2$. For example, the truncated mean becomes $\mathrm{E}[y_2|\mathbf{x}] = \mathbf{x}'\boldsymbol{\gamma}_2 + \sigma_{12}\lambda(\mathbf{x}'\boldsymbol{\gamma}_1)$. Note that $\boldsymbol{\gamma}_1$ and/or $\boldsymbol{\gamma}_2$ will have some zero entries if $\mathbf{x}_1 \neq \mathbf{x}_2$. Differentiating with respect to $\mathbf{x}$ yields the **marginal effects**

$$\begin{aligned}
&\text{uncensored:} && \partial \mathrm{E}[y_2^*|\mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\gamma}_2, \\
&\text{truncated (at 0):} && \partial \mathrm{E}[y_2|\mathbf{x}, y_1 = 1]/\partial \mathbf{x} = \boldsymbol{\gamma}_2 - \sigma_{12}\lambda(\mathbf{x}'\boldsymbol{\gamma}_1)(\mathbf{x}'\boldsymbol{\gamma}_1 + \lambda(\mathbf{x}'\boldsymbol{\gamma}_1)) \\
&\text{censored (at 0):} && \partial \mathrm{E}[y_2|\mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\gamma}_1\phi(\mathbf{x}'\boldsymbol{\gamma}_1)\mathbf{x}'\boldsymbol{\gamma}_2 + \Phi(\mathbf{x}'\boldsymbol{\gamma}_1)\boldsymbol{\gamma}_2 \\
&&& \quad - \sigma_{12}\mathbf{x}'\boldsymbol{\gamma}_1\phi(\mathbf{x}'\boldsymbol{\gamma}_1)\boldsymbol{\gamma}_1,
\end{aligned} \tag{16.43}$$

where $\lambda(z) = \phi(z)/\Phi(z)$, and we use $\partial\phi(z)/\partial z = -z\phi(z)$ and $\partial\lambda(z)/\partial z = -z\phi(z)/\Phi(z) - \phi(z)^2/\Phi(z)^2 = -\lambda(z)(z + \lambda(z))$. Interpretation of these three derivatives is similar to that discussed in some detail in Section 16.3.5. As already noted, analysis of the censored mean is appropriate only if y_2 takes the value of zero when $y_1 = 0$. In applications such as the log-normal health expenditures example discussed later there is no censored mean.

16.5.7. Selection on Observables and on Unobservables

There are many modeling situations that can be considered a two-part decision problem of first engaging in an activity and then determining the level of the activity. These decisions are intertwined and can be expected to depend on common factors. The natural model for such data is the bivariate selection model (16.29)–(16.31).

After inclusion of regressors any remaining error (ε_1 and ε_2) in the two processes may in some cases be uncorrelated. For example, for models of hospitalization it is possible that, after controlling for observed individual characteristics such as health status, there is no correlation between the error in the equation determining hospital admission and in the error in the equation determining length of hospital stay. In that case analysis is straightforward as selection is only based on observables since, for example, (16.37) simplifies when $\sigma_{12} = 0$. The two pieces can be modeled separately and the simpler two-part model of Section 16.4 can be used.

In other cases the errors may be correlated even after inclusion of the regressors. For example, in labor supply unobserved factors that make someone more likely to work may also make them more likely to work longer hours than would be predicted by the observable regressors. One can test whether there is such correlation between the errors. If there is correlation, then selection is on unobservables and the methods of this chapter come into play. Relatively strong distributional assumptions are needed, even with the Heckman two-step method.

The study by Duan et al. (1983) summarized in Section 16.4.2 was criticized for using the two-part model, which is more restrictive than the sample selection model. This led to considerable debate, with many of the relevant articles referenced in Leung

and Yu (1996), who emphasize the important role of potential correlation of the inverse Mills ratio term with the remaining regressors.

More generally, selection models such as the bivariate selection model permit **selection on both observables and unobservables**, as it permits selection on both observed regressors and unobserved errors. It is often more simply referred to as a model of **selection on unobservables**, with selection on observables implicit. This chapter emphasizes selection on unobservables.

If instead we have only **selection on observables**, analysis becomes much simpler. The two-part model of this chapter is an example. Chapter 25 on treatment evaluation emphasizes selection on observables (see the discussion in Section 25.3.3) and details methods such as propensity score matching.

16.6. Selection Example: Health Expenditures

For illustration we use data from the RAND Health Insurance Experiment (RHIE). The data extract comes from Deb and Trivedi (2002), who modeled the number of outpatient visits to a medical doctor and to all providers using count data models. Section 20.3 summarizes the data and Section 20.7 presents estimates of some standard count models.

Here instead we model annual health expenditures. The regressors are the same regressors as defined in detail in Table 20.4. They can be broken down into health insurance variables (LC, IDP, LPI, and FMDE), socioeconomic characteristics (LINC, LFAM, AGE, FEMALE, CHILD, FEMCHILD, BLACK, and EDUCDEC) and health status variables (PHYSLIM, NDISEASE, HLTHG, HLTHF, and HLTHP). The analysis in Chapter 20 uses four years of data whereas here we use only the second year of data, yielding 5,574 observations with summary statistics similar to but not exactly the same as those given in Table 20.4.

The dependent variable y is annual individual health expenditures. An econometric model needs to take account of two complications: (1) Health expenditures are zero for 23.2% of the sample and (2) the positive health expenditures are very right-skewed with a mean of $221 that is much larger than the median of $53. The logarithmic transformation eliminates this skewness, with a mean of 4.07 close to the median of 3.96 and the skewness statistic falls from 24.0 to 0.3. The kurtosis is 3.29, close to the normal value of 3.

We focus on modeling $\ln y$ for those with positive medical expenditures. Possible models include a two-part model, exposited for log medical expenditures in Section 16.4.2, and a bivariate sample selection model (see Section 16.5.2), where y_1 in (16.29) is an indicator for positive expenditures and y_2 in (16.30) is $\ln y$. Note that it is not meaningful to consider the value of y_2 when $y_1 = 0$ because $\ln 0$ is not defined. The two-part model is a special case of the bivariate sample selection model with $\sigma_{12} = 0$ in (16.32).

Table 16.1 presents results for the health insurance variables and health status regressors. Socioeconomic variables also included in the regression are omitted from the table for brevity.

Table 16.1. *Health Expenditure Data: Estimates from Two-Part and Selection Models*[a]

Model	Two-Part		Selection Two-Step		Selection MLE	
Equation	DMED	LNMED	DMED	LNMED	DMED	LNMED
LC	−0.119 (-4.41)	−0.016 (−0.52)	−0.119 (−4.41)	−0.028 (−0.70)	−0.107 (−4.03)	−0.076 (2.25)
IDP	−0.128 (−2.45)	−0.079 (−1.28)	−0.128 (−2.45)	−0.028 (−0.70)	−0.109 (−2.13)	−0.150 (−2.26)
LPI	0.028 (3.19)	0.003 (0.28)	0.028 (3.19)	0.005 (0.47)	0.029 (3.42)	0.015 (1.42)
FMDE	0.008 (0.47)	−0.031 (−1.69)	0.008 (0.47)	−0.030 (−1.62)	0.001 (0.05)	−0.024 (1.21)
PHYSLIM	0.273 (3.67)	0.262 (3.81)	0.273 (3.67)	0.281 (3.50)	0.285 (3.94)	0.355 (4.70)
NDISEASE	0.022 (6.25)	0.020 (5.78)	0.022 (6.25)	0.022 (4.29)	0.021 (6.03)	0.029 (7.54)
HLTHG	0.039 (0.88)	0.144 (2.97)	0.039 (0.88)	0.147 (3.01)	0.058 (1.35)	0.156 (2.99)
HLTHF	0.192 (2.29)	0.364 (4.13)	0.192 (2.29)	0.382 (3.98)	0.224 (2.75)	0.445 (4.66)
HLTHP	0.640 (3.01)	0.787 (4.63)	0.640 (3.01)	0.833 (4.22)	0.798 (3.90)	0.999 (5.32)
ρ		0.000		0.168		0.736
σ_2				1.401		1.570
$\sigma_{12} = \rho\sigma_2$		0.000		0.236 (0.47)		1.155 (16.43)
−ln L	10184.1				10170.1	

[a] The t-statistics are in parentheses. Regressors also include eight socioeconomic characteristics. DMED is an indicator for whether or not medical expenditures are positive and LNMED is the natural logarithm of expenditures if positive. The t-statistics for the second step of the two-step selection model are based on errors that correct for the first-step estimation used to obtain the fitted inverse Mills ratio term.

We first compare the two-part model estimates with the two-step estimates of the bivariate sample selection model. The DMED equation estimates are identical as they are obtained by probit regression of DMED on the same regressors. The LNMED equation estimates differ because for two-step sample selection the second-step OLS regression for LNMED additionally includes as a regressor the fitted value of the inverse Mills ratio term. This additional term is statistically insignificant ($t = 0.47$) and low in magnitude with implied $\widehat{\rho} = 0.168$ that is close to zero. As a result the two models lead to similar coefficient estimates in the LNMED equation.

As noted in Section 16.4.4 the two-step estimator can perform poorly if the inverse Mills ratio term is highly correlated with the other regressors. Here this does not appear to be the case as there is considerable range in the probit model predicted probabilities from 0.15 to 0.99 and the condition number (see Section 10.4.4) of the second-stage regressors at the second stage, although somewhat high, only doubles from 37 to 82 upon inclusion of the inverse Mills ratio. Although it is still preferable to have some exclusion restrictions, it is not clear in this application which regressors in the DMED equation might be reasonably excluded on a priori grounds from the LNMED equation.

The ML estimates of the bivariate sample selection model differ considerably from the previous estimates, in both DMED and LNMED equations. The errors in the

latent variable models for DMED and LNMED are highly correlated with estimate $\widehat{\rho} = 0.736$ that is highly statistically significant ($t = 16.43$). The big difference between the two-step estimates and the ML estimates of σ_{12} (or of ρ) is best viewed as signifying a problem with the bivariate sample selection model. Rejection of the null hypothesis that the estimates have the same probability limit, a Hausman test given in Section 8.4, can be interpreted as rejection of the additional joint normality assumption needed to go from two-step estimation to ML estimation of the bivariate selection model. However, there may be a more fundamental problem that the bivariate sample selection model with the weaker assumption (16.41) and ε_1 iid normal is also not reasonable. Such fragility of the bivariate sample selection model is not unusual, especially if the same regressors are being used in both parts of the model so that identification is being secured through model specification assumptions. It is compounded here by use of health expenditure data, which can have quite large outliers so that errors may not be normal. Even though LNMED has skewness close to 0 and kurtosis close to 3, as already noted, standard tests of heteroskedasticity, skewness, and kurtosis resoundingly reject (with p-value 0.000) the null hypothesis that LNMED is normally distributed.

The regressor of most interest is LC, the natural logarithm of the coinsurance rate where the coinsurance rate equals the percentage of health cost borne by the insured paid by the patient. The most statistically significant effect is in determining whether or not expenditures are positive, rather than on the size of positive expenditures. If all observations were positive then the coefficient of LC in regression on LNMED equals the price elasticity of demand for health care. In fact in predicting the effect of changes in price on the conditional truncated mean of log expenditure we need to control for the effect of those with zero expenditure, as in the second line of (16.43).

In some applications interest lies in prediction rather than estimation of marginal effects. This is complicated in this example by a desire to predict the level rather than the log of expenditure. Assuming log-normality, the expression for the two-part model is given in (16.28). Duan et al. (1983) present a method to make predictions without the log-normality assumption that can be viewed as a variant of a bootstrap. See also Mullahy (1998).

16.7. Roy Model

In the bivariate sample selection model the dependent variable for an individual might not be observed. Thus we observe y_2 for an individual if $y_1 = 1$ but may not observe y_2 at all if $y_1 = 0$. In this section we consider a model in which y_2 is observed for all individuals, but in only one of the two possible states. This important model emphasizes **counterfactuals** and connects with the program evaluation literature presented in Chapter 25.

16.7.1. Roy Model

An often-cited article by Roy (1951) considered the consequences for the occupational distribution of earnings (both mean and variance) when there is individual

latent variable models for DMED and LNMED are highly correlated with estimate $\widehat{\rho} = 0.736$ that is highly statistically significant ($t = 16.43$). The big difference between the two-step estimates and the ML estimates of σ_{12} (or of ρ) is best viewed as signifying a problem with the bivariate sample selection model. Rejection of the null hypothesis that the estimates have the same probability limit, a Hausman test given in Section 8.4, can be interpreted as rejection of the additional joint normality assumption needed to go from two-step estimation to ML estimation of the bivariate selection model. However, there may be a more fundamental problem that the bivariate sample selection model with the weaker assumption (16.41) and ε_1 iid normal is also not reasonable. Such fragility of the bivariate sample selection model is not unusual, especially if the same regressors are being used in both parts of the model so that identification is being secured through model specification assumptions. It is compounded here by use of health expenditure data, which can have quite large outliers so that errors may not be normal. Even though LNMED has skewness close to 0 and kurtosis close to 3, as already noted, standard tests of heteroskedasticity, skewness, and kurtosis resoundingly reject (with p-value 0.000) the null hypothesis that LNMED is normally distributed.

The regressor of most interest is LC, the natural logarithm of the coinsurance rate where the coinsurance rate equals the percentage of health cost borne by the insured paid by the patient. The most statistically significant effect is in determining whether or not expenditures are positive, rather than on the size of positive expenditures. If all observations were positive then the coefficient of LC in regression on LNMED equals the price elasticity of demand for health care. In fact in predicting the effect of changes in price on the conditional truncated mean of log expenditure we need to control for the effect of those with zero expenditure, as in the second line of (16.43).

In some applications interest lies in prediction rather than estimation of marginal effects. This is complicated in this example by a desire to predict the level rather than the log of expenditure. Assuming log-normality, the expression for the two-part model is given in (16.28). Duan et al. (1983) present a method to make predictions without the log-normality assumption that can be viewed as a variant of a bootstrap. See also Mullahy (1998).

16.7. Roy Model

In the bivariate sample selection model the dependent variable for an individual might not be observed. Thus we observe y_2 for an individual if $y_1 = 1$ but may not observe y_2 at all if $y_1 = 0$. In this section we consider a model in which y_2 is observed for all individuals, but in only one of the two possible states. This important model emphasizes **counterfactuals** and connects with the program evaluation literature presented in Chapter 25.

16.7.1. Roy Model

An often-cited article by Roy (1951) considered the consequences for the occupational distribution of earnings (both mean and variance) when there is individual

heterogeneity in skills and individuals self-select into occupations. The treatment was relatively general and nonmathematical, though it did assume that individual worker output in an occupation is log-normally distributed in the absence of selection, and it did not consider at all estimation of a formal model. During the 1970s a number of authors independently proposed models for similar situations that were estimable with cross-section data and considered selection on both observables and unobservables. Such models have become known as Roy models.

We define the prototypical **Roy model** as follows. A latent variable y_1^* determines whether the outcome observed is y_2^* or y_3^*. Specifically, we observe whether y_1^* is positive or negative,

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0, \\ 0 & \text{if } y_1^* \leq 0, \end{cases} \tag{16.44}$$

and observe exactly one of y_2^* and y_3^* according to

$$y = \begin{cases} y_2^* & \text{if } y_1^* > 0, \\ y_3^* & \text{if } y_1^* \leq 0. \end{cases} \tag{16.45}$$

It is customary to specify a linear model with additive errors for the latent variables, with

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \\ y_3^* &= \mathbf{x}_3'\boldsymbol{\beta}_3 + \varepsilon_3. \end{aligned} \tag{16.46}$$

A model with additive effect is the specialization $\mathbf{x}_3'\boldsymbol{\beta}_3 = \mathbf{x}_2'\boldsymbol{\beta}_2 + \alpha$. The simplest parametric model for correlated errors is the joint normal, with

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \right], \tag{16.47}$$

where as usual the normalization $\sigma_1^2 = 1$ is used as only the sign of y_1^* is observed.

The log-likelihood function is similar to that for the bivariate sample selection model of Section 16.5, except that now y_3^* is observed if $y_1^* \leq 0$, so the term $\Pr[y_{1i}^* \leq 0]$ in (16.33) is replaced by $f(y_{3i} \mid y_{1i}^* \leq 0) \times \Pr[y_{1i}{}^* \leq 0]$.

It is more common to estimate the model using Heckman's two-step method applied to the truncated means,

$$\begin{aligned} \mathrm{E}[y|\mathbf{x}, y_1^* > 0] &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1), \\ \mathrm{E}[y|\mathbf{x}, y_1^* \leq 0] &= \mathbf{x}_3'\boldsymbol{\beta}_3 - \sigma_{13}\lambda(-\mathbf{x}_1'\boldsymbol{\beta}_1), \end{aligned} \tag{16.48}$$

where $\lambda(z) = \phi(z)/\Phi(z)$ and we have used $\sigma_1^2 = 1$. First-stage probit estimation of whether or not $y_1^* > 0$ yields an estimate of $\boldsymbol{\beta}_1$ and hence $\lambda(\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1)$. Two separate OLS regressions then lead to direct estimates of $(\boldsymbol{\beta}_2, \sigma_{12})$ and $(\boldsymbol{\beta}_3, \sigma_{13})$. Estimates of σ_2^2 and σ_3^2 can then be obtained using the squared residuals from the regressions, similar to the technique used for the bivariate sample selection model after (16.40). Maddala (1983, p. 225) provides complete details for this model, which he calls a **switching**

16.8. Structural Models

Regression models for selected samples have the feature that the outcome of interest depends in part on a participation decision that will in turn depend on expected outcomes. The participation decision and outcomes are simultaneous decisions. The preceding presentations simplified this interdependence by giving a **reduced-form** version of the participation equation. In particular, see the exposition of Lee (1978) in Section 16.7.2. This is a valid approach though is less efficient than working with a fully structural version.

In this section we explicitly model the interdependence using **structural** economic models based on utility maximization, and using structural statistical models that extend linear simultaneous equations to cover censoring and truncation, including binary outcomes.

16.8.1. Structural Models Based on Utility Maximization

Initial **structural model** research considered female **labor supply**. The textbook model has consumers maximizing utility, a function of goods consumption and leisure time, subject to a budget constraint and a time constraint that available discretionary time be allocated between leisure time and working time. At an interior solution the marginal rate of substitution (MRS) between leisure and goods consumption equals the wage rate. However, a corner solution where the woman chooses not to work can arise if the MRS exceeds the offered wage. Gronau (1973) and Heckman (1974) presented econometric models consistent with utility maximization that led to Tobit-like models, accounting for the additional complication that the offered wage is not observed for women who do not work. Subsequent advances include incorporation of fixed costs of work, leading to sample selection models, and use of panel data, leading to panel Tobit models. Killingsworth and Heckman (1986) and Blundell and MaCurdy (2001) provide surveys and Mroz (1987) provides an application.

To illustrate the structural approach we summarize the following example. Dubin and McFadden (1984) modelled household **consumption of energy** (electricity or natural gas) and **choice of appliances** (such as electric heater or natural gas heater) as being interrelated decisions coming from the **same utility function**. Specifically, it is assumed that for the jth of m appliance portfolios household **indirect utility** is given by

$$V_j = \{\alpha_{0j} + \alpha_1/\beta + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}'\gamma + \beta(y - r_j) + \eta\}e^{-\beta p_1} + \varepsilon_j, \qquad (16.49)$$

where p_1 and p_2 denote the prices of electricity and gas, y denotes income, and r_j denotes the annualized total life-cycle cost of portfolio j with

$$r_j = p_1 q_{1j} + p_2 q_{2j} + \rho c_j,$$

where q_{1j} and q_{2j} denote the typical electricity and gas consumption by household with appliance portfolio j, c_j is the cost of appliance portfolio j, and ρ is the discount rate. Tastes differ across households owing to observable characteristics $\mathbf{w}$, unobservable error η, and an appliance portfolio specific error ε_j, which is assumed to be

independent over j but correlated with η. In addition, there is a common appliance specific taste factor α_{0j}.

Electricity demand x_1 given appliance portfolio j equals $-(\partial V_j/\partial p_1)/(\partial V_j/\partial y)$, by **Roy's identity**, yielding

$$x_1 - q_{1j} = \alpha_{0j} + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}'\gamma + \beta(y - r_j) + \eta.$$

To emphasize that choice of appliance portfolio j is **endogenous**, introduce m mutually exclusive indicator variables δ_{jk}, $k = 1, \ldots, m$, where

$$\delta_{jk} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j. \end{cases}$$

Then electricity demand x_1 given appliance portfolio j is given by

$$x_1 - q_{1j} = \sum_{k=1}^{m} \alpha_{0k}\delta_{jk} + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}'\gamma + \beta\left(y - \sum_{k=1}^{m} r_j\delta_{jk}\right) + \eta. \qquad (16.50)$$

Even though the model (16.50) is linear, OLS regression yields inconsistent estimates as the result of endogeneity of δ_{jk}. Dubin and McFadden (1984) present two alternative estimation procedures.

An **IV approach** estimates (16.50) using $\widehat{p}_k$ and $r_j\widehat{p}_k$ as instruments for δ_{jk} and $r_j\delta_{jk}$, $k = 1, \ldots, m$, where $\widehat{p}_k$ are the predicted probabilities of choosing the various appliance portfolios. Here V_j is being used to denote the indirect utility function. It includes both deterministic and stochastic components of utility and corresponds to U_j in the Section 15.5.1 presentation of the ARUM. A similar approach yields

$$\begin{aligned} p_k &= \Pr[V_k > V_l,\ l \neq k,\ l = 1, \ldots, m] \\ &= \Pr[\varepsilon_l - \varepsilon_k < \{(\alpha_{0k} - \alpha_{0l}) - \beta(r_k - r_l)\}e^{-\beta p_1},\ \text{all } l \neq k] \\ &= \frac{\exp[(\alpha_{0k} - \beta r_k)e^{-\beta p_1}\pi/\lambda\sqrt{3}]}{\sum_{l=1}^{m}\exp[(\alpha_{0l} - \beta r_l)e^{-\beta p_1}\pi/\lambda\sqrt{3}]}, \end{aligned}$$

under the assumption that the ε_j, $j = 1, \ldots, m$, are iid type II extreme value with cdf $F(\varepsilon) = \exp(-\exp(-\gamma - \varepsilon\pi/\lambda\sqrt{3}))$, where $\gamma \simeq 0.5772$ is Euler's constant. Note that here ε_j has mean zero and variance $\lambda^2/2$ that differ from those for the parameterization of the type II extreme value distribution used in Chapters 14 and 15. Estimation of a nonlinear multinomial logit model gives predicted probabilities $\widehat{p}_k$.

An alternative **sample selection approach** notes that E[η|portfolio j chosen] $\neq 0$ and uses assumptions on the distribution of η and $\varepsilon_1, \ldots, \varepsilon_m$ to obtain this expectation. Specifically, assume that $\eta|\varepsilon_1, \ldots, \varepsilon_m$ is iid with mean $(\sqrt{2}\sigma/\lambda)\sum_{k=1}^{m} R_k\varepsilon_k$ and variance $\sigma^2(1 - \sum_{k=1}^{m} R_k^2)$, where $\sum_{k=1}^{m} R_k = 0$ and $\sum_{k=1}^{m} R_k^2 < 1$ and the distribution of ε_k has already been given. Then performing some algebra given in Dubin and McFadden yields

$$\text{E}[\eta|\text{portfolio } j \text{ chosen}] = \sum_{k\neq j}^{m}(\sigma\sqrt{6}R_k/\pi)\left[\frac{p_k \ln p_k}{1 - p_k} + \ln p_k\right].$$

A Heckman two-step procedure then estimates by OLS

$$x_1 - q_{1j} = \sum_{k=1}^{m} \alpha_{0k}\delta_{jk} + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}'\gamma + \beta\left(y - \sum_{k=1}^{m} r_j \delta_{jk}\right) + \sum_{k\neq j}^{m} \gamma_k \left[\frac{\widehat{p}_k \ln \widehat{p}_k}{1-\widehat{p}_k} + \ln \widehat{p}_k\right] + \xi,$$

where p_k are predicted probabilities from the preceding model for p_k, and ξ is an error with asymptotic mean zero.

Dubin and McFadden estimated these models using data on 3,249 households with two possible appliance portfolios: electric for water and space heating and gas for water and space heating.

Related examples include those of Hanemann (1984), who modeled the consumption level of a branded good where consumers consume only one of the possible branded goods in the choice set, and of Cameron et al. (1988), who modeled health service demand conditional on choice of one of a number of mutually exclusive health insurance policies.

Much creativity, evident in the Dubin and McFadden example, can be required to specify a model that yields analytical solutions for both choice probabilities and demand conditional on choice. The advances in computational methods detailed in Chapters 12 and 13 permit estimation of such models even when analytical solutions are not obtained. Nonetheless, results will still be dependent on the assumed utility function and distribution of unobservables.

16.8.2. Simultaneous Equations Tobit and Probit Models

To illustrate the issues involved in extending the linear SEM approach of Section 2.4 we consider a selection model that depends on two latent variables and introduce **simultaneity** into the models for the latent variables. A quite general model is

$$\begin{aligned} y_1^* &= \alpha_1 y_2^* + \gamma_1 y_1 + \delta_1 y_2 + \mathbf{x}_1'\beta_1 + \varepsilon_1, \\ y_2^* &= \alpha_2 y_1^* + \gamma_2 y_1 + \delta_2 y_2 + \mathbf{x}_2'\beta_2 + \varepsilon_2, \end{aligned} \tag{16.51}$$

where y_1^* and y_2^* are not completely observed but do determine the observed variables y_1 and y_2, and the errors are assumed to be joint normally distributed. For example, we may observe the binary indicator $y_1 = 1$ if $y_1^* > 0$ and observe $y_2 = y_2^*$ if $y_1^* > 0$. Note that in principal either latent variables or observed outcomes or both may appear as regressors, though identification requires restrictions such as those given in the following.

Endogenous Latent Variables

It is simplest to permit only the latent variables to be regressors in (16.51). Then

$$\begin{aligned} y_1^* &= \alpha_1 y_2^* + \mathbf{x}_1'\beta_1 + \varepsilon_1, \\ y_2^* &= \alpha_2 y_1^* + \mathbf{x}_2'\beta_2 + \varepsilon_2. \end{aligned} \tag{16.52}$$

A Heckman two-step procedure then estimates by OLS

$$x_1 - q_{1j} = \sum_{k=1}^{m} \alpha_{0k}\delta_{jk} + \alpha_1 p_1 + \alpha_2 p_2 + \mathbf{w}'\gamma + \beta\left(y - \sum_{k=1}^{m} r_j \delta_{jk}\right) + \sum_{k\neq j}^{m} \gamma_k \left[\frac{\widehat{p}_k \ln \widehat{p}_k}{1-\widehat{p}_k} + \ln \widehat{p}_k\right] + \xi,$$

where p_k are predicted probabilities from the preceding model for p_k, and ξ is an error with asymptotic mean zero.

Dubin and McFadden estimated these models using data on 3,249 households with two possible appliance portfolios: electric for water and space heating and gas for water and space heating.

Related examples include those of Hanemann (1984), who modeled the consumption level of a branded good where consumers consume only one of the possible branded goods in the choice set, and of Cameron et al. (1988), who modeled health service demand conditional on choice of one of a number of mutually exclusive health insurance policies.

Much creativity, evident in the Dubin and McFadden example, can be required to specify a model that yields analytical solutions for both choice probabilities and demand conditional on choice. The advances in computational methods detailed in Chapters 12 and 13 permit estimation of such models even when analytical solutions are not obtained. Nonetheless, results will still be dependent on the assumed utility function and distribution of unobservables.

16.8.2. Simultaneous Equations Tobit and Probit Models

To illustrate the issues involved in extending the linear SEM approach of Section 2.4 we consider a selection model that depends on two latent variables and introduce **simultaneity** into the models for the latent variables. A quite general model is

$$\begin{aligned} y_1^* &= \alpha_1 y_2^* + \gamma_1 y_1 + \delta_1 y_2 + \mathbf{x}_1'\beta_1 + \varepsilon_1, \\ y_2^* &= \alpha_2 y_1^* + \gamma_2 y_1 + \delta_2 y_2 + \mathbf{x}_2'\beta_2 + \varepsilon_2, \end{aligned} \tag{16.51}$$

where y_1^* and y_2^* are not completely observed but do determine the observed variables y_1 and y_2, and the errors are assumed to be joint normally distributed. For example, we may observe the binary indicator $y_1 = 1$ if $y_1^* > 0$ and observe $y_2 = y_2^*$ if $y_1^* > 0$. Note that in principal either latent variables or observed outcomes or both may appear as regressors, though identification requires restrictions such as those given in the following.

Endogenous Latent Variables

It is simplest to permit only the latent variables to be regressors in (16.51). Then

$$\begin{aligned} y_1^* &= \alpha_1 y_2^* + \mathbf{x}_1'\beta_1 + \varepsilon_1, \\ y_2^* &= \alpha_2 y_1^* + \mathbf{x}_2'\beta_2 + \varepsilon_2. \end{aligned} \tag{16.52}$$

The bivariate sample selection model (16.31) is an example that additionally specifies $\alpha_2 = 0$ and directly specifies a reduced form rather than a structural form for the y_1^* equation. Model (16.52) is easily estimated because the reduced form for y_1^* and y_2^* can be obtained in exactly the same way as for regular linear simultaneous equations. This reduced form can then be estimated using methods such as probit or Tobit depending on the way that y_1 and y_2 are determined given y_1^* and y_2^*. The parameters of the structural model (16.52) can then be estimated by replacing the regressors y_2^* and y_1^* by the reduced-form predictions $\widehat{y}_2^*$ and $\widehat{y}_1^*$.

Models such as (16.52) are called **simultaneous equations Tobit models**. A **simultaneous equations probit model** arises if the observed dependent variables y_1 and y_2 are binary. Estimators are presented by Nelson and Olson (1978), Amemiya (1979), and Lee, Maddala, and Trost (1980) and a very general treatment for a range of models is given in L-F. Lee (1981). The standard errors of the estimators can be obtained using the results on sequential two-step m-estimators in Section 6.6. However, it is much simpler to obtain them using the bootstrap pairs procedure presented in Section 11.2. Identification requires exclusion restrictions in (16.51) similar to those for linear simultaneous equations.

Endogenous Regressors

A common specialization of the model (16.52) is to a Tobit model with **endogenous regressor** that is **completely observed**. Then y_2^* is fully observed, so $y_2 = y_2^*$, whereas we observe $y_1 = y_1^*$ if $y_1^* > 0$ and $y_1 = 0$ otherwise. The model becomes

$$\begin{aligned} y_1^* &= \alpha_1 y_2 + \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2 &= \mathbf{x}'\boldsymbol{\pi} + v, \end{aligned} \tag{16.53}$$

where the first equation is the structural equation of interest and the second equation is the reduced form for the endogenous regressor y_2. Again note that here y_2 is continuous, not discrete. For joint normal errors $\varepsilon_1 = \gamma v + \xi$, where ξ is an independent normal error (see Section 5.1), so $y_1^* = \alpha_1 y_2 + \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma v + \xi$.

A two-step estimation procedure calculates predicted residuals $\widehat{v} = y_2 - \mathbf{x}'\widehat{\boldsymbol{\pi}}$ from OLS regression of y_2 on $\mathbf{x}$ and then obtains Tobit estimates from the model

$$y_1^* = \alpha_1 y_2 + \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma\widehat{v} + e_1,$$

where the error e_1 is normally distributed. A test for endogeneity of y_2 can be implemented as a Wald test of $\gamma = 0$ using the usual standard errors from a Tobit package. This test is an extension of the auxiliary regression to implement the Hausman endogeneity test in the linear model (see Section 8.4.3). If the null hypothesis is rejected then the aforementioned second-step Tobit regression yields consistent estimates of α_1 and $\boldsymbol{\beta}_1$, but standard errors then need to be adjusted because of first-step estimation of the additional regressor $\widehat{v}$. See Smith and Blundell (1986) for details for the Tobit model and Rivers and Vuong (1988) for a similar procedure that estimates a probit model at the second step.

Endogenous Censored or Binary Variables

Analysis is more complicated if the observed **censored or binary endogenous variables** y_1 or y_2 appear as regressors in (16.51). Heckman (1978) considered the following model:

$$\begin{aligned} y_1^* &= \gamma_1 y_1 + \delta_1 y_2^* + \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \\ y_2^* &= \alpha_2 y_1^* + \gamma_2 y_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \end{aligned} \tag{16.54}$$

where we observe $y_1 = 1$ if $y_1^* > 0$ and $y_1 = 0$ if $y_1^* \leq 0$, and we observe $y_2 = y_2^*$ all the time. The complication here is that y_1 appears as a regressor. A meaningful reduced form for y_1^* can depend only on $\mathbf{x}_1$ and $\mathbf{x}_2$ and not y_1. This imposes the restriction that $\delta_1\gamma_2 + \gamma_1 = 0$, an example of what is called a **coherency condition** in this literature. Then the reduced form of the model becomes

$$\begin{aligned} y_1^* &= \mathbf{x}'\boldsymbol{\pi}_1 + v_1, \\ y_2 &= \gamma_2 y_1 + \mathbf{x}'\boldsymbol{\pi}_2 + v_2. \end{aligned}$$

This is a special case of the Roy model where participation ($y_1 = 1$) leads to only an intercept shift (via γ_2) in the outcome. In general, models with regressors that include censored or truncated endogenous variables are difficult to estimate. See, for example, Blundell and Smith (1989).

Example

Brooks, Cameron, and Carter (1998) applied a simultaneous equations Tobit model to explain the vote by congressional representatives on a pro-sugar amendment. The three observed outcomes y_1, y_2, and y_3 were, respectively, the vote (yes or no) and contributions to their campaign funds from sugar interests and (opposing) sweetener-user interests. The first outcome is a binary outcome and the other two outcomes are censored at zero. A simultaneous equations model for the associated latent variables y_1^*, y_2^*, and y_3^* was specified, so the structural model is of the simpler form (16.52).

How reasonable is this specification? Here campaign contributions y_2^* and y_3^* should depend on the latent variable y_1^* since the actual vote y_1 was made at a later date. For y_1^* however, an alternative and more difficult model is that y_1^*, the latent variable for the vote, depends on actual contributions received (y_2 and y_3) rather than on the latent contributions. However, if this is viewed as a game likely to be repeated in the future, a case can be made for using y_2^* and y_3^*. Clearly, the reasonableness of such assumptions will vary with the application. Parameter identification was secured by exclusion restrictions on the exogenous regressors. Consistent estimation relies on errors being joint normally distributed.

16.9. Semiparametric Estimation

Censoring, truncation, and sample selection lead to a sample that differs from the population. This is essentially a missing data problem, one that is complicated because data are missing on the dependent variable(s) rather than on exogenous regressors.

The preceding methods solved this missing data problem by making distributional assumptions to obtain either a likelihood function for the sample data or an appropriate censored, truncated, or selected conditional mean.

These methods are fragile to even very minor misspecification of error distributions. For example, both the MLE and the Heckman two-step estimator in the standard Tobit model are inconsistent if errors are normal but heteroskedastic, or if they are homoskedastic but nonnormal. See, for example, Paarsch (1982) and the references therein.

Considerable efforts have been devoted to developing semiparametric estimators that are consistent under weaker distributional assumptions. Before presenting leading examples, however, we note that an alternative is to continue to take a fully parametric approach that is based on richer, more flexible distributional assumptions.

16.9.1. Flexible Parametric Models

For simplicity begin with the classical Tobit model $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$. The assumption that $\varepsilon_i \sim \mathcal{N}[0, \sigma_i^2]$ can be relaxed in two ways. First, heteroscedasticity can be incorporated through an explicit model $\sigma_i^2 = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$, where now both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ need to be estimated. Second, more flexible distributions than the normal distribution might be used. For example, one might use a squared polynomial expansion of the normal (see Section 9.7.7).

For the bivariate sample selection model a similar approach may be taken, where now a more flexible joint distribution for $(\varepsilon_1, \varepsilon_2)$ is used. Lee (1983) proposed working with transformations $(\varepsilon_1^*, \varepsilon_2^*)$ of $(\varepsilon_1, \varepsilon_2)$ for which the bivariate normality assumption may be more reasonable.

Bayesian methods can also be applied to such models. Chib (1992) considered the censored Tobit model. The latent variables $\mathbf{y}^*$ are introduced as auxiliary variables and the data augmentation approach (see Section 13.7) is used. The Gibbs sampler cycles among (1) the conditional posterior for $\boldsymbol{\beta}|\mathbf{y}, \mathbf{y}^*, \sigma^2$, (2) the conditional posterior for $\sigma^2|\mathbf{y}, \mathbf{y}^*, \boldsymbol{\beta}$, and (3) the posterior for $\mathbf{y}^*|\mathbf{y}, \boldsymbol{\beta}, \sigma^2$.

A **flexible parametric approach** is particularly advantageous for handling censoring, truncation, and sample selection in nonlinear models such as those for counts and for duration data or mixed types of data, as semiparametric methods are less likely to be available then.

16.9.2. Semiparametric Estimation for Censored Models

We now move on to semiparametric estimation. We consider a linear model for the latent variable $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, which is left-censored at zero so that we observe $y_i = y_i^*$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$. The semiparametric literature usually expresses the model as

$$y_i = \max(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, 0). \tag{16.55}$$

This is the Tobit model (16.11)–(16.13), except the distribution of ε is unspecified. With some adaptation this model also covers left-censoring at known fixed point other than zero and to right-censoring such as for top-coded data. For example, if

$y = \min(\mathbf{x}'\beta + \varepsilon, U)$ then $U - y = \max(U - \mathbf{x}'\beta - \varepsilon, 0)$. The goal is to consistently estimate β without specifying a complete parametric distribution for ε_i. The estimators are called **semiparametric** as the uncensored mean $\mathbf{x}_i'\beta$ is parameterized but the error distribution is not. The methods presented in the following differ in the assumptions made on the distribution of ε.

From (16.8) ML estimation is possible given knowledge of the cdf of y^* and hence of ε. The cdf of ε can be nonparametrically estimated using the Kaplan–Meier product limit estimator for the cdf presented in Chapter 17 for the case of right-censored duration data. Alternatively, the distribution of ε can be nonparametrically determined using the series expansion of Gallant and Nychka (1987); see Section 9.7.7. These semiparametric ML estimation methods are rarely implemented.

Instead, the literature focuses on estimation based on conditional moments. From (16.20) the conditional censored mean $\mathrm{E}[y|\mathbf{x}]$ is clearly a single-index model with $\mathrm{E}[y|\mathbf{x}] = g(\mathbf{x}'\beta)$, where the function $g(\cdot)$ is unknown if the distribution of ε is not specified. The single-index methods of Section 9.7.4 can therefore be applied, though as noted there β can be estimated only up to location and scale.

A more popular approach considers alternative conditional censored moments that are less altered by censoring. Powell (1984) proposed using the **conditional median**. The key distributional assumption is that $\varepsilon|\mathbf{x}$ has median zero, in which case the conditional median of $y|\mathbf{x}$ equals the conditional mean $\mathbf{x}'\beta$. The intuition for Powell's estimator is most easily obtained by supposing y is iid. If less than half the sample is censored, so that less than half of the observations are zero and more than half are positive, then the censored sample median provides a consistent estimate of the population median. Powell (1984) extended this idea to the regression case, where the same logic follows for those observations for which less than half the observations on $\varepsilon|\mathbf{x}$ are censored, where $\varepsilon = y - \mathbf{x}'\beta$ depends on β, which needs to be estimated. The regression analogue of median estimation is LAD estimation (see Section 4.6). This leads to the **censored least absolute deviations (CLAD) estimator** $\widehat{\beta}_{\text{CLAD}}$, which minimizes

$$Q_N(\beta) = N^{-1} \sum_{i=1}^{N} |y_i - \max(\mathbf{x}_i'\beta, 0)|. \tag{16.56}$$

The essential assumption for consistency of this estimator is that $\varepsilon|\mathbf{x}$ has median zero. Given this assumption the estimator is consistent even if errors are conditionally heteroskedastic. The estimator for β is $\sqrt{N}$-consistent and asymptotically normal. More efficient estimators can be obtained by weighting the terms in sums by $f(0|\mathbf{x}_i)$, the conditional density of $\varepsilon_i|\mathbf{x}_i$ evaluated at zero. The method can also be extended to conditional quantiles.

An alternative procedure uses a **symmetrically trimmed mean**, rather than the median, that is also unaffected by censoring. Assume that the distribution of $\varepsilon|\mathbf{x}$ is symmetrically distributed. This implies that for observations with positive mean (i.e., $\mathbf{x}'\beta > 0$) $y|\mathbf{x}$ is symmetrically distributed on the interval $(0, 2\mathbf{x}'\beta)$. Then either $\mathbf{x}'\beta + \varepsilon < 0$ and $y = 0$ is observed or, with equal probability, $\mathbf{x}'\beta + \varepsilon > 2\mathbf{x}'\beta$ and the data are artificially set to $2\mathbf{x}'\beta$ to preserve the symmetry about $\mathbf{x}'\beta$. We have shown that

$$\mathrm{E}[\mathbf{1}(\mathbf{x}'\beta > 0)[\min(y, 2\mathbf{x}'\beta) - \mathbf{x}'\beta]|\mathbf{x}] = \mathbf{0}, \tag{16.57}$$

where $\mathbf{1}(\mathbf{x}'\boldsymbol{\beta} > 0)$ restricts attention to observations with positive mean, and the new dependent variable is $y = 0$, or $0 < y < 2\mathbf{x}'\boldsymbol{\beta}$, or $2\mathbf{x}'\boldsymbol{\beta}$ if $y > 2\mathbf{x}'\boldsymbol{\beta}$. The moment estimator based on (16.57) does not have unique solution for $\boldsymbol{\beta}$. Powell (1986b) proposed the **symmetrically censored least squares (SCLS) estimator** that minimizes

$$Q_N(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^{N}\{[y_i - \max(y_i/2, \mathbf{x}_i'\boldsymbol{\beta})]^2 + \mathbf{1}(y_i > 2\mathbf{x}_i'\boldsymbol{\beta})[y_i^2/4 - \max(0, \mathbf{x}_i'\boldsymbol{\beta})]^2\}, \tag{16.58}$$

which with some algebra can be shown to yield first-order conditions that are the sample analogue of moment condition (16.57). Chay and Honoré (1998) provide a graphical exposition of the trimming for the SCLS estimator, as well as for the related pairwise difference estimators of Honoré and Powell (1994).

Melenberg and Van Soest (1996), Chay and Honoré (1998), and Chay and Powell (2001) provide applications of some of these estimators. Pagan and Ullah (1999) provide additional methods and theory.

As an empirical example we applied CLAD estimation to the Section 16.2.1 data that were generated from a Tobit model with normal errors. The slope parameter (set to 1000) was estimated to be 956 (standard error 117) using ML compared to 838 (standard error 165) using CLAD. As expected the CLAD robustness to nonormality comes at the expense of some loss in efficiency.

16.9.3. Semiparametric Estimation for Selection Models

Semiparametric estimation of sample selection models is more challenging. We consider the most commonly studied model, the **bivariate sample selection model** defined in Section 16.5.2, where now we relax the assumption that the errors $(\varepsilon_1, \varepsilon_2)$ are joint normally distributed.

Semiparametric ML estimation is possible. In particular Gallant and Nychka (1987) explicitly considered the bivariate sample selection model as a suitable candidate for their series expansion estimator presented in Section 9.7.7.

The literature instead uses as starting point the expression for the truncated conditional mean, which from (16.34) is given by

$$\begin{aligned}\mathrm{E}[y_{2i}|\mathbf{x}_i, y_{1i}^* > 0] &= \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \mathrm{E}[\varepsilon_2|\varepsilon_1 > -\mathbf{x}_{1i}'\boldsymbol{\beta}_1] \\ &= \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + g(\mathbf{x}_{1i}'\boldsymbol{\beta}_1),\end{aligned} \tag{16.59}$$

where the second equality assumes that $\varepsilon_{2i}|\mathbf{x}_i, \varepsilon_{1i}$ has distribution that depends on just $\mathbf{x}_{1i}$ similar to assumption (16.41). The distribution of $(\varepsilon_1, \varepsilon_2)$ is left unspecified so the function $g(\cdot)$ is unknown, leading to a semiparametric estimation problem. Since it is possible that $g(\mathbf{x}_1'\boldsymbol{\beta}_1) = \mathbf{x}_1'\boldsymbol{\beta}_1$, identification in this model with $g(\cdot)$ unspecified requires an **exclusion restriction** that at least one component of $\mathbf{x}_1$ does not appear in $\mathbf{x}_2$. Moreover, the more uncorrelated $\mathbf{x}_1'\boldsymbol{\beta}_1$ is with $\mathbf{x}_2$ the better $\boldsymbol{\beta}_2$ and $g(\cdot)$ can be distinguished. The model (16.59) is a partially linear model, which can be estimated using methods presented in Section 9.7.3. Popular methods include the Robinson (1988a) differencing estimator and using a series expansion for $g(\mathbf{x}_1'\boldsymbol{\beta}_1)$. Since $\boldsymbol{\beta}_1$ is unknown the regression is of y_{2i} on $\mathbf{x}_{2i}'\boldsymbol{\beta}_2 + g(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)$, where $\widehat{\boldsymbol{\beta}}_1$ can be obtained by regression

where $\mathbf{1}(\mathbf{x}'\boldsymbol{\beta} > 0)$ restricts attention to observations with positive mean, and the new dependent variable is $y = 0$, or $0 < y < 2\mathbf{x}'\boldsymbol{\beta}$, or $2\mathbf{x}'\boldsymbol{\beta}$ if $y > 2\mathbf{x}'\boldsymbol{\beta}$. The moment estimator based on (16.57) does not have unique solution for $\boldsymbol{\beta}$. Powell (1986b) proposed the **symmetrically censored least squares (SCLS) estimator** that minimizes

$$Q_N(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^{N}\{[y_i - \max(y_i/2, \mathbf{x}_i'\boldsymbol{\beta})]^2 + \mathbf{1}(y_i > 2\mathbf{x}_i'\boldsymbol{\beta})[y_i^2/4 - \max(0, \mathbf{x}_i'\boldsymbol{\beta})]^2\}, \tag{16.58}$$

which with some algebra can be shown to yield first-order conditions that are the sample analogue of moment condition (16.57). Chay and Honoré (1998) provide a graphical exposition of the trimming for the SCLS estimator, as well as for the related pairwise difference estimators of Honoré and Powell (1994).

Melenberg and Van Soest (1996), Chay and Honoré (1998), and Chay and Powell (2001) provide applications of some of these estimators. Pagan and Ullah (1999) provide additional methods and theory.

As an empirical example we applied CLAD estimation to the Section 16.2.1 data that were generated from a Tobit model with normal errors. The slope parameter (set to 1000) was estimated to be 956 (standard error 117) using ML compared to 838 (standard error 165) using CLAD. As expected the CLAD robustness to nonormality comes at the expense of some loss in efficiency.

16.9.3. Semiparametric Estimation for Selection Models

Semiparametric estimation of sample selection models is more challenging. We consider the most commonly studied model, the **bivariate sample selection model** defined in Section 16.5.2, where now we relax the assumption that the errors $(\varepsilon_1, \varepsilon_2)$ are joint normally distributed.

Semiparametric ML estimation is possible. In particular Gallant and Nychka (1987) explicitly considered the bivariate sample selection model as a suitable candidate for their series expansion estimator presented in Section 9.7.7.

The literature instead uses as starting point the expression for the truncated conditional mean, which from (16.34) is given by

$$\begin{aligned}\mathrm{E}[y_{2i}|\mathbf{x}_i, y_{1i}^* > 0] &= \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \mathrm{E}[\varepsilon_2|\varepsilon_1 > -\mathbf{x}_{1i}'\boldsymbol{\beta}_1] \\ &= \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + g(\mathbf{x}_{1i}'\boldsymbol{\beta}_1),\end{aligned} \tag{16.59}$$

where the second equality assumes that $\varepsilon_{2i}|\mathbf{x}_i, \varepsilon_{1i}$ has distribution that depends on just $\mathbf{x}_{1i}$ similar to assumption (16.41). The distribution of $(\varepsilon_1, \varepsilon_2)$ is left unspecified so the function $g(\cdot)$ is unknown, leading to a semiparametric estimation problem. Since it is possible that $g(\mathbf{x}_1'\boldsymbol{\beta}_1) = \mathbf{x}_1'\boldsymbol{\beta}_1$, identification in this model with $g(\cdot)$ unspecified requires an **exclusion restriction** that at least one component of $\mathbf{x}_1$ does not appear in $\mathbf{x}_2$. Moreover, the more uncorrelated $\mathbf{x}_1'\boldsymbol{\beta}_1$ is with $\mathbf{x}_2$ the better $\boldsymbol{\beta}_2$ and $g(\cdot)$ can be distinguished. The model (16.59) is a partially linear model, which can be estimated using methods presented in Section 9.7.3. Popular methods include the Robinson (1988a) differencing estimator and using a series expansion for $g(\mathbf{x}_1'\boldsymbol{\beta}_1)$. Since $\boldsymbol{\beta}_1$ is unknown the regression is of y_{2i} on $\mathbf{x}_{2i}'\boldsymbol{\beta}_2 + g(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)$, where $\widehat{\boldsymbol{\beta}}_1$ can be obtained by regression

of the binary outcome y_{1i} on $\mathbf{x}_{1i}$, using one of the semiparametric binary model estimators given in Section 14.7. These methods provide consistent estimates of the slope parameters β_2. To additionally estimate the intercept, necessary for analysis of the levels rather than changes in y_2, see Andrews and Schafgens (1998).

Newey, Powell, and Walker (1990) applied this approach to female labor supply. The participation indicator model was estimated using several different methods and the equation for the outcome y_2 was estimated using the method of Robinson (1988a). Melenberg and Van Soest (1996) modeled vacation expenditures using a wide range of semiparametric methods for both the bivariate sample selection and censored regression models. A richer model is provided by Das, Newey and Vella (2003).

Manski (1989) considered **identification** in the bivariate sample selection model under relatively minimal assumptions and provided **bounds** for the mean and for marginal effects, conditional on both regressors and selection.

16.10. Derivations for the Tobit Model

16.10.1. Truncated Moments of Standard Normal

Consider $z \sim \mathcal{N}[0, 1]$, with density $\phi(z) = (1/\sqrt{2\pi})\exp(-z^2/2)$ and cdf $\Phi(z)$. Since $\Pr[z > c] = 1 - \Phi(c)$, the conditional density of $z|z > c$ is $\phi(z)/(1 - \Phi(c))$. It follows that

$$\begin{aligned}
\mathrm{E}[z|z > c] &= \int_c^\infty z\,(\phi(z)/[1 - \Phi(c)])\,dz \\
&= \int_c^\infty z(1/\sqrt{2\pi})\exp(-z^2/2)\,dz \Big/ [1 - \Phi(c)] \\
&= \int_c^\infty \frac{\partial}{\partial z}\left(-(1/\sqrt{2\pi})\exp(-z^2/2)\right) dz \Big/ [1 - \Phi(c)] \\
&= \left[-(1/\sqrt{2\pi})\exp(-z^2/2)\right]_c^\infty \Big/ [1 - \Phi(c)] \\
&= \phi(c)/[1 - \Phi(c)].
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}[z^2|z > c] &= \int_c^\infty z^2\,(\phi(z)/[1 - \Phi(c)])\,dz \\
&= \int_c^\infty z \times z \times (1/\sqrt{2\pi})\exp(-z^2/2)\,dz \Big/ [1 - \Phi(c)] \\
&= \int_c^\infty z \times \frac{\partial}{\partial z}\left(-(1/\sqrt{2\pi})\exp(-z^2/2)\right) dz \Big/ [1 - \Phi(c)] \\
&= \left[z \times (-1/\sqrt{2\pi})\exp(-z^2/2)\right]_c^\infty \Big/ [1 - \Phi(c)] \\
&\quad - \int_c^\infty \frac{\partial}{\partial z}(z) \times \left(-(1/\sqrt{2\pi})\exp(-z^2/2)\right) dz \Big/ [1 - \Phi(c)] \\
&= c\phi(c)/[1 - \Phi(c)] + (1 - \Phi(c))/[1 - \Phi(c)] \\
&= c\phi(c)/[1 - \Phi(c)] + 1.
\end{aligned}$$

It follows after a little algebra that

$$\begin{aligned} \mathrm{V}[z|z > c] &= \mathrm{E}[z^2|z > c] - (\mathrm{E}[z|z > c])^2 \\ &= 1 + c\phi(c)/[1 - \Phi(c)] - \phi(c)^2/[1 - \Phi(c)]^2. \end{aligned}$$

16.10.2. Asymptotic Theory for Heckman's Two-Step Estimator in the Tobit Model

The asymptotic variance matrix of the two-step Heckman estimator is complicated by its dependence on first-step parameter estimates. There are several ways to obtain the asymptotic variance, such as that in Amemiya (1985, pp. 369–370). Here we instead apply the general result for sequential two-step m-estimators given in Section 6.6. We consider the simplest case of the Tobit model (see Section 16.3.6). The methods can be adapted to two-step estimators for the bivariate sample selection model (Section 16.5.4) and simultaneous equations Tobit model (Section 16.8.2). A much simpler quite different approach is to use the bootstrap pairs procedure (see Section 11.2).

From (16.26) we wish to estimate the parameters $\gamma = [\beta' \quad \sigma]'$ in the equation for positive y_i:

$$\begin{aligned} y_i &= \mathbf{x}_i'\beta + \sigma\lambda(\mathbf{x}_i'\alpha) + \eta_i \\ &= \mathbf{w}_i(\alpha)'\gamma + \eta_i, \end{aligned}$$

where $\mathbf{w}_i(\alpha) = [\mathbf{x}_i' \quad \lambda_i(\mathbf{x}_i'\alpha)]'$ and $\eta_i = y_i - \mathbf{x}_i'\beta - \sigma\lambda(\mathbf{x}_i'\alpha)$ is heteroskedastic with variance $\sigma_{\eta i}^2$ defined in (16.24). The first step of the two-step procedure is to obtain an estimate $\widehat{\alpha}$ of the unknown parameter α by probit MLE. It follows that the normal equations for the two parts of the Heckman two-step estimator are

$$\sum_{i=1}^{N}(y_i - \Phi(\mathbf{x}_i'\alpha))\frac{\phi^2(\mathbf{x}_i'\alpha)}{\Phi(\mathbf{x}_i'\alpha)(1 - \Phi(\mathbf{x}_i'\alpha))}\mathbf{x}_i = \mathbf{0}, \tag{16.60}$$

$$-\sum_{i=1}^{N} d_i\mathbf{w}_i(\alpha)(y_i - \mathbf{w}_i(\alpha)'\gamma) = \mathbf{0},$$

where the first equation gives the probit first-order conditions for α, and the second equation gives first-order conditions for γ for OLS on positive y_i ($d_i = 1$).

These equations can be combined as $\sum_{i=1}^{N}\mathbf{h}(\mathbf{x}_i, \theta) = \mathbf{0}$ where $\theta = (\alpha', \gamma')'$. By the usual first-order Taylor series expansion $\widehat{\gamma} - \gamma \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{G}_0^{-1}\mathbf{S}_0(\mathbf{G}_0^{-1})']]$ where $\mathbf{G}_0 = \lim N^{-1}\mathrm{E}[\sum_{i=1}^{N}\partial\mathbf{h}(\mathbf{x}_i, \theta)/\partial\theta]$ and $\mathbf{S}_0 = \lim N^{-1}\mathrm{E}[\sum_{i=1}^{N}\mathbf{h}(\mathbf{x}_i, \theta)\mathbf{h}(\mathbf{x}_i, \theta)]'$. We are interested in the subcomponent corresponding to γ. Simplification occurs because $\partial\mathbf{h}(\mathbf{x}_i, \theta)/\partial\theta$ is block triangular because γ does not appear in the first set of equations. Partitioning yields the general result

$$\mathrm{V}[\widehat{\theta}_2] = \mathbf{G}_{22}^{-1}\left\{\mathbf{S}_{22} + \mathbf{G}_{21}[\mathbf{G}_{11}^{-1}\mathbf{S}_{11}\mathbf{G}_{11}^{-1}]\mathbf{G}_{21}' - \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{S}_{12} - \mathbf{S}_{21}\mathbf{G}_{11}^{-1}\mathbf{G}_{21}'\right\}\mathbf{G}_{22}^{-1},$$

where the matrices are defined in Section 6.6.

Specializing to the problem here, we first consider the terms in $\mathbf{G}_0$. Then

$$\mathbf{G}_{11} = \lim \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \dfrac{\phi^2(\mathbf{x}_i'\boldsymbol{\alpha})}{\Phi(\mathbf{x}_i'\boldsymbol{\alpha})(1-\Phi(\mathbf{x}_i'\boldsymbol{\alpha}))}\mathbf{x}_i\mathbf{x}_i',$$

$$\mathbf{G}_{21} = \lim \tfrac{1}{N} \textstyle\sum_{i=1}^{N} d_i\mathbf{w}_i \dfrac{\partial \boldsymbol{\lambda}(\mathbf{x}_i'\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}},$$

$$\mathbf{G}_{22} = \lim \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathrm{E}[d_i\mathbf{w}_i\mathbf{w}_i'].$$

The expression for $\mathbf{G}_{11}$ uses knowledge that $\mathbf{G}_{11}^{-1}$ is just the variance of the probit MLE. The expression for $\mathbf{G}_{21}$ uses

$$\begin{aligned}\mathrm{E}\left[\frac{\partial \mathbf{h}_{2i}}{\partial \boldsymbol{\theta}_1'}\right] &= \mathrm{E}\left[-\frac{\partial d_i\mathbf{w}_i(\boldsymbol{\alpha})(y_i - \mathbf{w}_i(\boldsymbol{\alpha})'\boldsymbol{\gamma})}{\partial \boldsymbol{\alpha}}\right] \\ &= \mathrm{E}\left[\mathbf{w}_i \frac{\partial d_i \mathbf{w}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'}\right] \\ &= \mathrm{E}\left[d_i\mathbf{w}_i \frac{\partial \boldsymbol{\lambda}(\mathbf{x}_i'\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right].\end{aligned}$$

The expression for $\mathbf{G}_{22}$ uses

$$\frac{\partial \mathbf{h}_{2i}}{\partial \boldsymbol{\theta}_2'} = \frac{\partial d_i\mathbf{w}_i(\boldsymbol{\alpha})(y_i - \mathbf{w}_i(\boldsymbol{\alpha})'\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = d_i\mathbf{w}_i\mathbf{w}_i'.$$

Turning to $\mathbf{S}_0$ we have

$$\begin{aligned}\mathbf{S}_{11} &= \mathbf{G}_{11}^{-1}, \\ \mathbf{S}_{21} &= \mathbf{0}, \\ \mathbf{S}_{22} &= \lim \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathrm{E}[d_i(\,y_i - \mathbf{w}_i(\boldsymbol{\alpha})'\boldsymbol{\gamma})^2].\end{aligned}$$

The expression for $\mathbf{S}_{11}$ follows by applying the information matrix equality. Taking expectations and some manipulation leads to $\mathbf{S}_{21} = \mathbf{0}$, and $\mathbf{S}_{22}$ is simply $\mathrm{V}[\eta_i]$.

Combining these results gives the Heckman two-step estimator $\widehat{\boldsymbol{\gamma}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\gamma}, \mathbf{V}_\gamma)$, where

$$\widehat{\mathbf{V}}_\gamma = \left(\widehat{\mathbf{W}}'\widehat{\mathbf{W}}\right)^{-1}\left(\widehat{\mathbf{W}}'\boldsymbol{\Sigma}_{\widehat{\eta}}\widehat{\mathbf{W}} + \widehat{\mathbf{W}}'\widehat{\mathbf{D}}\widehat{\mathbf{V}}_\alpha\widehat{\mathbf{D}}\widehat{\mathbf{W}}\right)\left(\widehat{\mathbf{W}}'\widehat{\mathbf{W}}\right)^{-1}, \tag{16.61}$$

and where $\widehat{\mathbf{W}}'\widehat{\mathbf{W}} = \sum_{i=1}^{N} d_i\widehat{\mathbf{w}}_i\widehat{\mathbf{w}}_i'$, $\widehat{\mathbf{D}} = \mathrm{Diag}\left[\partial\lambda(\mathbf{x}_i'\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\;|_{\widehat{\alpha}}\right]$, $\widehat{\mathbf{V}}_\alpha$ is the variance matrix for the first-stage probit MLE, and $\boldsymbol{\Sigma}_{\widehat{\eta}}$ is a diagonal matrix with ith entry $\widehat{\sigma}_{\eta_i}^2$. This estimate is straightforward to obtain if matrix commands are available. The hardest part can be analytically obtaining $\sigma_{\eta_i}^2 = \mathrm{V}[\eta_i]$ given in (16.24). If this is difficult we can instead use $\widehat{\sigma}_i^2 = (y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{\sigma}\lambda_i(\mathbf{x}_i'\widehat{\boldsymbol{\alpha}}))^2$ following the approach of White (1980).

16.11. Practical Considerations

Most major packages include ML estimation of the Tobit model under normality. The two-part model is easy to estimate as one can separately estimate each part. In principle

the bivariate sample selection model can be estimated by Heckman's two-step procedure using only a probit and OLS routine. However, the standard errors are difficult to compute owing to the two-step nature of the estimator, and it is much easier to obtain standard errors using a package with Heckman's two-step procedure built-in. Implementing semiparametric estimators generally requires specialized code in a programming language such as GAUSS. Some packages also permit ML estimation of censored and truncated variants of other models, such as the Poisson and negative binomial for count data.

Censoring and truncation are easily handled if one views as reasonable the specified distribution. For example, top-coded income data are easily handled if the log-normal distribution fits the data well. Censored LAD, which relies on much weaker distributional assumptions, can also be used in this situation.

Much more problematic is handling models with sample selection. The more parametric versions of these models can rely on distributional assumptions that are felt to be strong. Semiparametric versions still have to struggle with the identification requirement that a variable that determines participation does not also determine the outcome of interest. A more promising route, one often taken in the treatment effects literature, is to limit attention to cases where it may be reasonable to assume that selection is only on observables.

16.12. Bibliographic Notes

The literature on models from selected samples is vast. Book-length treatments are provided by Maddala (1983) and Gouriéroux (2000), and shorter summaries are provided by Amemiya (1984, 1985) and Greene (2003).

16.3 Tobit (1958) proposed and applied the Tobit model to expenditure data. Amemiya (1973) formally established its consistency and asymptotic normality. Heckman (1974) provides an excellent female labor supply application with detailed analysis of results.

16.4 The many studies of the Rand Health Insurance Experimant, such as that by Duan et al. (1983), are leading applications of the two-part model.

16.5 Heckman (1976, 1979) presented the two-step estimator of the bivariate sample selection model that is also the basis for many more recent semiparametric estimation procedures. Mroz (1987) provides an excellent application to female labor supply that places emphasis on the role of assumptions on wage exogeneity.

16.7 There are many variants on the ideas of Roy (1951), just as there are many variants of the Tobit model. L-F. Lee (1978) provides a good early application to the union–nonunion wage differential.

16.8 The work by Dubin and McFadden (1984) is a leading example of structural micreconometric analysis based on complete specification of utility function and distribution of unobservables.

16.9 Semiparametric estimation of binary choice models is presented in detail in the books by M-J. Lee (1996), Horowitz (1997), and Pagan and Ullah (1999) and in surveys by

Vella (1998) and L-F. Lee (2001). Chay and Honoré (1998) and Chay and Powell (2001) provide applications for censored models, and Melenberg and Van Soest (1996) additionally estimate bivariate sample selection models.

Exercises

16–1 This question considers the impact of different degrees of truncation in the Tobit model.

(a) Generate 200 draws of a latent variable $y^* = k + 3x + u$, where $u \sim \mathcal{N}[0, 3]$ and the regressor $x \sim$ uniform[0, 1]. Choose k such that you generate approximately 30% of y^* to be negative.

(b) Generate a censored or truncated subsample by excluding observations that correspond to $y^* < 0$.

(c) Estimate the model using all 2,000 observations, as if the latent variable were observable, by OLS. Evaluate your results in the light of the theoretical properties of OLS, keeping in mind that you have only one replication.

(d) Using the truncated subsample of $y > 0$ only, estimate the model by OLS.

(e) Use the truncated maximum likelihood option to estimate the parameters using all observations. Evaluate your results in light of the properties of the truncated MLE. Compare with the least-squares results from the previous two parts.

(f) Repeat all previous steps using a value of k so as to generate 20, 40, and 50% censored observations. Compare your results with those based on 30% censored observations. Hence suggest what is the consequence on the parameter estimates of higher levels of censoring. Reinforce your arguments using theory where possible.

16–2 Consider a latent variable modeled by $y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}[0, \sigma^2]$. Suppose y_i^* is censored from above so that we observe $y_i = y_i^*$ if $y_i^* < U_i$ and $y_i = U_i$ if $y_i^* \geq U_i$, where the upper limit U_i is a known constant for each individual (i.e., data) and may differ over individuals.

(a) Give the log-likelihood function for this model. [Hint: Note that this differs from the standard case both owing to presence of U_i and because the equalities are reversed with $y_i = y_i^*$ if $y_i^* < U_i$.]

(b) Obtain the expression for the truncated mean $\mathrm{E}[y_i|\mathbf{x}_i, y_i < U_i]$. [Hint: For $z \sim \mathcal{N}[0, 1]$, we have $\mathrm{E}[z|z > c] = \phi(c)/[1 - \Phi(c)]$. Also, $\mathrm{E}[z|z < c] = -\mathrm{E}[-z|-z > -c]$ and $-z \sim \mathcal{N}[0, 1]$.]

(c) Hence give Heckman's two-step estimator for this model.

(d) Obtain the expression for the censored mean $\mathrm{E}[y_i|\mathbf{x}_i]$. [Hint: An essential part is the answer in part (b).]

16–3 This question considers the consequences of misspecification in the Tobit model. The starting point is the model of Exercise 16.1.

(a) Generate y^* with heteroskedasticity by letting $u \sim \mathcal{N}[0, \sigma^2 z]$, where $z > 0$ is chosen to be a suitable positive-valued variable that is correlated with x, though not perfectly so. Again set k to obtain about 30% of censored observations. Use the MLE for censored normal to estimate this model and compare your results with the corresponding homoskedastic case.

(b) Now consider the impact of nonnormality in the sample. Use the simulation macro available in some packages to carry out a Monte Carlo evaluation based on a sample of 1,000 observations and 500 replications. In each replication generate a sample with censored observations such that the errors are drawn from a mixture of two normals: $\mathcal{N}[1, 9]$ or $\mathcal{N}[0.4, 1]$ with probabilities 0.4 and 0.6, respectively. Estimate the model using the censored Tobit MLE and compare your results with the normal case. Carry out an analysis of the Monte Carlo output for the two estimators. Draw appropriate conclusions about the impact of nonnormality on the distribution of the Tobit estimator.

16–4 Consider a Poisson regression model where y^* has density $f^*(y^*) = e^{-\mu}\mu^y/y^*!$, $y_i^* = 0, 1, 2, \ldots$, and we have independence over i. Because of coding error we only fully observe y^* when $y^* \geq 2$. When $y^* = 0$ or 1 we only observe that $y^* \leq 1$. Suppose this is coded as $y^* = 1$. Define the observed data $y = y^*$ for $y_i^* \geq 2$ and $y = 1$ for $y_i^* = 0$ or 1.

(a) Obtain the density $f(y)$ of the observed y.

(b) Obtain $\mathrm{E}[y]$. [There is some algebra here.]
Now introduce regressors with $\mathrm{E}[y^*|\mathbf{x}] = \exp(\mathbf{x}'\beta)$ and define the indicator variable $d = 1$ for $y^* \geq 2$ and $d = 0$ for $y^* = 0$ or 1.

(c) Give the exact formula for this example of the objective function of an estimator that provides a consistent estimator of β using data on y_i, d_i, and $\mathbf{x}_i$.

(d) Give the exact formula for this example of the objective function of an estimator that provides a consistent estimator of β using data on only d_i and $\mathbf{x}_i$.

(e) Is it possible to consistently estimate β using data on only d_i and $\mathbf{x}_i$? Explain your answer.

16–5 Using a 50% random subsample of the RAND data on medical expenditure over a 12-month period used in this chapter, and using a similar model specification, we wish to consider the following broad question: Which model is appropriate for modeling the expenditure data?

(a) Using the data summary of the expenditure variable, analyze the implications of the high proportion of zero expenditures observed. Is this a violation of the normality assumption? Is there a transformation of expenditure that would make the assumption of normality more appropriate?

(b) Three candidate models are considered, each with the same set of covariates. These covariates are the same as in the count data Exercise 20.6. The models are (i) the Tobit model, (ii) the two-part ("hurdle") model (TPM), and (iii) the selection model. Explain how each one of these will be set up, the relationship and connections among them, and how one might compare and choose among them. If you are likely to encounter any specific specifications or estimation problems, state them and suggest how you might handle them. Pay attention to the choice of exclusion restrictions.

(c) Estimate in turn the Tobit model, the TPM, and the selection models. For the TPM you have two equations, and the second is for those who have positive expenditures only. In the case of the selection model, use both the MLE and the two-step (Heckman) estimators. Discuss your reasons underlying

the exclusion restriction required in the estimation of the selection model. Is there evidence that the selection problem is a serious issue?

(d) How can we compare the statistical fit of the three models? Which model appears to provide the best fit to the data? By what criterion?

(e) Suppose our main interest is in the impact of two variables on expenditure, log income, and log of (1 + coinsurance rate). Use the results of your estimated Tobit model and TPM to make a comparison between the marginal impact of a change in these variables on expenditure. Given that there is considerable heterogeneity in the sample, suggest how to present the results of your analysis in the most informative manner.

(f) Briefly explain how quantile regression (see Section 4.6) provides an alternative method of analyzing the same data. What are the main advantages and disadvantages of this approach in the present data situation?