8
Overview of endogeneity

"A government study today revealed that 83% of statistics are misleading."
- Ziggy by Tom Wilson

As discussed in chapter 2, managers actively make production-investment, financing, and accounting choices. These choices are intertwined and far from innocuous. Design of accounting (like other information systems) is highly dependent on the implications and responses to accounting information in combination with other information. As these decisions are interrelated, their analysis is inherently endogenous (Demski [2004]). Endogeneity presents substantial challenges for econometric analysis. The behavior of unobservable (to the analyst) components and omitted, correlated variables are continuing themes.

In this chapter, we briefly overview econometric analysis of endogeneity, explore some highly stylized examples that motivate its importance, and lay some ground work for exploring treatment effects in the following chapters. A theme for this discussion is that econometric analysis of endogeneity is a three-legged problem: theory, data, and model specification (or logically consistent discovery of the *DGP*). Failure to support any leg and the entire inquiry is likely to collapse. Progress is impeded when authors fail to explicitly define the causal effects of interest or state what conditions are perceived for identification of the estimand of interest. As Heckman and Vytlacil [2007] argue in regards to the economics literature, this makes it difficult to build upon past literature and amass a coherent body of evidence. We explore various identifying conditions in the ensuing discussions of endogenous causal effects.

8.1 Overview

Many, perhaps all, endogeneity concerns can be expressed in the form of an omitted, correlated variable problem. We remind the reader (see chapter 3) that standard parameter estimators (such as *OLS*) are not asymptotically consistent in the face of omitted, correlated variables.

8.1.1 Simultaneous equations

When many of us think of endogeneity, simultaneous equations is one of the first settings that comes to mind. That is, when we have multiple variables whose behavior are interrelated such that they are effectively simultaneously determined, endogeneity is a first-order consideration. For instance, consider a simple example where the *DGP* is expressed as the following structural equations[1]

$$\begin{aligned} Y_1 &= \beta_1 X_1 + \beta_2 Y_2 + \varepsilon_1 \\ Y_2 &= \gamma_1 X_2 + \gamma_2 Y_1 + \varepsilon_2 \end{aligned}$$

Clearly, little can be said about either Y_1 or Y_2 without including the other (a form of omitted variable). It is not possible to speak of manipulation of only Y_1 or Y_2. Perhaps, this is most readily apparent if we rewrite the equations in reduced form:

$$\begin{bmatrix} 1 & -\beta_2 \\ -\gamma_2 & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \beta_1 & 0 \\ 0 & \gamma_1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

assuming $\beta_2\gamma_2 \neq 1$

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & -\beta_2 \\ -\gamma_2 & 1 \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \beta_1 & 0 \\ 0 & \gamma_1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \right\}$$

$$\begin{aligned} Y_1 &= \frac{\beta_1}{1-\beta_2\gamma_2} X_1 + \frac{\beta_2\gamma_1}{1-\beta_2\gamma_2} X_2 + \frac{1}{1-\beta_2\gamma_2}\varepsilon_1 + \frac{\beta_2}{1-\beta_2\gamma_2}\varepsilon_2 \\ Y_2 &= \frac{\beta_1\gamma_2}{1-\beta_2\gamma_2} X_1 + \frac{\gamma_1}{1-\beta_2\gamma_2} X_2 + \frac{\gamma_2}{1-\beta_2\gamma_2}\varepsilon_1 + \frac{1}{1-\beta_2\gamma_2}\varepsilon_2 \end{aligned}$$

which can be rewritten as

$$\begin{aligned} Y_1 &= \omega_{11} X_1 + \omega_{12} X_2 + \eta_1 \\ Y_2 &= \omega_{21} X_1 + \omega_{22} X_2 + \eta_2 \end{aligned}$$

where $Var \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{bmatrix}$. Since rank and order conditions are satisfied (assuming $\beta_2\gamma_2 \neq 1$), the structural parameters can be recovered from the reduced

[1] Goldberger [1972, p. 979] defines structural equations as an approach that employs "stochastic models in which each equation represents a causal link, rather than a mere empirical association."

form parameters as follows.

$$\begin{aligned}
\beta_1 &= \omega_{11} - \frac{\omega_{12}\omega_{21}}{\omega_{22}} \\
\beta_2 &= \frac{\omega_{12}}{\omega_{22}} \\
\gamma_1 &= \omega_{22} - \frac{\omega_{12}\omega_{21}}{\omega_{11}} \\
\gamma_2 &= \frac{\omega_{21}}{\omega_{11}} \\
Var\left[\varepsilon_1\right] &= v_{11} + \frac{\omega_{12}\left(v_{22}\omega_{12} - 2v_{12}\omega_{22}\right)}{\omega_{22}^2} \\
Var\left[\varepsilon_2\right] &= v_{22} + \frac{\omega_{21}\left(v_{11}\omega_{21} - 2v_{12}\omega_{11}\right)}{\omega_{11}^2} \\
Cov\left[\varepsilon_1, \varepsilon_2\right] &= \frac{v_{12}\left(\omega_{12}\omega_{21} + \omega_{11}\omega_{22}\right) - v_{11}\omega_{21}\omega_{22} - v_{22}\omega_{11}\omega_{12}}{\omega_{11}\omega_{22}}
\end{aligned}$$

Suppose the causal effects of interest are β_1 and γ_1. Examination of the reduced form equations reveals that ignoring simultaneity produces inconsistent estimates of β_1 and γ_1 even if X_1 and X_2 are uncorrelated (unless β_2 or $\gamma_2 = 0$).

More naively, suppose we attempt to estimate the structural equations directly (say, via *OLS*). Since the response variables are each a function of the other response variable, the regressors are correlated with the errors and the fundamental condition of regression $E\left[X^T\varepsilon\right] = 0$ is violated and *OLS* parameter estimates are inconsistent. A couple of recursive substitutions highlight the point. For illustrative purposes, we work with Y_1 but the same ideas obviously apply to Y_2.

$$\begin{aligned}
Y_1 &= \beta_1 X_1 + \beta_2 Y_2 + \varepsilon_1 \\
&= \beta_1 X_1 + \beta_2\left(\gamma_1 X_2 + \gamma_2 Y_1 + \varepsilon_2\right) + \varepsilon_1
\end{aligned}$$

Of course, if $E\left[\varepsilon_2^T\varepsilon_1\right] \neq 0$ then we've demonstrated the point; notice this is a standard endogenous regressor problem. Simultaneity bias (inconsistency) is illustrated with one more substitution.

$$Y_1 = \beta_1 X_1 + \beta_2\left(\gamma_1 X_2 + \gamma_2\left(\beta_1 X_1 + \beta_2 Y_2 + \varepsilon_1\right) + \varepsilon_2\right) + \varepsilon_1$$

Since Y_2 is a function of Y_1, inclusion of Y_2 as a regressor produces a clear violation of $E\left[X^T\varepsilon\right] = 0$ as we have $E\left[\varepsilon_1^T\varepsilon_1\right] \neq 0$.

Notice, we can think of simultaneity problems as arising from omitted, correlated unobservable variables. Hence, this simple example effectively identifies the basis — *omitted, correlated unobservable variables* — for most (perhaps all) endogeneity concerns. Further, this simple structural example readily connects to estimation of causal effects.

Definition 8.1 *Causal effects are the ceteris paribus response to a change in variable or parameter (Marshall [1961] and Heckman [2000]).*

As the simultaneity setting illustrates, endogeneity often makes it infeasible to "turn one dial at a time."

8.1.2 Endogenous regressors

Linear models with endogenous regressors are commonplace (see Larcker and Rusticus [2004] for an extensive review of the accounting literature). Suppose the *DGP* is

$$Y_1 = X_1\beta_1 + Y_2\beta_2 + \varepsilon_1$$

$$Y_2 = \gamma_1 X_2 + \varepsilon_2$$

where $E\left[X^T\varepsilon_1\right] = 0$ and $E\left[X_2^T\varepsilon_2\right] = 0$ but $E\left[\varepsilon_2^T\varepsilon_1\right] \neq 0$. In other words, $Y_1 = \beta_1 X_1 + \beta_2\left(\gamma_1 X_2 + \varepsilon_2\right) + \varepsilon_1$. Of course, *OLS* produces inconsistent estimates. Instrumental variables (*IV*) are a standard remedy. Suppose we observe variables X_2. Variables X_2 are clearly instruments as they are unrelated to ε_1 but highly correlated with the endogenous regressors Y_2 (assuming $\gamma_1 \neq 0$).

Two-stage least squares instrumental variable (*2SLS-IV*) estimation is a standard approach for dealing with endogenous regressors. In the first stage, project all of the regressors (endogenous plus exogenous) onto the instruments plus all other exogenous regressors (see chapter 3 on overidentifying restrictions and *IV*). Let $X = \left[\begin{array}{cc} X_1 & Y_2 \end{array}\right]$ and $Z = \left[\begin{array}{cc} X_1 & X_2 \end{array}\right]$

$$\hat{X} = Z\left(Z^TZ\right)^{-1}Z^TX = P_ZX = \left[\begin{array}{cc} P_ZX_1 & P_ZY_2 \end{array}\right]$$

In the second stage, replace the regressors with the predicted values from the first stage regression.

$$Y_1 = P_ZX_1\beta_1 + P_ZY_2\beta_2 + \varepsilon_1^{'}$$

The *IV* estimator for β (for convenience, we have reversed the order of the variables) is

$$\left(\left[\begin{array}{c} Y_2^TP_Z \\ X_1^TP_Z \end{array}\right]\left[\begin{array}{cc} P_ZY_2 & P_ZX_1 \end{array}\right]\right)^{-1}\left[\begin{array}{c} Y_2^TP_Z \\ X_1^TP_Z \end{array}\right]Y_1$$

The probability limit of the estimator is

$$\begin{aligned} & p\lim\left[\begin{array}{cc} Y_2^TP_ZY_2 & Y_2^TP_ZX_1 \\ X_1^TP_ZY_2 & X_1^TP_ZX_1 \end{array}\right]^{-1}\left[\begin{array}{c} Y_2^TP_Z \\ X_1^TP_Z \end{array}\right](X_1\beta_1 + Y_2\beta_2 + \varepsilon_1) \\ = & \left[\begin{array}{c} \beta_2 \\ \beta_1 \end{array}\right] \end{aligned}$$

To see this, recall the inverse of the partitioned matrix

$$\left[\begin{array}{cc} Y_2^TP_ZY_2 & Y_2^TP_ZX_1 \\ X_1^TP_ZY_2 & X_1^TP_ZX_1 \end{array}\right]^{-1}$$

via block "rank-one" LDL^T representation (see *FWL* in chapter 3) is

$$\left(\left[\begin{array}{cc} I & 0 \\ A & I \end{array}\right]\left[\begin{array}{cc} Y_2^TP_ZY_2 & 0 \\ 0 & X_1^TP_ZM_{P_ZY_2}P_ZX_1 \end{array}\right]\left[\begin{array}{cc} I & A^T \\ 0 & I \end{array}\right]\right)^{-1}$$

where $A = X_1^T P_Z Y_2 \left(Y_2^T P_Z Y_2\right)^{-1}$. Simplification gives

$$\begin{aligned} & \left[\begin{array}{cc} I & -A^T \\ 0 & I \end{array} \right] \left[\begin{array}{cc} \left(Y_2^T P_Z Y_2\right)^{-1} & 0 \\ 0 & B \end{array} \right] \left[\begin{array}{cc} I & 0 \\ -A & I \end{array} \right] \\ = & \left[\begin{array}{cc} \left(Y_2^T P_Z Y_2\right)^{-1} + A^T B A & -A^T B \\ -BA & B \end{array} \right] \end{aligned}$$

where

$$B = \left(X_1^T P_Z M_{P_Z Y_2} P_Z X_1\right)^{-1}$$

and

$$M_{P_Z Y_2} = I - P_Z Y_2 \left(Y_2^T P_Z Y_2\right)^{-1} Y_2^T P_Z$$

Now, focus on the second equation.

$$\begin{aligned} & \left(-BAY_2^T P_Z + BX_1^T P_Z\right) Y_1 \\ = & \left(\begin{array}{c} -\left(X_1^T P_Z M_{P_Z Y_2} P_Z X_1\right)^{-1} X_1^T P_Z Y_2 \left(Y_2^T P_Z Y_2\right)^{-1} Y_2^T P_Z \\ + \left(X_1^T P_Z M_{P_Z Y_2} P_Z X_1\right)^{-1} X_1^T P_Z \end{array} \right) Y_1 \\ = & \left(X_1^T P_Z M_{P_Z Y_2} P_Z X_1\right)^{-1} X_1^T P_Z \left(I - P_Z Y_2 \left(Y_2^T P_Z Y_2\right)^{-1} Y_2^T P_Z\right) \\ & \left(X_1 \beta_1 + Y_2 \beta_2 + \varepsilon_1\right) \\ = & \left(X_1^T P_Z M_{P_Z Y_2} P_Z X_1\right)^{-1} X_1^T P_Z M_{P_Z Y_2} \left(X_1 \beta_1 + Y_2 \beta_2 + \varepsilon_1\right) \end{aligned}$$

Since $P_Z M_{P_Z Y_2} = P_Z M_{P_Z Y_2} P_Z$, the second equation can be rewritten as

$$\begin{aligned} & \left(X_1^T P_Z M_{P_Z Y_2} P_Z X_1\right)^{-1} X_1^T P_Z M_{P_Z Y_2} P_Z \left(X_1 \beta_1 + Y_2 \beta_2 + \varepsilon_1\right) \\ = & \beta_1 + \left(X_1^T P_Z M_{P_Z Y_2} P_Z X_1\right)^{-1} X_1^T P_Z M_{P_Z Y_2} P_Z \left(Y_2 \beta_2 + \varepsilon_1\right) \end{aligned}$$

Since $M_{P_Z Y_2} P_Z Y_2 = 0$ (by orthogonality) and $p \lim \frac{1}{n} P_Z \varepsilon_1 = 0$, the estimator for β_1 is consistent. The derivation is completed by reversing the order of the variables in the equations again to show that β_2 is consistent.[2]

8.1.3 Fixed effects

Fixed effects models allow for time and/or individual differences in panel data. That is, separate regressions, say for m firms in the sample, are estimated with differences in intercepts but pooled slopes as illustrated in figure 8.1.

$$Y = X\beta + Z\gamma + \sum_{j=1}^{m} \alpha_j D_j + \varepsilon$$

[2]Of course, we could simplify the first equation but it seems very messy so why not exploit the effort we've already undertaken.

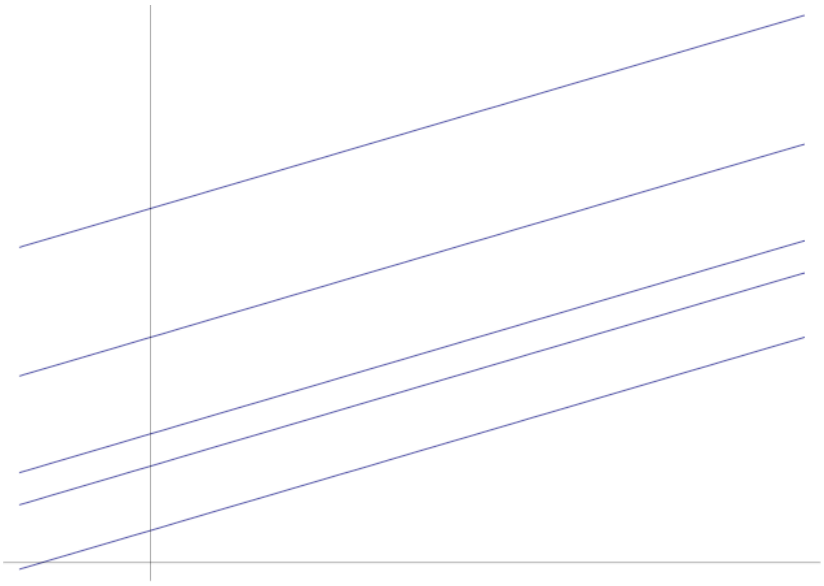

Figure 8.1: Fixed effects regression curves

where D_j is a firm indicator variable, X represents the experimental regressors and Z the control variables.[3] Geometrically, it's instructive to think of *FWL* (see chapter 3) where we condition on all control variables, then the experimental explanatory variables of interest are evaluated conditional on the control variables.[4]

$$M_Z Y = M_Z X\beta + \sum_{j=1}^{m} \alpha_j M_Z D_j + \epsilon$$

Of course, we can also consider semi- and non-parametric fixed effects regressions as well if we think of the nonparametric analog to *FWL* initiated by Robinson [1988] in the form of partial linear models and Stoker's [1991] partial index models (see chapter 6).

Causal effects are identified via a fixed effects model when there are constant, unobserved (otherwise they could be included as covariates) individual characteristics that because they are related to both outcomes and causing variables would be omitted, correlated variables if ignored. Differencing approaches such as fixed effects are simple and effective so long as individual fixed effects do not vary across periods and any correlation between treatment and unobserved outcome potential is described by an additive time-invariant covariate. Since this condition

[3]Clearly, time fixed effects can be accommodated in analogous fashion with time subscripts and indicator variables replacing the firm or individual variables.

[4]Of course, if an intercept is included in the fixed effects regression then the summation index is over $m - 1$ firms or individuals instead of m.

doesn't usually follow from economic theory or institutionally-relevant information, the utility of the fixed effects approach for identifying causal effects is limited.[5]

Nikolaev and Van Lent [2005] study variation through time in a firm's disclosure quality and its impact on the marginal cost of debt. In their setting, unobservable cross-firm heterogeneity, presumed largely constant through time, is accommodated via firm fixed effects. That is, firm-by-firm regressions that vary in intercept but have the same slope are estimated. Nikolaev and Van Lent argue that omitted variables and endogeneity plague evaluation of the impact of disclosure quality on cost of debt capital and the problem is mitigated by fixed effects.

Robinson [1989] concludes that fixed effects analysis more effectively copes with endogeneity than longitudinal, control function, or *IV* approaches in the analysis of the differential effects of union wages. In his setting, endogeneity is primarily related to worker behavior and measurement error. Robinson suggests that while there is wide agreement that union status is not exogenous, there is little consistency in teasing out the effect of union status on wages. While longitudinal analysis typically reports smaller effects than *OLS*, cross-sectional approaches such as *IV* or control function approaches (inverse Mills ratio) typically report larger effects than *OLS*. Robinson concludes that a simple fixed effects analysis of union status is a good compromise. (Also, see Wooldridge [2002], p. 581-590.)

On the other hand, Lalonde [1986] finds that regression approaches (including fixed effects) perform poorly compared with "experimental" methods in the analysis of the National Supported Work (NSW) training program. Dehejia and Wahba [1995] reanalyze the NSW data via propensity score matching and find similar results to Lalonde's experimental evidence. Once again we find no single approach works in all settings and the appropriate method depends on the context.

8.1.4 Differences-in-differences

Differences-in-differences (*DID*) is a close cousin to fixed effects. *DID* is a panel data approach that identifies causal effects when certain groups are treated and other groups are not. The treated are exposed to sharp changes in the causing variable due to shifts in the economic environment or changes in (government) policy. Typically, potential outcomes, in the absence of the change, are composed of the sum of a time effect that is common to all groups and a time invariant individual fixed effect, say,

$$E\left[Y_0 \mid t, i\right] = \beta_t + \gamma_i$$

Then, the causal effect δ is simply the difference between expected outcomes with treatment and expected outcomes without treatment

$$E\left[Y_1 \mid t, i\right] = E\left[Y_0 \mid t, i\right] + \delta$$

[5] It is well-known that fixed effects yield inconsistent parameter estimates when the model involves lagged dependent variables (see Chamberlain [1984] and Angrist and Krueger [1998]).

The key identifying condition for *DID* is the parameters associated with (treatment time and treatment group) interaction terms are zero in the absence of intervention.

Sometimes apparent interventions are themselves terminated and provide opportunities to explore the absence of intervention. Relatedly, R. A. Fisher (quoted in Cochran [1965]) suggested the case for causality is stronger when the model has many implications supported by the evidence. This emerges in terms of robustness checks, exploration of sub-populations in which treatment effects should not be observed (because the subpopulation is insensitive or immune to treatment or did not receive treatment), and comparison of experimental and non-experimental research methods (Lalonde [1986]). However, general equilibrium forces may confound direct evidence from such absence of intervention analyses. As Angrist and Krueger [1998, p. 56] point out, "Tests of refutability may have flaws. It is possible, for example, that a subpopulation that is believed unaffected by the intervention is indirectly affected by it."

8.1.5 Bivariate probit

A variation on a standard self-selection theme is when both selection and outcome equations are observed as discrete responses. If the unobservables are jointly normally distributed a bivariate probit accommodates endogeneity in the same way that a standard Heckman (inverse Mills ratio) control function approach works with continuous outcome response. Endogeneity is reflected in nonzero correlation among the unobservables. Dubin and Rivers [1989] provide a straightforward overview of this approach.

$$U_D = Z\theta + V, D = \begin{array}{ll} 1 & \text{if } U_D > 0 \\ 0 & \text{otherwise} \end{array}$$

$$Y^* = X\beta + \varepsilon, Y = \begin{array}{ll} 1 & \text{if } Y^* > 0 \\ 0 & \text{otherwise} \end{array}$$

$$E\left[\begin{array}{c} V \\ \varepsilon \end{array}\right] = 0, Var\left[\begin{array}{c} V \\ \varepsilon \end{array}\right] = \Sigma = \left[\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array}\right]$$

Following the shorthand of Greene [1997], let $q_{i1} = 2U_{iD} - 1$ and $q_{i2} = 2Y_i - 1$, so that $q_{ij} = 1$ or -1. The bivariate normal cumulative distribution function is

$$\Pr\left(X_1 < x_1, X_2 < x_2\right) = \Phi_2\left(z_1, z_2, \rho\right) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} \phi_2\left(z_1, z_2, \rho\right) dz_1 dz_2$$

where

$$\phi_2\left(z_1, z_2, \rho\right) = \frac{1}{2\pi\left(1-\rho^2\right)^{\frac{1}{2}}} \exp\left[-\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2\left(1-\rho^2\right)}\right]$$

denotes the bivariate normal (unit variance) density. Now let

$$\begin{array}{ll} z_{i1} = \theta^T Z_i & w_{i1} = q_{i1} z_{i1} \\ z_{i2} = \beta^T X_i & w_{i2} = q_{i2} z_{i2} \end{array}$$

$$\rho_{i*} = q_{i1}q_{i2}\rho$$

With this setup, the log-likelihood function can be written in a simple form where all the sign changes associated with D and Y equal to 0 and 1 are accounted for

$$lnL = \sum_{i=1}^{n} \Phi_2 \left(w_{i1}, w_{i2}, \rho_{i*} \right)$$

and maximization proceeds in the usual manner (see, for example, Greene [1997] for details).[6]

8.1.6 Simultaneous probit

Suppose we're investigating a discrete choice setting where an experimental variable (regressor) is endogenously determined. An example is Bagnoli, Liu, and Watts [2006] (BLW). BLW are interested in the effect of family ownership on the inclusion of covenants in debt contracts. Terms of debt contracts, such as covenants, are likely influenced by interest rates and interest rates are likely determined simultaneously with terms such as covenants. A variety of limited information approaches[7] have been proposed for estimating these models - broadly referred to as simultaneous probit models (see Rivers and Vuong [1988]). BLW adopted two stage conditional maximum likelihood estimation (*2SCML*; discussed below).

The base model involves a structural equation

$$y^* = Y\gamma + X_1\beta + u$$

where discrete D is observed

$$D_i = \begin{array}{ll} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{array}$$

The endogenous explanatory variables have reduced form

$$Y = \Pi X + V$$

where exogenous variable X and X_1 are related via matrix J, $X_1 = JX$, Y is an $n \times m$ matrix of endogenous variables, X_1 is $n \times k$, and X is $n \times p$. The following conditions are applied to all variations:

Condition 8.1 *(X_i, u_i, V_i) is iid with X_i having finite positive definite variance matrix Σ_{XX}, and $(u_i, V_i \mid X_i)$ are jointly normally distributed with mean zero and finite positive definite variance matrix* $\Omega = \begin{bmatrix} \sigma_{uu} & \Sigma_{uV} \\ \Sigma_{Vu} & \Sigma_{VV} \end{bmatrix}$.

---

[6] Evans and Schwab [1995] employ bivariate probit to empirically estimate causal effects of schooling.

[7] They are called limited information approaches in that they typically focus on one equation at a time and hence ignore information in other equations.

Condition 8.2 *rank*$(\Pi, J) = m + k$.

Condition 8.3 $(\gamma, \beta, \Pi, \Omega)$ *lie in the interior of a compact parameter space* Θ.

Identification of the parameters in the structural equation involves normalization. A convenient normalization is

$$Var\left[y_i^* \mid X_i, Y_i\right] = \sigma_{uu} - \lambda^T \Sigma_{VV} \lambda = 1,$$

where $\lambda = \Sigma_{VV}^{-1} \Sigma_{Vu}$, the structural equation is rewritten as

$$y^* = Y\gamma + X_1\beta + V\lambda + \eta$$

and $\eta_i = u_i - V_i^T \lambda \sim N\left(0, \sigma_{uu} - \lambda^T \Sigma_{VV} \lambda = 1\right)$.

Limited information maximum likelihood (*LIML*)

A limited information maximum likelihood (*LIML*) approach was adopted by Godfrey and Wickens [1982]. The likelihood function is

$$\prod_{i=1}^{n} (2\pi)^{-\frac{(m+1)}{2}} |\Omega|^{-\frac{1}{2}} \left[\int_{c_i}^{\infty} \exp\left\{-\frac{1}{2}\left(u, V_i^T\right) \Omega^{-1} \left(u, V_i^T\right)^T\right\} du\right]^{D_i}$$
$$\times \left[\int_{-\infty}^{c_i} \exp\left\{-\frac{1}{2}\left(u, V_i^T\right) \Omega^{-1} \left(u, V_i^T\right)^T\right\} du\right]^{1-D_i}$$

where $c_i = -\left(Y_i^T\gamma - X_{1i}^T\beta\right)$. Following some manipulation, estimation involves maximizing the log-likelihood with respect to $(\gamma, \beta, \lambda, \Pi, \Sigma_{VV})$. As *LIML* is computationally difficult in large models, it has received little attention except as a benchmark case.

Instrumental variables probit (*IVP*)

Lee [1981] proposed an instrumental variables probit (*IVP*). Lee rewrites the structural equation in reduced form

$$y_i^* = \left(\Pi^T X_i\right)\gamma + X_{1i}\beta + V_i\lambda + \eta_i$$

The log-likelihood for D given X is

$$\sum_{i=1}^{n} D_i \log \Phi\left[\left(\Pi^T X_i\right)\gamma_* + X_{1i}^T\beta_*\right]$$
$$+ (1 - D_i) \log\left\{1 - \Phi\left[\left(\Pi^T X_i\right)\gamma_* + X_{1i}^T\beta_*\right]\right\}$$

where $\Phi(\cdot)$ denotes a standard normal *cdf* and

$$\gamma_* = \frac{\gamma}{\omega} \quad \beta_* = \frac{\beta}{\omega}$$

$$\begin{aligned}\omega^2 &= Var\left[u_i + V_i^T\gamma\right] = \sigma_{uu}^2 + \gamma^T\Sigma_{VV}\gamma + \gamma^T\Sigma_{VV}\lambda + \lambda^T\Sigma_{VV}\gamma \\ &= 1 + \lambda^T\Sigma_{VV}\lambda + \gamma^T\Sigma_{VV}\gamma + \gamma^T\Sigma_{VV}\lambda + \lambda^T\Sigma_{VV}\gamma \\ &= 1 + (\gamma+\lambda)^T\,\Sigma_{VV}\,(\gamma+\lambda)\end{aligned}$$

Consistent estimates for Π, $\hat{\Pi}$, are obtained via *OLS*. Then, utilizing $\hat{\Pi}$ in place of Π, maximization of the log-likelihood with respect to γ_* and β_* is computed via m regressions followed by a probit estimation.

Generalized two-stage simultaneous probit (*G2SP*)

Amemiya [1978] suggested a general method for obtaining structural parameter estimates from reduced form estimates (*G2SP*). Heckman's [1978] two-stage endogenous dummy variable model is a special case of *G2SP*. Amemiya's proposal is a variation on *IVP* where the unconstrained log-likelihood is maximized with respect to τ_*

$$\sum_{i=1}^{n} D_i \log\Phi\left(X_i^T\tau_*\right) + (1-D_i)\log\left[1-\Phi\left(X_i^T\tau_*\right)\right]$$

$$\tau_* = \Pi\gamma_* + J\beta_*$$

In terms of the sample estimates we have the regression problem

$$\begin{aligned}\hat{\tau}_* &= \left[\begin{array}{cc}\hat{\Pi} & J\end{array}\right]\left[\begin{array}{c}\gamma_* \\ \beta_*\end{array}\right] + (\hat{\tau}_* - \tau_*) - \left(\hat{\Pi}-\Pi\right)\gamma_* \\ &= \hat{H}\left[\begin{array}{c}\gamma_* \\ \beta_*\end{array}\right] + e\end{aligned}$$

where $e = (\hat{\tau}_* - \tau_*) - \left(\hat{\Pi}-\Pi\right)\gamma_*$. *OLS* provides consistent estimates of γ_* and β_* but *GLS* is more efficient. Let $\hat{V}$ denote an asymptotic consistent estimator for the variance e. Then Amemiya's *G2SP* estimator is

$$\left[\begin{array}{c}\hat{\gamma}_* \\ \hat{\beta}_*\end{array}\right] = \left(\hat{H}^T\hat{V}^{-1}\hat{H}\right)^{-1}\hat{H}^T\hat{V}^{-1}\hat{\tau}_*$$

This last step constitutes one more computational step (in addition to the m reduced form regressions and one probit) than required for *IVP* (and *2SCML* described below).

Two-stage conditional maximum likelihood (*2SCML*)

Rivers and Vuong [1988] proposed two-stage conditional maximum likelihood (*2SCML*). Vuong [1984] notes when the joint density for a set of endogenous variables can be factored into a conditional distribution for one variable and a marginal distribution for the remaining variables, estimation can often be simplified by using conditional maximum likelihood methods. In the simultaneous

probit setting, the joint density for D_i and Y_i factors into a probit likelihood and a normal density.

$$\begin{aligned} & h\left(D_i, Y_i \mid X_i; \gamma, \beta, \lambda, \Pi, \Sigma_{VV}\right) \\ = \; & f\left(D_i \mid Y_i, X_i; \gamma, \beta, \lambda, \Pi\right) g\left(Y_i \mid X_i; \Pi, \Sigma_{VV}\right) \end{aligned}$$

where

$$\begin{aligned} & f\left(D_i \mid Y_i, X_i; \gamma, \beta, \lambda, \Pi\right) \\ = \; & \Phi\left(Y_i^T\gamma + X_{1i}^T\beta + V_i^T\lambda\right)^{D_i}\left[1 - \Phi\left(Y_i^T\gamma + X_{1i}^T\beta + V_i^T\lambda\right)\right]^{(1-D_i)} \end{aligned}$$

$$\begin{aligned} & g\left(Y_i \mid X_i; \Pi, \Sigma_{VV}\right) \\ = \; & (2\pi)^{-\frac{m}{2}}\left|\Sigma_{VV}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y_i - \Pi^T X_i\right)^T \Sigma_{VV}^{-1}\left(Y_i - \Pi^T X_i\right)\right\} \end{aligned}$$

Two steps are utilized to compute the *2SCML* estimator. First, the marginal log-likelihood for Y_i is maximized with respect to $\hat{\Pi}$ and $\hat{\Sigma}_{VV}$. This is computed by m reduced form regressions of Y on X to obtain $\hat{\Pi}$. Let the residuals be $\hat{V}_i = Y_i - \hat{\Pi}X_i$, then the standard variance estimator is $\hat{\Sigma}_{VV} = n^{-1}\sum_{i=1}^{n}\hat{V}_i\hat{V}_i^T$. Second, replacing Π with $\hat{\Pi}$, the conditional log-likelihood for D_i is maximized with respect to $\left(\hat{\gamma}, \hat{\beta}, \hat{\lambda}\right)$. This is computed via a probit analysis of D_i with regressors Y_i, X_{1i}, and $\hat{V}_i$.

2SCML provides several convenient tests of endogeneity. When Y_i and u_i are correlated, standard probit produces inconsistent estimators for γ and β. However, if $\Sigma_{Vu} = 0$, or equivalently, $\lambda = 0$, the $Y_i s$ are effectively exogenous. A modified *Wald* statistic is

$$MW = n\,\hat{\lambda}^T\hat{V}_0\left(\hat{\lambda}\right)^{-1}\hat{\lambda}$$

where $\hat{V}_0\left(\hat{\lambda}\right)$ is a consistent estimator for the lower right-hand block (corresponding to λ) of $V_0\left(\theta\right) = \left(\check{H}^T\check{\Sigma}\check{H}\right)^{-1}$ where

$$\check{H} = \begin{bmatrix} \Pi & J & 0 \\ I_m & 0 & I_m \end{bmatrix}$$

and

$$\begin{aligned} \check{\Sigma} &= \begin{bmatrix} \check{\Sigma}_{XX} & \check{\Sigma}_{XV} \\ \check{\Sigma}_{VX} & \check{\Sigma}_{VV} \end{bmatrix} \\ &= E\left[\frac{\phi\left(Z_i^T\delta + V_i^T\lambda\right)^2}{\Phi\left(Z_i^T\delta + V_i^T\lambda\right)\left[1 - \Phi\left(Z_i^T\delta + V_i^T\lambda\right)\right]}\begin{bmatrix} X_i \\ V_i \end{bmatrix}\begin{bmatrix} X_i \\ V_i \end{bmatrix}^T\right] \end{aligned}$$

with $Z_i = \begin{bmatrix} Y_i \\ X_{1i} \end{bmatrix}$, $\delta = \begin{bmatrix} \gamma \\ \beta \end{bmatrix}$, and $\phi(\cdot)$ is the standard normal density. Notice the modified *Wald* statistic draws from the variance estimator under the null. The conditional score statistic is

$$CS = \frac{1}{n} \frac{\partial L\left(\tilde{\gamma}, \tilde{\beta}, 0, \hat{\Pi}\right)}{\partial \lambda^T} \hat{V}_0\left(\hat{\lambda}\right) \frac{\partial L\left(\tilde{\gamma}, \tilde{\beta}, 0, \hat{\Pi}\right)}{\partial \lambda}$$

where $\left(\tilde{\gamma}, \tilde{\beta}\right)$ are the standard probit maximum likelihood estimators. The conditional likelihood ratio statistic is

$$CLR = 2\left[L\left(\hat{\gamma}, \hat{\beta}, \hat{\lambda}, \hat{\Pi}\right) - L\left(\hat{\gamma}, \hat{\beta}, 0, \hat{\Pi}\right)\right]$$

As is typical (see chapter 3), the modified *Wald*, conditional score, and conditional likelihood ratio statistics have the same asymptotic properties.[8]

8.1.7 Strategic choice model

Amemiya [1974] and Heckman [1978] suggest resolving identification problems in simultaneous probit models by making the model recursive. Bresnahan and Reiss [1990] show that this approach rules out interesting interactions in strategic choice models. Alternatively, they propose modifying the error structure to identify unique equilibria in strategic, multi-person choice models.

Statistical analysis of strategic choice extends random utility analysis by adding game structure and Nash equilibrium strategies (Bresnahan and Reiss [1990, 1991] and Berry [1992]). McKelvey and Palfrey [1995] proposed quantal response equilibrium analysis by assigning extreme value (logistic) distributed random errors to players' strategies. Strategic error by the players makes the model amenable to statistical analysis as the likelihood function does not degenerate. Signorino [2003] extends the idea to political science by replacing extreme value errors with assignment of normally distributed errors associated with analyst uncertainty and/or private information regarding the players' utility for outcomes. Since analyst error due to unobservable components is ubiquitous in business and economic data and private information problems are typical in settings where accounting plays an important role, we focus on the game setting with analyst error and private information.

A simple two player, sequential game with analyst error and private information (combined as π) is depicted in figure 8.2. Player *A* moves first by playing either left (*l*) or right (*r*). Player *B* moves next but player *A*'s choice depends on the anticipated response of player *B* to player *A*'s move. For simplicity, assume $\pi_i \sim N\left(0, \sigma^2 I\right)$ where

$$\pi_i^T = \begin{bmatrix} \pi_{lLi}^A & \pi_{lLi}^B & \pi_{lRi}^A & \pi_{lRi}^B & \pi_{rLi}^A & \pi_{rLi}^B & \pi_{rRi}^A & \pi_{rRi}^B \end{bmatrix}$$

[8] Rivers and Vuong also identify three Hausman-type test statistics for endogeneity but their simulations suggest the modified *Wald*, conditional score, and conditional likelihood ratio statistics perform at least as well and in most cases better.

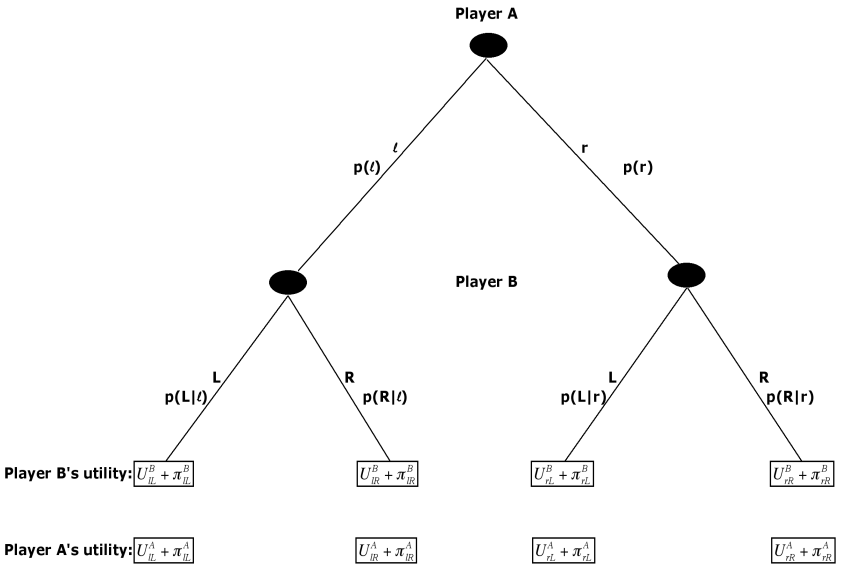

Figure 8.2: Strategic choice game tree

Since choice is scale-free (see chapter 5) maximum likelihood estimation proceeds with σ^2 normalized to 1.

The log-likelihood is

$$\sum_{i=1}^{n} Y_{lLi} \log\left(P_{lLi}\right) + Y_{lRi} \log\left(P_{lRi}\right) + Y_{rLi} \log\left(P_{rLi}\right) + Y_{rRi} \log\left(P_{rRi}\right)$$

where $Y_{jki} = 1$ if strategy j is played by *A* and k is played by *B* for sample i, and P_{jki} is the probability that strategy j is played by *A* and k is played by *B* for sample i. The latter requires some elaboration. Sequential play yields $P_{jk} = P_{(k|j)}P_j$. Now, only the conditional and marginal probabilities remain to be identified. Player *B*'s strategy depends on player *A*'s observed move. Hence,

$$\begin{aligned} P_{(L|l)} &= \Phi\left(\frac{U_{lL} - U_{lR}}{\sqrt{2\sigma^2}}\right) \\ P_{(R|l)} &= 1 - P_{(L|l)} \\ P_{(R|r)} &= 1 - P_{(L|r)} \\ P_{(L|r)} &= \Phi\left(\frac{U_{rL} - U_{rR}}{\sqrt{2\sigma^2}}\right) \end{aligned}$$

Player *A*'s strategy however depends on *B*'s response to *A*'s move. Therefore,

$$P_l = \Phi\left(\frac{P_{(L|l)}U_{lL} - P_{(L|r)}U_{rL} + P_{(R|l)}U_{lR} - P_{(R|r)}U_{rR}}{\sqrt{\left(P^2_{(L|l)} + P^2_{(L|r)} + P^2_{(R|l)} + P^2_{(R|r)}\right)\sigma^2}}\right)$$

and

$$P_r = 1 - P_l$$

Usually, the observable portion of expected utility is modeled as an index function; for Player *B* we have

$$U_{jk} - U_{jk'} = U_j^B = \left(X_{jk} - X_{jk'}\right)\beta_{jk}^B = X_j^B \beta_j^B$$

Since Player *B* moves following Player *A*, stochastic analysis of Player *B*'s utility is analogous to the simple binary discrete choice problem. That is,

$$\begin{aligned} P_{(L|l)} &= \Phi\left(\frac{U_{lL} - U_{lR}}{\sqrt{2\sigma^2}}\right) \\ &= \Phi\left(\frac{X_l^B \beta_l^B}{\sqrt{2}}\right) \end{aligned}$$

and

$$P_{(L|r)} = \Phi\left(\frac{X_r^B \beta_r^B}{\sqrt{2}}\right)$$

However, stochastic analysis of Player *A*'s utility is a little more subtle. Player *A*'s expected utility depends on Player *B*'s response to Player *A*'s move. Hence, Player *A*'s utilities are weighted by the conditional probabilities associated with Player *B*'s strategies. That is, from an estimation perspective the regressors X interact with the conditional probabilities to determine the coefficients in Player *A*'s index function.

$$U_{jk} - U_{j'k} = X_{jk}\beta_{jk}^A - X_{j'k}\beta_{j'k}^A$$

Consequently, Player *A*'s contribution to the likelihood function is a bit more complex than that representing Player *B*'s utilities.[9] Stochastic analysis of Player *A*'s strategy is

$$\begin{aligned} P_l &= \Phi\left(\frac{P_{(L|l)}U_{lL} - P_{(L|r)}U_{rL} + P_{(R|l)}U_{lR} - P_{(R|r)}U_{rR}}{\sqrt{\left(P_{(L|l)}^2 + P_{(L|r)}^2 + P_{(R|l)}^2 + P_{(R|r)}^2\right)\sigma^2}}\right) \\ &= \Phi\left(\frac{\begin{array}{c} P_{(L|l)}X_{lL}\beta_{lL}^A - P_{(L|r)}X_{rL}\beta_{rL}^A \\ +P_{(R|l)}X_{lR}\beta_{lR}^A - P_{(R|r)}X_{rR}\beta_{rR}^A \end{array}}{\sqrt{\left(P_{(L|l)}^2 + P_{(L|r)}^2 + P_{(R|l)}^2 + P_{(R|r)}^2\right)}}\right) \end{aligned}$$

[9] Recall the analysis is stochastic because the analyst doesn't observe part of the agents' utilities. Likewise, private information produces agent uncertainty regarding the other player's utility. Hence, private information produces a similar stochastic analysis. This probabilistic nature ensures that the likelihood doesn't degenerate even in a game of pure strategies.

Example 8.1 *Consider a simple experiment comparing a sequential strategic choice model with standard binary choice models for each player. We generated* 200 *simulated samples of size* $n = 2,000$ *with uniformly distributed regressors and standard normal errors. In particular,*

$$\begin{aligned} X_l^B &\sim U(-2,2) \\ X_r^B &\sim U(-5,5) \end{aligned}$$

$$X_{lL}^A, X_{lR}^A, X_{rL}^A, X_{rR}^A \sim U(-3,3)$$

and

$$\begin{aligned} \beta_l^B &= \begin{bmatrix} -0.5 & 1 \end{bmatrix}^T \\ \beta_r^B &= \begin{bmatrix} 0.5 & -1 \end{bmatrix}^T \\ \beta^A &= \begin{bmatrix} 0.5 & 1 & 1 & -1 & -1 \end{bmatrix}^T \end{aligned}$$

where the leading element of each vector is an intercept β_0.[10] *Results (means, standard deviations, and the* 0.01 *and* 0.99 *quantiles) are reported in tables 8.1 and 8.2. The standard discrete choice (DC) estimates seem to be more systemati-*

Table 8.1: Strategic choice analysis for player B

	β_{l0}^B	β_l^B	β_{r0}^B	β_r^B
parameter	-0.5	1	0.5	-1
SC mean	-0.482	0.932	0.460	-0.953
DC mean	-0.357	0.711	0.354	-0.713
SC std dev	0.061	0.057	0.101	0.059
DC std dev	0.035	0.033	0.050	0.030
SC $\begin{pmatrix} 0.01, \\ 0.99 \end{pmatrix}$ quantiles	$\begin{pmatrix} -0.62, \\ -0.34 \end{pmatrix}$	$\begin{pmatrix} 0.80, \\ 1.10 \end{pmatrix}$	$\begin{pmatrix} 0.22, \\ 0.69 \end{pmatrix}$	$\begin{pmatrix} -1.10, \\ 0.82 \end{pmatrix}$
DC $\begin{pmatrix} 0.01, \\ 0.99 \end{pmatrix}$ quantiles	$\begin{pmatrix} -0.43, \\ -0.29 \end{pmatrix}$	$\begin{pmatrix} 0.65, \\ 0.80 \end{pmatrix}$	$\begin{pmatrix} 0.23, \\ 0.47 \end{pmatrix}$	$\begin{pmatrix} -0.79, \\ -0.64 \end{pmatrix}$

cally biased towards zero. Tables 8.3 and 8.4 expressly compare the parameter estimate differences between the strategic choice model (SC) and the discrete choice models (DC). Hence, not only are the standard discrete choice parameter estimates biased toward zero but also there is almost no overlap with the $(0.01, 0.99)$ *interval estimates for the strategic choice model.*

As in the case of conditionally-heteroskedastic probit (see chapter 5), marginal probability effects of regressors are likely to be nonmonotonic due to cross agent

[10] The elements of β^A correspond to $\begin{bmatrix} intercept & \beta_{lL}^A & \beta_{lR}^A & \beta_{rL}^A & \beta_{rR}^A \end{bmatrix}$ where the intercept is the mean difference in observed utility (conditional on the regressors) between strategies l and r.

Table 8.2: Strategic choice analysis for player A

	β_0^A	β_{lL}^A	β_{lR}^A
parameter	0.5	1	1
SC mean	0.462	0.921	0.891
DC mean	0.304	0.265	0.360
SC std dev	0.044	0.067	0.053
DC std dev	0.032	0.022	0.021
SC $\begin{pmatrix} 0.01, \\ 0.99 \end{pmatrix}$ quantiles	$\begin{pmatrix} 0.34, \\ 0.56 \end{pmatrix}$	$\begin{pmatrix} 0.78, \\ 1.08 \end{pmatrix}$	$\begin{pmatrix} 0.78, \\ 1.01 \end{pmatrix}$
DC $\begin{pmatrix} 0.01, \\ 0.99 \end{pmatrix}$ quantiles	$\begin{pmatrix} 0.23, \\ 0.38 \end{pmatrix}$	$\begin{pmatrix} 0.23, \\ 0.32 \end{pmatrix}$	$\begin{pmatrix} 0.31, \\ 0.41 \end{pmatrix}$

	β_{rL}^A	β_{rR}^A
parameter	-1	-1
SC mean	-0.911	-0.897
DC mean	-0.352	-0.297
SC std dev	0.053	0.058
DC std dev	0.022	0.023
SC $\begin{pmatrix} 0.01, \\ 0.99 \end{pmatrix}$ quantiles	$\begin{pmatrix} -1.04, \\ -0.79 \end{pmatrix}$	$\begin{pmatrix} -1.05, \\ -0.78 \end{pmatrix}$
DC $\begin{pmatrix} 0.01, \\ 0.99 \end{pmatrix}$ quantiles	$\begin{pmatrix} -0.40, \\ -0.30 \end{pmatrix}$	$\begin{pmatrix} -0.34, \\ -0.25 \end{pmatrix}$

probability interactions. Indeed, comparison of marginal effects for strategic probit with those of standard binary probit helps illustrate the contrast between statistical analysis of strategic and single person decisions. For the sequential strategic game above, the marginal probabilities for player *A*'s regressors include

$$\frac{\partial P_{lLj}}{\partial X_{ikj}^A} = P_{(L|l)j} f_{lj}\left(sign_j\right) P_{(k|i)j} \beta_{ik}^A Den^{-\frac{1}{2}}$$

$$\frac{\partial P_{lRj}}{\partial X_{ikj}^A} = P_{(R|l)j} f_{lj}\left(sign_j\right) P_{(k|i)j} \beta_{ik}^A Den^{-\frac{1}{2}}$$

$$\frac{\partial P_{rLj}}{\partial X_{ikj}^A} = P_{(L|r)j} f_{rj}\left(sign_j\right) P_{(k|i)j} \beta_{ik}^A Den^{-\frac{1}{2}}$$

$$\frac{\partial P_{rRj}}{\partial X_{ikj}^A} = P_{(R|r)j} f_{rj}\left(sign_j\right) P_{(k|i)j} \beta_{ik}^A Den^{-\frac{1}{2}}$$

where $sign_j$ is the sign of the X_{ikj} term in P_{mnj}, f_{ij} and $f_{(k|i)j}$ is the standard normal density function evaluated at the same arguments as P_{ij} and $P_{(k|i)j}$,

$$Den = \left(P_{(L|l)j}^2 + P_{(L|r)j}^2 + P_{(R|l)j}^2 + P_{(R|r)j}^2\right)$$

Table 8.3: Parameter differences in strategic choice analysis for player B

SC-DC	β^B_{l0}	β^B_l	β^B_{r0}	β^B_r
parameter	-0.5	1	0.5	-1
mean	-0.125	0.221	0.106	-0.241
std dev	0.039	0.041	0.079	0.049
quantiles	$\begin{pmatrix} 0.01, \\ 0.99 \end{pmatrix}$	$\begin{pmatrix} -0.22, \\ -0.03 \end{pmatrix}$	$\begin{pmatrix} 0.13, \\ 0.33 \end{pmatrix}$	$\begin{pmatrix} -0.06, \\ 0.29 \end{pmatrix}$	$\begin{pmatrix} -0.36, \\ -0.14 \end{pmatrix}$

Table 8.4: Parameter differences in strategic choice analysis for player A

SC-DC	β^A_0	β^A_{lL}	β^A_{lR}
parameter	0.5	1	1
mean	0.158	0.656	0.531
std dev	0.027	0.056	0.044
$(0.01, 0.99)$ quantiles	$(0.10, 0.22)$	$(0.54, 0.80)$	$(0.43, 0.62)$
SC-DC	β^A_{rL}	β^A_{rR}	
parameter	-1	-1	
mean	-0.559	-0.600	
std dev	0.045	0.050	
$(0.01, 0.99)$ quantiles	$(-0.67, -0.46)$	$(-0.73, -0.49)$	

and

$$Num = \begin{pmatrix} P_{(L|l)j} X^A_{lLj} \beta^A_{lL} - P_{(L|r)j} X^A_{rLj} \beta^A_{rL} \\ +P_{(R|l)j} X^A_{lRj} \beta^A_{lR} - P_{(R|r)j} X^A_{rRj} \beta^A_{rR} \end{pmatrix}$$

Similarly, the marginal probabilities with respect to player *B*'s regressors include

$$\begin{aligned} \frac{\partial P_{lLj}}{\partial X^B_{lj}} &= f_{(L|l)j} \frac{\beta^B_l}{\sqrt{2}} P_{lj} + P_{(L|l)j} f_{lj} f_{(L|l)j} \frac{\beta^B_l}{\sqrt{2}} \\ &\quad \times \begin{Bmatrix} \left(X^A_{lLj} \beta^A_{lL} - X^A_{lRj} \beta^A_{lR} \right) Den^{-\frac{1}{2}} \\ -NumDen^{-\frac{3}{2}} \left(P_{(L|l)j} - P_{(R|l)j} \right) \end{Bmatrix} \end{aligned}$$

$$\begin{aligned} \frac{\partial P_{lLj}}{\partial X^B_{rj}} &= P_{(L|l)j} f_{lj} f_{(L|r)j} \frac{\beta^B_r}{\sqrt{2}} \\ &\quad \times \begin{Bmatrix} -\left(X^A_{rLj} \beta^A_{rL} - X^A_{rRj} \beta^A_{rR} \right) Den^{-\frac{1}{2}} \\ -NumDen^{-\frac{3}{2}} \left(P_{(L|r)j} - P_{(R|r)j} \right) \end{Bmatrix} \end{aligned}$$

$$\frac{\partial P_{lRj}}{\partial X^B_{lj}} = f_{(R|l)j}\frac{-\beta^B_l}{\sqrt{2}}P_{lj} + P_{(R|l)j}f_{lj}f_{(R|l)j}\frac{\beta^B_l}{\sqrt{2}} \times \left\{ \begin{array}{c} \left(X^A_{lLj}\beta^A_{lL} - X^A_{lRj}\beta^A_{lR}\right)Den^{-\frac{1}{2}} \\ -NumDen^{-\frac{3}{2}}\left(P_{(L|l)j} - P_{(R|l)j}\right) \end{array} \right\}$$

$$\frac{\partial P_{lRj}}{\partial X^B_{rj}} = P_{(R|l)j}f_{lj}f_{(R|r)j}\frac{-\beta^B_l}{\sqrt{2}} \times \left\{ \begin{array}{c} \left(X^A_{rLj}\beta^A_{rL} - X^A_{rRj}\beta^A_{rR}\right)Den^{-\frac{1}{2}} \\ +NumDen^{-\frac{3}{2}}\left(P_{(L|r)j} - P_{(R|r)j}\right) \end{array} \right\}$$

$$\frac{\partial P_{rLj}}{\partial X^B_{lj}} = P_{(L|r)j}f_{rj}f_{(L|l)j}\frac{\beta^B_l}{\sqrt{2}} \times \left\{ \begin{array}{c} -\left(X^A_{lLj}\beta^A_{lL} - X^A_{lRj}\beta^A_{lR}\right)Den^{-\frac{1}{2}} \\ +NumDen^{-\frac{3}{2}}\left(P_{(L|l)j} - P_{(R|l)j}\right) \end{array} \right\}$$

$$\frac{\partial P_{rLj}}{\partial X^B_{rj}} = f_{(L|r)j}\frac{\beta^B_r}{\sqrt{2}}P_{rj} + P_{(L|r)j}f_{rj}f_{(L|r)j}\frac{\beta^B_r}{\sqrt{2}} \times \left\{ \begin{array}{c} \left(X^A_{rLj}\beta^A_{rL} - X^A_{rRj}\beta^A_{rR}\right)Den^{-\frac{1}{2}} \\ +NumDen^{-\frac{3}{2}}\left(P_{(L|r)j} - P_{(R|r)j}\right) \end{array} \right\}$$

$$\frac{\partial P_{rRj}}{\partial X^B_{lj}} = P_{(R|r)j}f_{rj}f_{(R|l)j}\frac{-\beta^B_l}{\sqrt{2}} \times \left\{ \begin{array}{c} \left(X^A_{lLj}\beta^A_{lL} - X^A_{lRj}\beta^A_{lR}\right)Den^{-\frac{1}{2}} \\ -NumDen^{-\frac{3}{2}}\left(P_{(L|l)j} - P_{(R|l)j}\right) \end{array} \right\}$$

$$\frac{\partial P_{rRj}}{\partial X^B_{rj}} = f_{(R|r)j}\frac{-\beta^B_r}{\sqrt{2}}P_{rj} + P_{(R|r)j}f_{rj}f_{(R|r)j}\frac{\beta^B_r}{\sqrt{2}} \times \left\{ \begin{array}{c} \left(X^A_{rLj}\beta^A_{rL} - X^A_{rRj}\beta^A_{rR}\right)Den^{-\frac{1}{2}} \\ +NumDen^{-\frac{3}{2}}\left(P_{(L|r)j} - P_{(R|r)j}\right) \end{array} \right\}$$

Clearly, analyzing responses to anticipated moves by other agents who themselves are anticipating responses changes the game. In other words, endogeneity is fundamental to the analysis of strategic play.

Multi-person strategic choice models can be extended in a variety of ways including simultaneous move games, games with learning, games with private information, games with multiple equilibria, etc. (Bresnahan and Reiss [1990], Tamer [2003]). The key point is that strategic interaction is endogenous and standard (single-person) discrete choice models (as well as simultaneous probit models) ignore this source of endogeneity.

8.1.8 Sample selection

A common problem involves estimation of β for the model

$$Y^* = X\beta + \varepsilon$$

however sample selection results in Y being observed only for individuals receiving treatment (when $D = 1$). The data are censored but not at a fixed value (as in a Tobit problem; see chapter 5). Treating sample selection D as an exogenous variable is inappropriate if the unobservable portion of the selection equation, say V_D, is correlated with unobservables in the outcome equation ε.

Heckman [1974, 1976, 1979] addressed this problem and proposed the classic two stage approach. In the first stage, estimate the selection equation via probit. Identification in this model does not depend on an exclusion restriction (Z need not include variables appropriately excluded from X) but if instruments are available they're likely to reduce collinearity issues.

To fix ideas, identification conditions include

Condition 8.4 (X, D) *are always observed,* Y_1 *is observed when* $D = 1$ *(*$D^* > 1$*),*

Condition 8.5 (ε, V_D) *are independent of* X *with mean zero,*

Condition 8.6 $V_D \sim N(0, 1)$*,*

Condition 8.7 $E[\varepsilon \mid V_D] = \gamma_1 V_D$.[11]

The two-stage procedure estimates θ from a first stage probit.

$$D^* = Z\theta - V_D$$

These estimates $\widehat{\theta}$ are used to construct the inverse Mills ratio $\lambda_i = \frac{\phi(Z_i\widehat{\theta})}{\Phi(Z_i\widehat{\theta})}$ which is utilized as a covariate in the second stage regression.

$$Y_1 = X\beta + \gamma\lambda + \eta$$

where $E[\eta \mid X, \lambda] = 0$. Given proper specification of the selection equation (including normality of V_D), Heckman shows that the two-step estimator is asymptotically consistent (if not efficient) for β, the focal parameter of the analysis.[12]

---

[11] Bivariate normality of (ε, V_D) is often posed, but strictly speaking is not required for identification.

[12] It should be noted that even though Heckman's two stage approach is commonly employed to estimate treatment effects (discussed later), treatment effects are not the object of the sample selection model. In fact, since treatment effects involve counterfactuals and we have no data from which to identify population parameters for the counterfactuals, treatment effects in this setting are unassailable.

A semi-nonparametric alternative

Concern over reliance on normal probability assignment to unobservables in the selection equation as well as the functional form of the outcome equation, has resulted in numerous proposals to relax these conditions. Ahn and Powell [1993] provide an alternative via their semi-nonparametric two stage approach. However, nonparametric identification involves an exclusion restriction or, in other words, at least one instrument. That is, (at least) one variable included in the selection equation is properly omitted from the outcome equation. Intuitively, this is because the selection equation could be linear and the second stage would then involve colinear regressors. Ahn and Powell propose a nonparametric selection model coupled with a partial index outcome (second stage) model. The first stage selection index is estimated via nonparametric regression

$$\widehat{g}_i = \frac{\sum_{j=1}^{n} K\left(\frac{w_i - w_j}{h_1}\right) D_j}{\sum_{j=1}^{n} K\left(\frac{w_i - w_j}{h_1}\right)}$$

The second stage uses instruments Z, which are functions of W, and the estimated selection index.

$$\widehat{\beta} = \left[\widehat{S}_{XX}\right]^{-1} \widehat{S}_{XY}$$

where

$$\begin{aligned} \widehat{S}_{XX} &= \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \widehat{\omega}_{ij} \left(z_i - z_j\right) \left(x_i - x_j\right)^T \\ \widehat{S}_{XY} &= \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \widehat{\omega}_{ij} \left(z_i - z_j\right) \left(y_i - y_j\right) \end{aligned}$$

and

$$\widehat{\omega}_{ij} = \frac{1}{h_2} K\left(\frac{\widehat{g}_i - \widehat{g}_j}{h_2}\right) D_i D_j$$

Ahn and Powell show the instrumental variable density-weighted average derivative estimator for β achieves root-n convergence (see the discussion of nonparametric regression and Powell, Stock, and Stoker's [1989] instrumental variable density-weighted average derivative estimator in chapter 6).

8.1.9 Duration models

Sometimes the question involves how long to complete a task. For instance, how long to complete an audit (internal or external), how long to turn around a distressed business unit or firm, how long to complete custom projects, how long will a recession last, and so on. Such questions can be addressed via duration models.

The most popular duration models are proportional hazard models. Analysis of such questions can be plagued by the same challenges of endogeneity and unobservable heterogeneity as other regression models.

We'll explore a standard version of the model and a couple of relaxations. Namely, we'll look at Horowitz's [1999] semiparametric proportional hazard (classical) model with unobserved heterogeneity and Campolieti's [2001] Bayesian semiparametric duration model with unobserved heterogeneity.

Unconditional hazard rate

The probability that an individual leaves a state during a specified interval given the individual was previously in the particular state is

$$\Pr\left(t < T < t + h \mid T > t\right)$$

The hazard function, then is $\lambda(t) = \lim_{h \to 0} \frac{\Pr(t<T<t+h|T>t)}{h}$, the instantaneous rate of leaving per unit of time. To relate this to the hazard function write

$$\begin{aligned} \Pr\left(t < T < t + h \mid T > t\right) &= \frac{\Pr\left(t < T < t + h\right)}{\Pr\left(T > t\right)} \\ &= \frac{F\left(t + h\right) - F\left(t\right)}{1 - F\left(t\right)} \end{aligned}$$

where F is the probability distribution function and f is the density function for T. When F is differentiable, the hazard rate is seen as the limit of the right hand side divided by h as h approaches 0 (from above)

$$\begin{aligned} \lambda\left(t\right) &= \lim_{h \to 0} \frac{F\left(t + h\right) - F\left(t\right)}{h} \frac{1}{1 - F\left(t\right)} \\ &= \frac{f\left(t\right)}{1 - F\left(t\right)} \end{aligned}$$

To move this closer to a version of the model that is frequently employed define the integrated hazard function as[13]

$$\Lambda\left(t\right) \equiv \int_0^t \lambda\left(s\right) ds$$

Now,

$$\frac{d \int_0^t \lambda\left(s\right) ds}{dt} = \lambda\left(t\right)$$

[13]The lower limit of integration is due to $F\left(0\right) = 0$.

and

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = -\frac{d \ln S(t)}{dt}$$

Hence, $-\ln S(t) = \int_0^t \lambda(s)\, ds$ and the survivor function is

$$S(t) = \exp\left[-\int_0^t \lambda(s)\, ds\right]$$

Since $S(t) = 1 - F(t)$, the distribution function can be written

$$F(t) = 1 - \exp\left[-\int_0^t \lambda(s)\, ds\right]$$

and the density function (following differentiation) can be written

$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(s)\, ds\right]$$

And all probabilities can conveniently be expressed in terms of the hazard function. For instance,

$$\begin{aligned} \Pr(T \geq t_2 \mid T \geq t_1) &= \frac{1 - F(t_2)}{1 - F(t_1)} \\ &= \exp\left[-\int_{t_1}^{t_2} \lambda(s)\, ds\right] \end{aligned}$$

for $t_2 > t_1$. The above discussion focuses on unconditional hazard rates but frequently we're interested in conditional hazard rates.

Regression (conditional hazard rate) models

Conditional hazard rate models may be parametric or essentially nonparametric (Cox [1972]). Parametric models focus on $\lambda(t \mid x)$ where the conditional distribution is known (typically, Weibull, exponential, or lognormal). Much conditional duration analysis is based on the proportional hazard model. The proportional hazard model relates the hazard rate for an individual with characteristics x to some (perhaps unspecified) baseline hazard rate by some positive function of x. Since, as seen above, the probability of change is an exponential function it is convenient to also express this positive function as an exponential function. The proportional hazard model then is

$$\lambda(t \mid x, u) = \lambda_0(t) \exp[-(x\beta + u)]$$

where λ is the hazard that $T = t$ conditional on observables $X = x$ and unobservables $U = u$, λ_0 is the baseline hazard function, and β is a vector of (constant) parameters.

A common parameterization follows from a Weibull (α, γ) distribution. Then, the baseline hazard rate is

$$\lambda_0(t) = \frac{\alpha}{\gamma}\left(\frac{t}{\gamma}\right)^{\alpha-1}$$

and the hazard rate is

$$\lambda(t \mid x_1) = \frac{\alpha}{\gamma}\left(\frac{t}{\gamma}\right)^{\alpha-1}\exp\left[-x_1\beta_1\right]$$

The latter is frequently rewritten by adding a vector of ones to x_1 (denote this x) and absorbing γ (denote the augmented parameter vector β) so that

$$\lambda(t \mid x) = \alpha t^{\alpha-1}\exp\left[-x\beta\right]$$

This model can be estimated in standard fashion via maximization of the log-likelihood.

Since Cox's [1972] method doesn't require the baseline hazard function to be estimated, the method is essentially nonparametric in nature. Heterogeneity stems from observable and unobservable components of

$$\exp\left[-\left(x\beta + u\right)\right]$$

Cox's method accommodates observed heterogeneity but assumes unobserved homogeneity. As usual, unobservable heterogeneity can be problematic as conditional exchangeability is difficult to satisfy. Therefore, we look to alternative approaches to address unobservable heterogeneity.

Horowitz [1999] describes an approach for nonparametrically estimating the baseline hazard rate λ_0 and the integrated hazard rate Λ. In addition, the distribution function F and density function f for U, the unobserved source of heterogeneity with time-invariant covariates x, are nonparametrically estimated. The approach employs kernel density estimation methods similar to those discussed in chapter 6. As the estimators for F and f are slow to converge, the approach calls for large samples.

Campolieti [2001] addresses unobservable heterogeneity and the unknown error distribution via an alternative tack - Bayesian data augmentation (similar to that discussed in chapter 7). Discrete duration is modeled as a sequence of multi-period probits where duration dependence is accounted for via nonparametric estimation of the baseline hazard. A Dirichlet process prior supplies the nonparametric nature to the baseline hazard estimation.

8.1.10 Latent IV

Sometimes (perhaps frequently) it is difficult to identify instruments. Of course, this makes instrumental variable (*IV*) estimation unattractive. However, latent *IV*

methods may help to overcome this deficiency. If the endogenous data are nonnormal (exhibit skewness and/or multi-modality) then it may be possible to decompose the data into parts that are unrelated to the regressor error and the part that is related. This is referred to as latent *IV*. Ebbes [2004] reviews the history of latent *IV* related primarily to measurement error and extends latent *IV* via analysis and simulation to various endogeneity concerns, including self selection.

8.2 Selectivity and treatment effects

This chapter is already much too long so next we only briefly introduce our main thesis - analysis of treatment effects in the face of potential endogeneity. Treatment effects are a special case of causal effects which we can under suitable conditions address without a fully structural model. As such treatment effects are both simple and challenging at the same time. Discussion of treatment effects occupies much of our focus in chapters 9 through 12.

First, we describe a prototypical setting. Then, we identify some typical treatment effects followed by a brief review of various identification conditions.

Suppose the *DGP* is
outcome equations:

$$Y_j = \mu_j(X) + V_j, j = 0, 1$$

selection equation:[14]

$$D^* = \mu_D(Z) - V_D$$

observable response:

$$Y = DY_1 + (1 - D)Y_0$$

where

$$D = \begin{array}{ll} 1 & D^* > 0 \\ 0 & otherwise \end{array}$$

In the binary case, the treatment effect is the effect on outcome of treatment compared with no treatment, $\Delta = Y_1 - Y_0$. Typical average treatment effects include *ATE*, *ATT*, and *ATUT*.[15] *ATE* refers to the average treatment effect,

$$ATE = E[\Delta] = E[Y_1 - Y_0]$$

In other words, the average effect on outcome of treatment for a random draw from the population. *ATT* refers to the average treatment effect on the treated,

$$ATT = E[\Delta \mid D = 1] = E[Y_1 - Y_0 \mid D = 1]$$

[14] We'll stick with binary choice for simplicity, though this can be readily generalized to the multinomial case.

[15] Additional treatment effects are discussed in subsequent chapters.

In other words, the average effect on outcome of treatment for a random draw from the subpopulation selecting (or assigned) treatment. *ATUT* refers to the average treatment effect on the untreated,

$$ATUT = E\left[\Delta \mid D = 0\right] = E\left[Y_1 - Y_0 \mid D = 0\right]$$

In other words, the average effect on outcome of treatment for a random draw from the subpopulation selecting (or assigned) no treatment.

The simplest approaches (strongest data conditions) involve ignorable treatment (sometimes referred to as selection on observables). These approaches include exogenous dummy variable regression, nonparametric regression, propensity score, propensity score matching, and control function methods. Various conditions and relaxations are discussed in the next chapter.

Instrumental variables (*IV*) are a common treatment effect identification strategy when ignorability is ill-suited to the data at hand. *IV* strategies accommodate homogeneous response at their simplest (strongest conditions) or unobservable heterogeneity at their most challenging (weakest conditions). Various *IV* approaches including standard *IV*, propensity score *IV*, control function *IV*, local *IV*, and Bayesian data augmentation are discussed in subsequent chapters. Heckman and Vytlacil [2005] argue that each of these strategies potentially estimate different treatment effects under varying conditions including continuous treatment and general equilibrium treatment effects.

8.3 Why bother with endogeneity?

Despite great effort by analysts, experiments frequently fail to identify substantive endogenous effects (Heckman [2000, 2001]). Why then do we bother? In this section we present a couple of stylized examples that depict some of our concerns regarding ignoring endogeneity. A theme of these examples is that failing to adequately attend to the *DGP* may produce a Simpson's paradox result.

8.3.1 Sample selection example

Suppose a firm has two production facilities, *A* and *B*. Facility *A* is perceived to be more efficient (produces a higher proportion of non-defectives). Consequently, production has historically been skewed in favor of facility *A*. The firm is interested in improving production efficiency, and particularly, improving facility *B*. Management has identified new production technology and is interested in whether the new technology improves production efficiency. Production using the new technology is skewed toward facility *B*. This "experiment" generates the data depicted in table 8.5.

Is the new technology more effective than the old technology? What is the technology treatment effect? As management knows, the choice of facility is important. The facility is a sufficiently important variable that its inclusion illuminates

Table 8.5: Production data: Simpson's paradox

	Facility *A*		Facility *B*		Total	
Technology	New	Old	New	Old	New	Old
Successes	10	120	133	25	143	145
Trials	10	150	190	50	200	200
% successes	100	80	70	50	71.5	72.5

the production technology treatment effect but its exclusion obfuscates the effect.[16] Aggregate results reported under the "Total" columns are misleading. For facility *A*, on average, there is a 20% improvement from the new technology. Likewise, for facility *B*, there is an average 20% improvement from the new technology.

Now, suppose an analyst collects the data but is unaware that there are two different facilities (the analyst only has the last two columns of data). What conclusion regarding the technology treatment effect is likely to be reached? This level of aggregation results in a serious omitted variable problem that leads to inferences opposite what the data suggest. This, of course, is a classic Simpson's paradox result produced via a sample selection problem. The data are not generated randomly but rather reflect management's selective "experimentation" on production technology.

8.3.2 Tuebingen-style treatment effect examples

Treatment effects are the focus of much economic self-selection analyses. When we ask what is the potential outcome response (Y) to treatment? — we pose a treatment effect question. A variety of treatment effects may be of interest. To setup the next example we define a few of the more standard treatment effects that may be of interest.

Suppose treatment is binary ($D = 1$ for treatment, $D = 0$ for untreated), for simplicity. As each individual is only observed either with treatment or without treatment, the observed outcome is

$$Y = DY_1 + (1 - D)\, Y_0$$

where

$$Y_1 = \mu_1 + V_1$$

is outcome response with treatment,

$$Y_0 = \mu_0 + V_0$$

is outcome response without treatment, μ_j is observed outcome for treatment $j = 0$ or 1, and V_j is unobserved (by the analyst) outcome for treatment j. Now, the

[16] This is an example of ignorable treatment (see ch. 9 for additional details).

treatment effect is

$$\begin{aligned}\Delta &= Y_1 - Y_0 \\ &= \mu_1 + V_1 - \mu_0 - V_0 \\ &= (\mu_1 - \mu_0) + (V_1 - V_0)\end{aligned}$$

an individual's (potential) outcome response to a change in treatment from regime 0 to regime 1. Note $(\mu_1 - \mu_0)$ is the population level effect (based on observables) and $(V_1 - V_0)$ is the individual-specific gain. That is, while treatment effects focus on potential gains for an individual, the unobservable nature of counterfactuals often lead analysts to focus on population level parameters.

The average treatment effect

$$ATE = E[\Delta] = E[Y_1 - Y_0]$$

is the average response to treatment for a random sample from the population. Even though seemingly cumbersome, we can rewrite *ATE* in a manner that illuminates connections with other treatment effects,

$$\begin{aligned}E[Y_1 - Y_0] &= E[Y_1 - Y_0|D = 1]\Pr(D = 1) \\ &\quad + E[Y_1 - Y_0|D = 0]\Pr(D = 0)\end{aligned}$$

The average treatment effect on the treated

$$ATT = E[\Delta|D = 1] = E[Y_1 - Y_0|D = 1]$$

is the average response to treatment for a sample of individuals that choose (or are assigned) treatment. Selection (or treatment) is assumed to follow some *RUM* (random utility model; see chapter 5), $D^* = Z - V_D$ where D^* is latent utility index associated with treatment, Z is the observed portion, V_D is the part unobserved by the analyst, and $D = 1$ if $D^* > 0$ or $D = 0$ otherwise.

The average treatment effect on the untreated

$$ATUT = E[\Delta|D = 0] = E[Y_1 - Y_0|D = 0]$$

is the average response to treatment for a sample of individuals that choose (or are assigned) no treatment. Again, selection (or treatment) is assumed to follow some *RUM*, $D^* = Z - V_D$.

To focus attention on endogeneity, it's helpful to identify what is estimated by *OLS* (exogenous treatment). Exogenous dummy variable regression estimates

$$OLS = E[Y_1|D = 1] - E[Y_0|D = 0]$$

An important question is when and to what extent is OLS a biased measure of the treatment effect.

Bias in the OLS estimate for ATT is

$$\begin{aligned}OLS &= ATT + bias_{ATT} \\ &= E[Y_1|D = 1] - E[Y_0|D = 0] \\ &= E[Y_1|D = 1] - E[Y_0|D = 1] + \{E[Y_0|D = 1] - E[Y_0|D = 0]\}\end{aligned}$$

Hence,

$$bias_{ATT} = \{E[Y_0|D=1]] - E[Y_0|D=0]\}$$

Bias in the OLS estimate for $ATUT$ is

$$\begin{aligned} OLS &= ATUT + bias_{ATUT} \\ &= E[Y_1|D=1] - E[Y_0|D=0] \\ &= E[Y_1|D=0] - E[Y_0|D=0] + \{E[Y_1|D=1] - E[Y_1|D=0]\} \end{aligned}$$

Hence,

$$bias_{ATUT} = \{E[Y_1|D=1] - E[Y_1|D=0]\}$$

Since

$$\begin{aligned} ATE &= \Pr(D=1) E[Y_1 - Y_0|D=1] + \Pr(D=0) E[Y_1 - Y_0|D=0] \\ &= \Pr(D=1) ATT + \Pr(D=0) ATUT \end{aligned}$$

bias in the OLS estimate for ATE can be written as a function of the bias in other treatment effects

$$bias_{ATE} = \Pr(D=1) bias_{ATT} + \Pr(D=0) bias_{ATUT}$$

Now we explore some examples.

Case 1

The setup involves simple (no regressors), discrete probability and outcome structure. It is important for identification of counterfactuals that outcome distributions are not affected by treatment selection. Hence, outcomes Y_0 and Y_1 vary only between states (and not by D within a state) as described, for instance, in table 8.6. Key components, the treatment effects, and any bias for case 1 are reported in table 8.7. Case 1 exhibits no endogeneity bias. This, in part, can be attributed to the idea that Y_1 is constant. However, even with Y_1 constant, this is a knife-edge result as the next cases illustrate.

Case 2

Case 2, depicted in table 8.8, perturbs the state two conditional probabilities only. Key components, the treatment effects, and any bias for case 2 are reported in table 8.9. Hence, a modest perturbation of the probability structure produces endogeneity bias in both ATT and ATE (but of course not $ATUT$ as Y_1 is constant).

Table 8.6: Tuebingen example case 1: ignorable treatment

State (s)	*one*		*two*		*three*	
$\Pr(Y, D, s)$	0.0272	0.0128	0.32	0.0	0.5888	0.0512
D	0	1	0	1	0	1
Y	0	1	1	1	2	1
Y_0	0	0	1	1	2	2
Y_1	1	1	1	1	1	1

Table 8.7: Tuebingen example case 1 results: ignorable treatment

Results	Key components
$ATE = E[Y_1 - Y_0]$ $= -0.6$	$p = \Pr(D = 1) = 0.064$
$ATT = E[Y_1 - Y_0 \mid D = 1]$ $= -0.6$	$E[Y_1 \mid D = 1] = 1.0$
$ATUT = E[Y_1 - Y_0 \mid D = 0]$ $= -0.6$	$E[Y_1 \mid D = 0] = 1.0$
$OLS = E[Y_1 \mid D = 1]$ $-E[Y_0 \mid D = 0] = -0.6$	$E[Y_1] = 1.0$
$bias_{ATT} = E[Y_0 \mid D = 1]$ $-E[Y_0 \mid D = 0] = 0.0$	$E[Y_0 \mid D = 1] = 1.6$
$bias_{ATUT} = E[Y_1 \mid D = 1]$ $-E[Y_1 \mid D = 0] = 0.0$	$E[Y_0 \mid D = 0] = 1.6$
$bias_{ATE} = pbias_{ATT}$ $+ (1-p)\, bias_{ATUT} = 0.0$	$E[Y_0] = 1.6$

Table 8.8: Tuebingen example case 2: heterogeneous response

State (s)	*one*		*two*		*three*	
$\Pr(Y, D, s)$	0.0272	0.0128	0.224	0.096	0.5888	0.0512
D	0	1	0	1	0	1
Y	0	1	1	1	2	1
Y_0	0	0	1	1	2	2
Y_1	1	1	1	1	1	1

Table 8.9: Tuebingen example case 2 results: heterogeneous response

Results	Key components
$ATE = E[Y_1 - Y_0]$ $= -0.6$	$p = \Pr(D = 1) = 0.16$
$ATT = E[Y_1 - Y_0 \mid D = 1]$ $= -0.24$	$E[Y_1 \mid D = 1] = 1.0$
$ATUT = E[Y_1 - Y_0 \mid D = 0]$ $= -0.669$	$E[Y_1 \mid D = 0] = 1.0$
$OLS = E[Y_1 \mid D = 1]$ $-E[Y_0 \mid D = 0] = -0.669$	$E[Y_1] = 1.0$
$bias_{ATT} = E[Y_0 \mid D = 1]$ $-E[Y_0 \mid D = 0] = -0.429$	$E[Y_0 \mid D = 1] = 1.24$
$bias_{ATUT} = E[Y_1 \mid D = 1]$ $-E[Y_1 \mid D = 0] = 0.0$	$E[Y_0 \mid D = 0] = 1.669$
$bias_{ATE} = pbias_{ATT}$ $+ (1-p)\, bias_{ATUT} = -0.069$	$E[Y_0] = 1.6$

Table 8.10: Tuebingen example case 3: more heterogeneity

State (s)	*one*		*two*		*three*	
$\Pr(Y, D, s)$	0.0272	0.0128	0.224	0.096	0.5888	0.0512
D	0	1	0	1	0	1
Y	0	1	1	1	2	0
Y_0	0	0	1	1	2	2
Y_1	1	1	1	1	0	0

Table 8.11: Tuebingen example case 3 results: more heterogeneity

Results	Key components
$ATE = E[Y_1 - Y_0] = -1.24$	$p = \Pr(D = 1) = 0.16$
$ATT = E[Y_1 - Y_0 \mid D = 1] = -0.56$	$E[Y_1 \mid D = 1] = 0.68$
$ATUT = E[Y_1 - Y_0 \mid D = 0] = -1.370$	$E[Y_1 \mid D = 0] = 0.299$
$OLS = E[Y_1 \mid D = 1] - E[Y_0 \mid D = 0] = -0.989$	$E[Y_1] = 0.36$
$bias_{ATT} = E[Y_0 \mid D = 1] - E[Y_0 \mid D = 0] = -0.429$	$E[Y_0 \mid D = 1] = 1.24$
$bias_{ATUT} = E[Y_1 \mid D = 1] - E[Y_1 \mid D = 0] = 0.381$	$E[Y_0 \mid D = 0] = 1.669$
$bias_{ATE} = p bias_{ATT} + (1 - p)\, bias_{ATUT} = 0.251$	$E[Y_0] = 1.6$

Case 3

Case 3, described in table 8.10, maintains the probability structure of case 2 but alters the outcomes with treatment Y_1. Key components, the treatment effects, and any bias for case 3 are reported in table 8.11. A modest change in the outcomes with treatment produces endogeneity bias in all three average treatment effects (ATT, ATE, and $ATUT$).

Case 4

Case 4 maintains the probability structure of case 3 but alters the outcomes with treatment Y_1 as described in table 8.12. Key components, the treatment effects, and any bias for case 4 are reported in table 8.13. Case 4 is particularly noteworthy as OLS indicates a negative treatment effect, while all standard treatment effects, ATE, ATT,and $ATUT$ are positive. The endogeneity bias is so severe that it produces a Simpson's paradox result. Failure to accommodate endogeneity results in inferences opposite the *DGP*. Could this *DGP* represent earnings management? While these examples may not be as rich and deep as Lucas' [1976] critique of econometric policy evaluation, the message is similar — *endogeneity matters*!

Table 8.12: Tuebingen example case 4: Simpson's paradox

State (s)	*one*		*two*		*three*	
$\Pr(Y, D, s)$	0.0272	0.0128	0.224	0.096	0.5888	0.0512
D	0	1	0	1	0	1
Y	0	1	1	1	2	2.3
Y_0	0	0	1	1	2	2
Y_1	1	1	1	1	2.3	2.3

Table 8.13: Tuebingen example case 4 results: Simpson's paradox

Results	Key components
$ATE = E[Y_1 - Y_0] = 0.232$	$p = \Pr(D = 1) = 0.16$
$ATT = E[Y_1 - Y_0 \mid D = 1] = 0.176$	$E[Y_1 \mid D = 1] = 1.416$
$ATUT = E[Y_1 - Y_0 \mid D = 0] = 0.243$	$E[Y_1 \mid D = 0] = 1.911$
$OLS = E[Y_1 \mid D = 1] - E[Y_0 \mid D = 0] = -0.253$	$E[Y_1] = 1.832$
$bias_{ATT} = E[Y_0 \mid D = 1] - E[Y_0 \mid D = 0] = -0.429$	$E[Y_0 \mid D = 1] = 1.24$
$bias_{ATUT} = E[Y_1 \mid D = 1] - E[Y_1 \mid D = 0] = -0.495$	$E[Y_0 \mid D = 0] = 1.669$
$bias_{ATE} = p bias_{ATT} + (1 - p)\, bias_{ATUT} = -0.485$	$E[Y_0] = 1.6$

8.4 Discussion and concluding remarks

"All models are wrong but some are useful."
- G. E. P. Box

It's time to return to our theme. Identifying causal effects suggests close attention to the interplay between theory, data, and model specification. Theory frames the problem so that economically meaningful effects can be deduced. Data supplies the evidence from which inference is drawn. Model specification attempts to consistently identify properties of the *DGP*. These elements are interdependent and iteratively divined.

Heckman [2000,2001] criticizes the selection literature for periods of preoccupation with devising estimators with nice statistical properties (e.g., consistency) but little economic import. Heckman's work juxtaposes policy evaluation implications of the treatment effects literature with the more ambitious structural modeling of the Cowles commission. It is clear for policy evaluation that theory or framing is of paramount importance.

"Every econometric study is incomplete."
- Zvi Griliches

In his discussion of economic data issues, Griliches [1986] reminds us that the quality of the data depends on both its source and its use. This suggests that creativity is needed to embrace the data issue. Presently, it seems that creativity in the address of omitted correlated variables, unobservable heterogeneity, and identification of instruments is in short supply in the accounting and business literature.

Model specification receives more attention in these pages but there is little to offer if theory and data are not carefully and creatively attended. With our current understanding of econometrics it seems we can't say much about a potential specification issue (including endogeneity) unless we accommodate it in the analysis. Even so, it is typically quite challenging to assess the nature and extent of the problem. If there is a mismatch with the theory or data, then discovery of (properties of) the *DGP* is likely hopelessly confounded. Logical consistency has been compromised.

8.5 Additional reading

The accounting literature gives increasing attention to endogeneity issues. Larcker and Rusticus [2004] review much of this work. Thought-provoking discussions of accounting and endogeneity are reported in an issue of *The European Accounting Review* including Chenhall and Moers. [2007a,2007b], Larcker and Rusticus [2007], and Van Lent [2007].

Amemiya [1985], Wooldridge [2002], Cameron and Trivedi [2005], Angrist and Krueger [1998], and the volumes of *Handbook of Econometrics* (especially volumes 5 and 6b) offer extensive reviews of econometric analysis of endogeneity. Latent *IV* traces back to Madansky [1959] and is resurrected by Lewbel [1997]. Heckman and Singer [1985,1986] discuss endogeneity challenges in longitudinal studies or duration models. The treatment effect examples are adapted from Joel Demski's seminars at the University of Florida and Eberhard Karls University of Tuebingen, Germany.