Ec. 2120 Spring 2010

G. Chamberlain

LECTURE NOTE 14

INSTRUMENTAL VARIABLE MODEL

1. OMITTED VARIABLE BIAS

Suppose that we are interested in the long regression:

$$E(Y_i \,|\, FB_i, ED_i, A_i) = FB_i'\phi + ED_i\theta + A_i,$$

but data on A_i are not available, and we run a least-squares fit of Y on FB and ED. The least-squares coefficients will converge in probability to the coefficients in the following short linear predictor:

$$E^*(Y_i \,|\, FB_i, ED_i) = FB_i'\tilde{\phi} + ED_i\tilde{\theta}.$$

The relationship between the long coefficients ϕ, θ and the short coefficients $\tilde{\phi}$, $\tilde{\theta}$ is worked out in Note 3. We need to consider an auxiliary linear predictor of the omitted variable A_i on FB_i and ED_i:

$$E^*(A_i \,|\, FB_i, ED_i) = FB_i'\psi_1 + ED_i\psi_2.$$

The omitted variable formula gives

$$\tilde{\phi} = \phi + \psi_1, \quad \tilde{\theta} = \theta + \psi_2.$$

For example, Y_i is the log of earnings of individual i, FB_i consists of a constant and a set of family background variables, ED_i is years of schooling, and A_i is a measure of initial (prior to the schooling) ability. The scale of A_i is chosen so that its coefficient equals one in the long regression.

The short least-squares fit provides consistent estimates of the short linear predictor coefficients $\tilde{\phi}$ and $\tilde{\theta}$. But these differ from the long regression coefficients by the auxiliary coefficients ψ_1 and ψ_2. This is a classic problem of omitted variable bias. The instrumental variable model will provide a solution. This new model requires an additional variable (or set of variables) that satisfy certain exclusion restrictions.

2. EXCLUSION RESTRICTIONS AND RANDOM ASSIGNMENT

Now suppose that we observe an additional variable (or set of variables) SUB_i, so that we observe

$$(FB_i, SUB_i, ED_i, Y_i) \quad \text{for} \quad i = 1, \ldots, n.$$

A_i is not observed. As in Note 6, we assume random sampling. The first exclusion restriction is that SUB_i does not help to predict Y_i if it is added to the long regression:

$$E(Y_i \,|\, FB_i, SUB_i, ED_i, A_i) = FB_i'\phi + ED_i\theta + A_i.$$

The second exclusion restriction is that SUB_i does not help to predict A_i in a linear predictor that includes FB_i:

$$E^*(A_i \,|\, FB_i, SUB_i) = FB_i'\lambda.$$

For example, SUB_i is an education subsidy that provides encouragement to obtain additional schooling. So it is correlated with ED_i, but the first exclusion restriction is that once we control for ED_i (and the other regressors in the long regression), the amount of subsidy that the individual receives does not have any additional predictive power. The second exclusion restriction is satisfied if the subsidy is randomly assigned. Suppose that the subsidy takes on only two values, zero and one, and the value that is assigned to i is determined by a coin flip. Then SUB_i will not be correlated with A_i or FB_i, and so the partial correlation of A_i and SUB_i given FB_i will be zero. (See problem set 1 on partial correlation.)

Define the prediction errors

$$\epsilon_i = A_i - E^*(A_i \,|\, FB_i, SUB_i),$$

$$U_i = Y_i - E(Y_i \,|\, FB_i, SUB_i, ED_i, A_i),$$

and write the equations

$$A_i = FB_i'\lambda + \epsilon_i$$

$$Y_i = FB_i'\phi + Ed_i\theta + A_i + U_i.$$

Note that ϵ_i and U_i are orthogonal to FB_i and SUB_i. Substitute for A_i in the Y_i equation:

$$\begin{aligned} Y_i &= FB_i'(\phi + \lambda) + ED_i\theta + (\epsilon_i + U_i) \\ &= FB_i'\delta + ED_i\theta + V_i, \end{aligned}$$

with $\delta = \phi + \lambda$ and $V_i = \epsilon_i + U_i$. Note that FB_i and SUB_i are orthogonal to V_i:

$$E(FB_i \cdot V_i) = 0, \quad E(SUB_i \cdot V_i) = 0.$$

Now define

$$R_i = (\, FB_i' \quad ED_i \,), \quad W_i = \begin{pmatrix} FB_i \\ SUB_i \end{pmatrix}, \quad \gamma = \begin{pmatrix} \delta \\ \theta \end{pmatrix}.$$

Then the exclusion restrictions imply that

$$Y_i = R_i\gamma + V_i, \quad E(W_iV_i) = 0. \tag{1}$$

This fits in the framework developed in Note 9. We can use results from Note 9 to obtain a consistent estimator for γ (provided that $E(W_iR_i)$ satisfies a rank condition). A consistent estimate of γ provides a consistent estimate of the coefficient θ on ED. The coefficient ϕ on FB in the long regression is not, however, consistently estimated. Instead we obtain a consistent estimate of $\delta = \phi + \lambda$. So there is still omitted variable bias in the FB coefficient (if FB_i and A_i are correlated).

The next section uses the orthogonality condition in (1) to obtain estimates and inferences that are valid in large samples.

3. JUST-IDENTIFIED CASE

Since ED_i is a scalar, the dimension K of the coefficient vector γ in (1) is

$$K = \dim(FB_i) + 1.$$

The dimension L of W_i is

$$L = \dim(FB_i) + \dim(SUB_i).$$

So if there is a single variable in SUB, then $L = K$ and the number of orthogonality conditions equals the number of parameters to be estimated. This is the *just-identified* case. The estimation of γ is based on the L orthogonality conditions in $E(W_i V_i) = 0$. The resulting estimator is often called an instrumental variables (IV) estimator. In the estimation context, all the variables in W are instrumental variables; there is no distinction between FB and SUB in providing orthogonality conditions. But in terms of the underlying model, FB and SUB play very different roles. The exclusion restrictions at the core of the model only apply to SUB. The random assignment argument only applies to SUB. So if we do refer to FB as instrumental variables (in the sense of generating orthogonality conditions), we should keep in mind that it is the *excluded* instrumental variables in SUB that play the key role in an instrumental variable model.

We can exploit the orthogonality conditions in (1) by multiplying the Y_i equation by W_i:

$$W_i Y_i = (W_i R_i)\gamma + W_i V_i,$$

$$E(W_i Y_i) = [E(W_i R_i)]\gamma,$$

and so

$$\gamma = [E(W_i R_i)]^{-1} E(W_i Y_i)$$

if $E(W_i R_i)$ is nonsingular. Then we can obtain a consistent estimate of γ by replacing population expectations by sample averages:

$$\begin{aligned}\hat{\gamma} &= \left(\frac{1}{n}\sum_{i=1}^{n} W_i R_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} W_i Y_i\right)\\ &= S_{WR}^{-1} S_{WY}.\end{aligned}$$

Suppose that $FB_i = 1$, so that (1) becomes

$$Y_i = \delta + ED_i\theta + V_i, \quad E(V_i) = 0 \quad E(SUB_i \cdot V_i) = 0.$$

Then $\mathrm{Cov}(SUB_i, V_i) = 0$ and

$$\mathrm{Cov}(SUB_i, Y_i) = \mathrm{Cov}(SUB_i, ED_i)\theta.$$

We can solve for

$$\theta = \frac{\mathrm{Cov}(SUB_i, Y_i)}{\mathrm{Cov}(SUB_i, ED_i)}$$

if

$$\mathrm{Cov}(SUB_i, ED_i) \neq 0.$$

So in addition to the exclusion restrictions on SUB, we require that SUB be correlated with ED. Then we can obtain a consistent estimate of θ by replacing the population covariances by their sample counterparts:

$$\hat{\theta} = \frac{\text{sample}\,\mathrm{Cov}(SUB, Y)}{\text{sample}\,\mathrm{Cov}(SUB, ED)}.$$

4. OVER-IDENTIFIED CASE

Now suppose there are two or more variables in SUB, so that $L > K$. This is the *over-identified* case. We still have

$$E(W_i Y_i) = E(W_i R_i)\gamma. \tag{2}$$

The rank condition on $E(W_i R_i)$ is that this $L \times K$ matrix has rank $= K$ (full column rank). This ensures that (2) determines γ uniquely. But in general we will not be able to solve for $\hat{\gamma}$ in the sample counterpart to (2), since $S_{WY} = S_{WR}\hat{\gamma}$ would give L equations for K unknowns. So we use a minimum-distance estimator:

$$\begin{aligned}\hat{\gamma} &= \arg\min_a (S_{WY} - S_{WR}a)'\hat{C}(S_{WY} - S_{WR}a) \\ &= (S'_{WR}\hat{C}S_{WR})^{-1}S'_{WR}\hat{C}S_{WY}.\end{aligned}$$

The only requirements on the $L \times L$ weight matrix $\hat{C}$ is that it be positive definite, symmetric and converge to a nonrandom matrix C that is positive definite, symmetric.

5. OPTIMAL WEIGHT MATRIX

From Note 10, the optimal choice for C is a matrix that is proportional to Σ^{-1}, where

$$\Sigma = \mathrm{Cov}(W_i V_i) = E(W_i V_i V_i' W_i') = E(V_i^2 W_i W_i')$$

(since V_i is scalar). It is common to use a weight matrix that would be optimal under homoskedasticity. Then having chosen C, we use the general results in Note 9 for inference. So the standard errors, confidence sets, and p-values are valid in large samples without restricting the form of the heteroskedasticity.

The homoskedastic case has

$$\begin{aligned}&\text{(i) } E(V_i \,|\, W_i) = 0, \\ &\text{(ii) } \mathrm{Var}(V_i \,|\, W_i) = E(V_i^2 \,|\, W_i) = \sigma_v^2.\end{aligned}$$

So the orthogonality condition $E(W_i V_i) = 0$ is strengthened to V_i mean-independent of W_i, and the conditional variance of V_i given W_i is assumed to be constant. Then

$$\Sigma = \sigma_v^2 E(W_i W_i').$$

Since we only need C to be proportional to Σ^{-1}, we can ignore σ_v^2 and use

$$\hat{C} = \left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i'\right)^{-1} = S_{WW'}^{-1}.$$

Using this weight matrix gives

$$\hat{\gamma} = (S'_{WR} S^{-1}_{WW'} S_{WR})^{-1} S'_{WR} S^{-1}_{WW'} S_{WY}. \tag{3}$$

This is known as the two-stage least-squares estimator (TSLS or 2SLS). The two-stage interpretation comes from a different way of deriving the estimator, which is developed in the next section.

6. POPULATION: TWO-STAGE LINEAR PREDICTOR

Writing out the components of R_i in (1) gives

$$Y_i = FB'_i\delta + ED_i\theta + V_i, \quad E(W_iV_i) = 0. \tag{1$'$}$$

Use $(1')$ to form the linear predictor of Y_i given W_i:

$$E^*(Y_i \,|W_i) = FB'_i\delta + E^*(ED_i \,|\, W_i)\theta.$$

Define

$$ED^*_i = E^*(ED_i \,|\, W_i) = W'_i\tau.$$

Then the linear predictor of Y_i given FB_i and ED^*_i identifies δ and θ:

$$E^*(Y_i \,|\, FB_i, ED^*_i) = FB'_i\delta + ED^*_i\theta. \tag{4}$$

7. SAMPLE: TWO-STAGE LEAST SQUARES

From (4), a least-squares fit of Y on FB and ED^* would provide consistent estimates of δ and θ. The predicted value ED^*_i is orthogonal to the error V_i because ED^*_i is constructed from W_i, which is orthogonal to V_i. The TSLS estimator obtains a consistent estimate of τ in stage 1. This is a least-squares fit of ED on W, with fitted values $\widehat{ED}_i = W_i\hat{\tau}$. The second stage obtains consistent estimates of δ and θ from a least-squares fit of Y on FB and $\widehat{ED}$.

This two-stage least-squares estimator is in fact the same as the estimator in (3), based on an optimal weight matrix. In the first stage, we can form fitted values for each variable in R_i:

$$\hat{R}_i = \begin{pmatrix} FB_i' & \widehat{ED}_i \end{pmatrix} = W_i' S_{WW'}^{-1} S_{WR}.$$

We get a perfect fit for the FB variables, since they are included in W, but we can still use the formula for the least-squares fitted value. Then in the second stage, we have a least-squares fit of Y on $\hat{R}$:

$$\begin{aligned}
\hat{\gamma} &= \Big(\frac{1}{n}\sum_{i=1}^{n} \hat{R}_i' \hat{R}_i)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{R}_i' Y_i \\
&= [S_{WR}' S_{WW'}^{-1} \Big(\frac{1}{n}\sum_{i=1}^{n} W_i W_i'\Big) S_{WW'}^{-1} S_{WR}]^{-1} S_{WR}' S_{WW'}^{-1} \Big(\frac{1}{n}\sum_{i=1}^{n} W_i Y_i\Big) \\
&= (S_{WR}' S_{WW'}^{-1} S_{WR})^{-1} S_{WR}' S_{WW'}^{-1} S_{WY}.
\end{aligned}$$

This is our orthogonality condition estimator with weight matrix $\hat{C} = S_{WW'}^{-1}$.

8. POTENTIAL OUTCOME FUNCTION, TREATMENT EFFECTS, SELECTION BIAS, AND RANDOM ASSIGNMENT

We shall use a potential outcome function to define an average treatment effect. Then we shall see how random assignment of the treatment allows us to obtain the average treatment effect from a predictive effect.

For each individual i, there is a *potential outcome function* $Y_i(\cdot)$. It can be evaluated at any feasible level t of the treatment. Then $Y_i(t)$ is a random variable, whose realized value is the outcome for i at treatment level t. As t varies, we have a set of potential outcomes. Only one of these potential outcomes is actually observed. Let T_i denote the treatment level that is assigned to i. Then the observed outcome is the potential outcome corresponding to the assigned treatment level:

$$Y_i = Y_i(T_i).$$

The *average treatment effect* in comparing treatment level t_1 with treatment level t_2 is

$$\mathrm{ATE}(t_1, t_2) = E[Y_i(t_2) - Y_i(t_1)].$$

The corresponding predictive effect, as defined in Note 2, is

$$\begin{aligned}\mathrm{PE}(t_1, t_2) &= E(Y_i \,|\, T_i = t_2) - E(Y_i \,|\, T_i = t_1) \\ &= E[Y_i(t_2) \,|\, T_i = t_2] - E[Y_i(t_1) \,|\, T_i = t_1].\end{aligned}$$

The predictive effect does not, in general, equal the average treatment effect, because the assigned treatment T_i may be correlated with potential outcomes. The difference between the predictive effect and the average treatment effect is called *selection bias*.

For example, suppose that the outcome is blood pressure and there are two treatments: $t = 0$ does nothing (a placebo) and $t = 1$ is a new drug. Each individual has two potential outcomes, $Y_i(0)$ and $Y_i(1)$. Suppose that the individuals who are assigned $T_i = 1$ have high blood pressure. Then $t = 1$ may lower blood pressure for each individual: $Y_i(1) - Y_i(0) < 0$, but $E[Y_i(1) \,|\, T_i = 1]$ is higher than $E[Y_i(0) \,|\, T_i = 0]$. Then the average treatment effect of the new drug is to lower blood pressure, but the predictive effect shows higher blood pressure on average for the treated $T_i = 1$ individuals compared with the untreated $T_i = 0$.

A solution to selection bias is *random assignment*. Suppose that each individual is assigned $T_i = 0$ or $T_i = 1$ based on a coin flip. Then T_i will be independent of the potential outcomes. Let $\mathcal{T}$ denote the set of possible values for the treatment. Our general definition of a randomly assigned treatment is that

$$\{Y_i(t),\ t \in \mathcal{T}\} \perp\!\!\!\perp T_i$$

—the random variables corresponding to the potential outcomes are jointly independent of the treatment assignment T_i. In the case of the placebo and the new drug, the treatment is randomly assigned if

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i.$$

Under random assignment, for any treatment levels t_1, t_2, and t:

$$E[Y_i(t_2) \mid T_i = t] = E[Y_i(t_2)], \quad E[Y_i(t_1) \mid T_i = t] = E[Y_i(t_1)],$$

and so the predictive effect equals the average treatment effect:

$$\begin{aligned} \mathrm{PE}(t_1, t_2) &= E(Y_i \mid T_i = t_2) - E(Y_i \mid T_i = t_1) \\ &= E[Y_i(t_2) \mid T_i = t_2] - E[Y_i(t_1) \mid T_i = t_1] \\ &= E[Y_i(t_2)] - E[Y_i(t_1)] \\ &= \mathrm{ATE}(t_1, t_2). \end{aligned}$$

The next section develops an instrumental variable model in which the treatment is not randomly assigned but there is an instrumental variable that is randomly assigned. We shall develop an orthogonality condition estimator for the average treatment effect, but this will require strong restrictions on the potential outcome function.

9. INSTRUMENTAL VARIABLE MODEL

In the drug example, suppose that individuals are randomly assigned to $t = 0$ (no treatment) and $t = 1$ (new drug), but the new drug has side effects and some of the people assigned to take it in fact do not take it. So there is a randomly assigned "intent to treat" but the actual treatment that i receives depends also on choices made by i. So there is the possibility of selection bias. The individuals who take the new drug in spite of the side effects may have potential outcomes that differ on average from the potential outcomes of the people who drop out.

More generally, suppose that T_i is not randomly assigned, but there is a variable (or set of variables) S_i that is randomly assigned and that is correlated with T_i. In the drug example, S could be the randomly assigned intent to treat. In the earnings and education example, S could be a randomly assigned education subsidy. We shall refer to S as a subsidy.

For each individual i, there is a potential outcome function $Y_i(\cdot\,,\cdot\,)$. It can be evaluated at any level t of the treatment and level s of the subsidy. Then $Y_i(t, s)$ is a random variable, whose realized value is the outcome for i at treatment level t and subsidy level s. As t and s vary, we have a set of potential outcomes. Only one of these potential outcomes is actually observed. Let T_i denote the treatment level assigned to i, and let S_i denote the subsidy level assigned to i. Then the observed outcome is the potential outcome corresponding to the assigned treatment and subsidy:

$$Y_i = Y_i(T_i, S_i).$$

A key exclusion restriction is that the distribution of the potential outcome $Y_i(t, s)$ does not depend on s: for all (feasible) values of t, s_1, and s_2,

$$Y_i(t, s_1) \stackrel{d}{=} Y_i(t, s_2).$$

(Here $\stackrel{d}{=}$ means that two random variables have the same distribution.) So we can write the potential outcome function as a function just of the treatment level:

$$Y_i(t, s) = Y_i(t), \quad Y_i = Y_i(T_i).$$

The definition of the subsidy being randomly assigned mimics the definition in Section 8 of a randomly assigned treatment. The random variables corresponding to the potential outcomes are jointly independent of the subsidy assignment:

$$\{Y_i(t),\ t \in \mathcal{T}\} \perp\!\!\!\perp S_i.$$

So there are two key assumptions in the instrumental variable model: the instrumental variable (or variables) S is excluded from the potential outcome function and S is randomly assigned. In order for our orthogonality condition estimator (also known as an IV estimator) to provide a consistent estimate of the average treatment effect, we need, in addition to the two key assumptions, to restrict the form of the potential outcome function. Here is the restricted potential outcome function:

$$\begin{aligned} Y_i(t) &= Y_i(t_0) + \theta(t - t_0) \\ &= [Y_i(t_0) - \theta t_0] + \theta t \\ &= Y_{i0} + \theta t, \end{aligned}$$

where t_0 is some feasible treatment level and $Y_{i0} = Y_i(t_0) - \theta t_0$. So there is a linear response to the treatment level and the slope θ of the response does not vary across the individuals. The average treatment effect is

$$\text{ATE}(t_1, t_2) = \theta(t_2 - t_1),$$

which is the same as the treatment effect for each i. The heterogeneity across individuals is confined to the random intercept Y_{i0}.

Since S_i is randomly assigned,

$$E^*(Y_{i0} \,|\, 1, S_i) = E(Y_{i0}) \equiv \delta.$$

Define the prediction error

$$V_i = Y_{i0} - E^*(Y_{i0} \,|\, 1, S_i),$$

and note that V_i is orthogonal to 1 and to S_i. Then we have

$$Y_i(t) = \delta + \theta t + V_i,$$

and the observed outcome satisfies

$$Y_i = Y_i(T_i) = \delta + \theta T_i + V_i.$$

Just as with equation (1) in Section 2, we can put this in the form of the framework developed in Note 9:

$$Y_i = R_i \gamma + V_i, \quad E(W_i V_i) = 0,$$

with

$$R_i = \begin{pmatrix} 1 & T_i \end{pmatrix}, \quad \gamma = \begin{pmatrix} \delta \\ \theta \end{pmatrix}, \quad W_i = \begin{pmatrix} 1 \\ S_i \end{pmatrix}.$$

So we can use the orthogonality condition (IV) estimators developed in Note 9.

Even if S_i is not randomly assigned, we may be able to argue that, conditional on a set of variables Z_i, S_i is "as good as" randomly assigned. Given the restricted form of the potential outcome function, the assumption we need is that

$$E^*(Y_{i0} \mid Z_i, S_i) = Z_i'\delta,$$

so that S_i does not help to predict Y_{i0} in a linear predictor that includes Z_i. (Assume that Z_i includes a constant.) Define the prediction error

$$V_i = Y_{i0} - E^*(Y_{i0} \mid Z_i, S_i),$$

and note that V_i is orthogonal to Z_i and to S_i. Then we have

$$Y_i(t) = Z_i'\delta + \theta t + V_i,$$

and the observed outcome satisfies

$$Y_i = Y_i(T_i) = Z_i'\delta + \theta T_i + V_i.$$

So

$$Y_i = R_i\gamma + V_i, \quad E(W_iV_i) = 0, \tag{5}$$

with

$$R_i = \begin{pmatrix} Z_i & T_i \end{pmatrix}, \quad \gamma = \begin{pmatrix} \delta \\ \theta \end{pmatrix}, \quad W_i = \begin{pmatrix} Z_i \\ S_i \end{pmatrix}.$$

Once again we can use the orthogonality condition (IV) estimators developed in Note 9.

Note the similarity of (5) with equation (1) in Section 2. Suppose that Y_i is the log of earnings for individual i, Z_i consists of a constant and a set of family background variables, the treatment T_i is years of schooling, and S_i is an education subsidy. The selection bias arises because Y_{i0} contains A_i, a measure of initial ability that is not in the data set. V_i is the part of A_i that is not predictable from the family background variables. If V_i and T_i are correlated, then the coefficient on T_i in the linear predictor of Y_i given Z_i and T_i does not equal θ. Here the selection bias is equivalent to omitted variable bias.

10. REDUCED FORM

Another terminology for instrumental variables is *exogenous* variables. The variables Z and S in (5) are exogenous and the *endogenous* variables are the outcome Y and the assigned treatment T. The *reduced form* consists of either the conditional expectations or the linear predictors of the endogenous variables given the exogenous variables. In our IV model in (5), the linear predictors contain useful information:

$$E^*(T_i \mid Z_i, S_i) = Z_i'\alpha_1 + S_i'\pi_1, \tag{6}$$

$$\begin{aligned} E^*(Y_i \mid Z_i, S_i) &= Z_i'\delta + \theta(Z_i'\alpha_1 + S_i'\pi_1) \\ &= Z_i'(\delta + \theta\alpha_1) + S_i'(\theta\pi_1) \\ &= Z_i'\alpha_2 + S_i'\pi_2, \end{aligned} \tag{7}$$

where $\alpha_2 = \delta + \theta\alpha_1$ and

$$\pi_2 = \theta\pi_1. \tag{8}$$

The coefficients π_2 on S in predicting Y are proportional to the coefficients π_1 on S in predicting T, and the proportionality factor identifies θ.

Let $J = \dim(S_i)$. If $J = 2$,

$$\begin{pmatrix} \pi_{21} \\ \pi_{22} \end{pmatrix} = \theta \begin{pmatrix} \pi_{11} \\ \pi_{12} \end{pmatrix},$$

and the least-squares estimates of the linear predictors are

$$\hat{T}_i = Z_i'\hat{\alpha}_1 + S_{i1}\hat{\pi}_{11} + S_{i2}\hat{\pi}_{12},$$

$$\hat{Y}_i = Z_i'\hat{\alpha}_2 + S_{i1}\hat{\pi}_{21} + S_{i2}\hat{\pi}_{22}.$$

In this over-identified case, we can obtain two consistent estimates of θ from the least-squares estimates of the reduced form:

$$\hat{\theta}^{(1)} = \hat{\pi}_{21}/\hat{\pi}_{11}, \quad \hat{\theta}^{(2)} = \hat{\pi}_{22}/\hat{\pi}_{12}.$$

The two-stage least-squares estimator provides a way to combine consistent estimates in the over-identified case. There is an alternative minimum-distance estimator that directly imposes the key proportionality restriction in (8). Because of the proportionality restriction, we can express (π_1, π_2) as a function of a lower dimension, unrestricted parameter (θ, β):

$$\pi_1 = \beta, \quad \pi_2 = \theta\beta,$$

where π_1, π_2, and β are $J \times 1$ and θ is a scalar. The least-squares estimates $(\hat{\pi}_1, \hat{\pi}_2)$ will not satisfy the proportionality restriction in a finite sample, but the estimates are consistent and so converge in probability to (π_1, π_2), which do satisfy the restriction. The following minimum-distance estimator obtains consistent estimates of θ and β by imposing the proportionality restriction:

$$\begin{aligned}(\hat{\theta}, \hat{\beta}) &= \arg \min_{a \in \mathcal{R}, b \in \mathcal{R}^J} || \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - \begin{pmatrix} b \\ a \cdot b \end{pmatrix} ||^2 \\ &= \arg \min_{a \in \mathcal{R}, b \in \mathcal{R}^J} \left(\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - \begin{pmatrix} b \\ a \cdot b \end{pmatrix} \right)' \hat{C} \left(\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - \begin{pmatrix} b \\ a \cdot b \end{pmatrix} \right).\end{aligned}$$

Define

$$\hat{\pi} = \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix}, \quad \pi = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}.$$

The results from Note 9 can be used to show that

$$\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} \mathcal{N}(0, \Lambda).$$

An optimal choice for C is Λ^{-1}:

$$\hat{C} = \hat{\Lambda}^{-1} \xrightarrow{p} \Lambda^{-1} = C.$$

The estimate $(\hat{\theta}, \hat{\beta})$ can be used to form estimates of the reduced-form coefficients that impose the proportionality restriction:

$$\hat{\pi}_1^* = \hat{\beta}, \quad \hat{\pi}_2^* = \hat{\theta} \cdot \hat{\beta}.$$

11. DEMAND FUNCTION

Let $D_i(p)$ denote the quantity demanded in market i at price p. The demand function $D_i(\cdot)$ can be evaluated at any price. Only one of these quantities is actually observed. Let P_i denote the observed price in market i. Then, assuming that the observed quantity Q_i is on the demand curve,

$$Q_i = D_i(P_i).$$

We shall work with a restricted form of the demand function:

$$\begin{aligned} D_i(p) &= D_i(p_0) + \theta(p - p_0) \\ &= [D_i(p_0) - \theta p_0] + \theta p \\ &= D_{i0} + \theta p, \end{aligned}$$

where p_0 is some reference price and $D_{i0} = D_i(p_0) - \theta p_0$. ($D_i(p)$ could be the log of the quantity demanded and p could be the log of price.) The goal is to estimate θ, the slope (or, in logs, the elasticity) of the demand curve. This slope is assumed to be the same in all markets. The heterogeneity across markets is confined to the intercept D_{i0}, which represents shifts in the demand curve.

Let $SUP_i(p)$ denote the quantity supplied in market i at price p, and suppose that the supply function has the following form:

$$SUP_i(p) = SUP_{i0} + \lambda p.$$

The heterogeneity across markets is confined to the intercept SUP_{i0}, which represents shifts in the supply curve.

Suppose that the observed price P_i is assigned to clear the market, equating the quantity demanded at P_i with the quantity supplied at P_i:

$$D_i(P_i) = SUP_i(P_i) = Q_i.$$

Then we can solve for

$$P_i = \frac{D_{i0} - SUP_{i0}}{\lambda - \theta}.$$

The predictive effect of price on quantity, comparing the prices p_1 and p_2, is

$$\begin{aligned}\mathrm{PE}(p_1, p_2) &= E(Q_i \mid P_i = p_2) - E(Q_i \mid P_i = p_1) \\ &= E(D_{i0} \mid P_i = p_2) - E(D_{i0} \mid P_i = p_1) + \theta(p_2 - p_1).\end{aligned}$$

This does not, in general, equal $\theta(p_2 - p_1)$ if the demand shift D_{i0} is correlated with the market clearing price P_i. Because

$$\mathrm{Cov}(D_{i0}, P_i) = \frac{\mathrm{Var}(D_{i0}) - \mathrm{Cov}(SUP_{i0}, D_{i0})}{\lambda - \theta},$$

there will, in general, be a correlation between the demand shift and the market clearing price. So the predictive effect does not correspond to the slope of the demand curve (or, in logs, the demand elasticity). This is a form of selection bias, since the price P_i is not randomly assigned. For any p, the assigned price P_i is correlated with $D_i(p)$ through its correlation with the demand shift D_{i0}. (This bias is also called a simultaneity bias, because the the observed price P_i and quantity Q_i are simultaneously determined by the intersection of the demand and supply curves for market i.)

There is an instrumental variable solution to this bias problem. The key exclusion restriction is

$$E^*(D_{i0} \mid Z_i, S_i) = Z_i'\delta.$$

Here Z_i consists of observed demand shift variables (and a constant), and S_i consists of observed supply shift variables. The excluded instrumental variables S_i are assumed to be "as good as randomly assigned," in that they do not help to predict the demand shift D_{i0} in a linear predictor that includes Z_i.

Define the prediction error

$$V_i = D_{i0} - E^*(D_{i0} \mid Z_i, S_i),$$

and note that V_i is orthogonal to Z_i and S_i. Then we have

$$D_i(p) = Z_i'\delta + \theta p + V_i,$$

and the observed quantity satisfies

$$Q_i = D_i(P_i) = Z_i'\delta + \theta P_i + V_i.$$

So

$$Q_i = R_i\gamma + V_i, \quad E(W_i V_i) = 0, \tag{9}$$

with

$$R_i = \begin{pmatrix} Z_i & P_i \end{pmatrix}, \quad \gamma = \begin{pmatrix} \delta \\ \theta \end{pmatrix}, \quad W_i = \begin{pmatrix} Z_i \\ S_i \end{pmatrix}.$$

As with equation (5) in Section 9, we can use the orthogonality condition (IV) estimators developed in Note 9.