1 Structural Models are Causal Models

Answer all of the questions in the posted handout "Causality in Econometrics and Statistics: Structural Models are Causal Models."

Problem 1.1. Given the causal relationships

$$\begin{aligned} Y &= f_Y(X, U, \epsilon_Y) \quad Y \text{observed} \\ X &= f_X(V, \epsilon_X), \quad X \text{ observed} \\ U &= f_U(V, \epsilon_U) \quad U \text{ unobserved} \\ V &= f_V(\epsilon_V) \quad V \text{ unobserved} \end{aligned}$$

which statistical relationships are generated by this causal model? Is there an equivalence between statistical relationships and casual relationships?

Solution. The causal model generates the statistical relationship between Y and X. In general, the equivalence between statistical and causal relationships will not hold since the unobservables (U, V) have a causal relationship to X and Y. The equivalence wil hold once we are able to fix U and V.

■

Problem 1.2. What assumption does IV use to solve the identification problem, compared to the usage of the Sequential Ignorability assumption?

Solution. The sequential ignorability assumption in the setting of $T \to M \to Y$ is

$$[1] : Y(t', m), M(t) \perp T | X$$
$$[2] : Y(t', m) \perp M(t) | (T, X)$$

[1] says that T is exogenous conditioned on X and no unobserved variable causes T and Y or T and M. [2] says that M is exogenous conditioned on X and T, which is a stronger assumption than randomization. In constrast, the IV approach makes the following assumptions:

$$[3] : Y(t) \perp Z | X$$
$$[4] : Z \not\perp T$$

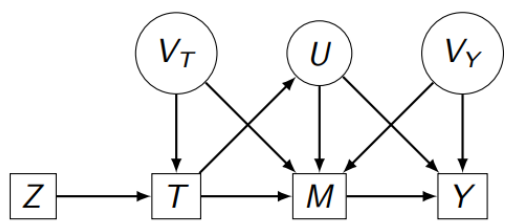
Note that [1] (which is used in matching) rules out a major identification problem that stems from the possibly asymmetry in information between the agents making the participation decisions and the observing economist.

Another way to see the contrast is to examine the opposite roles for T. In sequential ignorability, the variation in T that arises after conditioning on X provides the source of randomness that switches people across treatment status. In IV, however, the variation in $P(X, Z)$ produces variations in T that switches the treatment status. The components of variation in T not predictable by (X, Z) is a problem for IV, while for matching it is the source of identification.

■

Problem 1.3. Prove Property 3 that holds in the mediation model:

Figure 3: DAG for the Mediation Model with IV and Confounding Variables



Solution. Property 3 states that

$$Z \not\perp M|T \text{ and } Z \perp Y(m)|T$$

▷ In the model, T is caused by both Z and V_T and $V_T \perp Z$.

▷ Therefore, conditioning on T induces correlation between Z and V_T.

▷ Since V_T causes M and does not directly cause Y, it follows that conditioned on T, Z affects M. This shows that $Z \not\perp M|T$.

▷ Since it does not affect Y by any channel other than M, we have $Z \perp Y(m)|T$.

■

Problem 1.4. (Bonus) Prove the Graphoid Axioms.

Solution. The answers are below:

1. Symmetry: $X \perp Y|Z \implies Y \perp X|Z$:

   ▷ $X \perp Y|Z$ is equivalent to:
   $$p(x,y|z) = p(x|z)p(y|z)$$

   ▷ Therefore:
   $$p(y,x|z) = p(x,y|z) = p(y|z)p(x|z) \Leftrightarrow Y \perp X|Z.$$

   ■

2. Decomposition: $X \perp (W,Y)|Z \implies X \perp Y|Z$.

   ▷ $X \perp (W,Y)|Z$ implies
   $$p(x,w,y|z) = p(x|z)p(w,y|z)$$

   ▷ Therefore:
   $$\begin{aligned} p(x,y|z) &= \int p(x,y,w|z)dw \\ &= \int p(x|z)p(y,w|z)dw \\ &= p(x|z)\int p(y,w|z)dw \\ &= p(x|z)p(y|z) \end{aligned}$$

   which is equivalent to $X \perp Y|Z$. ■

3. Weak Union: $X \perp (W,Y)|Z \implies X \perp Y|(W,Z)$

   ▷ $X \perp (W,Y)|Z$ implies:
   $$p(x|y,w,z) = p(x|w,z) = p(x|y,z) = p(x|z)$$

   ▷ Therefore:
   $$\begin{aligned} p(x,y|w,z) &= p(x|y,w,z)p(y|w,z) && \text{(Bayes' Rule)} \\ &= p(x|w,z)p(y|w,z) \end{aligned}$$

   which is equivalent to $X \perp Y|(W,Z)$. ■

4. Contraction: $X \perp W|(Y, Z)$ and $X \perp Y|Z \implies X \perp (W, Y)|Z$.

   ▷ $X \perp (W, Y)|Z$ implies:

$$p(x|y, w, z) = p(x|w, z) = p(x|y, z) = p(x|z)$$

   ▷ $X \perp Y|Z$ implies:

$$p(x|y, z) = p(z)$$

   ▷ Therefore:

$$\begin{aligned} p(x, w, y|z) &= p(x|y, w, z)p(y, w|z) && \text{(Bayes' Rule)} \\ &= p(x|y, z)p(y, w|z) \\ &= p(x|z)p(y, w|z). \end{aligned}$$

■

5. Intersection: $X \perp W|(Y, Z)$ and $X \perp Y|(W, Z) \implies X \perp (W, Y)|Z$

   ▷ Our assumption gives us

$$p(x|w, z) = p(x|w, y, z) = p(x|y, z).$$

   ▷ Bayes rule tells us this equality is also

$$\frac{p(x, w|z)}{p(w|z)} = \frac{p(x, y|z)}{p(y|z)}$$

   which implies

$$p(x, w|z)p(y|z) = p(x, y|z)p(w|z).$$

   ▷ Integrating over w implies

$$\begin{aligned} p(y|z) \int p(x, w|z)dw &= p(x, y|z) \int p(w|z)dw \\ \implies p(y|z)p(x|z) &= p(x, y|z). \end{aligned}$$

   and therefore $X(W, Y)|Z$. ■

6. Redundancy: $X \perp Y|X$.

$$\begin{aligned} p(x, y|x) &= \frac{p(x, y, x)}{\int p(x, y, x)dx} \\ &= \frac{p(x, y)}{p(y)} \\ &= p(y|x) \\ &= p(x|x)p(y|x) \end{aligned}$$

and thus our proof is complete. ■

Problem 1.5. (Bonus) Prove the laws connecting hypothetical empirical models.

Solution. The notations are: $\mathcal{T}_E$ (variable set), Pa_E (parents), D_E (descendants), Ch_E (children). Replacing E with H gives the analog for a hypothetical model.

1. **L-1**: Let $\{W, Z\}$ be any disjoint set of variables in $\mathcal{T}_E \backslash D_H(\tilde{X})$ then:

$$P_H(W|Z) = P_H(W|Z, \tilde{X}) = P_E(W|Z), \quad \forall \{W, Z\} \subset \mathcal{T}_E \backslash D_H(\tilde{X}).$$

▷ Intuitively, the first equality follows from the fact that $\tilde{X}$ is an external variable with no parents in the hypothetical model and from our assumption has no descendants. The second equality follows from the fact when we remove $\tilde{X}$ from the hypothetical model, we now have the empirical model since $\{W, Z\} \subset \mathcal{T}_E$.Therefore, the distribution of non-descendants of $\tilde{X}$ are the same in hypothetical and empirical models.

▷ Mathematically:

$$\begin{aligned}
\Pr_{\mathrm{H}}\left(\mathcal{T}_{\mathrm{E}} \backslash D_{\mathrm{H}}(\tilde{X})\right) &= \prod_{T \in \mathcal{T}_{\mathrm{E}} \backslash D_{\mathrm{H}}(\tilde{X})} \Pr_{\mathrm{H}}\left(T | Pa_{\mathrm{H}}(T)\right) \\
&= \prod_{T \in \mathcal{T}_{\mathrm{E}} \backslash D_{\mathrm{H}}(\tilde{X})} \Pr_{\mathrm{E}}\left(T | Pa_{\mathrm{E}}(T)\right) = \Pr_{\mathrm{E}}\left(\mathcal{T}_{\mathrm{E}} \backslash D_{\mathrm{H}}(\tilde{X})\right)
\end{aligned}$$

▷ As a consequence, $P_H(W) = P_E(W)$ for all $W \subset \mathcal{T}_E \backslash D_H(\tilde{X})$, and therefore:

$$\begin{aligned}
\Pr_{\mathrm{H}}(W = w | Z = z) &= \frac{\Pr_{\mathrm{H}}(W = w, Z = z)}{\Pr_{\mathrm{H}}(Z = z)} \\
&= \frac{\Pr_{\mathrm{E}}(W = w, Z = z)}{\Pr_{\mathrm{E}}(Z = z)} = \Pr_{\mathrm{E}}(W = w | Z = z)
\end{aligned}$$

■

2. **T-1**: Let $\{W, Z\}$ be any disjoint set of variables in $\mathcal{T}_E$ then:

$$P_H(W|Z, X = x, \tilde{X} = x) = P_E(W|Z, X = x), \quad \forall \{W, Z\} \subset \mathcal{T}_E.$$

▷ Intuitively, this result follows from the fact that when we add $\tilde{X}$ into our empirical model and sever the relationship between X and Y, setting them equal to x produces the same relationship as the model prior to modification. Therefore, distributions of variables conditional on X and $\tilde{X}$ at the same value of x in empirical and in the hypothetical model is the same as the distribution of variables conditional on $X = x$ in the empirical model.

▷ Mathematically, start by partitioning the set of variables $\mathcal{T}_E$ into four sets:

$$\mathcal{T}_E = \{T_E \backslash D_E(X)\} \cup \{D_E(W) \backslash Ch_E(X)\} \cup \{Ch_H(X)\} \cup \{Ch_H(\tilde{X})\}$$

and examine the variables of each set separately.

* For the first set, the equality holds directly from the result in L-1.
* For the second set, it holds by LMC and weak union.
* For the third set, it holds from LMC applied to T and the fact that $\tilde{X}$ is external.
* For the fourth set, it holds because X must be a non-descendant of T due to the acyclical property of the empirical model, and applying LMC gives us the desired result.

▷ Grouping these four sets together, we arrive at:

$$\begin{aligned}\mathrm{Pr}_{\mathrm{H}}\left(\mathcal{T}_{\mathrm{E}}|X=x,\tilde{X}=x\right) &= \prod_{T\in\mathcal{T}_{\mathrm{E}}}\mathrm{Pr}_{\mathrm{H}}\left(T|Pa_{\mathrm{H}}(T),\tilde{X}=x,X=x\right)\\ &= \prod_{T\in\mathcal{T}_{\mathrm{E}}}\mathrm{Pr}_{\mathrm{E}}\left(T|Pa_{\mathrm{E}}(T),X=x\right)\\ &= \mathrm{Pr}_{\mathrm{E}}\left(\mathcal{T}_{\mathrm{E}}|X=x\right)\end{aligned}$$

■

3. **Matching**: Let $\{W,Z\}$ be any disjoint set of variables in $\mathcal{T}_E$ such that, in the hypothetical model, $X\perp W|(Z,\tilde{X})$, then:

$$P_H(W|Z,\tilde{X}=x)=P_E(W|Z,X=x),\quad \forall\{W,Z\}\subset\mathcal{T}_E.$$

▷ It holds since:

$$\begin{aligned}\mathrm{Pr}_{\mathrm{H}}(W|Z,\tilde{X}=x) &= \mathrm{Pr}_{\mathrm{H}}(W|Z,\tilde{X}=x,X=x)\quad \text{by assumption } X\perp W|(Z,\tilde{X}) \text{ in } G_{\mathrm{H}}\\ &= \mathrm{Pr}_{\mathrm{E}}(W|Z,X=x)\quad \text{by Theorem T}-1\end{aligned}$$

■

2 Principles Underlying Evaluation Estimators

Answer all the questions in the posted handout "The Principles Underlying Evaluation Estimators."

Problem 2.1. (Slide 4) Suppose $Y_1 - Y_0$ is a random variable that depends on X. Can you identify individual-level treatment effects?

Solution. No, we cannot identify individual-level treatment effect because for each individual only one state (treated or untreated) can take place, and the other is counterfactual. Even though we know $Y_1 - Y_0$ is a function of X and X is observable, we still don't have any information about that functional form. But we can identify the local average treatment effect in each matched cell $X = x$. If we further assume local (conditional) homogeneous treatment effect, then we can identify the individual-level treatment effects.

Problem 2.2. (Slide 10) In a Generalized Roy model in which agents have as much information as the observing economist, and both use the information in making decisions and forming estimates, show that conditional on(X, Z) (the assumed information set) (R-1)

$$(Y_1, Y_0) \perp D$$

is satisfied.

Solution. Since the agents have as much information as the observing economist, we know that conditional on the information set (X, Z), agents have no better idea about the probability of taking the treatment, i.e. there is no more information contributing to the probability of taking treatment:

$$\mathbb{P}(D = 1 \mid X, Z, Y_1, Y_0) = \mathbb{P}(D = 1 \mid X, Z)$$

So,

$$(Y_1, Y_0) \perp D \mid X, Z$$

Problem 2.3. (Slide 23) What causal parameter, if any, can be identified from an experiment with imperfect compliance?

Solution. Denote the randomized assignment status by R and denote the actual treatment status by A. And the observed outcomes are given by the switching regression:

$$Y = AY_1 + (1 - A)Y_0.$$

Under randomization:

$$(Y_0, Y_1) \perp R.$$

When we have imperfect compliance, we can identify ITT (*Intention to be Treated*):

$$\begin{aligned}
ITT &= E(Y|R=1, D=1) - E(Y|R=0, D=1) \\
&= \{E(Y|R=1, D=1, A=1)Pr(A=1|R=1, D=1) \\
&+ E(Y|R=1, D=1, A=0)Pr(A=0|R=1, D=1)\} \\
&- \{E(Y|R=0,, D=1A=1)Pr(A=1|R=0, D=1) \\
&+ E(Y|R=0,, D=1A=0)Pr(A=0|R=0, D=1)\} \\
&= \{E(Y_1|R=1, D=1, A=1)Pr(A=1|R=1, D=1) \\
&+ E(Y_0|R=1, D=1, A=0)Pr(A=0|R=1, D=1)\} \\
&- \{E(Y_1|R=0, D=1, A=1)Pr(A=1|R=0, D=1) \\
&+ E(Y_0|R=0, D=1, A=0)Pr(A=0|R=0, D=1)\}
\end{aligned}$$

Problem 2.4. (Slide 26) Assuming that you cannot compel program participation, show what a population-wise randomization of eligibility identifies.

Solution. Denote the randomized assignment status by R and denote the actual treatment status by D. Then, we have

$$\begin{aligned}
E(Y|R=1) &= E(Y|R=1, D=1)Pr(D=1|R=1) + E(Y|R=1, D=0)Pr(D=0|R=1) \\
&= E(Y_1|R=1, D=1)Pr(D=1|R=1) + E(Y_0|R=1, D=0)Pr(D=0|R=1) \\
E(Y|R=0) &= E(Y|R=0, D=1)Pr(D=1|R=0) + E(Y|R=0, D=0)Pr(D=0|R=0) \\
&= E(Y_1|R=0, D=1)Pr(D=1|R=0) + E(Y_0|R=0, D=0)Pr(D=0|R=0).
\end{aligned}$$

Therefore, we can identify

$$\begin{aligned}
E(Y|R=1) - E(Y|R=0) &= \{E(Y_1|R=1, D=1)Pr(D=1|R=1) \\
&+ E(Y_0|R=1, D=0)Pr(D=0|R=1)\} \\
&- \{E(Y_1|R=0, D=1)Pr(D=1|R=0) \\
&+ E(Y_0|R=0, D=0)Pr(D=0|R=0)\}.
\end{aligned}$$

3 ITT vs. TOT vs. PRTE

Compare ITT with TOT and PRTE. Consider two versions of ITT; conditional on $D = 1$ and unconditional on D.

Solution. Let R be the dummy for it treatment is assigned to an individual and D be the dummy for if the individual takes treatment. So here D determines actual treatment status.The unconditional ITT is:

$$\begin{aligned}\mathbb{E}[Y|R=1]-\mathbb{E}[Y|R=0] = &\Big(\mathbb{E}[Y_1|R=1,D=1]Pr(D=1|R=1)+\mathbb{E}[Y_0|R=1,D=0]Pr(D=0|R=1)\Big)\\ &-\Big(\mathbb{E}[Y_1|R=0,D=1]Pr(D=1|R=0)+\mathbb{E}[Y_0|R=0,D=0]Pr(D=0|R=0)\Big),\end{aligned}$$

where the equality follows by LIE. The unconditional ITT represents the average effect of assigning treatment to an individual. In general the unconditional ITT will include the treatment effect of non compliers ($Pr(D = 0|R = 1) > 0$ represents individuals assigned treatment who don't take it, $Pr(D = 1|R = 0) > 0$) represents individuals who are not assigned treatment but take it anyway).

Now assume perfect compliance and let R be the dummy for if treatment is assigned to this individual and D' be the dummy for if this individual selected to take treatment (but is potentially denied treatment based on R and complies with the denial). So here R determines actual treatment status. The ITT conditional on $D = 1$ is

$$\mathbb{E}[Y|R=1,D=1]-\mathbb{E}[Y|R=0,D=1]=\mathbb{E}[Y_1-Y_0|D=1],$$

where the last equality follows since whether or not the individual actually takes treatment is determined by R. The conditional ITT represents the average effect of treatment on those individuals who would have self-selected into treatment (i.e. we ask those who want treatment to show up for treatment, and then randomly deny some people). The conditional and unconditional ITTs are clearly measuring different quantities.

The TOT (average effect of treatment on the treated) is

$$\mathbb{E}[Y_1-Y_0|D=1]$$

and represents the effect of treatment on those who actually receive treatment, regardless of if they actually were assigned treatment or not (i.e. we allow for non-compliance here). Note that TOT is the same as the conditional ITT. Moreover, if we have full compliance ($Pr(D = 0|R = 1) = Pr(D = 1|R = 0) = 0$) and treatment is randomly assigned (so $R \perp (Y_1, Y_0)$), then we have for the unconditional ITT:

$$\begin{aligned}\mathbb{E}[Y|R=1]-\mathbb{E}[Y|R=0] &= \mathbb{E}[Y_1|R=1,D=1]-\mathbb{E}[Y_0|R=0,D=0], \text{ by full compliance}\\ &= \mathbb{E}[Y_1|R=1]-\mathbb{E}[Y_0|R=0], \text{ by full compliance}\\ &= \mathbb{E}[Y_1|R=1]-\mathbb{E}[Y_0|R=1], \text{ by random assignment}\\ &= \mathbb{E}[Y_1|D=1]-\mathbb{E}[Y_0|D=1], \text{ by full compliance}\end{aligned}$$

and so the unconditional ITT equals the TOT.

The PRTE is an entirely different object. Let p, p' be two policies, let ω index individuals, and let $s_p(\omega)$ represent the treatment individual ω takes under policy p. The PRTE is

$$\mathbb{E}[Y(s_p(\omega), \omega) - Y(s_{p'}(\omega), \omega)].$$

The PRTE represents the average effect of a policy that (potentially) changes the treatment level an individual selects but not the potential outcomes of an individual for fixed treatment level.

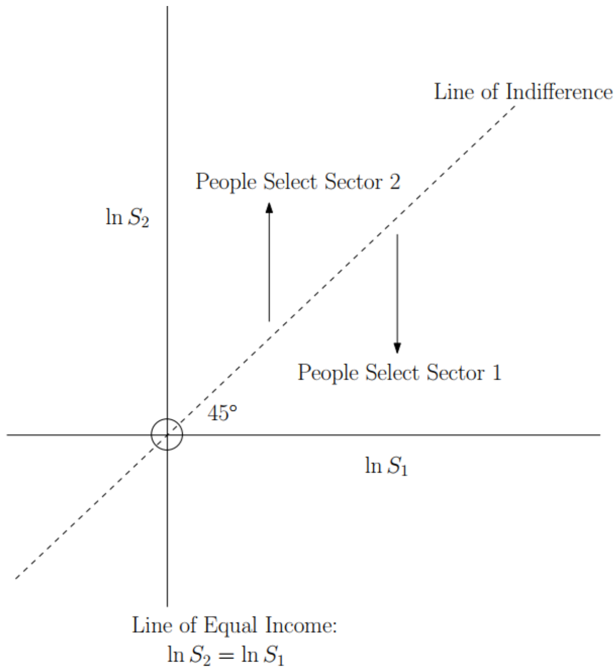
■

4 Roy Model and the Generalized Roy Model

Answer all of the questions in the posted handout "The Roy Model and the Generalized Roy Model."

Problem 4.1. (Slide 16) How does the sorting determine the distribution of observations on $\ln S_2$?

Solution. Recall that we had the following partition assuming $\pi_1 = \pi_2 = 1$



Agents choose sector 2 if his earnings are greater there:

$$W_1 = \pi_1 S_1 < \pi_2 S_2 = W_2 \Leftrightarrow \ln S_1 < \ln S_2$$

Therefore, if the true data generating process is of the following form:

$$(\ln S_1, \ln S_2) \sim N((\mu_1, \mu_2), \Sigma)$$

then we can write:

$$\begin{aligned}
\ln S_1 &= \mu_1 + U_1 \\
\ln S_2 &= \mu_2 + U_2 \\
(U_1, U_2)^\top &\sim N\left((0,0)^\top, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{pmatrix}\right)
\end{aligned}$$

Define $V := U_2 - U_1$. Then $V \sim N(0, \sigma_V^2)$ where

$$\sigma_V^2 = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

in which case:

$$(U_2, V)^\top \sim N\left((0,0)^\top, \begin{pmatrix} \sigma_{22}^2 & \sigma_{2V} \\ \sigma_{V2} & \sigma_V^2 \end{pmatrix}\right)$$

where

$$\sigma_{V2} = \sigma_{2V} = \mathrm{Cov}[U_2, V] = \mathrm{Cov}[U_2, U_2 - U_1] = \sigma_{22}^2 - \sigma_{21}$$

This allows us to finally write:

$$U_2 | V \sim N\left(\frac{\sigma_{22}}{\sigma_V^2} U_2, \sigma_{22}^2 - \frac{\sigma_{2V}}{\sigma_{22}^2}\right)$$

Therefore, our equation of interest becomes:

$$\begin{aligned} f(\ln S_2 | \ln S_2 > \ln S_1) &= P(\mu_2 + U_2 | \mu_2 - \mu_1 + U_2 - U_1 > 0) \\ &= P(\mu_2 + U_2 | V > \mu_1 - \mu_2) \end{aligned}$$

where the choice probability $P(\mu_2 + V > \mu_1 - \mu_2)$ can be modelled using the expressions above.

■

Problem 4.2. Specifically, how can the price of one skill rise but the wage associated with the skill decline if person i:

$$\text{(wages)} i = \pi \text{ (skill) } i$$

where π is the price of skill?

Solution. Using the setup in the lecture slides, the above case can arise if $\sigma_{11} > \sigma_{12}$.

▷ Recall that we have

$$\mathbb{E}(\ln W_1 | \ln W_1 > \ln W_2) = \ln \pi_1 + \mu_1 + \left(\frac{\sigma_{11} - \sigma_{12}}{\sigma^*}\right) \lambda(-c_1)$$

where

$$c_1 = \frac{\ln(\pi_1/\pi_2) + \mu_1 - \mu_2}{\sigma^*}, \quad \lambda(\cdot) = \mathbb{E}[Z | Z > \cdot], Z \sim N(0,1)$$

▷ The derivative of

$$\left(\frac{\sigma_{11}-\sigma_{12}}{\sigma^*}\right)\lambda(-c_1)$$

with respect to π_1 yields:

$$\frac{\partial\mathbb{E}(\ln W_1|\ln W_1>\ln W_2)}{\partial\ln\pi_1}=-\frac{(\sigma_{11}-\sigma_{12})}{(\sigma^*)^2}\lambda'(-c_1)$$

which is negative if $\sigma_{11}>\sigma_{12}$.

▷ Therefore, if σ_{11} is sufficiently larger than σ_{12}, then the wage associated with the skill may decline.

■

Problem 4.3. (Slide 54) Prove the following two claims:

▷ A 10% increase in π_1 produces a $<10\%$ increase in $\mathbb{E}[\ln W_1|\ln W_1>\ln W_2]$

▷ A 10% increase in π_2 produces a $>10\%$ increase in $\mathbb{E}[\ln W_2|\ln W_2>\ln W_1]$

Solution. Note that we are assuming $\sigma_{11}\geq\sigma_{12}\geq\sigma_{22}$.

1. A 10% increase in π_1 produces a $<10\%$ increase in $\mathbb{E}[\ln W_1|\ln W_1>\ln W_2]$

   ▷ Recall that we have

$$\mathbb{E}(\ln W_1|\ln W_1\leq\ln W_2)=\ln\pi_1+\mu_1+\left(\frac{\sigma_{11}-\sigma_{12}}{\sigma^*}\right)\lambda(-c_1)$$
$$\mathbb{E}(\ln S_1|\ln W_1>\ln W_2)=\mu_1+\left(\frac{\sigma_{11}-\sigma_{12}}{\sigma^*}\right)\lambda(-c_1)$$

   where

$$c_1=\frac{\ln(\pi_1/\pi_2)+\mu_1-\mu_2}{\sigma^*},\lambda(\cdot)=\mathbb{E}[Z|Z>\cdot],Z\sim N(0,1)$$

   ▷ Since

$$\mu_2+\left(\frac{\sigma_{22}-\sigma_{12}}{\sigma^*}\right)\lambda(-c_2)$$

   is decreasing in π_1 given our results in the previous part, and $\sigma_{11}>\sigma_{12}$, a 10% increase in π_1 produces a $<10\%$ increase in $\mathbb{E}[\ln W_1|\ln W_1\leq\ln W_2]$.

2. A 10% increase in π_2 produces a $>10\%$ increase in $\mathbb{E}[\ln W_2|\ln W_2>\ln W_1]$

▷ Note that both sectors cannot display negative selection since:

$$0 \leq \frac{\sigma_{12}}{\sigma_{11}} \frac{\sigma_{12}}{\sigma_{22}} \leq 1$$

which implies that

$$\sigma_{11} \leq \sigma_{12} \Rightarrow \sigma_{22} \leq \sigma_{12}$$

▷ Since we have:

$$\mathbb{E}(\ln W_2 | \ln W_1 \leq \ln W_2) = \ln \pi_2 + \mu_2 + \left(\frac{\sigma_{22} - \sigma_{12}}{\sigma^*} \right) \lambda(-c_2)$$

and

$$\mu_2 + \left(\frac{\sigma_{22} - \sigma_{12}}{\sigma^*} \right) \lambda(-c_2)$$

is increasing in π_2 given our results in the previous part, it follows that A 10% increase in π_2 produces a $> 10\%$ increase in $\mathbb{E}[\ln W_2 | \ln W_2 > \ln W_1]$.

■

5 LATE and the Generalized Roy Model: Some Relationships

Answer the questions embedded throughout the handout.

Problem 5.1. Derive the MTE from the sample selection model. What parameters are identified by the selection model that are not identified by MTE? Explain the advantages and disadvantages of each approach.

Solution. The MTE is defined as

$$E\left(Y_1 - Y_0 \mid U_D = u_D\right)$$

Sample selection model corrects the bias due to non-random sampling process, so it identifies the distribution of variables but MTE does not.Under MTE methodology, we first identify the ATE point-wise for the running variables, and then sum them up by appropriate weights. The advantage of MTE is that we can recover any wanted treatment effect by integrating MTE over some appropriate interval given that we have the right weights. One disadvantage is that we might misspecify the weights, and another drawback is that we might have small number of data points for each point estimate, which could lead to high variance.

Problem 5.2. In what sense is $E\left(Y_1 - Y_0 \mid P\left(z\right) \geq U_D\right)$ a measure of surplus of agents for whom $P\left(z\right) \geq U_D$?

Solution. Recall that

$$P\left(z\right) = F_{\left(\frac{V}{\sigma_V}\right)}\left(\frac{\mu_D\left(z\right)}{\sigma_V}\right)$$

$$U_D = F_{\left(\frac{V}{\sigma_V}\right)}\left(\frac{V}{\sigma_V}\right)$$

So,

$$\begin{aligned} E\left(Y_1 - Y_0 \mid P\left(z\right) \geq U_D\right) &= E\left(Y_1 - Y_0 \mid \frac{\mu_D\left(z\right)}{\sigma_V} \geq \frac{V}{\sigma_V}\right) \\ &= E\left(Y_1 - Y_0 \mid \mu_D\left(z\right) \geq V\right) \\ &= E\left(Y_1 - Y_0 \mid E\left(Y_1 - Y_0 \geq C \mid \mathcal{I}\right)\right) \end{aligned}$$

which is the expected benefit from treatment for those whose benefit is higher than the cost. So,

$$E\left(Y_1 - Y_0 \mid P\left(z\right) \geq U_D\right)$$

is the surplus plus the cost of treatment, which is closely related to the surplus.

Problem 5.3. Compare LATE and MTE for an instrument Z.

Solution. MTE($U_D = P(z)$) is the treatment effect for a particular value of instrument $Z = z$, while LATE is the average treatment effect for the compliers to switching treatment from z' to z. Mathematically,

$$\begin{aligned}\text{LATE} &= E\left(Y_1 - Y_0 \mid P(z) \geq U_D \geq P(z')\right) \\ &= \int_{P(z')}^{P(z)} \frac{E\left(Y_1 - Y_0 \mid U_D = u_D\right)}{P(z) - P(z')} du_D \\ &= \int_{P(z')}^{P(z)} \frac{\text{MTE}\left(u_D\right)}{P(z) - P(z')} du_D\end{aligned}$$

Problem 5.4. What is the role of $Pr(D = 1 \mid X; Z)$ in (i) LATE, (ii) MTE, (iii) selection bias models, and (iv) in propensity score matching estimators? Relate the parameters if you can.

Solution.

- ▷ In LATE, $Pr(D = 1 \mid X; Z = 1) = 1$ and $Pr(D = 1 \mid X; Z = 0) = 0$ pins down the group of compliers whose ATE we can identify
- ▷ In MTE, $Pr(D = 1 \mid X; Z = z) = P(z)$, which pins down each value for us to evaluate the MTE:

$$\begin{aligned}\text{MTE}\left(u_D\right) &= E\left(Y_1 - Y_0 \mid u_D = P(z)\right) \\ &= E\left(Y_1 - Y_0 \mid u_D = Pr(D = 1 \mid X; Z = z)\right)\end{aligned}$$

- ▷ In selection bias models, $Pr(D = 1 \mid X; Z)$ is the conditional probability of selecting a data point into the sample, which helps up recover the unconditional probability of selecting a data point into the sample:

$$Pr(D = 1) = \int_X \int_Z Pr(D = 1 \mid X; Z) dF_Z dF_X$$

- ▷ In propensity score matching estimators, we use $Pr(D = 1 \mid X; Z)$, the propensity score, as the matching variable to estimate the treatment effect in each matched cell, and then integrate across all cells.

Problem 5.5. How does $Pr(D = 1 \mid X; Z)$ naturally emerge as a key variable in LATE and MTE?

Solution.

- ▷ In LATE, $Pr(D = 1 \mid X; Z)$ is a key variable because we use $Pr(D = 1 \mid X; Z = 1) = 1$ and $Pr(D = 1 \mid X; Z = 0) = 0$ to find the group of compliers whose ATE we can identify.

- ▷ In MTE, $Pr(D = 1 \mid X; Z)$ is a key variable because $Pr(D = 1 \mid X; Z = z) = P(z)$ generates the value for us to evaluate the MTE by

$$\begin{aligned} \text{MTE}(u_D) &= E(Y_1 - Y_0 \mid u_D = P(z)) \\ &= E(Y_1 - Y_0 \mid u_D = Pr(D = 1 \mid X; Z = z)) \end{aligned}$$

6 Extension of 5

Under what conditions is the Generalized Roy model identified nonparametrically?

Solution. Following Heckman and HonorÃ© (1990), we can identify the population skill distribution under the Roy model from earnings data and data on sector choice under the following conditions.

▷ **Case 1:** The skill distribution is log-normal (parametric identification).

Then the joint lognormal skill distribution $LN(\mu, \Sigma)$ can be identified from a single cross section of earnings and sectoral choices.

▷ **Case 2:** The skill distribution is some general non-normal distribution.

Then in general we cannot identify the skill distribution from a single cross section of earnings and sectoral choices. However, we can still identify the population skill distribution if

- We have data on the earnings distributions in several different markets with the same skill distribution but different relative skill prices, **and**
- There is sufficient skill price variation:
  - If we have only one time period of data on multiple markets, then we need on of the skill prices to have support on $(0, \infty)$.
  - If we have a panel of data on multiple markets over time and the skill distribution is stationary, then we can weaken this condition. In the two skill case, we can identify the skill distribution $F(S_1, S_2)$ on $\pi_2 S_2 \leq S_1 \leq \pi_2' S_2$, where $\pi_2 < \pi_2'$ are observed skill prices in different markets for one skill.

Moreover, we can still identify the population skill distribution from a single cross-section of data if we have exogenous skill distribution shifters (and the associated exclusion restrictions). then we can still identify the population

7 Simultaneous Causality

Read the posted handout "Simultaneous Causality." Consider equations (2a) and 2(b):

$$[2a] : Y_1 = \alpha_1 + \gamma_{12} Y_2 + \beta_{11} X_1 + \beta_{12} X_2 + U_1$$
$$[2b] : Y_2 = \alpha_2 + \gamma_{21} Y_1 + \beta_{21} X_1 + \beta_{22} X_2 + U_2$$

Now suppose that $(X_1, X_2) \perp (U_1, U_2)$ with $U_1 \not\perp U_2$ and $\gamma_{12} \neq 0, \gamma_{21} \neq 0$.

Problem 7.1. Define the causal effect of Y_2 on Y_1 and of Y_1 on Y_2.

Solution. Applying Havelmo's (1943) analysis to [2a] and [2b], the causal effect of Y_2 on Y_1 is γ_{12}. We say this is the causal effect since this is the effect on Y_1 of fixing Y_2 at different values, holding constant the other variables in the equation. By symmetry, the causal effect of Y_1 on Y_2 is γ_{21}.

■

Problem 7.2. Are these causal effects identified?

Solution. The causal effects are not identified. This is because holding X_1, X_2, U_1, U_2 fixed in [2a] and [2b], it is not possible to vary Y_2 or Y_1 because they are exact functions of X_1, X_2, U_1, U_2. Note that this exact dependence holds true even if $U_1 = 0$ and $U_2 = 0$ (i.e. there are no unobservables.)

■

Problem 7.3. Suppose now that $\beta_{12} = 0$ and $\beta_{21} = 0$. Are the effects identified? Why or why not?

Solution. Under this assumption, we can identify both causal effects.

▷ To see this, consider the following model:

$$Y_1 = \pi_{10} + \pi_{11} X_1 + \pi_{12} X_2 + \epsilon_1$$
$$Y_2 = \pi_{20} + \pi_{21} X_1 + \pi_{22} X_2 + \epsilon_2$$

which yields:

$$\pi_{11} = \frac{\beta_{11} + \gamma_{12}\beta_{21}}{1 - \gamma_{12}\gamma_{21}}$$
$$\pi_{12} = \frac{\beta_{12} + \gamma_{12}\beta_{22}}{1 - \gamma_{12}\gamma_{21}}$$
$$\pi_{21} = \frac{\beta_{21} + \gamma_{21}\beta_{11}}{1 - \gamma_{12}\gamma_{21}}$$
$$\pi_{22} = \frac{\beta_{22} + \gamma_{12}\beta_{12}}{1 - \gamma_{12}\gamma_{21}}$$

▷ Plugging in $\beta_{12} = 0$ we obtain:

$$\frac{\pi_{12}}{\pi_{22}} = \gamma_{12}$$

and plugging in$\beta_{21} = 0$, we obtain:

$$\frac{\pi_{21}}{\pi_{11}} = \gamma_{21}$$

Therefore, we see that the causal effects are indeed identified.

■

Problem 7.4. Suppose instead of (c) that $U_1 \perp U_2$. Are the effects identified? Why or why not?

Solution. Yes. For example, we can exclude X_2 from $[2a]$, in which case we can vary Y_1 holding X_2 and U_2 constant.

■

Problem 7.5. Suppose $U_1 = U_2 = 0$ but $\beta_{12} \neq 0$ and $\beta_{21} \neq 0$. Is the model identified?

Solution. No, the model is not identified. Even if $U_1 = 0$ and $U_2 = 0$ so that there are no unobservables, we cannot simultaneously vary Y_2, X_1, X_2 since Y_1 is perfectly predictable by X_1 and X_2.

■