Potential Outcomes

- A "treatment" describes one of two states
- The "treatment status" for individual i is denoted by D_i which takes values of zero or one.
- Each individual has two counterfactual values for the outcome of interest
  - Y_{i0} is the outcome without treatment
  - Y_{i1} is the outcome with treatment
- The observed outcome is $Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}$
- Fisher (1951), Roy (1951), Rubin-Holland causal model

The Treatment Effect

- The effect of the treatment on individual i is $\Delta_i = Y_{i1} - Y_{i0}$
- But Δ_i is not directly observed!
- We might still be able to identify features of the distribution of Δ_i such as its moments or quantiles
- One can view the evaluation problem as one of missing data.

Parameters of Interest

- Average treatment effect (*ATE*): $E(\Delta_i)$.
- Average treatment effect on the treated (*ATET*): $E(\Delta_i | D_i = 1)$.
- Average treatment effect on the untreated (*ATEU*): $E(\Delta_i | D_i = 0)$.
- Notice that $ATE = Pr(D_i = 1)ATET + Pr(D_i = 0)ATEU$.
- Conditional versions: $ATE(x) = E(\Delta_i | X_i = x)$, $ATET(x) = E(\Delta_i | D_i = 1, X_i = x)$ and $ATEU(x) = E(\Delta_i | D_i = 0, X_i = x)$

Other Parameters of Interest

- Proportion of people benefiting from the program:

$$Pr(\Delta_i > 0 | D_i = 1)$$

- Distribution of treatment effects:

$$F(\Delta_i | D_i = 1)$$

- Selected quantile

$$\inf\{\Delta_i : F(\Delta_i | D_i = 1) > q\}$$

Main Identification Issues

The main difficulties here are:

- The effect Δ_i is (potentially) heterogenous. This implies that the various parameters may differ.

- The selection into treatment may depend on both Y_{i1} and Y_{i0} and consequently on the gains from the treatment (e.g., Roy model). This would render D_i endogenous.

Evaluation estimators are designed around assumptions that allow us to identify some feature of the distribution of Δ_i.

Roy Model

For example, let D_i denote one of two occupations chosen by person i (e.g., hunter or fisherman) and Y_{i0}, Y_{i1} are the wages in each occupation.

- Roy (1951) postulates that

$$D_i = 1[Y_{i1} > Y_{i0}]$$

- A generalized Roy model has

$$D_i = 1[Y_{i1} > Y_{i0} + C_i]$$

  where C_i is the direct cost of choosing 1 and is potentially heterogenous.

- Call it the *extended* Roy model if $C_i = C$ is constant.

Roy Model

- Re-write the model as

$$Y_{id} = \mu_d + U_d, \quad d = 0, 1$$

where $\mu_d = E(Y_{id})$ and $U_{di} = Y_{id} - \mu_d$.

- When covariates $X_i = x$ are present, let $\mu_d \equiv \mu_d(x) = E(Y_{id}|X_i = x)$.

- $ATE = E(\Delta_i) = E(\mu_1 - \mu_0 + U_{1i} - U_{0i}) = \mu_1 - \mu_0$

- $ATET = E(\Delta_i|D_i = 1) = ATE + E(U_{1i} - U_{0i}|D_i = 1)$

At this point it is worth representing outcomes and treatments as

$$
\begin{aligned}
Y_i &= D_i Y_{i1} + (1 - D_i) Y_{i0} \\
&= Y_{i0} + D_i (Y_{i1} - Y_{i0}) \\
&= \mu_0 + (\mu_1 - \mu_0 + U_{1i} - U_{0i}) D_i + U_{0i} \\
&= \alpha + \Delta_i D_i + \epsilon_i
\end{aligned}
$$

where $\alpha = \mu_0$, $\Delta_i = \mu_1 - \mu_0 + U_{1i} - U_{0i}$ and $\epsilon_i = U_{0i}$.

Let $\overline{\Delta} = \mu_1 - \mu_0 (= ATE)$, and $v_i = \Delta_i - \overline{\Delta} (= U_{1i} - U_{0i})$. (What is the ATET?)

This is a linear regression with a random coefficient.

Three Cases

1. The coefficient on D is fixed (given X_i) and is the same for everyone. This means that $U_{1i} = U_{0i}$ for every $i \Rightarrow \Delta_i = \overline{\Delta}$ for every i. ($ATE = ATET$.)
2. The coefficient on D is random (given X), but $U_{1i} - U_{0i}$ does not predict program participation.

   $$Pr(D_i = 1 | U_{1i} - U_{0i}) = Pr(D_i = 1) \Rightarrow E(U_{1i} - U_{0i} | D_i = 1) = 0$$

   There is heterogeneity ($v_i \neq 0$), but it is not acted upon ex ante. ($ATE = ATET$.)
3. The coefficient on D_i is random (given X_i) and $U_{1i} - U_{0i}$ predicts program participation: $E(U_{1i} - U_{0i} | D_i = 1) \neq 0$. (Roy Model.)

OLS

A natural impulse is to estimate the effect of D_i on Y_i via OLS. What does one obtain? Let $\{(Y_i, D_i); i = 1, \dots, N\}$ denote an iid sample.

$$\hat{\beta} = \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i D_i - \frac{1}{N}\sum_{i=1}^{N} Y_i \frac{1}{N}\sum_{i=1}^{N} D_i}{\frac{1}{N}\sum_{i=1}^{N} D_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} D_i\right)^2}$$

Let β_{OLS} denote the probability limit of β. For any $\epsilon > 0$,

$$Pr(|\hat{\beta} - \beta_{OLS}| > \epsilon) \to 0$$

as $N \to \infty$.

OLS

Because

$$\begin{aligned} E(D_i Y_i) &= E(Y_i|D_i = 1)Pr(D_i = 1) = E(Y_i|D_i = 1)E(D_i) \\ E(Y_i) &= E(Y_i|D_i = 1)Pr(D_i = 1) + E(Y_i|D_i = 0)Pr(D_i = 0) = \\ &= E(Y_i|D_i = 1)E(D_i) + E(Y_i|D_i = 0)(1 - E(D_i)) \end{aligned}$$

it is easy to show that

$$\begin{aligned} \beta_{OLS} &= E(Y_{i1}|D_i = 1) - E(Y_{i0}|D_i = 0) \\ &= ATET + E(Y_{i0}|D_i = 1) - E(Y_{i0}|D_i = 0) \end{aligned}$$

The second term is a selection effect. It still exists even if there is no impact heterogeneity (i.e. $v_i = 0$).

Potential Solutions to the Causal Inference Problem

A cadre of potential solutions to the evaluation of the problem is offered in the literature. Each suits different assumption, data and purpose scenarios:

1. Matching
2. Instrumental variables
3. DiD, RDD and quasi-natural experiments
4. Randomised control trials
5. Estimation of a structural economic model

(1-4) consider cases where treatment assignment D_i is (in some sense) independent of potential outcomes Y_{i0} and Y_{i1} (once conditioned on the relevant covariates). (5) seeks to model the selection into treatment.

Matching

- Matching estimators pair treated individuals ($D_i = 1$) with observably similarly untreated individuals ($D_i = 0$).
- To do that, we assume the *Conditional Independence Assumption* (CIA):

$$(Y_{i1}, Y_{i0}) \perp\!\!\!\perp D_i | X_i \quad \text{(i.e., } Pr[D_i | X_i, Y_{i0}, Y_{i1}] = Pr[D_i | X_i]\text{)}$$

- To justify this assumption, individuals cannot select into the program based on anticipated treatment impact.

Matching

Other assumptions in matching estimators are:

- *Common Support Assumption*: $0 < Pr(D_i = 1|X_i) < 1$ for any X_i. There is no x in the support of the covariates such that $D_i = 0$ or 1.
- *Stable Unit Treatment Value Assumption* (SUTVA): There are no spillovers. D_i has no impact on individual $j \neq i$.

  This rules out general equilibrium effects or social interactions.

Matching

- The CIA implies that

$$F(Y_{id}|D_i, X_i) = F(Y_{id}|X_i) \Rightarrow E(Y_{id}|D_i, X_i) = E(Y_{id}|X_i)$$

  for $d = 0, 1$.

- This implies that $ATE(X_i) = ATET(X_i) = ATEU(X_i)\ldots$

- …but does *not* imply $ATE = ATET = ATEU$ if $f(X_i|D_i = 1) \neq f(X_i)$.

- $ATET = E(ATET(X_i)|D_i = 1)$ and $ATE = E(ATE(X_i))$ by LIE.

Propensity Score Matching

- One immediate problem in implement a matching estimator is how to deal with high dimensional X_i.
- A typical dimension reducing strategy is to use Propensity Score Matching. The propensity score is defined as

$$P(x) = Pr(D_i = 1|X_i = x)$$

for all x in the support of X_i.

- **Theorem (Rosenbaum and Rubin):**

$$CIA \Rightarrow (Y_{i1}, Y_{i0}) \perp\!\!\!\perp D_i|P(X_i)$$

Propensity Score Matching

- **Proof:**

$$
\begin{aligned}
& Pr[D_i = 1 | Y_{i1}, Y_{i0}, P(X_i)] \\
= \quad & E\{Pr[D_i = 1 | Y_{i1}, Y_{i0}, X_i] | Y_{i1}, Y_{i0}, P(X_i)\} \quad \text{(by LIE)} \\
= \quad & E\{Pr[D_i = 1 | X_i] | Y_{i1}, Y_{i0}, P(X_i)\} \quad \text{(by CIA)} \\
= \quad & E\{P(X_i) | Y_{i1}, Y_{i0}, P(X_i)\} = P(X_i) = E\{P(X_i) | P(X_i)\} \\
= \quad & E\{Pr[D_i = 1 | X_i] | P(X_i)\} \quad \text{(by LIE)} \\
= \quad & Pr[D_i = 1 | P(X_i)]
\end{aligned}
$$

- So, we can reduce the problem to a unidimensional one if we know $P(\cdot)$! The catch is that if we do not know it we need to estimate it on a potentially highly dimensional X_i. This brings back the curse of dimensionality. (In practice one estimates a parametric propensity score.)

Propensity Score Matching

We can then estimate the *ATET* in two steps:

1. Estimate a model of program participation and obtain the propensity score $P(x_i)$ for each person

2. Select matches based on the estimated propensity score:

$$\widehat{ATET} = \frac{1}{\sum_{i=1}^{N} d_i} \sum_{i:d_i=1} [y_i - \hat{m}_0(x_i)]$$

where $\hat{m}_0(x_i)$ is an estimator for $E[Y_{0j}|P(X_j) = P(x_i), D_j = 0]$.

Propensity Score Matching

How does one estimate $\hat{m}_0(x_i)$?

Since the propensity score is unidimensional, it is easy to estimate it via kernel methods:

$$
\begin{aligned}
\hat{m}_0(x_i) &= \frac{\sum_{j:d_j=0} y_j \mathcal{K}(P(x_i) - P(x_j))}{\sum_{j:d_j=0} \mathcal{K}(P(x_i) - P(x_j))} \\
&= \frac{\widehat{E}\left\{y_j 1[D_j = 0, P(X_i) = P(X_j)]\right\}}{\widehat{P}(D_j = 0, P(X_i) = P(X_j))}
\end{aligned}
$$

Propensity Score Matching

1. Nearest-neighbor matching:

$$\mathcal{K}(P(x_i) - P(x_j)) = 1[j = \text{argmin}_{k \neq i} |P(x_i) - P(x_k)|].$$

2. Caliper matching: for some positive h

$$\mathcal{K}(P(x_i) - P(x_j)) = 1[|P(x_i) - P(x_j)| < h].$$

3. Kernel matching: $\mathcal{K}(u)$ is the density of a symmetric distribution such that $\mathcal{K}(u) > 0$, $\int \mathcal{K}(u)du = 1$, $\mathcal{K}(u) = \mathcal{K}(-u)$. The variance of this distribution controls the weight given to observations with similar propensity scores.
4. Local linear matching (Heckman, Ichimura and Todd (1997)).

- Remark: Estimation takes place only over the common support of X. If $P(x) = 1$ or $P(x) = 0$, this covariate value cannot be used.

Propensity Score Matching

To estimate *ATE*, notice that

$$
\begin{aligned}
& E\left[\frac{Y_i(D_i - P_i)}{P_i(1-P_i)}\right] \\
= \quad & E\left[\frac{E(Y_i|D_i = 1, P_i)P_i(1-P_i) + E(Y_i|D_i = 0, P_i)(-P_i)(1-P_i)}{P_i(1-P_i)}\right] \\
= \quad & E\left[E(Y_i|D_i = 1, P_i) - E(Y_i|D_i = 0, P_i)\right] \\
= \quad & ATE
\end{aligned}
$$

This suggests using

$$
\begin{aligned}
\widehat{ATE} &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i(d_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i d_i}{\hat{p}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{y_i(1-d_i)}{1-\hat{p}_i}
\end{aligned}
$$