Chapter 3. Selection on Observables. Matching

Joan Llull

Quantitative Statistical Methods II
Barcelona GSE

Selection Based on Observables and Matching

Experiments are often **expensive**, **unfeasible**, or **unethical** (smoking vs mortality) —or sometimes randomization is stratified $\Rightarrow$ observational data.

Independence is unlikely, but sometimes we can defend **conditional independence**:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i | X_i.$$

From Chapter 1:

$$\alpha_{ATE} = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i),$$

$$\alpha_{TT} = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i|D_i = 1).$$

Matching: compares individuals with the same characteristics and then integrates over the distribution of characteristics.

The common support condition

Essential condition: for each possible value of X, there are individuals in the treatment and control group for which we can average outcomes $\Rightarrow$ **common support condition**:

$$0 < P(D_i = 1|X_i) < 1 \quad \text{for all } X_i \text{ in its support.}$$

Counterexample (with a single covariate):

$$P(D_i = 1|X_i) = \begin{cases} 1 & \text{if } X_{min} \leq X < \underline{X} \\ p \in (0,1) & \text{if } \underline{X} \leq X \leq \overline{X} \\ 0 & \text{if } \overline{X} < X \leq X_{max} \end{cases}.$$

Implications:

- $\mathbb{E}[Y_i|D_i = 1, X_i]$ only identified for values of X_i in $(X_{min}, \overline{X})$,
- $\mathbb{E}[Y_i|D_i = 0, X_i]$ only identified for values of X_i in $(\underline{X}, X_{max})$,
- $\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$ only for values of X_i in the intersection range $(\underline{X}, \overline{X}) \Rightarrow \alpha_{ATE}$ and α_{TT} not identified.

Propensity Score Matching

Sometimes, set of variables X_i is too large or multivariate.

Not all info in X_i is relevant $\Rightarrow$ **propensity score matching**.

Rosenbaum and Rubin (1983) defined the **propensity score** as:

$$\pi(X_i) \equiv P(D_i = 1|X_i).$$

and note it is a **sufficient statistic** for the distribution of D_i.

Thus:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|X_i \quad \Leftrightarrow \quad Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|\pi(X_i),$$

$\Rightarrow$ match on the propensity score instead of the covariates.

Two-step methods: estimate propensity score, and then create the appropriate weighting.

Under (unconditional) **independence**:

$$\alpha_{ATE} = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \frac{\mathbb{E}[D_i Y_i]}{P(D_i = 1)} - \frac{\mathbb{E}[(1 - D_i)Y_i]}{P(D_i = 0)}.$$

Thus, under **conditional independence** we can write:

$$\begin{aligned}\mathbb{E}[Y_{1i} - Y_{0i}|X_i] &= \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i] \\ &= \mathbb{E}\left[Y_i \frac{D_i - \pi(X_i)}{\pi(X_i)(1 - \pi(X_i))}\middle| X_i\right],\end{aligned}$$

and:

$$\alpha_{ATE} = \mathbb{E}\left[\mathbb{E}[Y_{1i} - Y_{0i}|X_i]\right] = \mathbb{E}\left[Y_i \frac{D_i - \pi(X_i)}{\pi(X_i)[1 - \pi(X_i)]}\right].$$

Estimation: Discrete Low-Dimensional Case

Notation:

- X_i is discrete and takes on J possible values $\{x_j\}_{j=1}^J$,
- N observations $\{X_i\}_{i=1}^N$,
- N^j is the number of observations in cell j,
- N_ℓ^j be the number of observations in cell j with $D_i = \ell$,
- $\bar{Y}_\ell^j$ be the mean outcome in cell j for $D_i = \ell$.

Note $\bar{Y}_1^j - \bar{Y}_0^j$ is the sample counterpart of $\mathbb{E}[Y_i|D_i = 1, X_i = x_j] - \mathbb{E}[Y_i|D_i = 0, X_i = x_j]$, which can be used to get the following estimates:

$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N^j}{N}$$

$$\widehat{\alpha}_{TT} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N_1^j}{N_1} = \frac{1}{N_1} \sum_{i:D_i=1} \left(Y_i - \bar{Y}_0^{j(i)}\right).$$

where $j(i)$ indicates the cell to which X_i belongs (see the matching interpretation for $\widehat{\alpha}_{TT}$).

Estimation: Propensity Score Weighting

Using the sample analog of α_{ATE} in terms of the propensity score (**Hirano, Imbens, and Ridder, 2003**):

$$\widehat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i \left(\frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)[1 - \widehat{\pi}(X_i)]} \right),$$

where $\widehat{\pi}(X_i)$ is obtained in a **first stage** either nonparametrically, or by means of a flexibly specified Logit or Probit.

Estimation Methods: Weigthing

A **matching estimator** can be regarded as a way of constructing **imputations** for missing potential outcomes in a similar way, so that gains $Y_{1i} - Y_{0i}$ can be estimated for each unit.

In the **exact matching** we were doing:

$$\widehat{Y}_{0i} = \bar{Y}_0^{j(i)} \equiv \sum_{k:D_k=0} Y_k \frac{\mathbb{1}\{X_k = X_i\}}{\sum_{\ell:D_\ell=0} \mathbb{1}\{X_\ell = X_i\}}.$$

More generally we can compute:

$$\widehat{Y}_{0i} = \sum_{k:D_k=0} w(i,k) Y_k,$$

where **different weighting** $w(i,k)$ determine different estimators.

- **Nearest neighbor matching** (with replacement):

$$w(i,k) = \mathbb{1}\{X_k = \min_i ||X_k - X_i||\},$$

(picking the individual k in the control group with the closest observables to the individual i in the treated group).

- **Radius matching** (with replacement):

$$w(i,k) = \frac{\mathbb{1}\{||X_k - X_i|| < \varepsilon\}}{\sum_{\ell:D_\ell=0} \mathbb{1}\{||X_\ell - X_i|| < \varepsilon\}},$$

for some threshold ε (averages the observations from the control group with covariates within a window centered at X_i).

- **Kernel matching** (with replacement):

$$w(i,k) = \frac{\kappa\left(\frac{X_k - X_i}{\gamma_{N_0}}\right)}{\sum_{\ell:D_\ell=0} \kappa\left(\frac{X_\ell - X_i}{\gamma_{N_0}}\right)},$$

where $\kappa(\cdot)$ is a kernel function that downweights distant observations, and γ_{N_0} is a bandwidth parameter.

They can also be used for the **propensity score** $\pi(X_i)$ rather than X_i.

Matching vs Regression

Regression coefficient:

$$\beta_R = \mathbb{E}\left[\frac{\pi(X_i)(1-\pi(X_i))}{\mathbb{E}[\pi(X_i)(1-\pi(X_i))]}\delta_X\right] \neq \alpha_{ATE} = \mathbb{E}[\delta_X].$$

However, consistent **weighted average** treatment effect.

Weights based on **conditional variance** of D_i given X_i, namely $\pi(X_i)(1-\pi(X_i))$.

Advantages and Disadvantages of Matching

Advantages of matching:

- It avoids functional form assumptions.
- It emphasizes the common support condition.
- Focuses on a single parameter at a time, obtained through explicit aggregation.

Disadvantages of matching:

- Works under the presumption that for $X_i = x$ there is random variation in D_i, so that we can observe both Y_{0i} and $Y_{1i} \Rightarrow$ fails if D_i is a deterministic function of X_i (i.e. if $\pi(X_i)$ is 0 or 1).
- Good enough X_i may not have within-cell variation in D_i, but too much variation in D_i may be too little X_i.

Inference: Bootstrap Standard Errors

In matching, no straightforward way to compute **standard errors** $\Rightarrow$ **Bootstrap**.

Based on the intuition of why an estimator is a **random variable** for which we obtain a distribution

1. Obtain J size-N samples **resampling with replacement** from the original sample (intuition?)
2. For each of the J samples, **apply the matching procedure**, as we did to obtain our point estimates and store the results.
3. The bootstrap standard error is obtained as the **standard deviation** of our J stored matching estimates.