Chapter 3. Selection on Observables. Matching

Joan Llull

Quantitative & Statistical Methods II — Part I
Barcelona School of Economics

I. Selection Based on Observables and (Exact) Matching

There are many situations where experiments are too expensive, unfeasible, or unethical. A classical example is the analysis of the effects of smoking on mortality. Also, in experimental settings, often randomization is implemented conditional on observable characteristics. In any of these situations, we rely on observational data, which is unlikely to satisfy independence. In some situations, however, we can arguably defend the assumption of conditional independence, which is also referred to as selection based on observables:

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

When there is selection based on observables, the simple comparison of treatment and control averages does not deliver our treatment effects of interest, as the selection bias is not equal to zero. The problem is that the controls are not a counterfactual of treated in the absence of treatment, because the two groups differ in characteristics that are correlated with the outcome. As we discussed in Chapter 1, the average treatment effect is given by:

$$\alpha_{ATE} = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i), \tag{2}$$

and the average treatment effect on the treated is:

$$\alpha_{TT} = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i|D_i = 1). \tag{3}$$

What the above expressions do is to compare average outcomes for individuals with the same characteristics, and then integrate over the distribution of characteristics. In other words, for each treated (or control) unit, it imputes a counterfactual potential outcome when untreated (treated) obtained from individuals in the control (treatment) group that share the same characteristics. This imputation is called (exact) ***matching***, as it links each group of individuals in the treatment group with their counterparts in the control group (the "exact" qualifier is associated to the exercise of matching observations for every single value of X_i —below we review some alternatives that are more feasible when samples are

not very large or the number of different combinations of covariate values is large or infinite). Following the discussion in Chapter 1, the reason why Equations (**??**) and (**??**) are unbiased representations of α_{ATE} and α_{TT} is that, since the selection is based on observables, for a given X_i the assignment to treatment and control groups is random, and $\mathbb{E}[Y_i|D_i = 1, X_i] = \mathbb{E}[Y_{1i}|D_i = 1, X_i] = \mathbb{E}[Y_{1i}|X_i]$ and analogously $\mathbb{E}[Y_i|D_i = 0, X_i] = \mathbb{E}[Y_{0i}|D_i = 0, X_i] = \mathbb{E}[Y_{0i}|D_i = 1, X_i] = \mathbb{E}[Y_{0i}|X_i]$.

II. The Common Support Condition

An essential condition for matching is that there is some observation to match. In other words, for each possible value of X_i, there should be individuals in the treatment and control group for which we can average outcomes. This requirement is called the ***common support condition***. Formally, this condition is stated as:

$$0 < P(D_i = 1|X_i) < 1 \quad \text{for all } X_i \text{ in its support.} \tag{4}$$

For example, assume that X_i is a single covariate. Denote the support of X_i by (X_{min}, X_{max}). Assume that the support for the subpopulation of treated subjects is $(X_{min}, \overline{X})$, and the support for the controls is $(\underline{X}, X_{max})$, with $\overline{X} > \underline{X}$. Then:

$$P(D_i = 1|X_i) = \begin{cases} 1 & \text{if } X_{min} \leq X < \underline{X} \\ p \in (0,1) & \text{if } \underline{X} \leq X \leq \overline{X} \\ 0 & \text{if } \overline{X} < X \leq X_{max} \end{cases}. \tag{5}$$

Given these assumptions, $\mathbb{E}[Y_i|D_i = 1, X_i]$ is only identified for values of X_i in the range $(X_{min}, \overline{X})$, and $\mathbb{E}[Y_i|D_i = 0, X_i]$ is only identified for values of X_i in the range $(\underline{X}, X_{max})$. Thus, we can only compute the difference $\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$ for values of X_i in the intersection range $(\underline{X}, \overline{X})$, which implies that α_{ATE} and α_{TT} are not identified.

III. Propensity Score Matching

Sometimes, the set of variables on which we need to do the matching is too large or multivariate. However, not all information included in X_i is relevant to obtain independence. Rosenbaum and Rubin (1983) introduced the ***propensity score matching***, which is a method for reducing dimensionality based in the information that is relevant for independence. They define the propensity score, $\pi(X_i)$, as:

$$\pi(X_i) \equiv P(D_i = 1|X_i). \tag{6}$$

Then, they note that $\pi(X_i)$ is a sufficient statistic for the distribution of D_i by construction, as:

$$\begin{aligned} P(D_i = 1|Y_{1i}, Y_{0i}, \pi(X_i)) &= \mathbb{E}[D_i|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \mathbb{E}[\mathbb{E}[D_i|Y_{1i}, Y_{0i}, X_i]|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \mathbb{E}[\mathbb{E}[D_i|X_i]|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \mathbb{E}[P(D_i = 1|X_i)|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \mathbb{E}[\pi(X_i)|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \pi(X_i), \end{aligned} \tag{7}$$

where the third equality is obtained by using the conditional independence assumption. As a result, conditional independence given X_i is equivalent to conditional independence given the propensity score $\pi(X_i)$:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|X_i \quad \Leftrightarrow \quad Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|\pi(X_i). \tag{8}$$

Thus, instead of matching exactly (based on the different values of X_i), we can match all observations with the same propensity score, whether or not they share the same covariates X_i. That is the propensity score matching.

Intuitively, we can bunch all X_i that share the same propensity score together because then treated and control groups are not going to overrepresented in any of these characteristics. For example, consider the analysis of a training program offered to disadvantaged unemployed workers. Let X_i be a vector of race and gender. Let the propensity scores and distribution of characteristics be given by:

	Propensity score		Probability mass function	
	black	white	black	white
male	0.3	0.1	0.1	0.4
female	0.8	0.3	0.1	0.4

In this example, black male and white female have the same probability of receiving treatment. The fraction of treated black male in the population is $0.3 \times 0.1 = 0.03$ and the one of treated white female is $0.3 \times 0.4 = 0.12$. For controls, these fractions are 0.07 and 0.28. Thus, if we restrict the sample to black male and white female, we observe that $0.03/(0.03 + 0.12) = 20\%$ of treated individuals in this subsample are black male, and 80% are white female, and the same in the control group, $0.07/(0.07 + 0.28) = 20\%$ and 80%. Thus, within this subsample, there is no selection bias, as the treated and control groups are representative of

the same subpopulation (this is, also $0.1/(0.1+0.4) = 20\%$ of individuals in the subpopulation are black male, and 80% are white female).

This result suggests two-step procedures to estimate the treatment effects where first we estimate the propensity score, and then create the appropriate weighting. To do so, we rewrite α_{ATE} in terms of the propensity score. Under (unconditional) independence:

$$\alpha_{ATE} = \beta = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \frac{\mathbb{E}[D_i Y_i]}{P(D_i = 1)} - \frac{\mathbb{E}[(1 - D_i)Y_i]}{P(D_i = 0)}, \quad (9)$$

where the first equality has been discussed in Chapter 1 and last equality is obtained noting that $\mathbb{E}[D_i Y_i] = \mathbb{E}[Y_{1i}|D_i = 1]P(D_i = 1)$ and analogously for $\mathbb{E}[(1 - D_i)Y_i]$. Thus, under conditional independence we can write:

$$\begin{aligned}
\mathbb{E}[Y_{1i} - Y_{0i}|X_i] &= \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i] \\
&= \frac{\mathbb{E}[D_i Y_i|X_i]}{P(D_i = 1|X_i)} - \frac{\mathbb{E}[(1 - D_i)Y_i|X_i]}{P(D_i = 0|X_i)} \\
&= \frac{\mathbb{E}[D_i Y_i|X_i]}{\pi(X_i)} - \frac{\mathbb{E}[(1 - D_i)Y_i|X_i]}{1 - \pi(X_i)} \\
&= \mathbb{E}\left[\frac{D_i Y_i}{\pi(X_i)} - \frac{(1 - D_i)Y_i}{1 - \pi(X_i)}\middle| X_i\right] \\
&= \mathbb{E}\left[Y_i \frac{D_i - \pi(X_i)}{\pi(X_i)(1 - \pi(X_i))}\middle| X_i\right], \quad (10)
\end{aligned}$$

and:

$$\alpha_{ATE} = \mathbb{E}\left[\mathbb{E}[Y_{1i} - Y_{0i}|X_i]\right] = \mathbb{E}\left[Y_i \frac{D_i - \pi(X_i)}{\pi(X_i)[1 - \pi(X_i)]}\right]. \quad (11)$$

This expression constitute an estimand to make inference on by one of the estimation methods described below.

To gain intuition on this expression, note that observations with $D_i = 1$ have a contribution of $Y_i/\pi(X_i)$, whereas observations with $D_i = 0$ have a contribution of $-Y_i/(1 - \pi(X_i))$. In practice, what we are computing is a weighted average difference between observations in the treated group and in the control group. In our example before, we relatively underweight observations of black female in the treated group (their weight is $1/0.8 = 1.25$) because they are overrepresented (the overall fraction of treated is $0.1 \times 0.4 + 0.3 \times (0.1 + 0.4) + 0.8 \times 0.1 = 0.27$, and the fraction of them that are black female is $(0.8 \times 0.1)/0.27 = 29.6\%$, much larger than the 10% they represent overall), whereas white male are overweighted in this group (their weight is $1/0.1 = 10$), as they are underrepresented ($(0.1 \times 0.4)/0.27 = 14.8\% < 40\%$). The reverse is true for the control group.

IV. Estimation methods

The first and simplest method for matching estimation only works if X_i is discrete and relatively low-dimensional. Suppose X_i is indeed discrete and takes on J possible values $\{x_j\}_{j=1}^J$, and we have a sample of N observations $\{X_i\}_{i=1}^N$. Let N^j be the number of observations in cell j, N_ℓ^j be the number of observations in cell j with $D_i = \ell$, and $\bar{Y}_\ell^j$ be the mean outcome in cell j for $D_i = \ell$. With this notation, $\bar{Y}_1^j - \bar{Y}_0^j$ 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 obtain the following estimates:

$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N^j}{N} \tag{12}$$

$$\widehat{\alpha}_{TT} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N_1^j}{N_1}. \tag{13}$$

Note that the formula for $\widehat{\alpha}_{TT}$ can also be written in the form:

$$\widehat{\alpha}_{TT} = \frac{1}{N_1} \sum_{i:D_i=1} \left(Y_i - \bar{Y}_0^{j(i)}\right), \tag{14}$$

where $j(i)$ indicates the cell of X_i. Thus, $\widehat{\alpha}_{TT}$ matches the outcome of each treated unit with the mean of untreated units in the same cell. In practice, this is a way of imputing the missing potential outcome for the treated individuals, and compute the average treatment effect for them. Note that this expression is the sample analog of Equation (**??**). We can proceed analogously with the propensity score $\pi(X_i)$ instead of the regressors X_i.

Alternatively, a straightforward way to perform propensity score matching estimation was proposed by Hirano, Imbens, and Ridder (2003). This method essentially estimates a sample analog of Equation (**??**), which we implement in two stages. In a first stage, we estimate $\widehat{\pi}(X_i)$ either non-parametrically or by means of a flexible parametric model like a Logit or Probit with polynomials, interactions, and the alike. In a second stage, we estimate the following quantity:

$$\widehat{\alpha}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i \left(\frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)[1 - \widehat{\pi}(X_i)]} \right). \tag{15}$$

More generally, 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. For example, in Equation (**??**), the imputation is:

$$\widehat{Y}_{0i} = \bar{Y}_0^{j(i)} \equiv \sum_{k:D_k=0} Y_k \frac{\mathbb{1}\{X_k = X_i\}}{\sum_{\ell:D_\ell=0} \mathbb{1}\{X_\ell = X_i\}}. \tag{16}$$

More generally we compute:

$$\widehat{Y}_{0i} = \sum_{k:D_k=0} w(i,k) Y_k, \tag{17}$$

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

The ***nearest neighbor matching*** uses the following weighting function:

$$w(i,k) = \mathbb{1}\{X_k = \min_i ||X_k - X_i||\}, \tag{18}$$

which, in words, means picking the individual k in the control group with the closest observables to the individual i in the treated group. Alternatively, the ***radius matching*** uses:

$$w(i,k) = \frac{\mathbb{1}\{||X_k - X_i|| < \varepsilon\}}{\sum_{\ell:D_\ell=0} \mathbb{1}\{||X_\ell - X_i|| < \varepsilon\}}, \tag{19}$$

for some threshold ε. In words, this procedure averages the observations from the control group with covariates within a window centered at X_i. And finally, the ***kernel matching*** uses:

$$w(i,k) = \frac{\kappa\left(\frac{X_k - X_i}{\gamma_{N_0}}\right)}{\sum_{\ell:D_\ell=0} \kappa\left(\frac{X_\ell - X_i}{\gamma_{N_0}}\right)}, \tag{20}$$

where $\kappa(\cdot)$ is a kernel function that downweights distant observations, and γ_{N_0} is a bandwidth parameter. These procedures are generally implemented with replacement, meaning that each individual in the control group can be selected as a counterfactual for more than one individual in the treated group. Also they are typically applied to compute α_{TT}, but they are also applicable to α_{ATE}. And, furthermore, they can also be implemented on the propensity score $\pi(X_i)$ rather than the covariates X_i.

V. Matching versus Regression

Matching can be seen as an alternative to linear regression. Given the conditional independence assumption, the regression:

$$Y_i = \beta_0 + \beta_R D_i + \beta_X X_i + U_i, \qquad \text{with } \mathbb{E}[U_i | X_i, D_i] = 0, \tag{21}$$

would provide, by assumption, a consistent estimate of β_R (here we introduce X_i linearly without loss of generality, as we could redefine our vector of regressors to fully saturate the model with a set of dummies for each of the possible values of X_i). To prove it, let $\widetilde{D}_i$ denote the regression residual of D_i on X_i, defined as $\widetilde{D}_i \equiv D_i - \mathbb{E}[D_i|X_i]$. Likewise, let $\widetilde{Y}_i$ denote the corresponding residual for the observed outcome, defined as $\widetilde{Y}_i \equiv Y_i - \mathbb{E}[Y_i|X_i]$. Then:

$$\beta_R = \frac{\text{Cov}(\widetilde{Y}_i, \widetilde{D}_i)}{\text{Var}(\widetilde{D}_i)}. \tag{22}$$

As in Chapter 1, we operate this expression in pieces. First, we note that $\mathbb{E}[D_i|X_i] = \pi(X_i)$ by definition. Then, we operate the denominator:

$$\begin{aligned}
\text{Var}(\widetilde{D}_i) &= \mathbb{E}[(D_i - \pi(X_i))^2] - \mathbb{E}[D_i - \pi(X_i)]^2 \\
&= \mathbb{E}[(D_i - \pi(X_i))^2] \\
&= \mathbb{E}[D_i - 2\pi(X_i)D_i + \pi(X_i)^2] \\
&= \mathbb{E}[\mathbb{E}[D_i|X_i] - 2\pi(X_i)\,\mathbb{E}[D_i|X_i] + \pi(X_i)^2] \\
&= \mathbb{E}[\pi(X_i) - \pi(X_i)^2] \\
&= \mathbb{E}[\pi(X_i)(1 - \pi(X_i))],
\end{aligned} \tag{23}$$

where the second equality uses that $\mathbb{E}[D_i - \pi(X_i)] = \mathbb{E}[\mathbb{E}[D_i|X_i] - \pi(X_i)] = 0$, the third equality uses $D_i^2 = D_i$, and the fourth equality uses the law of iterated expectations in a similar way. And finally, we operate the numerator:

$$\begin{aligned}
\text{Cov}(\widetilde{Y}_i, \widetilde{D}_i) &= \mathbb{E}[\{D_i - \pi(X_i)\}\{Y_i - \mathbb{E}[Y_i|X_i]\}] - \mathbb{E}[D_i - \pi(X_i)]\,\mathbb{E}[Y_i - \mathbb{E}[Y_i|X_i])] \\
&= \mathbb{E}[\{D_i - \pi(X_i)\}\{Y_i - \mathbb{E}[Y_i|X_i]\}] \\
&= \mathbb{E}[\{D_i - \pi(X_i)\}Y_i] \\
&= \mathbb{E}[\{D_i - \pi(X_i)\}\,\mathbb{E}[Y_i|D_i, X_i]],
\end{aligned} \tag{24}$$

where, as before, the second equality uses $\mathbb{E}[D_i - \pi(X_i)] = \mathbb{E}[Y_i - \mathbb{E}[Y_i|D_i])] = 0$, the third one uses the law of iterated expectations to get $\mathbb{E}[\{D_i - \pi(X_i)\}\,\mathbb{E}[Y_i|X_i]] = \mathbb{E}[\{\pi(X_i) - \pi(X_i)\}\,\mathbb{E}[Y_i|X_i]] = 0$, and the last one uses again the law of iterated expectation. To simplify further, note that:

$$\mathbb{E}[Y_i|D_i, X] = \mathbb{E}[Y_i|D_i = 0, X] + \delta_X D_i, \tag{25}$$

where $\delta_X \equiv \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$. Thus:

$$\begin{aligned}\mathbb{E}[\{D_i - \pi(X_i)\}\,\mathbb{E}[Y_i|D_i, X_i]] &= \mathbb{E}[\{D_i - \pi(X_i)\}\,\mathbb{E}[Y_i|D_i = 0, X_i]] \\ &\quad + \mathbb{E}[\{D_i - \pi(X_i)\}D_i\delta_X] \\ &= \mathbb{E}[\{D_i - \pi(X_i)\}D_i\delta_X] \\ &= \mathbb{E}[\{D_i - \pi(X_i)D_i\}\delta_X] \\ &= \mathbb{E}[\pi(X_i)(1 - \pi(X_i))\delta_X], \end{aligned} \tag{26}$$

where the second equality is obtained by noting that $\mathbb{E}[Y_i|D_i = 0, X_i]$ only depends on X_i and, thus, it is once again orthogonal to $(D_i - \pi(X_i))$, and the last equality makes use of the law of iterated expectations. Thus:

$$\beta_R = \mathbb{E}\left[\frac{\pi(X_i)(1 - \pi(X_i))}{\mathbb{E}[\pi(X_i)(1 - \pi(X_i))]}\delta_X\right] \neq \alpha_{ATE} = \mathbb{E}[\delta_X]. \tag{27}$$

Thus, β_R provides a consistent average treatment effect, as weights sum to one, but this average is weighted. Noting that $\pi(X_i)(1 - \pi(X_i))$ is the variance of D_i given X_i, β_R provides a conditional variance-weighted average treatment effect. Thus, regression and matching provide, in general, different estimands, even if both are consistent estimates of the average treatment effect, and it is natural to compare the two.

The main advantages of matching are that it avoids functional form assumptions and it emphasizes the common support condition. Matching focuses on a single parameter at a time, which is obtained through explicit aggregation. On the downside, matching works under the presumption that for X_i there is random variation in D_i, so that we can observe both Y_{0i} and Y_{1i}. Hence, it fails if D_i is a deterministic function of X_i, that is, if $\pi(X_i)$ is either 0 or 1. Additionally, there is a tension between the thought that if X_i is good enough then there may not be within-cell variation in D_i, and the suspicion that seeing enough variation in D_i given X_i is an indication that exogeneity is at fault.

VI. Inference: Bootstrap Standard Errors

In the context on matching, it is not straightforward how to compute standard errors. There are no general asymptotic formulas to apply here, and the general practice is to use some ***bootstrap*** procedure to obtain standard errors and make inference. Thus, in the remaining of this section, we introduce the main intuition about bootstrap and the computation of bootstrapped standard errors.

First, recall that the reason why we compute standard errors is because an estimator, as a combination of random variables, is itself a random variable. Thus,

it has a distribution. What this means in practice is that if we were able to go back to the population and draw a different sample with the same procedure we would obtain a different estimate. If we could repeat this process infinite times, and we draw a histogram with the estimates we obtain each time we would obtain the distribution of our estimator.

Bootstrap standard errors (and other bootstrapped statistics) are computed following this notion closely. Unfortunately, it is costly (and in general, not possible) to go back to the population and obtain other samples. Thus, we re-sample (with replacement) from within our sample. To gain intuition about it, consider two discrete random variables (Y_i, D_i). Assume in our sample, the proportions of $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ observations is 25% each. If our sample is large enough, this probably means that the proportions in the population are close to 25% each. Thus, if we draw J samples of N observations with replacement (without replacement we would always trivially obtain the same sample!), the probability that each of the pairs is drawn is 25%, but some samples will have, say, more observations with $(0,0)$ observations and others more with $(1,1)$ observations. Thus, this resampling procedure would provide us with J different samples obtained from the same population.

Following this argument, this is how bootstrap works in practice in our context. First, we obtain J different samples of (Y_i, D_i, X_i) obtained from redrawing from our sample. Then, for each of this sample we apply the whole matching procedure, as we did to obtain our point estimates. With each sample we obtain an estimate, which we store. Finally, the bootstrap standard error is obtained as the standard deviation of our J stored matching estimates.