7 Matching

1. Regression as a matching estimator: theory.

   In the first exercise, you formally establish that regression is a special kind of matching estimator. Throughout, you can assume that the CIA holds.

   (a) First, show that the matching estimator for the ATT can be written as follows:

   $$\delta_M = \frac{\sum_x \delta_x Pr(D=1|X=x)Pr(X=x)}{\sum_x Pr(D=1|X=x)Pr(X=x)}, \tag{1}$$

   with $\delta_x \equiv \mathbb{E}[Y_i|X_i, D_i=1] - \mathbb{E}[Y_i|X_i, D_i=0]$.

   Now assume that X_i is a single variable taking on a finite number of discrete values, which means that we can formulate a 'saturated' model for the outcome variable regarding the regressor X_i [1].

   (b) First, show that the regression estimator can be written as follows:

   $$\delta_{\text{R}} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])\mathbb{E}[Y_i|D_i, X_i]]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2]}, \tag{2}$$

   with $\tilde{D}_i$ denoting the residual from a regression of D on X.
   Hint: Start noting that $\delta_R = \frac{Cov(Y_i, \tilde{D}_i)}{\mathbb{V}(\tilde{D}_i)} = \frac{\mathbb{E}[\tilde{D}_i Y_i]}{\mathbb{V}(\tilde{D}_i)}$. *Further note that* $\tilde{D}_i = D_i - \mathbb{E}[D_i|X_i]$ *because the model is saturated in* X.

   (c) Second, show that the nominator of equation (2) can be written as follows:

   $$\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])\mathbb{E}[Y_i|D_i, X_i]] = \mathbb{E}\{(D_i - \mathbb{E}[D_i|X_i])^2 \delta_x\}, \tag{3}$$

   with δ_x denoting the conditional treatment effect.
   Hint: Note that $\mathbb{E}[Y_i|D_i, X_i]$ *can be written as* $\mathbb{E}[Y_i|D_i=0, X_i] + \delta_x D_i$.

   (d) Third, show that this further simplifies to:

   $$\delta_{\text{R}} = \frac{\mathbb{E}[\sigma_D^2(x)\delta_x]}{\mathbb{E}[\sigma_D^2(x)]}, \tag{4}$$

   where $\sigma_D^2(x) \equiv \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]$ is the conditional variance of D_i given X_i.
   Hint: Use the law of iterated expectations.

   (e) Finally, show that:

   $$\delta_R = \frac{\sum_x \delta_x [Pr(D=1|X=x)(1 - Pr(D=1|X=x))]Pr(X=x)}{\sum_x [Pr(D=1|X=x)(1 - Pr(D=1|X=x))]Pr(X=x)} \tag{5}$$

[1] In other words, in this setting we define 'saturated' a regression model where, instead of X_i, we include a dummy for each value of X_i.

2. Regression as a matching estimator: application.

   The data is given in the Figure below, and also available on OLAT as an Excel spreadsheet `Berkeley.xls` for convenience.

Data from Berkeley 1973: Top 6 departments by enrollment

Major	Applicants	Admit	Deny	Total	Pct. Admitted
A	Men	512	314	825	62%
B	Men	353	207	560	63%
C	Men	120	205	325	37%
D	Men	138	279	417	33%
E	Men	53	138	191	28%
F	Men	22	351	373	6%
A	Women	89	19	108	82%
B	Women	17	8	25	68%
C	Women	202	391	593	34%
D	Women	131	244	375	35%
E	Women	94	299	393	24%
F	Women	24	317	341	7%

Figure 1: The outcome variable is *admit* (yes/no) and the treatment is the gender of the applicant, i.e., *male* (yes/no)

   (a) Explain the sense in which gender can be a treatment.

   (b) Assuming full independence between treatment and potential outcomes, what is the average treatment effect of gender on admission?

   (c) Do you find the assumption of full independence plausible? Why, or why not?

   (d) A weaker assumption is conditional independence. Assume that treatment and potential outcomes are independent conditional on the field of study and compute the unconditional average treatment effect (ATE), as well as the treatment effect on the treated (ATT).

   (e) Why do your answers in b) and d) differ?

   (f) Define five dummy-variables $fieldx$ that are 1 if the field $x = B, \ldots, E$ and run the regression

$$accept = \beta_0 + \beta_1 male + \beta_2 fieldB + \cdots + \beta_6 fieldE + \epsilon \tag{6}$$

   It is useful to start with a dataset of $6 \times 2 \times 2 = 24$ "observations", where each line represents a unique combination of *accept*, *male*, and $field$. In a next step, you can use the STATA `expand` command to generate n identical copies of each line. The final dataset should have 4526 observation.

   Alternatively you can use the frequency `fw` option without expanding the data.

   (g) Are your results from OLS and matching the same? If not, why?