Problem Set 2

Due October 7th, 6:40 pm

1. Neymanian inference and OLS

Consider a completely randomized experiment with n units. For unit i, let Z_i be the binary treatment indicator and Y_i be the observed outcome. We have discussed the Neymanian inference for the average causal effect in class, i.e, the unbiased estimator $\hat{\tau}$ and a conservative estimator for its variance $\hat{V}$.

Practitioners often use regression-based inference for the average causal effect, i.e., running OLS of the outcomes on the treatment indicators and using the coefficient of the treatment as the estimator. Show that

(1) this point estimator is identical to Neyman's unbiased estimator $\hat{\tau}$;

(2) the variance estimator from the classical OLS (for example, the one reported by the lm function of R) can be asymptotically different from $\hat{V}$, i.e., their ratio does not converge to 1 in probability even with a large sample size;

(3) the Eicker–Huber–White variance estimator from the OLS is asymptotically equivalent to $\hat{V}$, i.e., their ratio converges to 1 in probability. Note that if you run OLS of Y_i on $W_i = (1, Z_i)$ with coefficient $\hat{\beta}$, the Eicker–Huber–White variance estimator for the coefficient is

$$(W'W)^{-1}W'\text{diag}(\hat{e}_1^2, \dots, \hat{e}_n^2)W(W'W)^{-1},$$

where W is the matrix with rows of $(1, Z_i)$ and $\hat{e}_i = Y_i - W_i\hat{\beta}$ is the residual from the OLS.

2. FRT for the Project STAR data in the Imbens–Rubin book

Reanalyze the Project STAR data in Table 9.1 (at the end of this file) using the Fisher randomization test. Note that I use Z for the treatment indicator but the book uses W. Use $\hat{\tau}_{\text{S}}$, V and the aligned rank statistic in the Fisher randomization test. Compare the p-values.

3. Data re-analyses

(1) Re-analyze the LaLonde data used in `Neymanlalonde.R`. Conduct both Fisherian and Neymanian inferences.

The original experiment is a completely randomized experiment. Now we pretend that the original experiment is a stratified randomized experiment. First, re-analyze the data pretending that the experiment is stratified on the race (black, Hispanic or other). Second, re-analyze the data pretending that the experiment is stratified on marital status. Third, re-analyze the data pretending that the experiment is stratified on the indicator of high school diploma.

Compare with the results obtained under a completely randomized experiments.

(2) Re-analyze the data used in `SRE_Neyman_penn.R`. The analysis in class uses the treatment indicator, the outcome and the block indicator. Now we want to use all other covariates.

Conduct regression adjustments within strata of the experiment, and then combine these adjusted estimators to estimate the average causal effect. Report the point estimator, estimated standard error and 95% confidence interval. Compare them with those without regression adjustments.

4. Regression adjustment / post-stratification of CRE

In a completely randomized experiment, we have a binary treatment indicator Z, an outcome Y, and a discrete covariate $X \in \{1, \dots, K\}$. We can use two methods to use covariates to improve estimation efficiency.

First, we can use the post-stratification estimator, pretending that the experiment is a stratified randomized experiment, conditional on X. This estimator is $\hat{\tau}_{\text{PS}}$.

Second, we can create $K-1$ dummy variables $\delta_{1i} = I(X_i = 1), \dots, \delta_{K-1,i} = I(X_i = K-1)$, and use Lin's estimator with covariates $\delta_1, \dots, \delta_{K-1}$. This estimator is $\hat{\tau}_{\mathrm{L}}$.

Show that $\hat{\tau}_{\mathrm{PS}} = \hat{\tau}_{\mathrm{L}}$. Note that sometimes $\hat{\tau}_{\mathrm{PS}}$ or $\hat{\tau}_{\mathrm{L}}$ may not be well-defined. In those cases, we treat $\hat{\tau}_{\mathrm{PS}}$ and $\hat{\tau}_{\mathrm{L}}$ as equal.

For students in Stat 157, you can assume that $K = 2$.

5. Additional comments on the Neymanian inference under an SRE (Problem (2) only for Stat 260)

(1) Show that if $e_{[k]} = e$ for all k, then $\hat{\tau} = \hat{\tau}_{\mathrm{S}}$.

(2) Assume that the individual causal effect is constant $\tau_i = \tau$ for all $i = 1, \dots, n$. Consider the following class of weighted estimator for τ:

$$\hat{\tau}_w = \sum_{k=1}^{K} w_{[k]} \hat{\tau}_{[k]},$$

where $w_{[k]} \geq 0$ for all k.

When will $\hat{\tau}_w$ be unbiased for τ? Among all unbiased estimators, find the one with minimum variance.

Table 9.1. ***Class Average Mathematics Scores from Project Star***

School/ Stratum	No. of Classes	Regular Classes ($W_i = 0$)	Small Classes ($W_i = 1$)
1	4	−0.197, 0.236	0.165, 0.321
2	4	0.117, 1.190	0.918, −0.202
3	5	−0.496, 0.225	0.341, 0.561, −0.059
4	4	−1.104, −0.956	−0.024, −0.450
5	4	−0.126, 0.106	−0.258, −0.083
6	4	−0.597, −0.495	1.151, 0.707
7	4	0.685, 0.270	0.077, 0.371
8	6	−0.934, −0.633	−0.870, −0.496, −0.444, 0.392
9	4	−0.891, −0.856	−0.568, −1.189
10	4	−0.473, −0.807	−0.727, −0.580
11	4	−0.383, 0.313	−0.533, 0.458
12	5	0.474, 0.140	1.001, 0.102, 0.484
13	4	0.205, 0.296	0.855, 0.509
14	4	0.742, 0.175	0.618, 0.978
15	4	−0.434, −0.293	−0.545, 0.234
16	4	0.355, −0.130	−0.240, −0.150