Problem Set 3

Due October 18th, 11pm

Note that October 21st will be our in-class mid-term.

1. A predictive estimator and Lin's estimator

Consider a completely randomized experiment. Let Z_i, x_i and Y_i be the binary treatment, centered covariates, and outcome for unit i, $i = 1, \ldots, n$. We can use Lin's estimator $\hat{\tau}_{\mathrm{L}}$ to estimate the average treatment effect.

We also discussed a strategy to impute all missing potential outcomes. From the treatment group, we can use the OLS to fit a linear predictor for the potential outcome under treatment: $\hat{\mu}_1(x_i) = \hat{\gamma}_1 + \hat{\beta}_1^{\mathsf{T}} x_i$. From the control group, we can use the OLS to fit a linear predictor for the potential outcome under control: $\hat{\mu}_0(x_i) = \hat{\gamma}_0 + \hat{\beta}_0^{\mathsf{T}} x_i$. Then we can use these predictors to impute the missing potential outcome, leading to a predictive estimator

$$\hat{\tau}_{\mathrm{pre}} = \frac{1}{n} \left\{ \sum_{Z_i=1} Y_i + \sum_{Z_i=0} \hat{\mu}_1(x_i) - \sum_{Z_i=1} \hat{\mu}_0(x_i) - \sum_{Z_i=0} Y_i \right\}.$$

In class, I claimed that

$$\hat{\tau}_{\mathrm{L}} = \hat{\tau}_{\mathrm{pre}} = \hat{\gamma}_1 - \hat{\gamma}_0 = \left\{ \hat{\bar{Y}}(1) - \hat{\beta}_1^{\mathsf{T}} \hat{\bar{x}}(1) \right\} - \left\{ \hat{\bar{Y}}(0) - \hat{\beta}_0^{\mathsf{T}} \hat{\bar{x}}(0) \right\}.$$

Show the above identities using the properties of the OLS.

2. Data re-analyses

Re-analyze three datasets from matched-pair designs.

(1) In `FRTDarwinMP.R`, I analyze Darwin's data using the FRT based on the test statistic $\hat{\tau}$.

Re-analyze this dataset using the FRT with the Wilcoxon signed rank sum statistic.

Re-analyze this dataset based on the Neymanian inference: unbiased point estimator, conservative variance estimator, 95% confidence interval.

(2) In `NeymanMPstar.R`, I analyze the data from based on Neymanian inference.

Re-analyze this dataset using the FRT with different test statistics.

Re-analyze this dataset using the FRT with covariate adjustment, e.g., you can define test statistics based on residuals from the OLS fit of the observed outcome on covariates. Will the conclusion change if you do not include an intercept in your OLS fit?

(3) Use the data from Angrist and Lavy (2009). The original analysis is quite complicated. We focus only on Table A1 viewing the schools as experimental units. Then we have a matched-pair design on the schools. For simplicity, we drop pair 6 and all the pairs with noncompliance. This results in 14 complete pairs. The outcome is the Bagrut passing rates in 2001 and 2002, with the Bagrut passing rates in 1999 and 2000 as pretreatment covariates.

Re-analyze the data using the FRT with and without covariate adjustment.

Re-analyze the data based on the Neymanian inference with and without covariates.

3. Covariance estimator in matched-pair designs

In a matched-pair design, we define the within-pair differences of outcome and covariate as

$$\hat{\tau}_i = (2Z_i - 1)(Y_{i1} - Y_{i2}), \quad \hat{\tau}_{xi} = (2Z_i - 1)(x_{i1} - x_{i2}),$$

and the averages of them as

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i, \quad \hat{\tau}_x = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_{xi}.$$

Show that an unbiased estimator of $\mathrm{cov}(\hat{\tau}, \hat{\tau}_x)$ is

$$\hat{\theta} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\tau}_{xi} - \hat{\tau}_x)(\hat{\tau}_i - \hat{\tau}).$$

4. Data analysis: stratification and regression

Use the dataset `homocyst` in the R package `senstrat`. The outcome is `homocysteine`, the homocysteine level, and the treatment is `z`, where $z = 1$ for a daily smoker and $z = 0$ for a never smoker. Covariates are `female, age3, ed3, bmi3, pov2` with detailed explanations in the R package. `st` is a stratum indicator, defined by all the combinations of the discrete covariates.

(1) How many strata have only treated or control units? What is the proportion of the units in these strata? Drop these strata and perform a stratified analysis of the observational study. Report the point estimator, variance estimator and 95% confidence interval for the average treatment effect.

(2) Run OLS of the outcome on the treatment indicator and covariates without interactions. Report the result.

(3) Apply Lin's estimator of the average treatment effect. Report the result.

(4) Compare the results in the above three analyses. Which one is more credible?

5. More results on observational studies

The Hajek estimator differs from the Horvitz–Thompson estimator in the numerators. Show

$$E\left\{\sum_{i=1}^{n} \frac{Z_i}{e(X_i)}\right\} = n, \quad E\left\{\sum_{i=1}^{n} \frac{1 - Z_i}{1 - e(X_i)}\right\} = n.$$

6. Re-analysis of Rosenbaum and Rubin (1983)

Use Table 1 of this paper. If you are interested, you can read the whole paper. It is a canonical paper. But for this problem, you only need Table 1.

Rosenbaum and Rubin (1983) fitted a logistic regression model for the propensity score and stratified the data into 5 subclasses. Because the treatment (Surgical versus Medical) is binary and the outcome is also binary (improved or not), they represented the data by a table.

Based on this table, estimate the average treatment effect, and report the 95% confidence interval.

REFERENCES

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *The American Economic Review*, 99:1384–1414.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B (Methodological)*, 45:212–218.