Midterm and Problem Set 1 Solution

Juan C. Yamin
Brown University

October 15, 2024

Roadmap of Talk

Midterm Review

Problem Set 1

Regression *vs.* Matching

10.* Consider estimation of ATE under unconfoundedness. Consider the linear projection of observed outcome Y_i onto (D_i, X_i) in an additive way, i.e., run OLS of

$$Y_i = \delta D_i + X_i'\beta + \epsilon_i,$$

where I included the intercept in X_i. We denote the OLS estimator for δ by $\hat{\delta}$.

(a) Express δ, the estimand that $\hat{\delta}$ converges to, in terms of the distribution of observables (Y, D, X). [Hint: apply the partialling out argument of Frisch-Waugh-Lovelle theorem]

(b) Assume that the propensity score is linear in X, $p(X) = E[D|X] = X'\gamma$. Show

$$\delta = \frac{E[Var(D|X)\tau(X)]}{E[Var(D|X)]},$$

where $\tau(X) = E[Y(1) - Y(0)|X]$ is the conditional average treatment effect.

(c) Based on the expression of δ obtained in (b), discuss under what scenarios, δ can be interpreted as the average treatment effect.

(d) Viewing the expression of δ as a weighted average of $\tau(X)$ with weights proportional to $Var(D|X)$, discuss which subpopulation defined by the value of X gets a larger weight.

Regression *vs.* Matching

Before we start, crash course on regression mechanics:

- **Saturated regression**: Regression coincides with the CEF when all values of X_i are "dummied out"

- **Frisch-Waugh-Lovell Theorem**: The k-th non-constant slope coefficient is

$$\beta_k = \frac{Cov(\tilde{X}_{ki}, Y_i)}{Var(\tilde{X}_{ki})},$$

  where $\tilde{X}_{ki}$ is the residual from regression X_{ki} on all other elements of X_i.

- **A regression of Y_i on D_i and X_i is the same as a regression of $E[Y_i|D_i, X_i]$ on D_i, X_i**

  An application of the Regression CEF Theorem/ LIE :

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i] = E[X_i X_i']^{-1} E[X_i E[Y_i \mid X_i]]$$

Stata code

Regression *vs.* Matching

10.* Consider estimation of ATE under unconfoundedness. Consider the linear projection of observed outcome Y_i onto (D_i, X_i) in an additive way, i.e., run OLS of

$$Y_i = \delta D_i + X_i'\beta + \epsilon_i,$$

where I included the intercept in X_i. We denote the OLS estimator for δ by $\hat{\delta}$.

(a) Express δ, the estimand that $\hat{\delta}$ converges to, in terms of the distribution of observables (Y, D, X). [Hint: apply the partialling out argument of Frisch-Waugh-Lovelle theorem]

(b) Assume that the propensity score is linear in X, $p(X) = E[D|X] = X'\gamma$. Show

$$\delta = \frac{E[Var(D|X)\tau(X)]}{E[Var(D|X)]},$$

where $\tau(X) = E[Y(1) - Y(0)|X]$ is the conditional average treatment effect.

(c) Based on the expression of δ obtained in (b), discuss under what scenarios, δ can be interpreted as the average treatment effect.

(d) Viewing the expression of δ as a weighted average of $\tau(X)$ with weights proportional to $Var(D|X)$, discuss which subpopulation defined by the value of X gets a larger weight.

Regression *vs.* Matching

- The regression identifies a convex weighted average of the conditional average treatment effects! (as long as pscores are linear)

- In general, $\delta \neq ATE$. In this setting, regression only identifies ATE when:
  1. $CATE_i$ are constant
  2. The weights are constant

- Regression avoids the IPW problem of having pscores close to 0 or 1 $\Rightarrow$ regression just puts zero weight on such groups!

- Key of the proof: As always, FWL Theorem and LIE

Bottom line: With Selection on observables, regression is a straightforward approach to estimate treatment effects. But be careful... it does NOT estimate the ATE!

Weak Doubly Robust Property

(a) Show that under unconfoundedness,

$$\theta_{ATE} = E\left[\frac{(Y-\mu_1(X))D}{p(X)} - \frac{(Y-\mu_0(X))(1-D)}{1-p(X)} + (\mu_1(X)-\mu_0(X))\right]$$

holds, where $\mu_1(X) = E(Y|D=1,X)$ and $\mu_0(X) = E(Y|D=0,X)$.

(b) Let $m_1(X)$ and $m_0(X)$ be functions of X that are different from $\mu_1(X)$ and $\mu_0(X)$. Show that the identity shown in part (a) remains to hold even if you replace $\mu_1(X)$ and $\mu_0(X)$ with $m_1(X)$ and $m_0(X)$. This result shows that the moment condition for θ_{ATE} shown above is robust to misspecification in estimating $\mu_1(X)$ and $\mu_0(X)$.

(c) Let $\tilde{p}(X) \in (0,1)$ be a function of X that is different from $p(X)$. Show that the identity shown in part (a) remains to hold even if you replace $p(X)$ with $\tilde{p}(X)$. This result shows that the moment condition for θ_{ATE} shown above is robust to misspecification in estimating $p(X)$.

(d) Based on the identity shown in part (a), propose an estimator for θ_{ATE}. This estimation approach is called *doubly robust estimation for ATE.*

Roadmap of Talk

Midterm Review

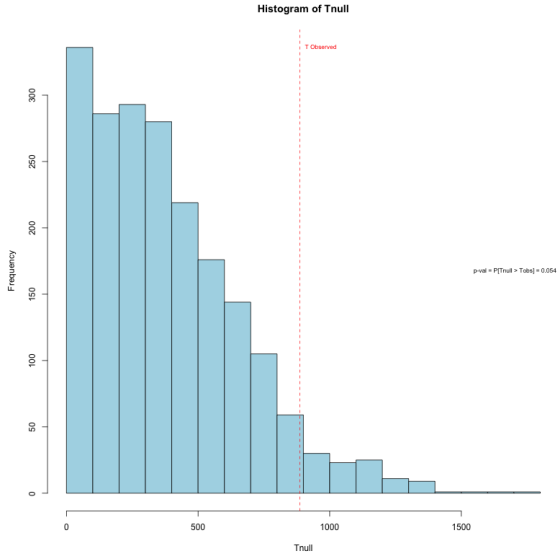
Problem Set 1

Question 2

1. Generate random assignment
2. Compute the statistic under the null
3. Repeat 2,000 times
4. Compute p-value:

$$\text{p-value} = \frac{1}{2000} \sum_{i=1}^{2000} 1\{T_i > T^{obs}\}.$$

Question 2


Histogram of Tnull
T Observed
p-val = P[Tnull > Tobs] = 0.054
Frequency
0
50
100
150
200
250
300
0
500
1000
1500
Tnull

Question 5

- t-test:

$$T^{obs} = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}}$$

  where s_1^2 and s_0^2 are the sample variances.

- p-value of the T-test (two-tailed):

$$2 \times P(T^{obs} > |t|) \approx 2(1 - \Phi(|T^{obs}|))$$

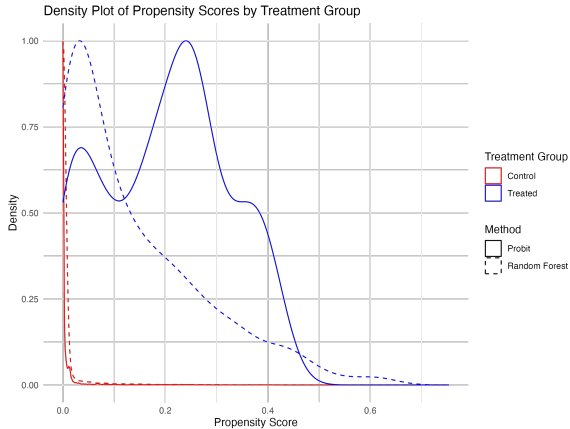
Question 5

Table: Balance of Treated and Control Groups

Variable	Treated	Control	ASMD	T	T p-value	KS	KS p-value
age	24.63	24.45	0.03	0.36	0.72	0.61	0.86
education	10.38	10.19	0.11	1.46	0.14	1.10	0.17
black	0.80	0.80	0.00	0.04	0.96	0.02	1.00
hispanic	0.09	0.11	0.06	-0.81	0.42	0.25	1.00
married	0.17	0.16	0.03	0.38	0.70	0.14	1.00
nodegree	0.73	0.81	0.20	-2.61	0.01	1.10	0.17
re75	3066.10	3026.68	0.01	0.10	0.92	0.60	0.79

Do you think is the randomization assumption reasonable?

Question 8



Are the propensity score distributions similar among methods? Discuss if the overlap is limited or not.

Question 9

Table: Balance of Treated and Control Groups

Variable	Treated	Control	ASMD	T	T p-value	KS	KS p-value
age	24.39	24.32	0.01	0.15	0.88	12.93	0.00
education	10.53	10.24	0.13	2.00	0.05	12.37	0.00
black	0.95	0.96	0.04	-0.53	0.60	12.53	0.00
hispanic	0.05	0.04	0.04	0.53	0.60	0.41	1.00
married	0.15	0.11	0.13	1.60	0.11	1.35	0.05
nodegree	0.72	0.75	0.06	-0.72	0.47	10.06	0.00
re75	2261.47	2556.50	0.08	-1.16	0.24	2.32	0.00

Question 11

Why the Doubly robust estimator might be preferable to IPW?

- The doubly robust estimator is consistent if either the estimates of the conditional response or propensity score are consistent (*i.e.*, *weak* double robustness property).

- Intuition:
  - The Doubly robust estimator combines the two ways to estimate the average treatment effect: (i) estimate the propensity score, and (ii) estimate the conditional response functions.
  - The DR estimator can be seen as first trying to estimate μ_1 and μ_0 and then it deals with any biases of the estimators by applying IPW to the regression residuals
  - Inherits robustness properties from both the regression and IPW estimators—it improves on both!

- Why *weak*? We not only care about consistency... we also care about rates of convergence and confidence intervals!

Question 11

What assumptions about the accuracy of machine learning models are needed to reduce the bias in the Doubly Robust estimator compared to IPW?

- Under the assumption that the machine learning methods have RMSE that decay at $n^{-1/4}$ (*i.e.*, the predictions are pretty accurate but not as accurate as the parametric models in well-specified parametric designs), the doubly robust construction succeeds in making bias equal or smaller than what either the regression or IPW estimators could achieve on their own.

- Formally,

$$\sqrt{n}(\hat{\theta}^{DR} - \theta) \rightarrow N(0, V)$$

- The estimator is so carefully designed that even if the ML methods do not have parametric convergence rates, we get parametric rates!

- We can build CI in the standard way

Question 12

Explain why cross-fitting is necessary when estimating $p(x)$, $\mu_1(x)$, and $\mu_0(x)$ nonparametrically by machine learning methods.

- Cross-fitting avoids bias due to overfitting (same intuition as cross-validation in predictive machine learning) $\Rightarrow$ Cross-fitting prevents the artificial shrinking of residuals if the same data is used for both estimation and evaluation

- In theory, cross-fitting is crucial for obtaining the asymptotic normality