ECMA 31380 - Causal Machine Learning - Homework 2

Fernando Rocha Urbano

Autumn 2024

---

1 Multiple Testing and Heterogeneous Treatment Effects

We have i.i.d. data from a randomized experiment with a binary treatment $T \in \{0, 1\}$, a continuous outcome of interest Y, a set of binary covariates $X = (X_1, X_2, \ldots, X_d)' \in \{0, 1\}^d$ and a set of continuous covariates $W \in \mathbb{R}^l$. Both X and W are pre-treatment and all our usual assumptions are met.

1.a Conditional Average Treatment Effects

(a) Define the univariate conditional average treatment effects with respect to each X_j as $\tau_j(x) = \mathbb{E}[Y(1) - Y(0) \mid X_j = x]$ for $j \in \{1, \ldots, d\}$ and $x \in \{0, 1\}$. Use a linear regression to propose an estimator for $\tau_j(x)$ and establish its asymptotic distribution. Provide all necessary regularity conditions. Construct your estimation so that the estimators for $\tau_j(x)$ and $\tau_k(x')$ are based on independent data if $j \neq k$ or $x \neq x'$.

Definition of Univariate Conditional Average Treatment Effects

For each $j \in \{1, \ldots, d\}$ and $x \in \{0, 1\}$, define the univariate conditional average treatment effect (CATE) as:
$$\tau_j(x) = \mathbb{E}[Y(1) - Y(0) \mid X_j = x].$$

This represents the expected difference in potential outcomes between treated and untreated individuals, conditional on $X_j = x$.

Estimator Using Linear Regression

To estimate $\tau_j(x)$, a linear regression model is applied to observations where $X_{ij} = x$:

$$Y_i = \alpha_{jx} + \tau_j(x)T_i + \beta'_{jx}W_i + \varepsilon_i, \quad \text{for all } i \text{ such that } X_{ij} = x.$$

In this model:

- α_{jx} is the intercept term specific to $X_j = x$.

1

- $\tau_j(x)$ is the coefficient on the treatment indicator T_i, representing the CATE.

- W_i is the vector of continuous covariates.

- β_{jx} is the vector of coefficients associated with W_i.

- ε_i is the error term with zero conditional mean: $\mathbb{E}[\varepsilon_i \mid T_i, W_i] = 0$.

The Ordinary Least Squares (OLS) estimator $\hat{\tau}_j(x)$ is obtained by fitting this regression model to the data subset where $X_{ij} = x$.

Asymptotic Distribution of the Estimator

Under suitable regularity conditions, the estimator $\hat{\tau}_j(x)$ is consistent and asymptotically normal:

$$\sqrt{n_{jx}}\left(\hat{\tau}_j(x) - \tau_j(x)\right) \xrightarrow{d} N\left(0, \sigma_{jx}^2\right),$$

where:

- n_{jx} is the number of observations with $X_{ij} = x$.

- σ_{jx}^2 is the asymptotic variance of $\hat{\tau}_j(x)$, which can be consistently estimated from the regression.

Regularity Conditions

The asymptotic result relies on the following regularity conditions:

1. Independent and Identically Distributed Sampling: The data $\{(Y_i, T_i, X_i, W_i)\}_{i=1}^n$ are i.i.d. draws from the population.

2. Random Assignment of Treatment: The treatment T_i is randomly assigned, independent of potential outcomes $Y(1), Y(0)$ and covariates X_i, W_i.

3. Finite Fourth Moments: All variables Y_i, T_i, and components of W_i have finite fourth moments.

4. Full Rank Condition: The matrix

$$\mathbb{E}\left[\begin{pmatrix} 1 \\ T_i \\ W_i \end{pmatrix} \begin{pmatrix} 1 & T_i & W_i' \end{pmatrix} \middle| X_j = x \right]$$

   is positive definite.

5. Zero Conditional Mean of Errors: The error term satisfies $\mathbb{E}[\varepsilon_i \mid T_i, W_i] = 0$.

Ensuring Independence Across Estimators

To ensure that estimators $\hat{\tau}_j(x)$ and $\hat{\tau}_k(x')$ are based on independent data when $j \neq k$ or $x \neq x'$, we partition the dataset into mutually exclusive subsets $\{D_{jx}\}$. Specifically, for each combination of j and x, we define:

$$D_{jx} = \{i : X_{ij} = x, \text{ and } i \in S_{jx}\},$$

where S_{jx} is a randomly assigned subset of indices such that:

$$D_{jx} \cap D_{kx'} = \emptyset \quad \text{whenever} \quad (j, x) \neq (k, x').$$

By estimating $\tau_j(x)$ using only data from D_{jx}, different estimators are based on disjoint samples, ensuring their independence.

1.b 5% Level Test for $H_0 : \tau_1(1) = 0$

(b) Construct a 5% level test of $H_0 : \tau_1(1) = 0$ versus $H_1 : \tau_1(1) \neq 0$ based on the t-statistic from the asymptotic distribution above. Call the test statistic t_{11}. Give the test statistic, its distribution, the critical region, and describe when you reject the null and when you fail to reject. Prove that your test is consistent.

Consider testing the null hypothesis $H_0 : \tau_1(1) = 0$ against the alternative hypothesis $H_1 : \tau_1(1) \neq 0$.

Test Statistic

The test statistic is constructed using the estimator $\hat{\tau}_1(1)$ and its estimated standard deviation $\hat{\sigma}_{11}$:

$$t_{11} = \frac{\sqrt{n_{11}}\, \hat{\tau}_1(1)}{\hat{\sigma}_{11}},$$

where:

- n_{11} is the number of observations with $X_{i1} = 1$.

- $\hat{\sigma}_{11}$ is a consistent estimator of the asymptotic standard deviation of $\sqrt{n_{11}}\, \hat{\tau}_1(1)$.

Distribution Under H_0

Under the null hypothesis H_0 and the regularity conditions, the test statistic t_{11} has an asymptotic standard normal distribution:

$$t_{11} \xrightarrow{d} N(0,1) \quad \text{under } H_0.$$

Critical Region

For a two-sided test at the 5% significance level, the critical values are the 2.5th and 97.5th percentiles of the standard normal distribution:

$$\text{Reject } H_0 \quad \text{if} \quad |t_{11}| > z_{0.975},$$

where $z_{0.975} \approx 1.96$.

Decision Rule

- Reject H_0 if $|t_{11}| > 1.96$.
- Fail to reject H_0 if $|t_{11}| \leq 1.96$.

Consistency of the Test

To prove that the test is consistent, we need to show that under the alternative hypothesis $H_1 : \tau_1(1) \neq 0$, the probability of rejecting H_0 approaches 1 as $n_{11} \to \infty$.

Under H_1, the test statistic can be expressed as:

$$t_{11} = \frac{\sqrt{n_{11}}\left(\hat{\tau}_1(1) - \tau_1(1) + \tau_1(1)\right)}{\hat{\sigma}_{11}} = \frac{\sqrt{n_{11}}\left(\hat{\tau}_1(1) - \tau_1(1)\right)}{\hat{\sigma}_{11}} + \frac{\sqrt{n_{11}}\,\tau_1(1)}{\hat{\sigma}_{11}}.$$

The first term converges in distribution to a standard normal random variable:

$$\frac{\sqrt{n_{11}}\left(\hat{\tau}_1(1) - \tau_1(1)\right)}{\hat{\sigma}_{11}} \xrightarrow{d} N(0, 1).$$

The second term diverges to infinity because $\tau_1(1) \neq 0$ and $\hat{\sigma}_{11}$ is bounded away from zero:

$$\frac{\sqrt{n_{11}}\,\tau_1(1)}{\hat{\sigma}_{11}} \rightarrow \begin{cases} +\infty, & \text{if } \tau_1(1) > 0, \\ -\infty, & \text{if } \tau_1(1) < 0. \end{cases}$$

Therefore, under H_1, the test statistic t_{11} diverges in the same direction as $\tau_1(1)$:

$$t_{11} \xrightarrow{p} \begin{cases} +\infty, & \text{if } \tau_1(1) > 0, \\ -\infty, & \text{if } \tau_1(1) < 0. \end{cases}$$

As a result, the probability of $|t_{11}| > 1.96$ approaches 1 under H_1:

$$\lim_{n_{11} \to \infty} P\left(|t_{11}| > 1.96 \mid H_1\right) = 1.$$

This demonstrates that the test is consistent, as it correctly rejects the null hypothesis with probability approaching 1 when $\tau_1(1) \neq 0$.

1.c   5% Level Test for $H_0 : \tau_1(0) = 0$

(c) Construct a 5% level test of $H_0 : \tau_1(0) = 0$ versus $H_1 : \tau_1(0) \neq 0$ based on the t-statistic from the asymptotic distribution above. Call the test statistic t_{10}. Give the test statistic, its distribution, the critical region, and describe when you reject the null and when you fail to reject. Prove that your test is consistent.

Constructing a 5% level test of $H_0 : \tau_1(0) = 0$ versus $H_1 : \tau_1(0) \neq 0$, we proceed as follows.

Test Statistic

The test statistic t_{10} is defined using the estimator $\hat{\tau}_1(0)$ and its estimated standard error $\hat{\sigma}_{10}$:

$$t_{10} = \frac{\sqrt{n_{10}}\,\hat{\tau}_1(0)}{\hat{\sigma}_{10}},$$

where:

- n_{10} is the number of observations with $X_{i1} = 0$. - $\hat{\sigma}_{10}$ is a consistent estimator of the asymptotic standard deviation of $\sqrt{n_{10}}\,\hat{\tau}_1(0)$.

Distribution Under H_0

Under the null hypothesis H_0 and given the regularity conditions, the test statistic t_{10} has an asymptotic standard normal distribution:

$$t_{10} \xrightarrow{d} N(0,1) \quad \text{under } H_0.$$

Critical Region

For a two-sided test at the 5% significance level, the critical values correspond to the 2.5th and 97.5th percentiles of the standard normal distribution:

$$\text{Reject } H_0 \quad \text{if} \quad |t_{10}| > z_{0.975},$$

where $z_{0.975} \approx 1.96$.

Decision Rule

- Reject H_0 if $|t_{10}| > 1.96$. - Fail to reject H_0 if $|t_{10}| \leq 1.96$.

Consistency of the Test

To prove that the test is consistent, we need to show that under the alternative hypothesis $H_1 : \tau_1(0) \neq 0$, the probability of rejecting H_0 approaches 1 as $n_{10} \to \infty$.

Under H_1, the test statistic can be expressed as:

$$t_{10} = \frac{\sqrt{n_{10}}\left(\hat{\tau}_1(0) - \tau_1(0) + \tau_1(0)\right)}{\hat{\sigma}_{10}} = \frac{\sqrt{n_{10}}\left(\hat{\tau}_1(0) - \tau_1(0)\right)}{\hat{\sigma}_{10}} + \frac{\sqrt{n_{10}}\,\tau_1(0)}{\hat{\sigma}_{10}}.$$

The first term converges in distribution to a standard normal random variable:

$$\frac{\sqrt{n_{10}}\left(\hat{\tau}_1(0) - \tau_1(0)\right)}{\hat{\sigma}_{10}} \xrightarrow{d} N(0,1).$$

The second term diverges to infinity because $\tau_1(0) \neq 0$ and $\hat{\sigma}_{10}$ is bounded away from zero:

$$\frac{\sqrt{n_{10}}\,\tau_1(0)}{\hat{\sigma}_{10}} \to \begin{cases} +\infty, & \text{if } \tau_1(0) > 0, \\ -\infty, & \text{if } \tau_1(0) < 0. \end{cases}$$

Therefore, under H_1, the test statistic t_{10} diverges to infinity in the direction of the sign of $\tau_1(0)$:

$$t_{10} \xrightarrow{p} \begin{cases} +\infty, & \text{if } \tau_1(0) > 0, \\ -\infty, & \text{if } \tau_1(0) < 0. \end{cases}$$

As a result, the probability of $|t_{10}| > 1.96$ approaches 1 under H_1:

$$\lim_{n_{10} \to \infty} P\left(|t_{10}| > 1.96 \mid H_1\right) = 1.$$

Thus, the test is consistent, as it correctly rejects the null hypothesis with probability approaching 1 when $\tau_1(0) \neq 0$.

1.d   Probability of False Positives in 5% Level Tests

(d) By your own argument from part (a), the tests (b) and (c) are independent, and you just established that they are 5% level tests. Find the probability that at least one of the tests gives a false positive. What does this tell you about how often you will make mistakes?

Given that tests (b) and (c) are independent 5% level tests, we can compute the probability that at least one test yields a false positive under the global null hypothesis (i.e., both $H_0 : \tau_1(1) = 0$ and $H_0 : \tau_1(0) = 0$ are true).

Probability of No False Positives

For each test, the probability of not committing a Type I error (i.e., correctly failing to reject H_0 when it is true) is 95%:

$$P(\text{No False Positive in a Single Test}) = 1 - \alpha = 0.95,$$

where $\alpha = 0.05$ is the significance level of the test.

Since the two tests are independent, the joint probability that neither test yields a false positive is:

$$P(\text{No False Positives in Both Tests}) = P(\text{No False Positive in Test (b)}) \times P(\text{No False Positive in Test (c)}) = 0.95 \times 0.95$$

Probability of At Least One False Positive

Therefore, the probability that at least one test yields a false positive is:

$$P(\text{At Least One False Positive}) = 1 - P(\text{No False Positives in Both Tests}) = 1 - 0.9025 = 0.0975.$$

Interpretation

This calculation shows that when conducting two independent tests at the 5% significance level, the probability of committing at least one Type I error increases to 9.75%. In other words, even though each test individually has a 5% chance of yielding a false positive, the combined chance of making a mistake in at least one test is higher due to the multiplicity of tests.

Implications

This result highlights the multiple testing problem: as the number of independent tests increases, the overall probability of making at least one Type I error also increases. Specifically, for m independent tests at the same significance level α, the probability of making at least one Type I error is:

$$P(\text{At Least One False Positive in } m \text{ Tests}) = 1 - (1 - \alpha)^m.$$

Applying this to our case with $m = 2$ and $\alpha = 0.05$:

$$P(\text{At Least One False Positive}) = 1 - (0.95)^2 = 1 - 0.9025 = 0.0975.$$

Conclusion

The probability of 9.75% indicates that when performing these two independent 5% level tests, there is nearly a 10% chance of incorrectly rejecting at least one true null hypothesis purely by chance. This increased error rate suggests that, in practice, one should be cautious about the potential for false positives when conducting multiple tests and consider methods to adjust for multiple comparisons, such as the Bonferroni correction or other techniques, to control the overall Type I error rate.

1.e   Testing Null Hypothesis for Both Groups

(e) I would like to test the null hypothesis that the treatment is not effective in both groups,

$$H_0 : \tau_1(1) = 0 \text{ and } \tau_1(0) = 0,$$

against the alternative that the treatment is effective for at least one group,

$$H_0 : \tau_1(1) \neq 0 \text{ or } \tau_1(0) \neq 0.$$

I will reject the null if $|t_{11}| > c$ or $|t_{10}| > c$. For what c is a 5% level test?

1.f   Testing Two Restrictions with Fixed Level Test

(f) The previous part refers to testing two restrictions. What happens to the value of c as the number of restrictions grows, but the level stays fixed at 5%? Give an expression for c (as a function of d) such that we can test the null that $\tau_j(x) = 0$ for all $j \in \{1, \ldots, d\}$ and $x \in \{0, 1\}$.

Notice that we have not even begun to explore subgroup effects in earnest, because the above does not consider any interactions.

To determine how the critical value c changes as the number of restrictions grows while maintaining the overall test level at 5%, we consider testing the null hypothesis:

$$H_0 : \tau_j(x) = 0 \quad \text{for all} \quad j \in \{1, \ldots, d\}, \quad x \in \{0, 1\},$$

against the alternative:

$$H_1 : \text{There exists } (j, x) \text{ such that } \tau_j(x) \neq 0.$$

We have a total of $m = 2d$ independent test statistics t_{jx}, each corresponding to one of the $\tau_j(x)$. Each t_{jx} follows an asymptotic standard normal distribution under H_0:

$$t_{jx} \xrightarrow{d} N(0, 1) \quad \text{under } H_0.$$

Test Procedure

We reject H_0 if any of the test statistics exceed the critical value c in absolute value:

$$\text{Reject } H_0 \quad \text{if} \quad \max_{j,x} |t_{jx}| > c.$$

Determining the Critical Value c

Under H_0, the test statistics t_{jx} are independent standard normal variables. We need to find c such that the overall Type I error rate is 5%:

$$P\left(\max_{j,x} |t_{jx}| > c \mid H_0\right) = 0.05.$$

The probability that a single test statistic does not exceed c in absolute value is:

$$P\left(|t_{jx}| \leq c\right) = 2\Phi(c) - 1,$$

where $\Phi(c)$ is the cumulative distribution function (CDF) of the standard normal distribution.

Since the test statistics are independent, the probability that none of them exceeds c is:

$$P\left(\max_{j,x} |t_{jx}| \leq c \mid H_0\right) = [P\left(|t_{jx}| \leq c\right)]^m = (2\Phi(c) - 1)^m.$$

Therefore, the probability of rejecting H_0 is:

$$P\left(\text{Reject } H_0 \mid H_0\right) = 1 - (2\Phi(c) - 1)^m = 0.05.$$

Solving for c

We solve for c using the equation:

$$(2\Phi(c) - 1)^m = 1 - 0.05 = 0.95.$$

Taking the m-th root:

$$2\Phi(c) - 1 = (0.95)^{1/m}.$$

Solving for $\Phi(c)$:

$$\Phi(c) = \frac{1 + (0.95)^{1/m}}{2}.$$

Therefore, the critical value c as a function of d (since $m = 2d$) is:

$$c = \Phi^{-1}\left(\frac{1 + (0.95)^{1/(2d)}}{2}\right).$$

Behavior of c as d Increases

As the number of restrictions d increases, the exponent $1/(2d)$ decreases, and $(0.95)^{1/(2d)}$ approaches 1 from below. Consequently, $\Phi(c)$ approaches 1, and c increases.

To illustrate, consider the limit as $d \to \infty$:

$$\lim_{d\to\infty} (0.95)^{1/(2d)} = (0.95)^0 = 1,$$

$$\lim_{d\to\infty} \Phi(c) = \frac{1+1}{2} = 1,$$

$$\lim_{d\to\infty} c = \Phi^{-1}(1) = \infty.$$

This indicates that as the number of restrictions grows, the critical value c increases without bound, making it increasingly difficult to reject H_0.

Example Calculation

For a specific value of d, we can compute c. Suppose $d = 5$ (so $m = 10$):

$$(0.95)^{1/10} \approx 0.995,$$

$$\Phi(c) = \frac{1 + 0.995}{2} = 0.9975$$

$$c = \Phi^{-1}(0.9975) \approx 2.81.$$

Therefore, with $d = 5$, the critical value c is approximately 2.81.

Conclusion

As the number of restrictions d increases, the critical value c required to maintain the overall test level at 5% increases according to:

$$c = \Phi^{-1}\left(\frac{1 + (0.95)^{1/(2d)}}{2}\right).$$

This relationship shows that controlling the family-wise Type I error rate in multiple testing leads to more stringent criteria for rejection, reflecting the need to account for the increased chance of false positives when conducting many tests simultaneously.

2 Propensity Score Weighting and ATT Estimation

Assume that the random variables $(Y_1, Y_0, T, X')' \in \mathbb{R} \times \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ obey $\{Y_1, Y_0\} \perp T \mid X$. The researcher observes $(Y, T, X')'$, where $Y = Y_1 T + Y_0(1 - T)$. Define the propensity score

$p(x) = \mathbb{P}[T = 1 \mid X = x]$ and assume it is bounded inside $(0, 1)$. Define $\mu_t = \mathbb{E}[Y(t) \mid T = 1]$ and $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$. The average treatment effect on the treated (ATT) is $\tau = \mu_1 - \mu_0$.

2.a   Balancing Score and Propensity Score

(a) A function $f(X)$ is called a balancing score if $X \perp T \mid f(X)$. Prove that the propensity score is a balancing score.

To prove that the propensity score $p(X) = \mathbb{P}[T = 1 \mid X]$ is a balancing score, we need to show that conditioning on $p(X)$ renders the covariates X independent of the treatment assignment T:

$$X \perp T \mid p(X).$$

Proof

We aim to show that the conditional distribution of X given T and $p(X)$ is the same as the conditional distribution of X given $p(X)$ alone. That is, for all measurable subsets A of the support of X:

$$\mathbb{P}[X \in A \mid T = t, p(X) = e] = \mathbb{P}[X \in A \mid p(X) = e], \quad \text{for } t \in \{0, 1\}, \ e \in (0, 1).$$

Step 1: Evaluate $\mathbb{P}[X \in A \mid T = t, p(X) = e]$

Using Bayes' theorem, we have:

$$\mathbb{P}[X \in A \mid T = t, p(X) = e] = \frac{\mathbb{P}[T = t \mid X \in A, p(X) = e]\mathbb{P}[X \in A \mid p(X) = e]}{\mathbb{P}[T = t \mid p(X) = e]}.$$

Step 2: Simplify $\mathbb{P}[T = t \mid X \in A, p(X) = e]$

Since $p(X)$ is a function of X, conditioning on X implies that $p(X)$ is known. Therefore:

$$\mathbb{P}[T = t \mid X, p(X) = e] = \mathbb{P}[T = t \mid X] = \begin{cases} p(X) = e, & \text{if } t = 1, \\ 1 - p(X) = 1 - e, & \text{if } t = 0. \end{cases}$$

Thus, for all x such that $p(x) = e$:

$$\mathbb{P}[T = t \mid X = x, p(X) = e] = \mathbb{P}[T = t \mid X = x] = \begin{cases} e, & \text{if } t = 1, \\ 1 - e, & \text{if } t = 0. \end{cases}$$

Therefore, for $X \in A$:

$$\mathbb{P}[T = t \mid X \in A, p(X) = e] = \begin{cases} e, & \text{if } t = 1, \\ 1 - e, & \text{if } t = 0. \end{cases}$$

Step 3: Compute $\mathbb{P}[T = t \mid p(X) = e]$

Again, since $p(X) = \mathbb{P}[T = 1 \mid X]$, we have:

$$\mathbb{P}[T = 1 \mid p(X) = e] = e, \quad \mathbb{P}[T = 0 \mid p(X) = e] = 1 - e.$$

Step 4: Substitute back into the original expression

Now, the numerator and denominator of the expression in Step 1 are equal:

$$\frac{\mathbb{P}[T = t \mid X \in A, p(X) = e]}{\mathbb{P}[T = t \mid p(X) = e]} = \frac{e}{e} = 1 \quad \text{or} \quad \frac{1 - e}{1 - e} = 1.$$

Thus:

$$\mathbb{P}[X \in A \mid T = t, p(X) = e] = \mathbb{P}[X \in A \mid p(X) = e].$$

Conclusion

Since the conditional distribution of X given T and $p(X)$ equals the conditional distribution of X given $p(X)$ alone, we have established that:

$$X \perp T \mid p(X).$$

Therefore, the propensity score $p(X)$ is indeed a balancing score.

2.b   Conditional Independence with Propensity Score

(b) Prove that $\{Y_1, Y_0\} \perp T \mid X$ implies $\{Y(1), Y(0)\} \perp T \mid p(X)$.

To prove that $\{Y_1, Y_0\} \perp T \mid p(X)$ given $\{Y_1, Y_0\} \perp T \mid X$, we proceed as follows.

Proof

We are given:

1. **Conditional Independence:** $\{Y_1, Y_0\} \perp T \mid X$. This means that, conditional on X, the potential outcomes Y_1 and Y_0 are independent of the treatment assignment T.

2. **Balancing Score Property:** From part (a), the propensity score $p(X)$ is a balancing score, so $X \perp T \mid p(X)$.

Our goal is to show that $\{Y_1, Y_0\} \perp T \mid p(X)$.

Step 1: Express the Joint Distribution

Consider the joint conditional probability:

$$P(Y_1 = y_1, Y_0 = y_0, T = t \mid p(X) = e).$$

Step 2: Use the Law of Total Probability

We can write this probability as an integral over X:

$$P(Y_1 = y_1, Y_0 = y_0, T = t \mid p(X) = e) = \int P(Y_1 = y_1, Y_0 = y_0, T = t \mid X = x) \, P(X = x \mid p(X) = e) \, dx.$$

Step 3: Apply Conditional Independence

Given $\{Y_1, Y_0\} \perp T \mid X$, we have:

$$P(Y_1 = y_1, Y_0 = y_0, T = t \mid X = x) = P(Y_1 = y_1, Y_0 = y_0 \mid X = x) \, P(T = t \mid X = x).$$

Step 4: Substitute Back into the Integral

Substituting into the integral, we get:

$$P(Y_1 = y_1, Y_0 = y_0, T = t \mid p(X) = e) = \int P(Y_1 = y_1, Y_0 = y_0 \mid X = x) \, P(T = t \mid X = x) \, P(X = x \mid p(X) = e) \, dx.$$

Step 5: Utilize the Balancing Score Property

From $X \perp T \mid p(X)$, we have:

$$P(T = t \mid X = x) = P(T = t \mid p(X) = e).$$

Since $p(X) = e$ for all X in the integration, $P(T = t \mid X = x)$ is constant with respect to x:

$$P(T = t \mid X = x) = \begin{cases} e, & \text{if } t = 1, \\ 1 - e, & \text{if } t = 0. \end{cases}$$

Step 6: Simplify the Integral

Pulling $P(T = t \mid p(X) = e)$ out of the integral:

$$P(Y_1 = y_1, Y_0 = y_0, T = t \mid p(X) = e) = P(T = t \mid p(X) = e) \int P(Y_1 = y_1, Y_0 = y_0 \mid X = x) \, P(X = x \mid p(X) = e) \, dx.$$

Step 7: Recognize the Marginal Distribution of Y_1, Y_0

The integral represents $P(Y_1 = y_1, Y_0 = y_0 \mid p(X) = e)$:

$$P(Y_1 = y_1, Y_0 = y_0 \mid p(X) = e) = \int P(Y_1 = y_1, Y_0 = y_0 \mid X = x) \, P(X = x \mid p(X) = e) \, dx.$$

Step 8: Conclude Independence

Therefore, the joint distribution simplifies to:

$$P(Y_1 = y_1, Y_0 = y_0, T = t \mid p(X) = e) = P(Y_1 = y_1, Y_0 = y_0 \mid p(X) = e) \, P(T = t \mid p(X) = e).$$

This equality holds for all y_1, y_0, t, e.

Conclusion

The factorization of the joint probability into the product of $P(Y_1, Y_0 \mid p(X))$ and $P(T \mid p(X))$ implies that:

$$\{Y_1, Y_0\} \perp T \mid p(X).$$

Thus, given $\{Y_1, Y_0\} \perp T \mid X$ and that the propensity score $p(X)$ is a balancing score, we have proven that $\{Y_1, Y_0\} \perp T \mid p(X)$.

2.c   Consistent Estimator for μ_1

(c) Prove that

$$\mu_1 = \mathbb{E}[Y(1) \mid T = 1] = \mathbb{E}\left[\frac{TY}{\mathbb{E}[T]}\right]$$

and use this to propose a consistent estimator of μ_1. Notice that estimation of the μ_1 half of the ATT does not require estimation of $p(X)$ or $\mu_0(X)$. Explain why.

Proof of the Equality

We aim to prove that:

$$\mu_1 = \mathbb{E}[Y(1) \mid T = 1] = \mathbb{E}\left[\frac{TY}{\mathbb{E}[T]}\right].$$

Step 1: Express μ_1 Using the Law of Iterated Expectations

By definition, μ_1 is the expected potential outcome under treatment for treated individuals:

$$\mu_1 = \mathbb{E}[Y(1) \mid T = 1].$$

Step 2: Relate $\mathbb{E}[T]$ and $\mathbb{E}[TY]$

Consider the expectation $\mathbb{E}[TY]$:

$$\mathbb{E}[TY] = \mathbb{E}[TY(1)] = \mathbb{E}[Y(1)T].$$

Since $T \in \{0, 1\}$, T acts as an indicator function selecting treated units.

Step 3: Use the Law of Total Probability

We can write $\mathbb{E}[Y(1)T]$ as:

$$\mathbb{E}[Y(1)T] = \mathbb{E}\left[\mathbb{E}[Y(1)T \mid T]\right] = \mathbb{E}\left[Y(1)\mathbb{E}[T \mid T]\right] = \mathbb{E}\left[Y(1)T\right].$$

Step 4: Express $\mathbb{E}[Y(1)T]$ in Terms of μ_1

Using the definition of conditional expectation:

$$\mathbb{E}[Y(1)T] = \mathbb{E}\left[\mathbb{E}[Y(1)T \mid T]\right] = \mathbb{E}\left[\mathbb{E}[Y(1) \mid T]T\right] = \mathbb{E}\left[\mu_1 T\right] = \mu_1\mathbb{E}[T].$$

Step 5: Solve for μ_1

From the above, we have:

$$\mathbb{E}[TY] = \mu_1 \mathbb{E}[T] \quad \Rightarrow \quad \mu_1 = \frac{\mathbb{E}[TY]}{\mathbb{E}[T]} = \mathbb{E}\left[\frac{TY}{\mathbb{E}[T]}\right].$$

Consistent Estimator of μ_1

We can estimate μ_1 using the sample counterparts of $\mathbb{E}[TY]$ and $\mathbb{E}[T]$. Given a random sample $\{(Y_i, T_i, X_i)\}_{i=1}^n$, the estimator is:

$$\hat{\mu}_1 = \frac{\frac{1}{n}\sum_{i=1}^n T_i Y_i}{\frac{1}{n}\sum_{i=1}^n T_i} = \frac{\sum_{i=1}^n T_i Y_i}{\sum_{i=1}^n T_i}.$$

This estimator simplifies to the sample average of Y_i among treated units:

$$\hat{\mu}_1 = \frac{1}{n_T}\sum_{i:T_i=1} Y_i,$$

where $n_T = \sum_{i=1}^n T_i$ is the number of treated units.

Why Estimation of μ_1 Does Not Require $p(X)$ or $\mu_0(X)$

The estimation of μ_1 relies solely on observed outcomes Y_i for individuals with $T_i = 1$. Under the assumption of unconfoundedness ($\{Y_1, Y_0\} \perp T \mid X$) and the fact that treatment assignment is independent of potential outcomes given X, the observed Y_i for treated units are unbiased estimates of $Y(1)$ for those units.

Since $\mu_1 = \mathbb{E}[Y(1) \mid T = 1]$, and we observe $Y_i = Y(1)$ when $T_i = 1$, the sample average of Y_i among treated units consistently estimates μ_1 without requiring any adjustment for covariates or estimation of the propensity score $p(X)$.

Explanation

- **No Need for $p(X)$:** The propensity score is used to adjust for differences in covariate distributions between treated and control groups. However, when estimating μ_1, we are only averaging over the treated group, so there is no imbalance to adjust for. - **No Need for $\mu_0(X)$:** The function $\mu_0(X) = \mathbb{E}[Y(0) \mid X]$ pertains to the control potential outcomes. Since μ_1 concerns only the treated potential outcomes $Y(1)$, knowledge of $\mu_0(X)$ is unnecessary for estimating μ_1.

Consistency of the Estimator

Under the assumption that the treated units are a random sample from the population of treated individuals, the Law of Large Numbers ensures that:

$$\hat{\mu}_1 = \frac{1}{n_T}\sum_{i:T_i=1} Y_i \xrightarrow{p} \mathbb{E}[Y \mid T = 1] = \mu_1.$$

Therefore, $\hat{\mu}_1$ is a consistent estimator of μ_1.

2.d   Estimator for μ_0

(d) Prove that

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \mathbb{E}\left[\frac{(1-T)p(X)Y}{(1-p(X))}\right].$$

Assume you have access to $\hat{p}(x)$ that is a uniformly consistent estimator of $p(x)$. Propose a consistent estimator of μ_0.

We aim to prove that

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \mathbb{E}\left[\frac{(1-T)\,p(X)\,Y}{1-p(X)}\right].$$

Proof of the Equality

Step 1: Express μ_0 in Terms of $\mathbb{E}[Y(0) \mid X]$

By definition,

$$\mu_0 = \mathbb{E}[Y(0) \mid T = 1].$$

Using the Law of Iterated Expectations,

$$\mu_0 = \mathbb{E}_X\left[\mathbb{E}[Y(0) \mid T = 1, X]\right].$$

Given the unconfoundedness assumption $\{Y_1, Y_0\} \perp T \mid X$, we have

$$\mathbb{E}[Y(0) \mid T = 1, X] = \mathbb{E}[Y(0) \mid X].$$

Therefore,

$$\mu_0 = \mathbb{E}_X\left[\mathbb{E}[Y(0) \mid X]\right] = \mathbb{E}[\mathbb{E}[Y(0) \mid X]].$$

Step 2: Relate $\mathbb{E}[Y(0) \mid X]$ to Observable Quantities

We can express $\mathbb{E}[Y(0) \mid X]$ using the observed data from untreated units:

$$\mathbb{E}[Y(0) \mid X] = \mathbb{E}[Y \mid T = 0, X].$$

However, we need to relate μ_0 to an expectation over the entire sample, involving both treated and untreated units.

Step 3: Construct Weighted Expectation

Consider the weighted expectation:

$$\mathbb{E}\left[\frac{(1-T)\,p(X)}{1-p(X)}\,Y\right].$$

We can compute its conditional expectation given X:

$$\mathbb{E}\left[\frac{(1-T)\,p(X)}{1-p(X)}\,Y\,\Big|\,X\right] = \frac{p(X)}{1-p(X)}\,\mathbb{E}\left[(1-T)Y \mid X\right].$$

Since $(1-T)Y = (1-T)Y(0)$, we have:

$$\mathbb{E}\left[(1-T)Y \mid X\right] = \mathbb{E}[(1-T)Y(0) \mid X] = \mathbb{E}[Y(0) \mid X]\,\mathbb{P}[T = 0 \mid X] = \mathbb{E}[Y(0) \mid X]\,[1 - p(X)].$$

Substituting back,

$$\mathbb{E}\left[\frac{(1-T)\,p(X)}{1-p(X)}\,Y\Big|X\right] = p(X)\,\mathbb{E}[Y(0)\mid X].$$

Step 4: Take Expectation over X

Now, take the expectation over X:

$$\mathbb{E}\left[\frac{(1-T)\,p(X)}{1-p(X)}\,Y\right] = \mathbb{E}\left[p(X)\,\mathbb{E}[Y(0)\mid X]\right].$$

Step 5: Relate to μ_0 and $\mathbb{E}[T]$

Recall that $\mathbb{E}[T] = \mathbb{E}[p(X)]$. Therefore,

$$\mathbb{E}\left[\frac{(1-T)\,p(X)}{1-p(X)}\,Y\right] = \mathbb{E}[p(X)\,\mathbb{E}[Y(0)\mid X]] = \mathbb{E}[T]\,\mu_0.$$

Step 6: Solve for μ_0

Solving for μ_0, we obtain:

$$\mu_0 = \frac{1}{\mathbb{E}[T]}\,\mathbb{E}\left[\frac{(1-T)\,p(X)}{1-p(X)}\,Y\right].$$

Consistent Estimator of μ_0

Given a uniformly consistent estimator $\hat{p}(X)$ of $p(X)$, we can construct an estimator $\hat{\mu}_0$ as follows:

$$\hat{\mu}_0 = \frac{1}{\hat{\pi}_T}\frac{1}{n}\sum_{i=1}^n \frac{(1-T_i)\,\hat{p}(X_i)}{1-\hat{p}(X_i)}\,Y_i,$$

where $\hat{\pi}_T = \frac{1}{n}\sum_{i=1}^n T_i$ is the sample proportion of treated units, which is a consistent estimator of $\mathbb{E}[T]$.

Explanation of Consistency

Uniform Consistency of $\hat{p}(X)$

Since $\hat{p}(X)$ is a uniformly consistent estimator of $p(X)$, we have:

$$\sup_X |\hat{p}(X) - p(X)| \xrightarrow{p} 0 \quad \text{as } n \to \infty.$$

Convergence of $\hat{\mu}_0$

Under regularity conditions (e.g., boundedness of Y and $p(X)$ away from 0 and 1), the Law of Large Numbers implies:

$$\hat{\mu}_0 = \frac{1}{\hat{\pi}_T}\frac{1}{n}\sum_{i=1}^n \frac{(1-T_i)\,\hat{p}(X_i)}{1-\hat{p}(X_i)}\,Y_i \xrightarrow{p} \frac{1}{\mathbb{E}[T]}\,\mathbb{E}\left[\frac{(1-T)\,p(X)\,Y}{1-p(X)}\right] = \mu_0.$$

Justification

- **Consistency of $\hat{\pi}_T$:** The sample proportion $\hat{\pi}_T$ converges in probability to $\mathbb{E}[T]$ by the Law of Large Numbers. - **Consistency of the Weighted Sum:** The weighted sum converges to its expected

value because the weights and Y_i are bounded and the estimator $\hat{p}(X_i)$ converges uniformly to $p(X_i)$.
- **Uniform Convergence Ensures Stability of Weights:** Uniform consistency of $\hat{p}(X)$ ensures that the denominators $1 - \hat{p}(X_i)$ are bounded away from zero, preventing instability in the weights.

Conclusion

Therefore, $\hat{\mu}_0$ is a consistent estimator of μ_0, utilizing the observed data (Y_i, T_i, X_i) and the estimated propensity scores $\hat{p}(X_i)$.

2.e   Doubly Robust Estimator

(e) Prove that

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \mathbb{E}\left[T\mu_0(X) + \frac{(1-T)p(X)(Y - \mu_0(X))}{(1 - p(X))}\right].$$

Further, prove that this moment condition is "doubly robust," meaning that it still holds even if one of $p(X)$ or $\mu_0(X)$ is not correctly specified (or cannot be estimated consistently). Replace $p(X)$ in the above by some other function $\tilde{p}(X)$ and show that the equality still holds. Do the same for $\mu_0(X)$.

Proof of the Equality

We aim to prove that

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \mathbb{E}\left[T\mu_0(X) + \frac{(1-T)\,p(X)\,(Y - \mu_0(X))}{1 - p(X)}\right].$$

Step 1: Start from the Expression of μ_0 Derived Previously

From part (d), we have established that

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \mathbb{E}\left[\frac{(1-T)\,p(X)\,Y}{1 - p(X)}\right].$$

Step 2: Decompose Y into $\mu_0(X)$ and the Residual

Note that

$$Y = \mu_0(X) + [Y - \mu_0(X)].$$

Substituting this into the expression, we get

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \mathbb{E}\left[\frac{(1-T)\,p(X)\,(\mu_0(X) + [Y - \mu_0(X)])}{1 - p(X)}\right].$$

Step 3: Separate the Terms

Split the expectation into two parts:

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \left\{\mathbb{E}\left[\frac{(1-T)\,p(X)\,\mu_0(X)}{1 - p(X)}\right] + \mathbb{E}\left[\frac{(1-T)\,p(X)\,[Y - \mu_0(X)]}{1 - p(X)}\right]\right\}.$$

Step 4: Recognize an Equality for the First Term

We claim that

$$\mathbb{E}\left[\frac{(1-T)\,p(X)\,\mu_0(X)}{1-p(X)}\right] = \mathbb{E}[T\,\mu_0(X)].$$

Proof of the Claim:

Since $T \in \{0,1\}$ and $\mathbb{P}[T = 1 \mid X] = p(X)$, we have

$$\mathbb{E}[T\,\mu_0(X)] = \mathbb{E}\left[\mathbb{E}[T \mid X]\,\mu_0(X)\right] = \mathbb{E}[p(X)\,\mu_0(X)].$$

Similarly,

$$\mathbb{E}\left[\frac{(1-T)\,p(X)\,\mu_0(X)}{1-p(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{(1-T)\,p(X)}{1-p(X)}\Big|X\right]\mu_0(X)\right].$$

But

$$\mathbb{E}\left[\frac{(1-T)\,p(X)}{1-p(X)}\Big|X\right] = \frac{p(X)}{1-p(X)}\,\mathbb{E}[1-T \mid X] = \frac{p(X)}{1-p(X)}\,(1-p(X)) = p(X).$$

Therefore,

$$\mathbb{E}\left[\frac{(1-T)\,p(X)\,\mu_0(X)}{1-p(X)}\right] = \mathbb{E}[p(X)\,\mu_0(X)] = \mathbb{E}[T\,\mu_0(X)].$$

Step 5: Substitute Back into the Expression for μ_0

Using the result from Step 4, we have

$$\mu_0 = \frac{1}{\mathbb{E}[T]}\left\{\mathbb{E}[T\,\mu_0(X)] + \mathbb{E}\left[\frac{(1-T)\,p(X)\,[Y-\mu_0(X)]}{1-p(X)}\right]\right\}.$$

Step 6: Combine the Terms

The expression now becomes

$$\mu_0 = \frac{1}{\mathbb{E}[T]}\,\mathbb{E}\left[T\,\mu_0(X) + \frac{(1-T)\,p(X)\,[Y-\mu_0(X)]}{1-p(X)}\right].$$

This completes the proof of the equality.

Proof of Double Robustness

We will show that the above equality holds even if either $p(X)$ or $\mu_0(X)$ is misspecified.

Case 1: $p(X)$ is Replaced by an Arbitrary Function $\tilde{p}(X)$

Suppose we replace $p(X)$ with any function $\tilde{p}(X)$ (not necessarily equal to the true propensity score). We need to show that

$$\mu_0 = \frac{1}{\mathbb{E}[T]}\,\mathbb{E}\left[T\,\mu_0(X) + \frac{(1-T)\,\tilde{p}(X)\,[Y-\mu_0(X)]}{1-\tilde{p}(X)}\right].$$

holds as long as $\mu_0(X)$ is correctly specified.

Proof:

Consider the expectation

$$E = \mathbb{E}\left[T\,\mu_0(X) + \frac{(1-T)\,\tilde{p}(X)\,[Y-\mu_0(X)]}{1-\tilde{p}(X)}\right].$$

Since $T \in \{0, 1\}$, we can write E as

$$E = \mathbb{E}\left[T \, \mu_0(X)\right] + \mathbb{E}\left[\frac{(1 - T) \, \tilde{p}(X) \, [Y - \mu_0(X)]}{1 - \tilde{p}(X)}\right].$$

First Term:

We have

$$\mathbb{E}\left[T \, \mu_0(X)\right] = \mathbb{E}\left[\mathbb{E}[T \mid X] \, \mu_0(X)\right] = \mathbb{E}\left[p(X) \, \mu_0(X)\right].$$

Second Term:

Since $\tilde{p}(X)$ is arbitrary, we need to analyze

$$\mathbb{E}\left[\frac{(1 - T) \, \tilde{p}(X) \, [Y - \mu_0(X)]}{1 - \tilde{p}(X)}\right].$$

We can write

$$\mathbb{E}\left[\frac{(1 - T) \, \tilde{p}(X) \, [Y - \mu_0(X)]}{1 - \tilde{p}(X)}\right] = \mathbb{E}_X\left[\frac{\tilde{p}(X)}{1 - \tilde{p}(X)} \, \mathbb{E}\left[(1 - T)[Y - \mu_0(X)] \mid X\right]\right].$$

But

$$\mathbb{E}\left[(1 - T)[Y - \mu_0(X)] \mid X\right] = \mathbb{E}\left[(1 - T)[Y(0) - \mu_0(X)] \mid X\right].$$

Given that $\mathbb{E}[Y(0) \mid X] = \mu_0(X)$, and $\mathbb{E}[Y(0) - \mu_0(X) \mid X] = 0$, we have

$$\mathbb{E}\left[(1 - T)[Y - \mu_0(X)] \mid X\right] = (1 - p(X)) \times 0 = 0.$$

Therefore, the second term is zero regardless of $\tilde{p}(X)$:

$$\mathbb{E}\left[\frac{(1 - T) \, \tilde{p}(X) \, [Y - \mu_0(X)]}{1 - \tilde{p}(X)}\right] = 0.$$

Combining the Terms:

Thus,

$$E = \mathbb{E}\left[p(X) \, \mu_0(X)\right].$$

Recall that $\mu_0 = \mathbb{E}[\mu_0(X)]$, since $\mu_0(X) = \mathbb{E}[Y(0) \mid X]$ and $\{Y_0, Y_1\} \perp T \mid X$. Therefore,

$$\mu_0 = \mathbb{E}\left[\mu_0(X)\right] = \mathbb{E}\left[p(X) \, \mu_0(X) + [1 - p(X)] \, \mu_0(X)\right] = \mathbb{E}\left[\mu_0(X)\right].$$

Since $E = \mathbb{E}\left[p(X) \, \mu_0(X)\right]$, we have

$$E = \mathbb{E}\left[p(X) \, \mu_0(X)\right] = \mathbb{E}\left[T \, \mu_0(X)\right].$$

But we need to account for the scaling factor $\frac{1}{\mathbb{E}[T]}$. Since $\mathbb{E}[T] = \mathbb{E}\left[p(X)\right]$, we have

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \, \mathbb{E}\left[T \, \mu_0(X)\right] = \mu_0.$$

Thus, the equality holds even when $\tilde{p}(X)$ is arbitrary, provided $\mu_0(X)$ is correctly specified.

Case 2: $\mu_0(X)$ is Replaced by an Arbitrary Function $\tilde{\mu}_0(X)$

Suppose we replace $\mu_0(X)$ with any function $\tilde{\mu}_0(X)$, while $p(X)$ is correctly specified. We need to show that

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \, \mathbb{E}\left[T\,\tilde{\mu}_0(X) + \frac{(1-T)\,p(X)\,[Y - \tilde{\mu}_0(X)]}{1 - p(X)} \right].$$

Proof:

Again, consider

$$E = \mathbb{E}\left[T\,\tilde{\mu}_0(X) + \frac{(1-T)\,p(X)\,[Y - \tilde{\mu}_0(X)]}{1 - p(X)} \right].$$

First Term:

$$\mathbb{E}\left[T\,\tilde{\mu}_0(X) \right] = \mathbb{E}\left[\mathbb{E}[T \mid X]\,\tilde{\mu}_0(X) \right] = \mathbb{E}\left[p(X)\,\tilde{\mu}_0(X) \right].$$

Second Term:

$$\mathbb{E}\left[\frac{(1-T)\,p(X)\,[Y - \tilde{\mu}_0(X)]}{1 - p(X)} \right] = \mathbb{E}_X\left[\frac{p(X)}{1 - p(X)} \, \mathbb{E}\left[(1-T)[Y - \tilde{\mu}_0(X)] \mid X \right] \right].$$

Since

$$\mathbb{E}\left[(1-T)[Y - \tilde{\mu}_0(X)] \mid X \right] = [1 - p(X)]\left(\mathbb{E}[Y(0) \mid X] - \tilde{\mu}_0(X) \right) = [1 - p(X)]\left(\mu_0(X) - \tilde{\mu}_0(X) \right).$$

Therefore,

$$\mathbb{E}\left[\frac{(1-T)\,p(X)\,[Y - \tilde{\mu}_0(X)]}{1 - p(X)} \right] = p(X)\,\mathbb{E}\left[\mu_0(X) - \tilde{\mu}_0(X) \right] = p(X)\left(\mathbb{E}[\mu_0(X)] - \mathbb{E}[\tilde{\mu}_0(X)] \right).$$

Combining the Terms:

Adding the two terms,

$$E = \mathbb{E}\left[p(X)\,\tilde{\mu}_0(X) \right] + p(X)\left(\mathbb{E}[\mu_0(X)] - \mathbb{E}[\tilde{\mu}_0(X)] \right) = p(X)\,\mathbb{E}[\mu_0(X)].$$

Since $\mathbb{E}[T] = \mathbb{E}[p(X)]$, we have

$$\frac{1}{\mathbb{E}[T]}\,E = \frac{p(X)}{\mathbb{E}[p(X)]}\,\mathbb{E}[\mu_0(X)] = \mu_0.$$

Thus, the equality holds even when $\tilde{\mu}_0(X)$ is arbitrary, provided $p(X)$ is correctly specified.

Conclusion:

The moment condition

$$\mu_0 = \frac{1}{\mathbb{E}[T]} \, \mathbb{E}\left[T\,\tilde{\mu}_0(X) + \frac{(1-T)\,\tilde{p}(X)\,[Y - \tilde{\mu}_0(X)]}{1 - \tilde{p}(X)} \right]$$

holds if either $\tilde{p}(X) = p(X)$ or $\tilde{\mu}_0(X) = \mu_0(X)$. This demonstrates the double robustness of the estimator: it remains consistent if either the propensity score model or the outcome model is correctly specified.

---

3   Application – Pricing Experiment

We have data from a pricing experiment from an online recruiting service. The unit of observation is a customer of this service, which is a firm looking to hire (applicants use the service for free). The firms are charged a fixed price for access to the online recruiting system and its tools. Currently, the price is 99. But they are concerned this price is too low, so they ran an experiment. Arriving customers were randomly assigned a price of either 99 or 249. We observe the decision to either buy the service or not and we observe the `customerSize` for each firm. The data is in the file `priceExperiment.csv`.

3.a   Regression of Purchase Decision on Price

(a) Run a regression of the binary outcome `buy` on the `price`. Is this regression causal? What do the intercept and slope in this regression represent? Use potential outcomes.

3.b   Dummy Variable Regression on Price

(b) Create a dummy variable indicating the different prices. Regress `buy` on this variable. Is this regression causal? What do the intercept and slope in this regression represent? Use potential outcomes.

3.c   Regression of Revenue on Price Dummy

(c) Create a variable that measures revenue. Regress this outcome on the dummy variable you just created. Is this regression causal? What do the intercept and slope in this regression represent? Compare explicitly to the previous question. Use potential outcomes.

3.d   Statistical Significance of Price Effects

(d) At the 95% level, are the effects in parts (b) and (c) statistically significant? Justify your choice of standard errors.

3.e   Price Recommendation for the Service

(e) Should the firm stick with 99 or switch to 249? Justify your answer using the results from what you've done so far.

The data includes a variable `customerSize` that gives the size of the customer firm (remember, the customers of this business are themselves firms). The sizes are ranked 0, 1, 2, for small, medium, and large firms.

3.f CATE Estimation by Firm Size

(f) Using a *single* regression (e.g., one `lm()` command), estimate the revenue effect for each firm size individually. That is, obtain estimates of the CATEs $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, for $x = 0, 1, 2$. Verify your answer manually using a difference in means for each group. Does this pattern make sense to you?

3.g Optimal Pricing Strategy

(g) Using these results, decide on the optimal pricing strategy to maximize revenue when the service can charge different prices to different customers based on their size. We are imagining that when a firm goes to the service, they first fill out several questions, including their firm size, and then are shown a price based on these answers. (This is known as third-degree price discrimination.)

3.h Revenue Maximization Potential

(h) Can the recruiting service improve its revenue? By how much? (*Careful computing the revenue from your strategy. When using the data, think about which observations were exposed to which price, and how many of each type of firm you have.*)

3.i Computing the Plug-in Estimator of the ATE

(i) The law of iterated expectations says that the ATE obeys

$$\tau = \mathbb{E}[\tau(X)] = \mathbb{E}[Y(1) - Y(0) \mid X = 0]\mathbb{P}[X = 0] + \mathbb{E}[Y(1) - Y(0) \mid X = 1]\mathbb{P}[X = 1] + \mathbb{E}[Y(1) - Y(0) \mid X = 2]\mathbb{P}[X = 2].$$

Use this and your single regression to compute plug-in estimator of the ATE:

$$\hat{\tau} = \hat{\tau}(0)\hat{\mathbb{P}}[X = 0] + \hat{\tau}(1)\hat{\mathbb{P}}[X = 1] + \hat{\tau}(2)\hat{\mathbb{P}}[X = 2].$$

Why does this value not match what you found in part (c)? Explain rigorously and propose a different way of aggregating data from each value of X so that you obtain exactly the answer in part (c).

4   Application – NSW

The National Supported Work (NSW) Demonstration was a randomized experiment done in the 1970s to study the impact of job training on earnings. We will use the data to study subgroup effects and to illustrate selection on observables.

Randomized Experiment

The data from the experiment is in the file `nsw_RCT.csv`. We observe the following variables:

- `income.after` = earnings after training, the outcome,

- `treat` = 1 if you had job training, 0 if not,

- `age, education` = demographics measured in years (continuous),

- `black, hispanic, married, hsdegree` = binary demographic variables,

- `income.before1, income.before2` = two years of data on earnings prior to the study.

We will use this data to study the effect of job training on average and for subgroups.

4.a   Estimating the ATT

(a) Estimate the ATT using the difference in means. Is job training (on average) beneficial?

4.b   Distribution of Earnings for Treatment and Control

(b) Plot/display the distribution of earnings for the treatment and control groups. What does this tell you about the effect of job training? How does this inform how you would use the ATT estimate for policy making? What type of uncertainty is shown here?

4.c   Statistical Uncertainty Around the ATT

(c) We wish to assess the statistical uncertainty around the ATT estimate. Do this by (i) obtaining an influence function representation for the difference in means estimate, (ii) using this result to prove that the estimator is asymptotically normal and characterize the asymptotic variance, and (iii) propose consistent standard errors. Compute a 90% confidence interval.

4.d   Maximizing Welfare Through Targeting Rules

(d) We want to maximize welfare using targeting rules. To make the resulting policy easy to implement and transparent, it must be a threshold policy based on a single covariate, so search for rules of the form $d(x) = 1\{x_j > c\}$ or $d(x) = 1\{x_j < c\}$ for a specific covariate x_j and some cutoff value c. What is the welfare maximizing rule?

4.e   Role of Group Size in Targeting

(e) What role does the size of the group flagged by $d(x)$ play in your conclusions?

4.f   Testing Targeting Rules for Welfare Improvement

(f) Explain how you would statistically test the effectiveness of your targeting rule on improving welfare.

4.g   Demographic Characteristics of Targeted Individuals

(g) Examine the demographic characteristics of who your program targets compared to who is not targeted. What do you find and is this pattern concerning? (*A real study should compare the targeted demographics to the relevant/eligible population, e.g., the whole city.*)

Observational Data

Suppose there was no experiment. We have data from the same 185 men that received job training but we do not have access to the NSW control sample. For a comparison sample we found 2490 men from the Panel Study of Income Dynamics (PSID) that did not have training. The data is in `nsw_PSID.csv`.

4.h   Concerns with Observational Data

(h) Before beginning the analysis, summarize the main concern when it comes to using observational data for the analysis. Why might the PSID *comparison* group not be a good control group?

4.i   Estimation Using PSID Control Sample

(i) Using the PSID control sample as though it were the control group for a randomized trial, estimate the average treatment effect. Explain what you find and why you found it.

4.j   Suitability of the PSID as a Control Group

(j) Does the PSID sample appear to be a good control group for this purpose? That is, using the covariates, does the treatment appear to be randomly assigned?

4.k   Controlling for Nonrandomization

(k) Using the above to guide you, build a linear regression that attempts to control for any sources of nonrandomization. Does your regression-based treatment effect estimate recover the experimental benchmark treatment effect estimate? Discuss the uncertainty of your regression-based estimate and how this relates to the experimental benchmark.