# PLSC 30600 — Homework 2

**Due:** Sun, February 1 (11:59pm)

**General instructions**

- **Show your work** for all hand calculations. A final numeric answer without intermediate steps, formulas, and/or justification is not sufficient.
- For coding problems, submit a reproducible script (`.R/.rmd/.rnw`) along with the fully compiled pdf. Set a random seed and report it.
- Include a brief note describing how you used AI tools (if at all), consistent with the course AI policy.
- For each proof question, start with one sentence naming the technique you are using (e.g., direct proof, Law of Iterated Expectations, bounding, counterexample, construction, contradiction).

---

## Problem 0: Matrix algebra warm-up

For matrix *addition*, add corresponding entries. For matrix *multiplication*, use row-by-column dot products. For example, if

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad B = \begin{bmatrix} e & f \\ g & h \end{bmatrix},$$

then

$$A + B = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}, \quad AB = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}.$$

Write out at least one entry (e.g., the $(1, 2)$ entry) step-by-step, such as

$$(AB)_{12} = a \cdot f + b \cdot h.$$

**0.a Commutative law of addition.** Let

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}.$$

Compute $A + B$ and $B + A$ and verify that $A + B = B + A$.

**0.b Multiplication is not commutative.** Using the same $A$ and $B$, compute $AB$ and $BA$ and show that $AB \neq BA$.

**0.c Associative laws.** Let

$$C = \begin{bmatrix} 1 & -1 \\ 4 & 0 \end{bmatrix}.$$

Verify $(A + B) + C = A + (B + C)$ and $(AB)C = A(BC)$.

**0.d Distributive laws.** Verify $A(B + C) = AB + AC$ and $(A + B)C = AC + BC$.

**0.e Identity matrix.** Let $I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Verify $AI_2 = A$ and $I_2A = A$.

---

# Problem 1: Weighting estimators

**1.a IPW formula.** Let $p_D(X_i) \in (0, 1)$ be the propensity score for unit $i$. The IPW estimator for the ATE is:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{D_i Y_i}{p_D(X_i)} - \frac{(1 - D_i) Y_i}{1 - p_D(X_i)} \right).$$

**1.b Numerical check.** Suppose $n = 4$ with

$$\boldsymbol{D} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} 5 \\ 3 \\ 2 \\ 1 \end{bmatrix}, \quad p_D(\boldsymbol{X}) = \begin{bmatrix} 0.6 \\ 0.6 \\ 0.4 \\ 0.4 \end{bmatrix}.$$

Compute $\hat{\tau}_{\text{IPW}}$.

**1.c Compare to difference in means.** Using the same four observations, compute the unweighted difference in means $\bar{Y}_{D=1} - \bar{Y}_{D=0}$ and compare it to your IPW estimate.

---

# Problem 2: IPW proofs and identification

**2.a Cell-balancing interpretation (discrete $X$).** Assume $X$ takes finitely many values $x \in \mathcal{X}$. Let

$$\mu_1(x) = \text{E}[Y \mid D = 1, X = x], \quad \mu_0(x) = \text{E}[Y \mid D = 0, X = x], \quad p(x) = \Pr[D = 1 \mid X = x].$$

Show that

$$\text{E}\left[ \frac{YD}{p(X)} \right] = \sum_{x \in \mathcal{X}} \Pr[X = x] \; \mu_1(x),$$

and

$$\text{E}\left[ \frac{Y(1 - D)}{1 - p(X)} \right] = \sum_{x \in \mathcal{X}} \Pr[X = x] \; \mu_0(x).$$

Conclude that the IPW estimand equals

$$\sum_{x \in \mathcal{X}} \Pr[X = x] \big( \mu_1(x) - \mu_0(x) \big),$$

and explain briefly why this corresponds to reweighting so treated/control have the same $X$-distribution. (You may cite the identity from Theorem 7.2.5 in Aronow-Miller.)

**2.b** **Optional extension (ATT via weighting).** Derive an analogous identity for the ATT:

$$\text{E}[Y(1) - Y(0) \mid D = 1].$$

Using weighting, show how to express the ATT in terms of observable quantities. (The weights and normalization differ from the ATE case.)

**2.c Where does it break?** Construct a simple data-generating process with binary $X \in \{0, 1\}$, binary $D$, and potential outcomes $Y(1), Y(0)$ such that:

(i) positivity holds: $0 < p(X) < 1$ for both $X = 0, 1$,

(ii) ignorability fails: $(Y(1), Y(0)) \not\perp D \mid X$.

You may use a finite “science table” counterexample or a probability model.

Compute both

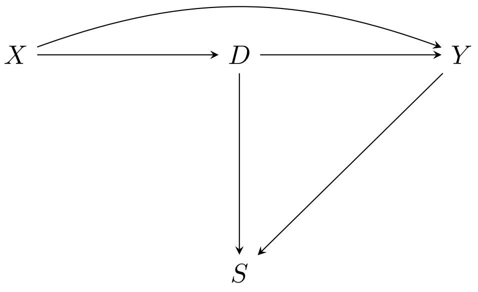
$$\mathrm{E}[\tau] = \mathrm{E}[Y(1) - Y(0)], \qquad \mathrm{E}\left[\frac{YD}{p(X)} - \frac{Y(1-D)}{1-p(X)}\right],$$

and show they differ. Refer to where the proof of Theorem 7.2.5 fails under your DGP.

---

## Problem 3: DAGs and post-treatment selection

Consider the DAG with observed pre-treatment covariates $X$, treatment $D$, outcome $Y$, and a post-treatment selection variable $S$:



**3.a Conditioning on a post-treatment collider.** Suppose you estimate

$$\mathrm{E}[Y \mid D = 1, S = 1] - \mathrm{E}[Y \mid D = 0, S = 1].$$

Does this identify the average treatment effect (ATE) $\mathrm{E}[Y(1) - Y(0)]$? Answer *yes* or *no* and justify briefly using DAG language (e.g., d-separation / backdoor paths). If your answer is *no*, name the noncausal path(s) that become(s) open when conditioning on $S$.

**3.b What estimand (if any) is being targeted?** Let $S(d)$ denote the potential selection status under treatment $d \in \{0, 1\}$. Consider the “always-selected” principal stratum $\{S(1) = 1,\ S(0) = 1\}$. Is the quantity

$$\mathrm{E}[Y \mid D = 1, S = 1] - \mathrm{E}[Y \mid D = 0, S = 1]$$

equal to the principal-stratum causal effect

$$\mathrm{E}\big[\,Y(1) - Y(0) \;\big|\; S(1) = 1,\ S(0) = 1\,\big]\ ?$$

Explain in 2–4 sentences. (If you think it can be given a causal interpretation only under additional assumptions, state one such assumption.)

---

## Problem 4: LaLonde data

**Data.** Use the experimental LaLonde dataset: Download the file from github. \

`https://raw.githubusercontent.com/xuyiqing/lalonde/master/data/lalonde/nsw.dta`.

The experimental NSW sample provides a randomized benchmark before introducing aditional adjustment methods. The outcome is `re78` and the treatment indicator is `treat`.

| Variable | Description |
|---|---|
| `treat` | Treatment indicator (NSW program) |
| `re78` | Earnings in 1978 (outcome) |
| `re75` | Earnings in 1975 (pre-treatment) |
| `age` | Age |
| `education` | Years of education |
| `black` | Indicator for Black |
| `hispanic` | Indicator for Hispanic |
| `married` | Indicator for married |
| `nodegree` | Indicator for no high school degree |

**Packages you may need:** `haven`, `hot.deck`, `estimatr`.

```
# You may need to run `install.packages(...)` first
library(haven) # to read in Stata .dta files
library(estimatr) # for lm_robust and lm_lin
library(hot.deck) # for hot-deck imputation

nsw_url <- "https://raw.githubusercontent.com/xuyiqing/lalonde/master/data/lalonde/nsw.dta"
nsw <- read_dta(nsw_url)
nsw <- as.data.frame(nsw)
```

**4.a Read and inspect.** Load the dataset and report:

(i) the number of treated and control units,

(ii) the mean of $Y$ in each group.

**4.b Difference in means.** Compute the unadjusted difference-in-means estimate $\widehat{\tau}_{\text{DM}} = \bar{Y}_{D=1} - \bar{Y}_{D=0}$.

**4.c Propensity score estimation and overlap.** Estimate the propensity score $\widehat{p}_D(X_i) = \Pr[D_i = 1 \mid X_i]$ using a logit model with `glm(..., family = binomial())`. Use the covariates `age`, `education`, `black`, `hispanic`, `married`, `nodegree`, and `re75`. After fitting the model, use `predict(..., type = "response")` to get propensity scores. Then plot treated vs control distributions (histograms or density plots) with the same x-axis limits to assess overlap. Comment on overlap.

**4.d IPW estimator.** Using the estimated propensity scores, compute the IPW ATE:

$$\widehat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\widehat{p}_D(X_i)} - \frac{(1-D_i)Y_i}{1-\widehat{p}_D(X_i)}\right).$$

Compute an approximate 95% confidence interval using the nonparametric bootstrap. *Hint:* include propensity score estimation inside each bootstrap resample.

**4.e Hot-deck matching (propensity score).** Use the `hot.deck` package to impute counterfactual outcomes using the opposite treatment group as donors. Make sure to set a random seed. Use the estimated propensity score as the matching variable (instead of the full covariate set). One simple way is to set `re78` to `NA` for treated units (so controls are the only donors) and run hot-deck imputation, then repeat with `re78` set to `NA` for controls. Use the completed datasets to compute the ATE.

**4.f Linear model with covariates.** Fit a linear regression of $Y$ on $D$ and the covariates used in the propensity score model. Report the coefficient on $D$ and interpret it as an adjusted ATE.

**4.g** **Linear model with propensity score (estimatr).** Fit a linear regression of $Y$ on $D$ and $\widehat{p}_D(X)$ using `estimatr::lm_robust`. Report the coefficient on $D$ and compare it to your earlier estimates.

**4.h** **Lin estimator vs. by-hand.** Use `estimatr::lm_lin` to compute the regression-adjusted ATE. Then demean the covariates, add treatment interactions, and estimate the same model using `lm_robust`. Compare the two estimates and confirm they match.

**4.i** **Diagnostics.** Briefly summarize what you learned from comparing the estimators in this problem (2–4 sentences).