PLSC 30600 — Homework 1

Due: Fri, January 16 (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: Proof techniques mini-lab

For each part below: (i) state the technique you are using (e.g., bounding, counterexample, construction, Law of Iterated Expectations), and (ii) write a clean argument. You may use the Law of Iterated Expectations without re-proving it. Use Aronow–Miller proofs as *direct* templates for identification arguments; the goal is to practice working through the steps of these proofs, not to come up with new techniques.

0.a Bounding argument. Let $\tilde{Y}_i$ be any completion of the missing outcomes such that $\tilde{Y}_i = Y_i$ when $R_i = 1$ and $\tilde{Y}_i \in [0, 1]$ when $R_i = 0$. Let $\bar{\tilde{Y}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{Y}_i$. Prove that

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} R_i Y_i}_{\text{impute missing as 0}} \;\leq\; \bar{\tilde{Y}} \;\leq\; \underbrace{\frac{1}{n}\sum_{i=1}^{n} \big(R_i Y_i + (1 - R_i)\big)}_{\text{impute missing as 1}}.$$

(Interpret these two endpoints as the plug-in Manski bounds.)

0.b Law of Iterated Expectations proof (MAR identification). Assume MAR given X:

$$Y \perp R \mid X, \qquad \Pr[R = 1 \mid X = x] > 0 \text{ for each } x \in \{A, B\}.$$

Prove that

$$\mathrm{E}[Y] = \sum_{x \in \{A,B\}} \mathrm{E}[Y^* \mid R = 1, X = x] \Pr[X = x].$$

(Your proof should use the Law of Iterated Expectations and the MAR assumption at the appropriate step.)

0.c Counterexample (independence vs conditional independence). Consider binary random variables $X, Y, R \in \{0, 1\}$ with $\Pr[X = 0] = \Pr[X = 1] = 1/2$ and

$$f_{Y,R|X}(y, r \mid x = 0) = \begin{cases} 1/2 & \text{if } (y, r) = (1, 1), \\ 1/2 & \text{if } (y, r) = (0, 0), \\ 0 & \text{otherwise,} \end{cases} \qquad f_{Y,R|X}(y, r \mid x = 1) = \begin{cases} 1/2 & \text{if } (y, r) = (1, 0), \\ 1/2 & \text{if } (y, r) = (0, 1), \\ 0 & \text{otherwise.} \end{cases}$$

(i) Show that $Y \perp R$ (unconditional independence).

(ii) Show that $Y \not\perp R \mid X$ (conditional independence fails).

Conclude: unconditional independence does *not* imply conditional independence.

Setup and toy dataset (used in Problem 1 and Coding A–B)

We observe i.i.d. draws of (Y_i^*, R_i, X_i) for $i = 1, \ldots, n$, where:

- $Y_i \in [0, 1]$ is the outcome of interest (bounded).
- $R_i \in \{0, 1\}$ is the **response/observation indicator**: $R_i = 1$ means Y_i is observed, $R_i = 0$ means Y_i is missing.
- Y_i^* is the recorded outcome:

$$Y_i^* = \begin{cases} Y_i, & R_i = 1, \\ -99, & R_i = 0. \end{cases}$$

- $X_i \in \{A, B\}$ is a covariate observed for **all** units (including nonrespondents).

The toy dataset is:

i	X_i	R_i	Y_i^*
1	A	1	0.20
2	A	0	-99
3	B	1	0.80
4	B	0	-99
5	A	1	0.25
6	A	0	-99
7	B	1	0.85
8	B	1	0.90
9	A	0	-99
10	A	1	0.30

Throughout, let $\mu \equiv \mathrm{E}[Y]$ denote the population mean of Y.

Problem 1: Manski bounds, MCAR, and MAR on the same dataset

1.a Manski bounds. Assume only that $Y \in [0, 1]$. Compute the **plug-in** lower and upper bounds for $\mu = \mathrm{E}[Y]$ using the estimators from Aronow & Miller Theorem 6.1.4.

1.b MCAR (point identification). Now assume **MCAR**:

$$Y \perp R \quad \text{and} \quad \Pr[R = 1] > 0.$$

Compute the MCAR plug-in estimator $\widehat{\mu}_{\mathrm{MCAR}}$ from the toy dataset.

1.c MAR given X (point identification by stratification). Now assume **MAR given X**:

$$Y \perp R \mid X, \qquad \Pr[R = 1 \mid X = x] > 0 \text{ for each } x \in \{A, B\}.$$

Compute the MAR plug-in estimator $\widehat{\mu}_{\mathrm{MAR}}$ from the toy dataset.

1.d Interpretation (short answer).

(i) Why is MCAR typically regarded as *stronger* than MAR?

(ii) How does the positivity condition differ in spirit between MCAR and MAR?

1.e Sanity checks via bounding (proof). Using the result from Problem 0(a), prove that both $\hat{\mu}_{\text{MCAR}}$ and $\hat{\mu}_{\text{MAR}}$ (from this dataset) must lie inside the plug-in Manski bounds you computed. (Do *not* argue by plugging in numbers; give an inequality argument.)

Problem 2: Propensity scores for missing data

Let $p_R(X) \equiv \Pr[R = 1 \mid X]$ and assume MAR given X: $Y \perp R \mid X$.

2.a Balance (conditioning on the propensity score). Show that

$$R \perp X \mid p_R(X).$$

Hint: because R is binary, it suffices to show $\mathrm{E}[R \mid X, p_R(X)] = \mathrm{E}[R \mid p_R(X)]$.

2.b Outcome–response independence given the propensity score. Show that

$$Y \perp R \mid p_R(X).$$

Hint: again use the binary-R trick and the Law of Iterated Expectations; under MAR, $\mathrm{E}[R \mid Y, X] = \mathrm{E}[R \mid X] = p_R(X)$.

Problem 3: MCAR vs. MAR diagnostic thinking (vignette)

A city surveys residents to estimate turnout in the most recent election. Let $Y \in \{0, 1\}$ indicate whether a person voted, $R \in \{0, 1\}$ indicate survey response, and let X be age group (recorded for everyone sampled from administrative data, including nonrespondents).

The director believes:

- older residents are more likely to respond (so R depends on X),
- turnout differs by age (so Y depends on X).

3.a Is MCAR ($Y \perp R$) plausible? Why or why not? (2–4 sentences)

3.b Give a realistic story under which MAR given age ($Y \perp R \mid X$) might be plausible. What does the story rule out? (3–5 sentences)

3.c Suppose age is *only* observed for respondents (you do not observe X when $R = 0$). Which step in the MAR identification argument fails, and why? (3–5 sentences)

Problem 4: Potential outcomes and the basic identification gap (Lemma 1)

Let $D \in \{0, 1\}$ denote a treatment indicator. Let $Y(1)$ and $Y(0)$ be potential outcomes, and let the observed outcome be

$$Y = Y(1)D + Y(0)(1 - D).$$

Lemma 1 (Identification gap). In general,

$$\mathrm{E}[Y(1)] \neq \mathrm{E}[Y \mid D=1] \quad \text{and} \quad \mathrm{E}[Y(0)] \neq \mathrm{E}[Y \mid D=0].$$

4.a Exercise: Prove Lemma 1. Provide a proof by constructing a counterexample distribution for $(Y(0), Y(1), D)$ where at least one inequality holds. (You may use a finite "science table" counterexample or a probability model.)

4.b When does equality hold? State a sufficient condition under which

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

Briefly justify your answer (a short argument is sufficient).

4.c Selection bias decomposition. Starting from $\mathrm{E}[Y \mid D=1] - \mathrm{E}[Y \mid D=0]$, show that

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

Interpret the two terms in 2–4 sentences.

Problem 5: Continuous treatment and average marginal causal effects

Let D be a continuous treatment with support $[0,1]$ and suppose each unit has potential outcomes $\{Y(d) : d \in [0,1]\}$, with $Y(d)$ differentiable in d. Assume strong ignorability:

$$\{Y(d)\}_{d\in[0,1]} \perp D \mid X, \qquad 0 < \Pr[D \in \mathcal{N} \mid X] \text{ for neighborhoods } \mathcal{N} \subset [0,1].$$

5.a Define the conditional response function $m(d,x) = \mathrm{E}[Y \mid D=d, X=x]$. Show that under strong ignorability,

$$\mathrm{E}[Y(d)] = \mathrm{E}\big[m(d,X)\big].$$

5.b Suppose

$$m(d,x) = \alpha + \beta d + \gamma d^2 + \delta x + \eta(d \cdot x),$$

where $x \in \{0,1\}$. Compute $\frac{\partial}{\partial d} m(d,x)$.

5.c The (population) average marginal causal effect (AMCE) is defined as

$$\mathrm{AMCE} \equiv \mathrm{E}\left[\frac{\partial}{\partial d} m(D,X)\right].$$

Assume $D \sim \mathrm{Uniform}(0,1)$ and $\Pr[X=1] = p$ with $D \perp X$. Compute AMCE in terms of (β, γ, η, p). (Show your work, including any integrals you use.)

Coding section

Your code should be readable and reproducible. You may use Claude Code for assistance, but you are responsible for correctness and interpretation.

Coding A: Replicate Problem 1 numerically

A1. Enter the toy dataset into your code (as a data frame with columns `i`, `X`, `R`, `Ystar`).

```
toy <- data.frame(
  i = 1:10,
  X = c("A","A","B","B","A","A","B","B","A","A"),
  R = c(1,0,1,0,1,0,1,1,0,1),
  Ystar = c(0.20,-99,0.80,-99,0.25,-99,0.85,0.90,-99,0.30)
)
```

A2. Write a function that takes (Y^*, R) and returns the plug-in Manski bounds for $\mathrm{E}[Y]$ when $Y \in [0, 1]$.

A3. Write a function that computes $\widehat{\mu}_{\mathrm{MCAR}}$.

A4. Write a function that computes $\widehat{\mu}_{\mathrm{MAR}}$ by stratification on $X \in \{A, B\}$.

A5. Print the three results (bounds, MCAR estimate, MAR estimate) and verify they match your hand calculations.

Coding B: Bootstrap confidence intervals

Use the nonparametric bootstrap (resample rows with replacement). Letting B be the number of bootstrap resamples, use $B = 100$.

B1. For each bootstrap sample, compute:

- the lower and upper Manski bounds;
- $\widehat{\mu}_{\mathrm{MCAR}}$;
- $\widehat{\mu}_{\mathrm{MAR}}$.

(Do *not* print out all 100, just show your code.)

B2. Positivity diagnostics in the bootstrap (MAR). In some bootstrap resamples, one of the strata (A or B) may have no respondents ($R = 1$), making $\widehat{\mu}_{\mathrm{MAR}}$ undefined.

- Report the fraction of bootstrap resamples in which $\widehat{\mu}_{\mathrm{MAR}}$ is undefined.
- For the resamples where it is defined, compute a 95% percentile CI for μ under MAR.

B3. Compute 95% percentile bootstrap CIs for:

$$\text{Lower bound}, \quad \text{Upper bound}, \quad \widehat{\mu}_{\mathrm{MCAR}}, \quad \widehat{\mu}_{\mathrm{MAR}}.$$