Problem Set 1 Answer Key

STAT II (Spring 2025)

Disclaimer: *These answers are not necessarily the only correct way to answer the question. Some problems have multiple ways to answer the question and receive full credit. Even if your answer differs from the answers here, this does not mean you will necessarily receive points off for the answer.*

Problem 1 (10 points)

Suppose that X, Y, and Z are random variables. Prove the following:

$$\text{cov}(X+Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$$

Answer (5 points for recalling covariance + 5 points for showing the equality):

Recall the formula for the covariance of two random variables:

$$\begin{aligned}
\text{cov}(X, Z) &= \mathbb{E}[(X - \mathbb{E}[X])(Z - \mathbb{E}[Z])] \\
&= \mathbb{E}[XZ - X\mathbb{E}[Z] - Z\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Z]] \\
&= \mathbb{E}[XZ] - \mathbb{E}[X\mathbb{E}[Z]] - \mathbb{E}[Z\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]\mathbb{E}[Z]] \\
&= \mathbb{E}[XZ] - 2\mathbb{E}[X]\mathbb{E}[Z] + \mathbb{E}[X]\mathbb{E}[Z] \\
&= \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z]
\end{aligned}$$

Similarly, $\text{cov}(Y, Z) = \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z]$. Together,

$$\text{cov}(X, Z) + cov(Y, Z) = \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] + \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z].$$

Now, using the same formula for the covariance, we can show:

$$\begin{aligned}
\text{cov}((X+Y), Z) &= \mathbb{E}[(X+Y)Z] - (\mathbb{E}[X+Y])\mathbb{E}[Z] \\
&= \mathbb{E}[XZ + YZ] - (\mathbb{E}[X] + \mathbb{E}[Y])\mathbb{E}[Z] \\
&= \mathbb{E}[XZ] + \mathbb{E}[YZ] + \mathbb{E}[X]\mathbb{E}[Z] + \mathbb{E}[Y]\mathbb{E}[Z] \\
&= \text{cov}(X+Z) + \text{cov}(Y+Z) \qquad \square
\end{aligned}$$

Problem 2 (10 points)

Suppose that you flip two fair coins. Define the random variables A, B, and C as follows:

$$A = \begin{cases} 1 & \text{if the first coin is heads} \\ 0 & \text{otherwise} \end{cases}$$

$$B = \begin{cases} 1 & \text{if the second coin is heads} \\ 0 & \text{otherwise} \end{cases}$$

$$C = \begin{cases} 1 & \text{if one coin is heads and the other is tails} \\ 0 & \text{otherwise} \end{cases}$$

Note that $A \perp\!\!\!\perp B$. Is it true that $A \perp\!\!\!\perp B \mid C$? Make sure to use the definition of conditional independence to offer a formal justification for your answer.

Answer (5 points for stating what conditional independence implies + 5 points for showing that conditional expectation does not hold):

We can use a proof by contradiction to demonstrate that this is not true.

Assume that conditional independence holds, then we know that:

$$\Pr(A, B \mid C = c) = \Pr(A \mid C = c) \Pr(B \mid C = c)$$

Note that since the coins are fair we have

$$\Pr(A = 1) = 0.5, \qquad \Pr(B = 1) = 0.5, \qquad \Pr(C = 1) = 0.5 \cdot 0.5 + 0.5 \cdot 0.5 = 0.5.$$

Next, using Bayes' rule and the probabilities above, we can express the conditional probabilities as follows:

$$\Pr(A = 1 \mid C = 1) = \frac{\Pr(C = 1 \mid A = 1) \Pr(A = 1)}{\Pr(C = 1)} = \frac{0.5 \cdot 0.5}{0.5} = 0.5$$
$$\Pr(B = 1 \mid C = 1) = \frac{\Pr(C = 1 \mid B = 1) \Pr(B = 1)}{\Pr(C = 1)} = \frac{0.5 \cdot 0.5}{0.5} = 0.5$$

Hence we know that:

$$= \Pr(A = 1 \mid C = 1) \Pr(B = 1 \mid C = 1)$$
$$= (0.5) \cdot (0.5) = 0.25$$

However, $p(A = 1, B = 1 \mid C = 1) = 0 \neq 0.25$. $A \perp\!\!\!\perp B$ but not $A \perp\!\!\!\perp B \mid C$. $\square$

Problem 3 (30 points)

This problem aims to help you get familiar with the potential outcomes notation. Throughout the problem, you can assume that the Stable Unit Treatment Value Assumption (SUTVA) holds.

(a) Explain in words what the notation $Y_i(0)$ represents.

Answer (3 points): $Y_i(0)$ denotes the potential outcome for unit i under no treatment, i.e., when $t_i = 0$.

(b) Contrast the meaning of $Y_i(0)$ with the meaning of Y_i.

Answer (3 points): The first is the potential outcome for subject i if this subject were untreated. The second is simply the observed outcome for subject i.

(c) Contrast the meaning of $Y_i(0)$ with the meaning of $Y_i(1)$. Is it ever possible to observe both at the same time for unit i? Why?

Answer (4 points): The first is the potential outcome for i under control. The second is the potential outcome for i under treatment. In any moment, only one of the two potential outcomes for i can be realized. A subject cannot be under control and treatment simultaneously, so observing both potential outcomes is not possible. This is known as the "fundamental problem of causal inference."

(d) Explain the notation $\mathbb{E}[Y_i(0) \mid T_i = 1]$, where T_i is a binary variable that gives the treatment status for subject i, 1 if treated, 0 if control.

Answer (5 points): $\mathbb{E}[Y_i(0) \mid T_i = 1]$ is the expected value of the potential outcome for subject i if the subject were untreated, given that this subject actually receives treatment. Another way to put it: the expected value of the untreated potential outcome for a subject in the treatment group.

(e) Contrast the meaning of $\mathbb{E}[Y_i(0)]$ with the meaning of $\mathbb{E}[Y_i \mid T_i = 0]$.

Answer (5 points): The difference between these two terms is that the former term is the expectation of untreated potential outcome for all units, regardless of assignment to treatment, while the latter, $\mathbb{E}[Y_i \mid T_i = 0]$, is the expectated observed outcome for among untreated units. The two terms may differ if there is selection into *not* receiving (and by extension into receiving) treatment.

(f) Contrast the meaning of $\mathbb{E}[Y_i(0) \mid T_i = 1]$ with the meaning of $\mathbb{E}[Y_i(0) \mid T_i = 0]$.

Answer (5 points): $\mathbb{E}[Y_i(0) \mid T_i = 1]$ is the expected untreated potential outcome among treated units ($T_i = 1$). Conversely, $\mathbb{E}[Y_i(0) \mid T_i = 0]$ is the expectated untreated potential outcome among untreated units ($T_i = 0$). Similarly to the previous question the two expectations might differe when there is selection into treatment.

(g) Which of the following four expectations (that you explained in the previous questions) can be identified from the observed data?

Note: Identification means that an expression can be reduced to a function that contains only observed variables and no potential outcomes. Do not make any additional assumptions about the distributions of the variables, except that there is at least one observation with $T_i = 1$, and at least one with $T_i = 0$ in the observed data.

$$\mathbb{E}[Y_i(0) \mid T_i = 1]$$
$$\mathbb{E}[Y_i(0)]$$
$$\mathbb{E}[Y_i \mid T_i = 0]$$
$$\mathbb{E}[Y_i(0) \mid T_i = 0]$$

Answer (5 points): We can identify (3) $\mathbb{E}[Y_i \mid T_i = 0]$ and (4) $\mathbb{E}[Y_i(0) \mid T_i = 0]$. Note also that under the Stable Unit Treatment Value Assumption (SUTVA), these two quantities are equal.

Problem 4 (10 points)

Consider the following population of six individuals, with respective treatment assignments and potential outcomes represented in the table below:

i	T_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	1	4	4	3
2	0	4	8	4
3	1	2	2	0
4	1	3	3	0
5	0	8	7	8
6	0	7	6	7

Answer the following questions about this population.

(a) If you naively compare *observed* differences between treatment and control groups (difference in means estimator, τ_{DiM}), what is your estimate of the average treatment effect? In other words, compute the value of:

$$\mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0]$$

Answer (3 points):

$$\begin{aligned}\tau &= \mathbb{E}[Y_i \mid T_i = 1] - \mathbb{E}[Y_i \mid T_i = 0] \\ &= \frac{4+2+3}{3} - \frac{4+8+7}{3} \\ &= 3 - 6\frac{1}{3} = -3\frac{1}{3}\end{aligned}$$

(b) What is the value of selection bias with respect to $Y_i(1)$ in this example? *Hint: recall the formula of the difference in means estimator with respect to Average Treatment Effect on the Control (ATC)*

Answer (3 points):

$$
\begin{aligned}
SB &= E[Y_i(1)|T_i = 1] - E[Y_i(1)|T_i = 0] \\
&= E\left[\frac{4+2+3}{3}\right] - E\left[\frac{8+7+6}{3}\right] \\
&= 3 - 7 = -4
\end{aligned}
$$

(c) Calculate value of the average treatment effect. Is it the same as the naive comparison of observed means you calculated in part (a)? Explain why or why not.

Answer (4 points):

$$
\begin{aligned}
\tau_{ATE} = \mathbb{E}[\tau_i] &= \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) \\
&= \frac{1}{N}[(4-3) + (8-4) + (2-0) + (3-0) + (7-8) + (6-7)] \\
&= \frac{1}{N}[1 + 4 + 2 + 3 - 1 - 1] \\
&= \frac{8}{6} = 1\frac{1}{3}
\end{aligned}
$$

This is different from the naïve difference-in-means estimate calculated in part (a). With that estimator we cannot calculate the true average treatment effect since we cannot observe both potential outcomes for any subject i. Thus, the true average treatment effect at the individual level is unattainable. Naïve difference-in-means estimate can return the true average treatment effect in expectation if, for example we know that potential outcomes are independent of the treatment assignment.

Problem 5 (10 points)

Recall the strong ignorability assumption we discussed in class, $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i$. Now suppose that we weaken it instead assuming that only $Y_i(0) \perp\!\!\!\perp T_i$. In this case, it is not possible to identify the ATE by naively comparing the difference in means of Y_i between treatment and control (untreated) groups. Can you identify the average treatment effect on the treated (ATT) with a naive comparison of means? Offer formal justification for your answer.

Answer (3 points for substituting POs + 4 points for $\pm$ trick + 3 for the rest of the proof):

With only the assumption that $Y_i(0) \perp\!\!\!\perp T_i$, we can derive the average treatment effect on the treated ($\mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1]$) with a naïve difference-in-means estimator. To do so we can recall first that under $X \perp\!\!\!\perp Y$ we have $\mathbb{E}[X \mid Y = y] = \mathbb{E}[X]$.

$$
\begin{aligned}
\mathbb{E}[Y_i \mid T_i = 1] - \mathbb{E}[Y_i \mid T_i = 0] &= \mathbb{E}[Y_i(1) - Y_i(0) + Y_i(0) \mid T_i = 1] - \mathbb{E}[Y_i(0) \mid T_i = 0] \\
&= \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1] + \mathbb{E}[Y_i(0) \mid T_i = 1] - \mathbb{E}[Y_i(0) \mid T_i = 0] \\
&= \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1] + \mathbb{E}[Y_i(0)] - \mathbb{E}[Y_i(0)] \\
&= \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1] \qquad \square
\end{aligned}
$$

Problem 6 (30 points)

Suppose you are interested in understanding the causal relationship between incumbency (binary) and vote share (share on a scale between 0 and 1) of a candidate in local elections.

(a) Identify at least one covariate that should be included in a causal regression of vote share on incumbency status. Explain why this variable should be included?

Answer (5 points): For this problem, let the election take place at time t and the incumbent politician was elected in period *t-1*. One covariate that should be included in a causal regression of vote share on incumbency status is vote share for the election at $t-1$. This should be included because it affects incumbency and it also affects vote share in election t independently of incumbency status. Specifically, vote share in election $t-1$ can serve as a proxy measure for candidate quality, where higher quality candidates routinely win a larger vote margin, making them more likely to be incumbents and have a higher vote share in elections. Conditioning on this variable can allow us to causally measure the effect of incumbency on vote share in an election at time t (if we assume there are no other confounding variables).

(b) Consider a variable that measures media coverage on candidate (e.g., number of stories about candidate per week), which can be influenced by both incumbency status and the vote share of a candidate. Discuss whether you should control for this variable in your analysis. Explain the potential implications of including or excluding this variable in your model, and whether it may introduce any biases.

Answer (5 points): Media coverage in the current election should not be included as a covariate in a causal regression of vote share on incumbency status. As mentioned in the question, media coverage is influenced by incumbency status and by the vote share of a candidate in the election at time period t, which makes it a collider. Controlling for a post-treatment variable could induce collider (selection) bias.

(c) Should you include a control variable for campaign spending in the election prior to the one being studied? Explain whether including this variable might help or hinder causal identification, and offer reasoning to support your decision.

Answer (5 points): Yes, you should include a control variable for campaign spending in the election prior to the one being studied. This is because previous campaign spending can be affect both incumbency status through the share in the past elections as well as the current vote share (e.g. through current campaign spendings). In other words, the past campaign spending could be another confounder.

(d) Draw a Directed Acyclic Graph (DAG) that illustrates the main relationship between incumbency status and vote share as well as relationships between these variables and the variables you discussed in parts (a)-(c) of this problem. Describe the directions (positive or negative) of causal relationships implied by your DAG.

Answer (4 points):

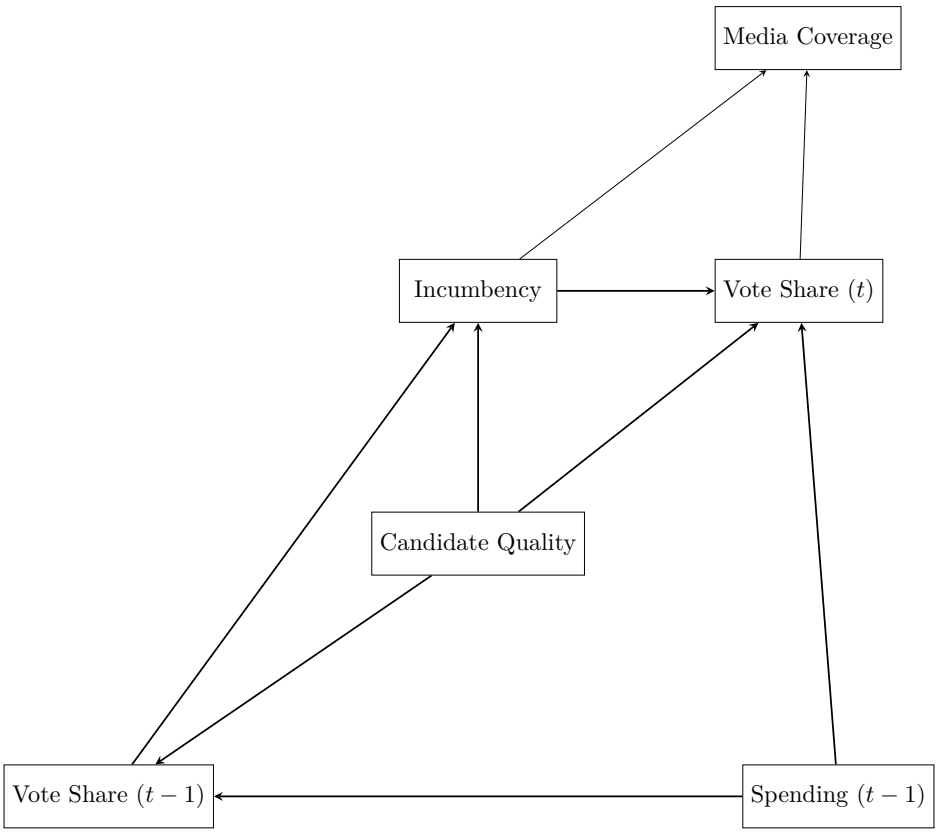

Figure 1: DAG

Voteshare $(t-1)$ and campaign spending $(t-1)$ both have a positive relationship with candidate quality and incumbency. Candidate quality has a positive relationship with incumbency and voteshare (t). Incumency has a positive relationship with voteshare (t).

(e) In R simulate a dataset with 1000 observations that reflects the DAG structure you presented in part (d), including potential variables such as campaign spending, electoral success, other relevant covariates you described in part (a), and variables discussed in parts (b)-(c).

Answer (5 points):

```
### Setup
# Set seed
set.seed(20250208)

# Number of observations
n <- 1000

### Step 1: Simulate candidate quality (U) as a confounder
# Candidate quality influences both spending at t-1 and vote share at t-1
can_qual <- truncnorm::rtruncnorm(n, a = 0, b = 1, mean = 0.5, sd = 0.5)

### Step 2: Simulate spending in t-1 (not used directly later, but a potential covariate)
spend_t1 <- truncnorm::rtruncnorm(n, a = 0, b = 100, mean = 10, sd = 5)

### Step 3: Simulate vote share (t-1)
```

```r
# Candidate quality and spending affect vote share at t-1
vs_t1 <- truncnorm::rtruncnorm(n,
  a = 0, b = 1,
  mean = 0.5 + 0.01 * can_qual + 0.05 * spend_t1,
  sd = 0.2
)

### Step 4: Simulate incumbency (collider variable)
# Incumbency is determined by candidate quality and previous vote share.
incumbency <- rbinom(n, size = 1, prob = 0.5 + 0.2 * can_qual + 0.3 * vs_t1)

### Step 5: Simulate vote share at t
# Vote share in the current period is influenced by candidate quality and incumbency.
vs_t <- truncnorm::rtruncnorm(n, a = 0, b = 1, mean = 0.5 + 0.1 * can_qual + 0.1 * incumbency + 0

### Step 6: Simulate Media Coverage (Collider)
# Media coverage is influenced by both vote share and incumbency.
media <- 0.5 * incumbency + 0.5 * vs_t + rnorm(n, sd = .5)

# This makes media a **collider variable**: if we condition on media coverage in analysis,
# we could induce spurious relationships between incumbency and vote share (t)
# since media is affected by both.

### Step 7: Create the final simulated dataset
simulated_data <- data.frame(
  candidate_quality = can_qual, # Confounder
  spending_t_minus_1 = spend_t1,
  voteshare_t_minus_1 = vs_t1, # Confounder
  incumbency = incumbency, # Collider
  media_coverage = media, # Collider
  voteshare_t = vs_t # Outcome
)
```

(f) Using `lm()` function in R estimate the base regression model of incumbency status on the vote share. Next, run the regression that you think will perform the best in terms of getting estimates close to the true effect (that you assumed when simulating the data). Finally run regression that includes media coverage variable and campaign spending in the election prior (parts (b) and (c)). Briefly discuss differences or similarities in estimates between campaign spending and the vote share you observe across these models.

Answer (5 points):

```r
# Baseline model
model_1 <- lm(
  voteshare_t ~ incumbency,
  data = simulated_data
)

# Best perfomring model (under assumptions from above)
model_2 <- lm(
```

```r
  voteshare_t ~
    incumbency +
    spending_t_minus_1 +
    candidate_quality,
  data = simulated_data
)

# The "kitchen sink" model
model_3 <- lm(
  voteshare_t ~
    incumbency +
    candidate_quality +
    spending_t_minus_1 +
    voteshare_t_minus_1 +
    media_coverage,
  data = simulated_data
)

# Let's see our model outputs!
summary(model_1)
```

```
Call:
lm(formula = voteshare_t ~ incumbency, data = simulated_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.303744 -0.068417 -0.001546  0.070103  0.296823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.559553   0.007803   71.71   <2e-16 ***
incumbency  0.143601   0.008575   16.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1023 on 998 degrees of freedom
Multiple R-squared:  0.2194,    Adjusted R-squared:  0.2186
F-statistic: 280.4 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
summary(model_2)
```

```
Call:
lm(formula = voteshare_t ~ incumbency + spending_t_minus_1 +
    candidate_quality, data = simulated_data)

Residuals:
```

```
      Min        1Q    Median        3Q       Max
-0.304488 -0.066280  0.001267  0.066815  0.274553

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.479543   0.010272  46.685  < 2e-16 ***
incumbency         0.120363   0.008301  14.499  < 2e-16 ***
spending_t_minus_1 0.004581   0.000641   7.147 1.71e-12 ***
candidate_quality  0.107193   0.011463   9.351  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09612 on 996 degrees of freedom
Multiple R-squared:  0.3126,    Adjusted R-squared:  0.3105
F-statistic:   151 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
summary(model_3)
```

```
Call:
lm(formula = voteshare_t ~ incumbency + candidate_quality + spending_t_minus_1 +
    voteshare_t_minus_1 + media_coverage, data = simulated_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.289140 -0.065396  0.001223  0.066600  0.274379

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.4525515  0.0174270  25.968  < 2e-16 ***
incumbency          0.1084570  0.0088847  12.207  < 2e-16 ***
candidate_quality   0.1042612  0.0114107   9.137  < 2e-16 ***
spending_t_minus_1  0.0038979  0.0007625   5.112 3.83e-07 ***
voteshare_t_minus_1 0.0361736  0.0226852   1.595 0.111122
media_coverage      0.0212613  0.0059941   3.547 0.000408 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09547 on 994 degrees of freedom
Multiple R-squared:  0.3232,    Adjusted R-squared:  0.3198
F-statistic: 94.94 on 5 and 994 DF,  p-value: < 2.2e-16
```

The three models produce fairly different results. While it is hard to say for sure based on the single draw of data (to prove presense of the bias we would need to run the Monte Carlo simulations and compare the resulting estimates with the true τ) we would expect that the first model would over- and the third model would under-estimate the *true* relationship between incumbency and vote share in period t ($\tau = 0.1$). The former difference would be induced by the confounder bias due to candidate quality and past campaign spending, while the latter would be due to collider bias induced by controlling for media coverage.

(g) (BONUS - 10 points) Which regression model will provide estimates of the relationship between incumbency and vote share that are close to the true relationship you assumed but also have highest precision of those estimates (i.e. lowest standard error). Show this regression model output and briefly explain why you think the standard errors will be the lowest in this case.

Answer (10 points):

A model in which there is some Z term which affects Y but is not causally connected to X, such as economic conditions at election t could increase the precision of the regression. This is because adding covariates that reduce variation in Y results in higher precision of the regression and thereby reduces the size of the standard errors. Such a model would look like the one below. It would be expected to have a lower standard error compared to model 2 for the reason listed above.

```
model_4 <- lm(
  voteshare_t ~
    incumbency +
    candidate_quality +
    spending_t_minus_1 +
    voteshare_t_minus_1 +
    incumbency +
    econ_condition_t,
  data = simulated_data
)
```