Homework 3

Yuki Joyama

```r
# libraries
library(tidyverse)
library(ggplot2)

# setup plot theme
theme_set(
  theme_bw() +
    theme(legend.position = "top")
  )

###CODE FOR HW3#####
set.seed(124)
n <- 16
p_C <- 1/5
C <- rbinom(n,1,p_C)
theta0 <- 1/2
theta1 <- -1/5
p_A <- theta0+theta1*C
A <- rbinom(n,1,p_A)
beta0 <- 110
beta1 <- 20
beta2 <- 5
sigma_Y <- 1
mu_Y <- beta0+beta1*C+beta2*A
Y <- rnorm(n,mu_Y, sigma_Y)
```

1. Interpret parameters

C: Obesity ($C = 1$: obese, $C = 0$: not obese)
A: Exposure to light ($A = 1$: bright light, $A = 0$: dark light)
Y: Glucose outcome such that $Y \sim N(\mu_Y, \sigma)$, where $\mu_Y = f(\beta_0 + \beta_1 * obesity + \beta_2 * light)$, $\beta_0 = 110, \beta_1 = 20$, and $\beta_2 = -5$

p: The probability of a mouse to be obese at baseline. Theoretically, 3.2 out of 16 mice are obese. In the simulation by the above R code, 3 out of 16 mice are set to be obese at baseline.
θ_0: The probability for a non-obese mouse ($C = 0$) to be exposed to light. $\theta_0 = 1/2$ means that there's a 50% chance for a non-obese mouse to be exposed to light.
θ_1: Describes how more (or less) likely it is for an obese mouse ($C = 1$) to be exposed to light. $\theta_1 = -1/5$ indicates that obese mice are 20% less likely to be exposed to the light. In other words, there's a 30% chance for an obese mouse to be exposed to light.
β_0: The baseline mean glucose level for non-obese mice that are not exposed to light, which was set to be 110 mg/dL.
β_1: The coefficient of obesity on glucose level. $\beta_1 = 20$ indicates that obese mice have 20 mg/dL higher

average glucose level from the baseline compared to non-obese mice, holding other variables constant.
β_2: The coefficient of light on glucose level. $\beta_2 = -5$ suggests that mice with exposure to light have 5 mg/dL lower average glucose level from the baseline compared to non-exposed mice, holding other variables constant.

2. PACE

Let Y_1 be the glucose outcome when $A = 1$, and Y_0 be the glucose outcome when $A = 0$.
Marginal PACE: $E[Y_1] - E[Y_0]$
This holds under consistency, SUTVA, exchangeability, and positivity assumption.
Conditional PACE: $E[Y_1|C = c] - E[Y_0|C = c]$
This holds under consistency, SUTVA, exchangeability, positivity, and NUCA ($Y_a \perp A|C$) assumption.

3. g-formula (randomized vs. observational study)

g-formula for the randomized study: $\sum_c E[Y|A = 1, C = c]Pr(C = c) - \sum_c E[Y|A = 0, C = c]Pr(C = c) = E[Y|A = 1] - E[Y|A = 0]$
g-formula for the observational study: $\sum_c E[Y|A = 1, C = c]Pr(C = c) - \sum_c E[Y|A = 0, C = c]Pr(C = c)$ under NUCA.
Randomization enforces unconfoundedness. The distribution of C does not depend on A, so g-formula simplifies to $E[Y|A = 1] - E[Y|A = 0]$. In an observational study, A is influenced by C, so the formula cannot be simplified without accounting for the confounders.

4. Estimate and confidence interval of $E[Y|A = 1] - E[Y|A = 0]$

I will use the following code to calculate the estimate of $E[Y|A = 1] - E[Y|A = 0]$:

```
# split Y into Y_1 and Y_0
Y_1 = Y[A==1]
Y_0 = Y[A==0]

# calculate E[Y|A=a]
eY1 = mean(Y[A==1])
eY0 = mean(Y[A==0])

# calculate the estimate of causal effect
diff = eY1 - eY0
```

$E[Y|A = 1] - E[Y|A = 0]$ = -2.57

95% confidence interval can be calculated by $E[Y|A = 1] - E[Y|A = 0] \pm z\sqrt{\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}}$, where z is the critical value from the standard normal distribution, S_1^2 and S_0^2 are the variances of the potential outcomes in $A = 1$ and $A = 0$ groups, and N_1 and N_0 are the sizes of $A = 1$ and $A = 0$ groups.

```
# critical value
z = qnorm(1 - 0.05 / 2)

# size
n_0 = 8
```

```
n_1 = 8

# variance
var_0 = var(Y_0)
var_1 = var(Y_1)

# standard error
se = sqrt(var_0/n_0 + var_1/n_1)

# 95% CI
ci_lower = diff - z*se
ci_upper = diff + z*se
```

The 95% confidence interval is (-9.68, 4.55).
The point estimate suggests that on average, the outcome Y is 2.57 mg/dL lower in bright light group compared to dark light group in the randomized study from previous homework.
The 95% confidence interval includes zero, so we cannot rule out the possibility that there is no causal effect between the bright and dark light groups (the result is not statistically significant with 95% confidence).

5. Estimate and confidence interval of $E[Y_1] - E[Y_0]$

$E[Y_1] - E[Y_0] = \sum_c E[Y|A = 1, C = c]Pr(C = c) - \sum_c E[Y|A = 0, C = c]Pr(C = c)$ given g-formula from (3). I will calculate the estimate by the following code:

```
# probability of C
p_C0 = mean(C)
p_C1 = 1 - mean(C)

# estimates of E[Y | A = a, C = c]
# if null, set to 0
mY_0_0 = mean(Y[A == 0 & C == 0])
mY_0_1 = mean(Y[A == 0 & C == 1])
mY_1_0 = mean(Y[A == 1 & C == 0])
mY_1_1 = mean(Y[A == 1 & C == 1])
```

Here we notice that no obese mouse was assigned to bright light group. This violates the probabilistic assumption, so I will remove obese mouse group and only consider the non-obese mice.

```
# g-formula estimates
eY1 = p_C0 * mY_1_0
eY0 = p_C0 * mY_0_0

# difference E[Y1] - E[Y0]
diff = eY1 - eY0
```

$E[Y_1] - E[Y_0] = 0.91$

95% condifence interval can be obtained using the following code:

```
# critical value
z = qnorm(1 - 0.05 / 2)
```

```r
# size
n1_0 = sum(A == 1 & C == 0) # A=1, C=0
n0_0 = sum(A == 0 & C == 0) # A=0, C=0

# variance
var1_0 = var(Y[A == 1 & C == 0])
var0_0 = var(Y[A == 0 & C == 0])

# standard error
se = sqrt(var1_0/n1_0 + var0_0/n0_0)

# 95% CI
ci_lower = diff - z*se
ci_upper = diff + z*se
```

The 95% confidence interval is (-0.27, 2.08).
The point estimate suggests that the outcome Y is 0.91 mg/dL higher on average in the bright light group compared to the dark light group among non-obese mice in this observational study. We cannot draw inferences about the causal effect in obese mice due to the violation of probabilistic assumptions. Since the 95% confidence interval includes zero, the point estimate is not statistically significant at the 95% confidence level.
Point estimates from (4) and (5) are different because (5) reflects the effect of adjustment for confounding bias.

6. Assumptions of estimate $E[Y_1]-E[Y_0]$ using linear regression with continuous covariates

In addition to consistency, SUTVA, exchangeability, positivity, and NUCA assumption, we need to assume the following:
- No measurement error: All variables need to be correctly measured. - No model misspecification: All models need to be correctly specified. In the linear regression, the relationship between Y, A, and covariates should be adequately approximated by a linear model.