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 ($p = 1/5$ of the mice are obese at baseline)
A: Exposure to light
Y: Glucose outcome such that $Y \sim N(\mu_Y, \sigma)$, where $E(\mu_Y) = f(\beta_0 + \beta_1 * obesity + \beta_2 * light)$, $\beta_0 = 110, \beta_1 = 20,$ 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: $E[Ya] = \sum_c E[Y|A = a, C = c]Pr(C = c) = E[Y|A = a]$
g-formula for the observational study: $E[Ya] = \sum_c E[Y|A = a, C = c]Pr(C = c)$ under NUCA.
Randomization enforces unconfoundedness. The distribution of C does not depend on A, so g-formula simplifies to $E[Y|A = a]$. 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 $E[Y|A = 1] - E[Y|A = 0]$

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

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