Homework 5

Yuki Joyama

```
# libraries
library(tidyverse)
library(ggplot2)
library(dagitty)
library(patchwork)
library(ipw)
library(knitr)
library(tableone)
library(boot)
library(survey)
library(personalized)

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

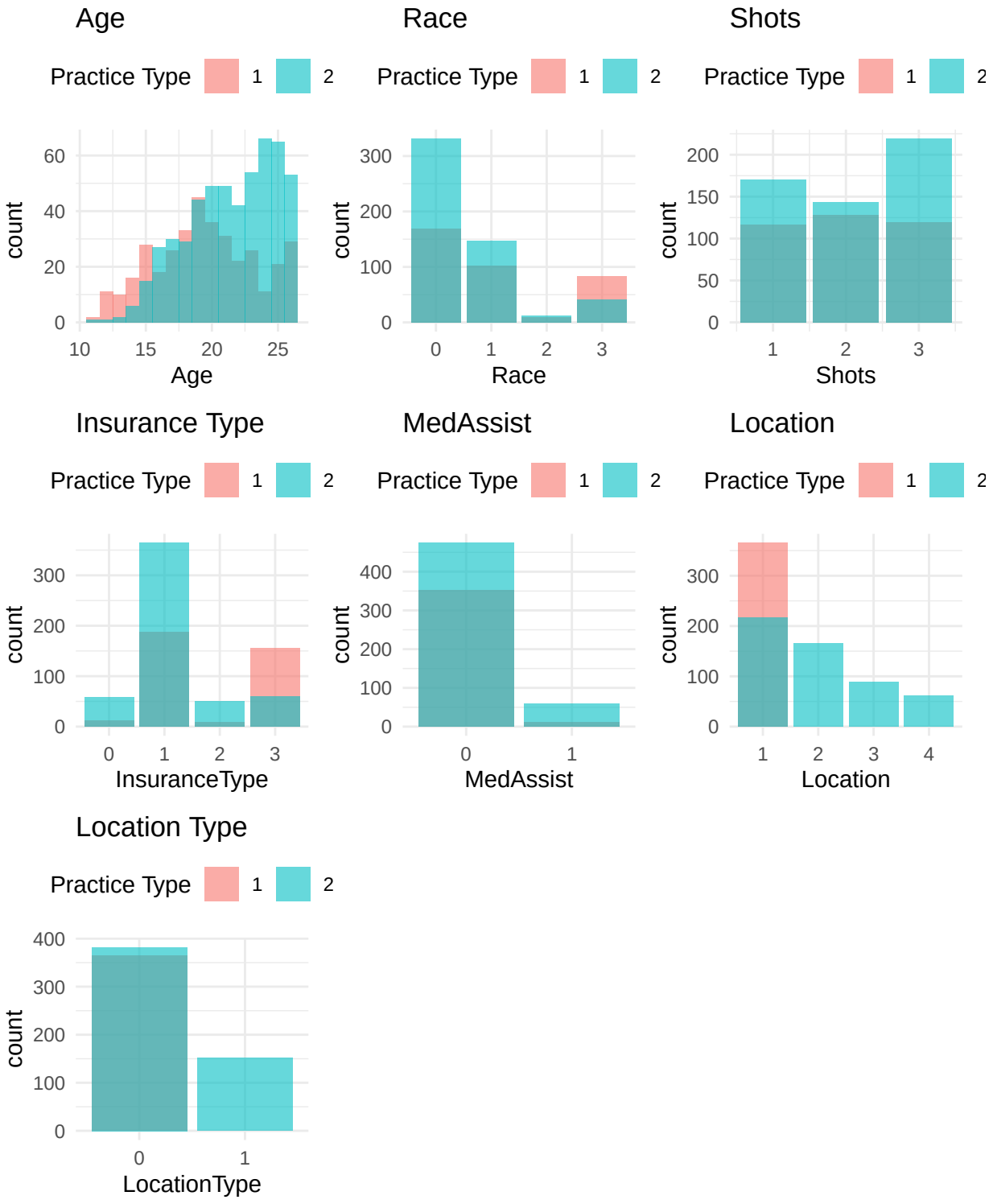
VARIABLE DESCRIPTIONS:

- `Age` = the patient's age in years
- `AgeGroup` = the age group in which the patient falls (0 = 11-17 years, 1 = 18-26 years)
- `Race` = the patient's race (0 = white, 1 = black, 2 = Hispanic, 3 = other/unknown)
- `Shots` = the number of shots that the patients completed during a period of 12 months from the time of the first shot
- `Completed` = did the patient complete the three-shot regimen within the recommended period of 12 months (0 = no, 1 = yes)
- `InsuranceType` = the type of insurance that the patient had (0 = medical assistance, 1 = private payer [Blue Cross Blue Shield, Aetna, Cigna, United, Commercial, CareFirst], 2 = hospital based [EHF], 3 = military [USFHP, Tricare, MA])
- `MedAssist` = did the patient have some type of medical assistance (0 = no, 1 = yes)
- `Location` = the clinic that the patient attended (1 = Odenton, 2 = White Marsh, 3 = Johns Hopkins Outpatient Center, 4 = Bayview)
- `LocationType` = was the clinic in a suburban or an urban location (0 = suburban, 1 = urban)

- `PracticeType` = the type of practice that the patient visited (0 = pediatric, 1 = family practice, 2 = OB-GYN)

```
# import data
df = read.delim("./data/gardasil.txt") |>
  select(-c(X, X.1))

df1 = df |> filter(PracticeType %in% c(1, 2)) |>
  mutate(
    Race = factor(Race),
    Completed = factor(Completed),
    InsuranceType = factor(InsuranceType),
    MedAssist = factor(MedAssist),
    Location = factor(Location),
    LocationType = factor(LocationType),
    PracticeType = factor(PracticeType)
  )

# EDA
plot_age <- ggplot(df1, aes(x = Age, fill = PracticeType)) +
  geom_histogram(position = "identity", alpha = 0.6, binwidth = 1) +
  labs(fill = "Practice Type", title = "Age")

plot_race <- ggplot(df1, aes(x = Race, fill = PracticeType)) +
  geom_bar(position = "identity", alpha = 0.6) +
  labs(fill = "Practice Type", title = "Race")

plot_shots <- ggplot(df1, aes(x = Shots, fill = PracticeType)) +
  geom_bar(position = "identity", alpha = 0.6) +
  labs(fill = "Practice Type", title = "Shots")

plot_insurance <- ggplot(df1, aes(x = InsuranceType, fill = PracticeType)) +
  geom_bar(position = "identity", alpha = 0.6) +
  labs(fill = "Practice Type", title = "Insurance Type")

plot_medassist <- ggplot(df1, aes(x = MedAssist, fill = PracticeType)) +
  geom_bar(position = "identity", alpha = 0.6) +
  labs(fill = "Practice Type", title = "MedAssist")

plot_location <- ggplot(df1, aes(x = Location, fill = PracticeType)) +
  geom_bar(position = "identity", alpha = 0.6) +
  labs(fill = "Practice Type", title = "Location")

plot_locationtype <- ggplot(df1, aes(x = LocationType, fill = PracticeType)) +
  geom_bar(position = "identity", alpha = 0.6) +
  labs(fill = "Practice Type", title = "Location Type")

# Combine plots using patchwork
combined_plot <- (
  plot_age + plot_race + plot_shots +
  plot_insurance + plot_medassist +
  plot_location + plot_locationtype
) +
  plot_layout(ncol = 3)
```

```
plot(combined_plot)
```



1 Protocol of the RCT

In this homework, I will investigate whether receiving Gardasil vaccine in an OB-GYN clinic improves the rates of completion compared to a family practice.

(i) Treatment arm: Participants receiving the vaccine series at an OB-GYN practice
Control arm: Participants receiving the vaccine series at a family practice

(ii) Eligibility (inclusion criteria)

- Girls aged 15-25 years including all race
- Those who have insurance of private payer [Blue Cross Blue Shield, Aetna, Cigna, United, Commercial, CareFirst] or military [USFHP, Tricare, MA]
- Does not need medical assistance
- Those who visited clinic in Odenton, in a suburban location

This eligibility is based on the plots above to ensure the probabilistic assumption between the two practice types (attempting to exclude variables that have no or little overlap between the two groups).

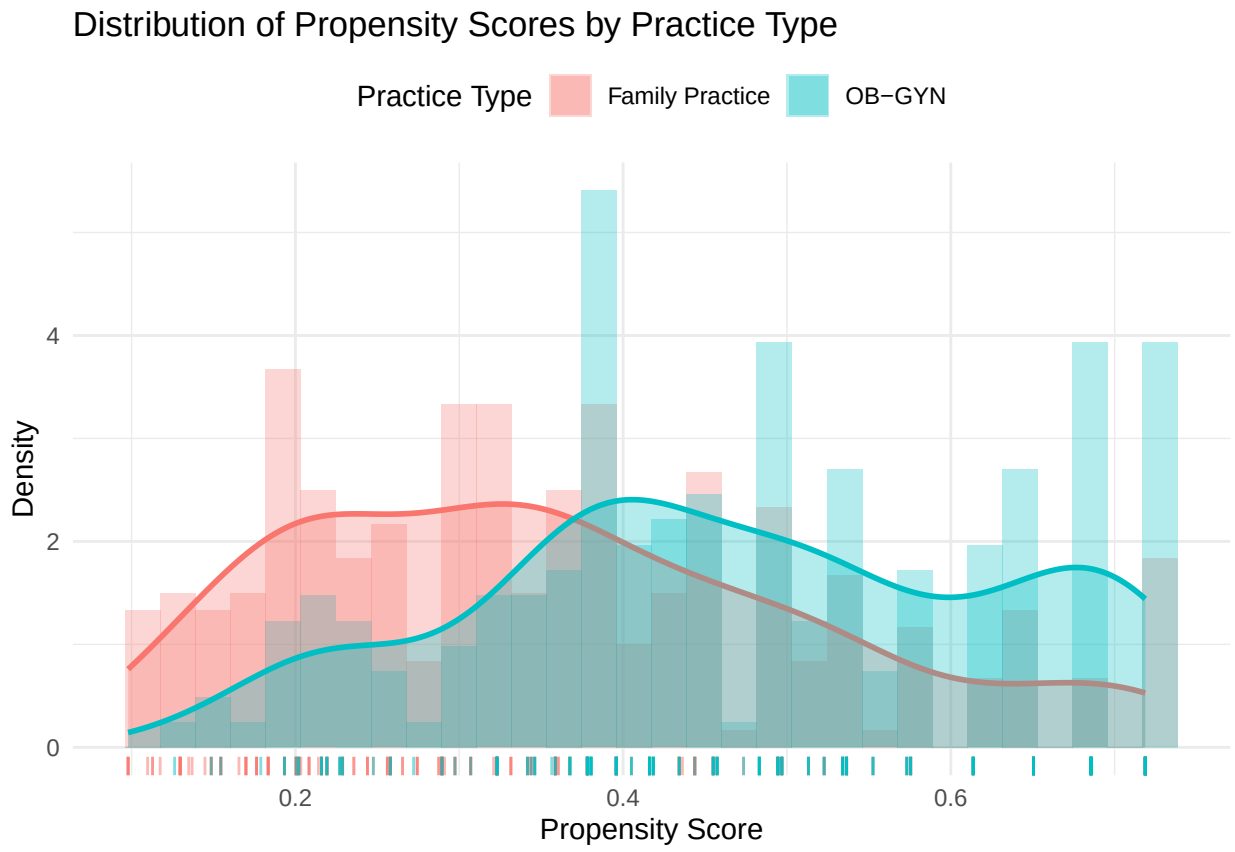
```
# filter by the above criteria
df1 <- df1 |>
  filter(
    Age >= 15 & Age <= 25,
    InsuranceType %in% c(1, 3),
    MedAssist == 0,
    Location == 1,
    LocationType == 0
  )
```

2 Propensity score estimation

I will use generalized logistic regression to estimate the propensity scores. The covariates included in the model are `Age`, `Race`, and `InsuranceType`.

```
ps_est <- glm(PracticeType ~ Age + Race + InsuranceType, df1, family = binomial(logit))

# propensity scores
df1$ps <- predict(ps_est, type = "response")

# visualize propensity scores
ggplot(df1, aes(x = ps, fill = as.factor(PracticeType))) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, alpha = 0.3, position = "identity") +  # Hist
  geom_density(aes(color = as.factor(PracticeType)), alpha = 0.3, size = 1) +  # Density plot
  scale_fill_manual(
    values = c("#F8766D", "#00BFC4"),
    name = "Practice Type",
    labels = c("Family Practice", "OB-GYN"),
    guide = guide_legend(override.aes = list(color = NA))
  ) +
  scale_color_manual(
    values = c("#F8766D", "#00BFC4"),
```

```
    guide = "none"
  ) +
  geom_rug(aes(color = as.factor(PracticeType)), sides = "b", alpha = 0.5) +  # Rug plot
  labs(
    title = "Distribution of Propensity Scores by Practice Type",
    x = "Propensity Score",
    y = "Density"
  )
```



Propensity scores largely overlap between the two groups except at the extreme ends. Probabilistic assumptions are mostly satisfied.

```
# Covariate balance
vars <- c("Age", "Race", "InsuranceType")
table1 <- CreateTableOne(vars = vars, strata = "PracticeType", data = df1, test = FALSE)

# SMD before adjustment
print(table1, smd = TRUE)
```

```
##                     Stratified by PracticeType
##                      1             2             SMD
##   n                    280           190
##   Age (mean (SD))    19.76 (2.86)  21.33 (2.87)   0.549
##   Race (%)                                        0.363
##      0                 127 (45.4)    118 (62.1)
```

```
##      1                 78 (27.9)     43 (22.6)
##      2                 10 ( 3.6)      3 ( 1.6)
##      3                 65 (23.2)     26 (13.7)
##   InsuranceType (%)                              0.425
##      0                  0 ( 0.0)      0 ( 0.0)
##      1                147 (52.5)    138 (72.6)
##      2                  0 ( 0.0)      0 ( 0.0)
##      3                133 (47.5)     52 (27.4)
```

We can see that the covariates are significantly imbalanced between the two groups.

```
ps_est |> broom::tidy()
```

```
## # A tibble: 6 x 5
##   term           estimate std.error statistic   p.value
##   <chr>             <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      -3.02     0.779      -3.88 0.000105
## 2 Age               0.158    0.0360      4.41 0.0000105
## 3 Race1            -0.643    0.244      -2.63 0.00845
## 4 Race2            -1.19     0.692      -1.73 0.0841
## 5 Race3            -0.847    0.278      -3.05 0.00228
## 6 InsuranceType3   -0.727    0.220      -3.31 0.000934
```

Model interpretation:
The logistic regression model indicates that age significantly increases the likelihood of selecting OB-GYN practices over family practices, with each additional year of age associated with higher odds. Race also plays a role: Black (Race1) and Other/Unknown (Race3) racial groups are significantly less likely to visit OB-GYN practices compared to White patients (Race0). Hispanic patients (Race2) show a similar trend, but the effect is not statistically significant ($p = 0.084$). Additionally, patients with military insurance (InsuranceType3) are significantly less likely to visit OB-GYN practices compared to those with private insurance. These findings highlight age, race, and insurance type as important factors influencing the likelihood of selecting OB-GYN practices over family practices.

3 Outcome regression model for g-formula approach for confounding adjustment

Below is the outcome regression model:

```
outcome_model <- glm(Completed ~ PracticeType + Age + Race + InsuranceType,
                     data = df1,
                     family = binomial(logit))

outcome_model |> broom::tidy()
```

```
## # A tibble: 7 x 5
##   term           estimate std.error statistic p.value
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)      0.727     0.780     0.932   0.351
## 2 PracticeType2    0.114     0.223     0.512   0.609
## 3 Age             -0.0728    0.0369   -1.97    0.0484
```

```
## 4 Race1           -0.299     0.255    -1.17    0.243
## 5 Race2           -0.0174    0.623    -0.0280  0.978
## 6 Race3           -0.579     0.297    -1.95    0.0510
## 7 InsuranceType3  -0.122     0.226    -0.540   0.589
```

Among all the variables in the dataset (`Age, AgeGroup, Race, Shots, Completed, InsuranceType, MedAssist, Location, LocationType, PracticeType`), I chose `Completed` as an outcome, `PracticeType` as an exposure, and the other three as covariates because they are thought to affect both exposure and outcome. I did not include other variables because `AgeGroup` is highly correlated with `Age`, and I though `Shots` is a mediator rather than a confounder.

Model interpretation:
The model indicates that age is the only significant predictor of vaccine regimen completion, with older patients being less likely to complete the regimen ($p = 0.048$). Other factors, including practice type (OB-GYN vs. family practice), race, and insurance type, did not show statistically significant effects, though the "Other/Unknown" race category (Race3) had a marginally significant negative association ($p = 0.051$). These results suggest that age is a key factor influencing vaccine completion, while the effects of other variables may require further investigation.

4 Marginal average causal effect

```r
# Make copies of original data
df.a2 <- df1 |>
  mutate(Completed  = NA,
  PracticeType = factor(2)) # Assign PracticeType = 2 to everyone

df.a1 <- df1 |>
  mutate(Completed  = NA,
  PracticeType = factor(1)) # Assign PracticeType = 1 to everyone

df.combined <- bind_rows(df.a2, df.a1)

# fit an outcome model to data and predict outcome values
gcomp.fit <-
  glm(Completed ~ PracticeType + Age + Race + InsuranceType,
                    data = df1,
                    family = binomial(logit))

df.combined$pred <- predict(gcomp.fit, newdata = df.combined, type = "response")

# calculate point estimate
df.combined |>
  group_by(PracticeType) |>
  summarise(
    mean.Y = mean(pred)
  ) |>
  pivot_wider(
    names_from = PracticeType,
    names_glue = "mean.Y.{PracticeType}", values_from = mean.Y
  ) |>
  mutate(
    ACE = mean.Y.2-mean.Y.1
```

```
  ) |>
  kable()
```

mean.Y.2	mean.Y.1	ACE
0.2966808	0.2737692	0.0229116

```
# CI via bootstrapping
std.boot <- function(data, indices) {
  df <- data[indices,]
  df.a2 <- df |>
    mutate(
      Completed = NA,
      PracticeType= factor(2)
    )

  df.a1 <- df |>
    mutate(
      Completed = NA,
      PracticeType = factor(1)
    )

  df.combined <- bind_rows(df.a2, df.a1)

  gcomp.fit <- glm(Completed ~ PracticeType + Age + Race + InsuranceType,
                   data = df, family = binomial(link = "logit"))

  df.combined$pred <- predict(gcomp.fit, newdata = df.combined, type = "response")

  output <- df.combined |>
    group_by(PracticeType) |>
    summarise(
      mean.Y = mean(pred), .groups = "drop"
    ) |>
    pivot_wider(
      names_from = PracticeType,
      names_glue = "mean.Y.{PracticeType}",
      values_from = mean.Y
    ) |>
    mutate(
      ACE = mean.Y.2 - mean.Y.1
    )

  return(output$ACE)
}

# bootstrap
set.seed(2024)
results <- boot(data = df1, statistic = std.boot, R = 100) # 100 bootstrapped samples

# generating confidence intervals
empirical.se <- sd(results$t) # get empirical standard error estimate
estimate <- results$t0
```

```r
lower_ci <- estimate - qnorm(0.975)*empirical.se # normal approximation
upper_ci <- estimate + qnorm(0.975)*empirical.se
data.frame(cbind(estimate, empirical.se, lower_ci, upper_ci)) |> kable()
```

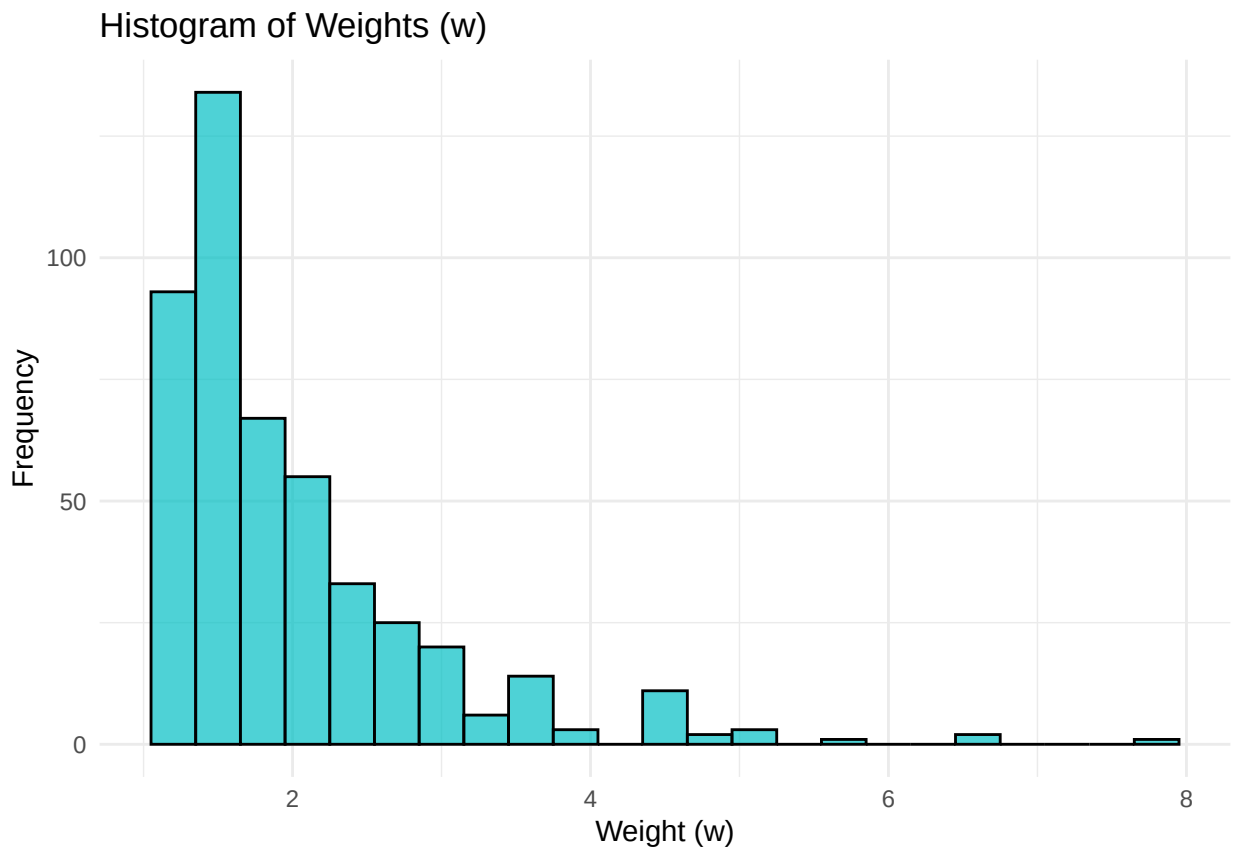
estimate	empirical.se	lower_ci	upper_ci
0.0229116	0.0455962	-0.0664552	0.1122784

Patients visiting an OB-GYN had an estimated 0.023 increase in the probability of completing the outcome compared to those visiting a family practice. However, the 95% confidence interval [-0.066, 0.112] includes zero, indicating the effect is not statistically significant at the 95% confidence level.

5

Below is the codes to calculate the inverse probability of being in the treatment arm

```r
# Estimate propensity scores
ps_est <- glm(PracticeType ~ Age + Race + InsuranceType, data = df1, family = binomial(link = "logit"))

# Calculate propensity scores
df1$ps <- predict(ps_est, type = "response")

# Calculate ipw
df1$w <- ifelse(df1$PracticeType == 2, 1 / df1$ps, 1 / (1 - df1$ps))

# plot
ggplot(df1, aes(x = w)) +
  geom_histogram(binwidth = 0.3, fill = "#00BFC4", color = "black", alpha = 0.7) +
  labs(
    title = "Histogram of Weights (w)",
    x = "Weight (w)",
    y = "Frequency"
  )
```



6

Bootstrap to estimate the MACE of receiving an HPV vaccine at OB-GYN facilities compared to other facilities on the rates of vaccination regimen completion using a MSM:

```
# Bootstrapping parameters
boots <- 100
b.holder <- rep(NA, boots)

n <- nrow(df1)

set.seed(2024)
for (i in 1:boots) {
  # Resample data without replacement
  S.b <- sample(1:n, size = n, replace = TRUE)
  data.b <- df1[S.b, ]

  # Propensity score model
  pprobs <- predict(
    glm(PracticeType ~ Age + Race + InsuranceType,
        data = data.b,
        family = binomial(link = "logit")),
    type = "response"
  )
```

```r
  # Inverse probability weights
  est.w <- ifelse(data.b$PracticeType == 1, 1 / pprobs, 1 / (1 - pprobs))

  # MSM without covariate adjustment (marginal effect)
  psw <- svyglm(Completed ~ PracticeType,
                design = svydesign(~1, weights = ~est.w, data = data.b),
                family = quasibinomial(link = "logit"))
  b.holder[i] <- psw$coefficients[2]
}

# Results
# Marginal effect
mean_marginal <- mean(as.vector(b.holder))

# Confidence intervals
ci_marginal <- quantile(b.holder, probs = c(0.025, 0.975))

# Display results
list(
  Marginal_Effect = list(Estimate = mean_marginal, CI = ci_marginal)
)
```

```
## $Marginal_Effect
## $Marginal_Effect$Estimate
## [1] 0.1248931
##
## $Marginal_Effect$CI
##       2.5%      97.5%
## -0.3040127  0.5690486
```

7

The mean marginal effect estimate (0.1249) suggests that attending an OB-GYN practice (PracticeType = 2) compared to a family practice (PracticeType = 1) may have a modest positive impact on the likelihood of completing the HPV vaccine regimen (Completed = 1). However, the 95% CI [-0.304, 0.569] includes 0, indicating that this effect is not statistically significant. This means we cannot confidently conclude that attending an OB-GYN practice has a true positive or negative impact on vaccine completion rates compared to a family practice. Limitations include potential instability of weights, as extreme weights can inflate variance and reduce precision, and possible violations of the positivity assumption. Addressing these limitations, such as trimming extreme weights or ensuring adequate overlap, could improve the reliability of the findings.

8

The inferences from the MSM and the g-formula show similar trends but differ slightly in their interpretation and methodological approaches. Both methods suggest that attending an OB-GYN practice (compared to a family practice) may positively impact HPV vaccine regimen completion rates. However, neither method yields statistically significant results, as their confidence intervals include zero.