Final

Yuki Joyama

```r
# libraries
library(tidyverse)
library(ggplot2)
library(readr)
library(dagitty)
library(tableone)
library(knitr)
library(boot)
library(survey)

# install.packages("remotes")
# remotes::install_github("LindaValeri/CMAverse")
library(CMAverse)

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

```r
# import data
df = read.table("./data/datafinal.txt", header = T) |>
  dplyr::mutate(
    gender = as.factor(gender),
    plural = as.factor(plural),
    race = as.factor(race),
    parity = as.factor(parity),
    married = as.factor(married),
    firstep = as.factor(firstep),
    welfare = as.factor(welfare),
    smoker = as.factor(smoker),
    drinker = as.factor(drinker)
  )
```

1

a

The estimand of interest is the causal effect of smoking during pregnancy on gestational age. In other words, it is an average difference in gestational age between pregnant individuals who smoke and those who do not smoke.

Assumptions:

- Consistency: The observed gestational age for each mother corresponds to the actual smoking status during pregnancy (i.e., the recorded smoking status is correct and there is no error in measurement)

- No unmeasured confounding: All confounding factors affecting both smoking and gestational age are measured and accounted for

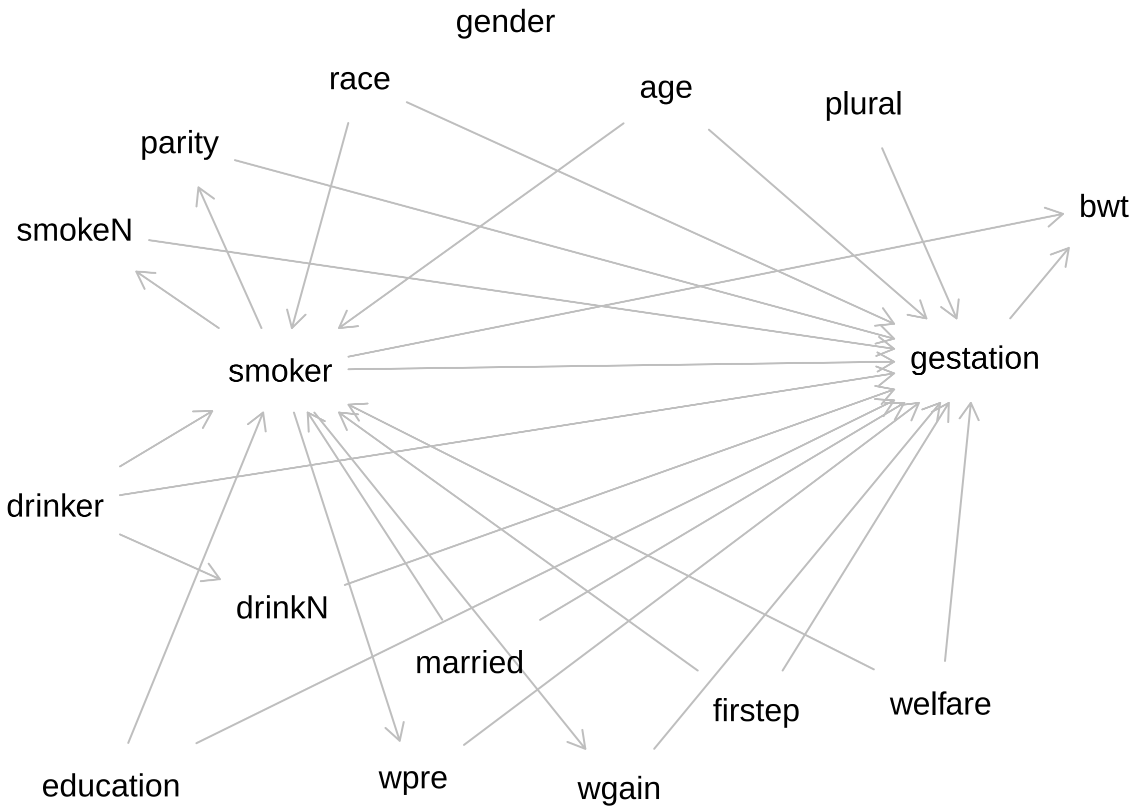
- Positivity: All individuals in the data have a positive probability of belonging to either the smoking or nonsmoking group, given the measured covariates

- SUTVA: Exposure (smoking status) is well defined and there is only one version of potential; one mother's smoking status does not affect another mother's gestational age

To evaluate these assumptions, I will first use domain knowledge to identify the confounders and adjust them in the statistical analysis. I will also check the distribution of covariates by smoking status to ensure that there are enough smokers and nonsmokers at each level of the confounding variable.

b

```
# DAG
g = dagitty('dag {
age [pos="-0.106,-0.624"]
bwt [pos="0.967,-0.413"]
drinkN [pos="-1.047,0.336"]
drinker [pos="-1.604,0.146"]
education [pos="-1.467,0.668"]
firstep [pos="0.115,0.533"]
gender [pos="-0.500,-0.752"]
gestation [outcome,pos="0.651,-0.124"]
married [pos="-0.588,0.438"]
parity [pos="-1.299,-0.526"]
plural [pos="0.378,-0.599"]
race [pos="-0.857,-0.646"]
smokeN [pos="-1.556,-0.369"]
smoker [exposure,pos="-1.052,-0.106"]
welfare [pos="0.567,0.515"]
wgain [pos="-0.221,0.679"]
wpre [pos="-0.726,0.664"]
age -> gestation
age -> smoker
drinkN -> gestation
drinker -> drinkN
drinker -> gestation
drinker -> smoker
education -> gestation
education -> smoker
firstep -> gestation
firstep -> smoker
gestation -> bwt
married -> gestation
married -> smoker
parity -> gestation
plural -> gestation
race -> gestation
race -> smoker
smokeN -> gestation
smoker -> bwt
smoker -> gestation
smoker -> parity
smoker -> smokeN
smoker -> wgain
smoker -> wpre
welfare -> gestation
welfare -> smoker
wgain -> gestation
wpre -> gestation
}')

plot(g)
```



`gender`: Maternal smoking status does not influence the infant's gender, and the infant's gender does not affest gestational age.
`plural`: Maternal smoking status does not influence singleton/twin/triplet pregnancy, and the number of child carriage affect gestational age.
`age`: Potential confounder. Depending on the mother's generation, smoking prevalence varies and maternal age also influence the gestational age.
`race`: Potential confounder. There is a racial difference in smoking prevalence and maternal race also affects gestational age.
`parity`: Potential mediator. Smokers may be more likely to have multiple parity and parity also affects

gestational age.
`married`: Potential confounder. Married people may be more likely to be a non-smoker, and marriage status may impact gestational age.
`bwt`: Potential collider. Both smoking status and gestational age impacts birth weight.
`smokeN`: Potential mediator. Smoking status impacts number of cigarettes smoked per day during pregnancy, and `smokeN` affects gestational age.
`drinkN`: This is in the backdoor path. Drinking state affects number of alcoholic drinks per week during pregnancy, and `drinkN` also affects gestational age.
`firstep`: Potential confounder. The participation influences both smoking status and gestational age.
`welfare`: Potential confounder. The participation influences both smoking status and gestational age.
`smoker`: Exposure.
`drinker`: Potential confounder. Drinkers are more likely to smoke and drinking status affects gestational age.
`wpre`: Potential mediator. Smoking status affects mother's weight in pounds prior to pregnancy, and `wpre` also affects gestational age.
`wgain`: Potential mediator. Smoking status affects mother's weight gain in pounds during pregnancy, and `wgain` also affects gestational age.
`education`: Potential confounder.
`gestation`: Outcome.

In addition, I think potential confounders includes income, residential area, working place, etc.
In order to satisty the first and second principles of causal inference, in this data, we need to block all backdoor paths to eliminate confounding (adjust for `age`, `drinker`, `education`, `firstep`, `married`, `race`, `welfare`), and avoid adjusting for mediators (`smokeN`, `parity`, `drinkN`, `wpre`, `wgain`) and colliders (`bwt`).

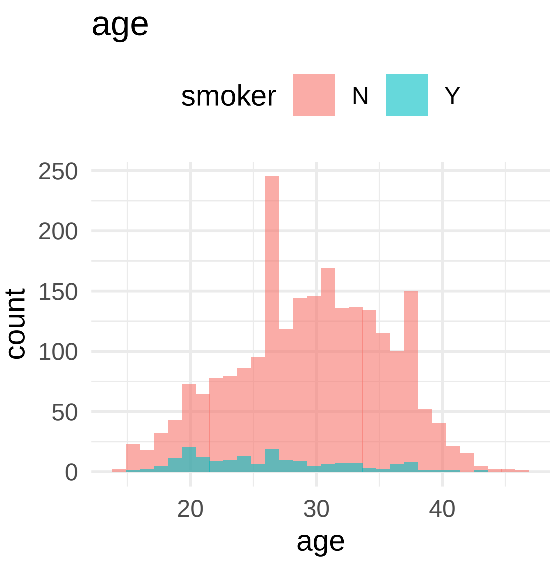
c

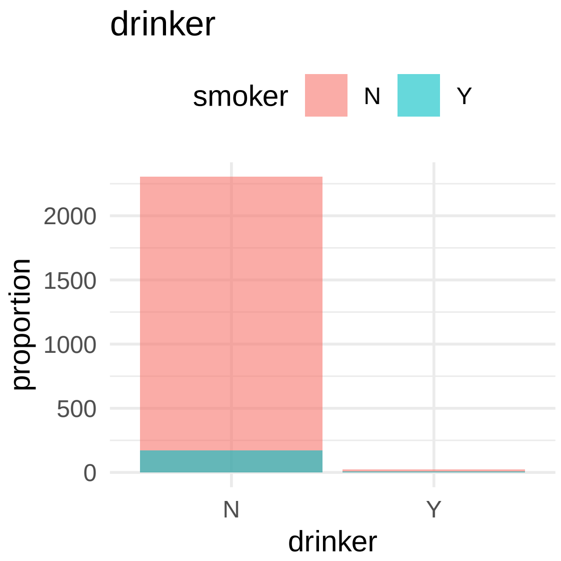
```r
covariates <- c("age", "drinker", "education", "firstep", "married", "race", "welfare")
exposure <- "smoker"
outcome <- "gestation"

# create table to check balance
tb <- CreateTableOne(vars = covariates, strata = exposure, data = df, test = FALSE)
print(tb, smd = TRUE) # Show standardized mean differences
```

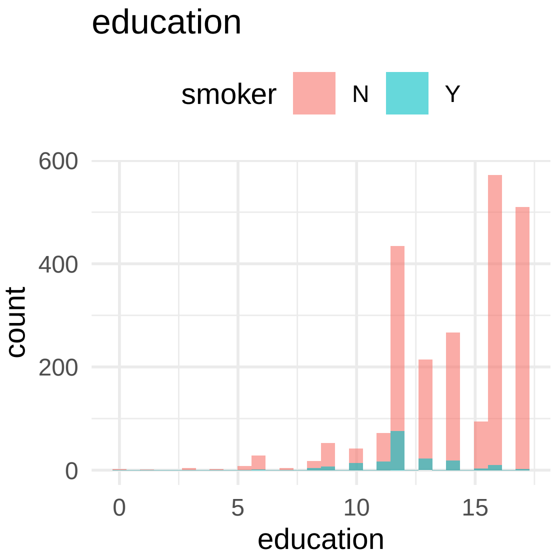
```
##                         Stratified by smoker
##                          N             Y             SMD
##   n                       2325           175
##   age (mean (SD))        29.54 (5.94)  26.16 (6.01)   0.566
##   drinker = Y (%)           23 ( 1.0)      6 ( 3.4)   0.167
##   education (mean (SD)) 14.22 (2.62)  12.18 (1.80)   0.906
##   firstep = 1 (%)          352 (15.1)     51 (29.1)   0.342
##   married = 1 (%)         1897 (81.6)     59 (33.7)   1.108
##   race (%)                                            0.487
##      asian                 384 (16.5)      8 ( 4.6)
##      black                 164 ( 7.1)     14 ( 8.0)
##      hispanic              212 ( 9.1)      8 ( 4.6)
##      other                  24 ( 1.0)      7 ( 4.0)
##      white                1541 (66.3)    138 (78.9)
##   welfare = 1 (%)           28 ( 1.2)     14 ( 8.0)   0.329
```

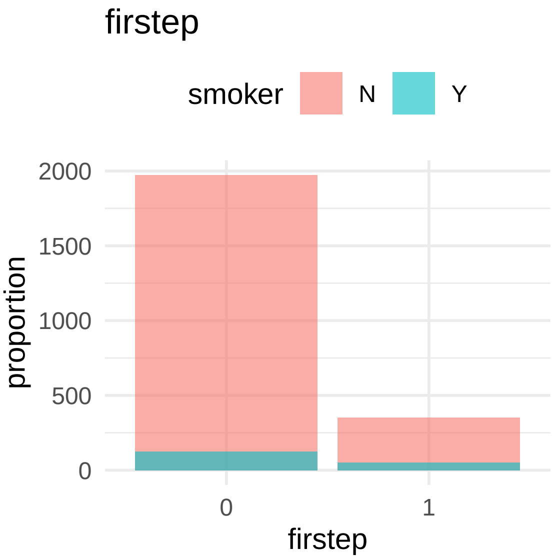
Given SMD >0.1, all the selected covariates are imbalanced between the two smoking groups, indicating that adjustment of these covariates are required in the later analysis to estimate smoking's causal effect on gestational age.

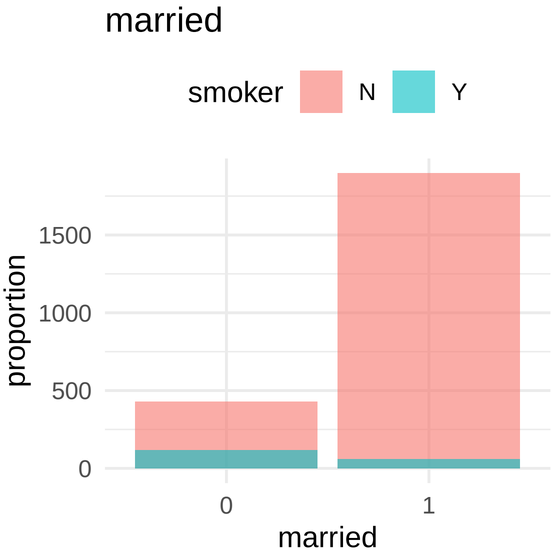
```r
# check overlap in covariates
for (cov in covariates) {
  if (is.numeric(df[[cov]])) {  # for numeric covariates, create histogram
    p <- ggplot(df, aes_string(x = cov, fill = "smoker")) +
      geom_histogram(position = "identity", alpha = 0.6) +
      labs(title = paste(cov),
           x = cov, fill = "smoker")
  } else {  # for categorical covariates, create bar plots
    p <- ggplot(df, aes_string(x = cov, fill = "smoker")) +
      geom_bar(position = "identity", alpha = 0.6) +
      labs(title = paste(cov),
           x = cov, y = "proportion", fill = "smoker")
  }
  print(p)
}
```

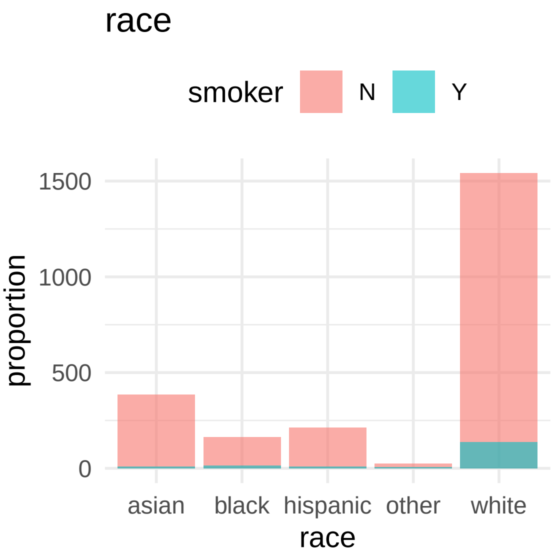


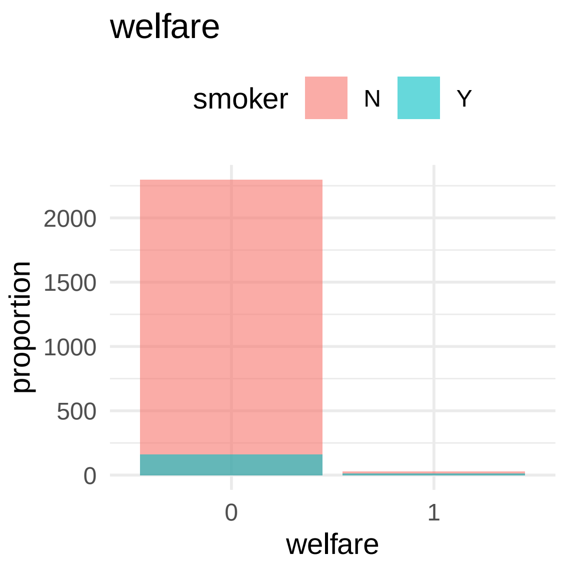



education
smoker
N
Y
count
600
400
200
0
0
5
10
15
education


firstep
smoker
N
Y
proportion
2000
1500
1000
500
0
0
1
firstep


married
smoker
N
Y
proportion
1500
1000
500
0
0
1
married


race
smoker
N
Y
proportion
1500
1000
500
0
asian
black
hispanic
other
white
race



There is sufficient overlap between smokers and non-smokers in `firstep`, `married`, `race`, `welfare` and the positivity assumption appears to be met for those covariates. We see some potential violation in `age` <18 and >38, `education` <8 and >18, and `drinker`.
In addition the SUTVA assumption may be violated. Because of the high participation rates in the "First Steps" program throughout King County, WA, it is possible that those who did not participate in the program were also surrounded by participants with better care knowledge that influenced their preterm birth behavior (i.e., spillover effects).

d

Given (c), I will set the eligibility as below:

- Mother aged 17-37
- Those who had 7-17 years of education

- Does not drink

```r
# filter by the above criteria
df_d = df |>
  filter(
    age >= 17 & age <= 27,
    education >= 7 & education <= 17,
    drinker == "N"
  )
```

Direct regression adjustment:

```r
reg <- lm(gestation ~ smoker + age + education + firstep + married + race + welfare, data = df_d)

summary(reg)
```

```
##
## Call:
```

```
## lm(formula = gestation ~ smoker + age + education + firstep +
##     married + race + welfare, data = df_d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5888  -1.0269   0.1976   1.1981   6.4029
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.143783   0.820121  46.510  < 2e-16 ***
## smokerY      -0.093242   0.266281  -0.350  0.72630
## age          -0.004135   0.033231  -0.124  0.90099
## education     0.046562   0.046028   1.012  0.31201
## firstep1      0.185716   0.186802   0.994  0.32041
## married1      0.341729   0.188654   1.811  0.07042 .
## raceblack     0.386082   0.316731   1.219  0.22319
## racehispanic  0.200507   0.320936   0.625  0.53229
## raceother    -0.675788   0.618673  -1.092  0.27499
## racewhite    -0.002055   0.230258  -0.009  0.99288
## welfare1     -1.392578   0.488612  -2.850  0.00447 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.35 on 879 degrees of freedom
## Multiple R-squared:  0.02153,    Adjusted R-squared:  0.0104
## F-statistic: 1.934 on 10 and 879 DF,  p-value: 0.03757
```

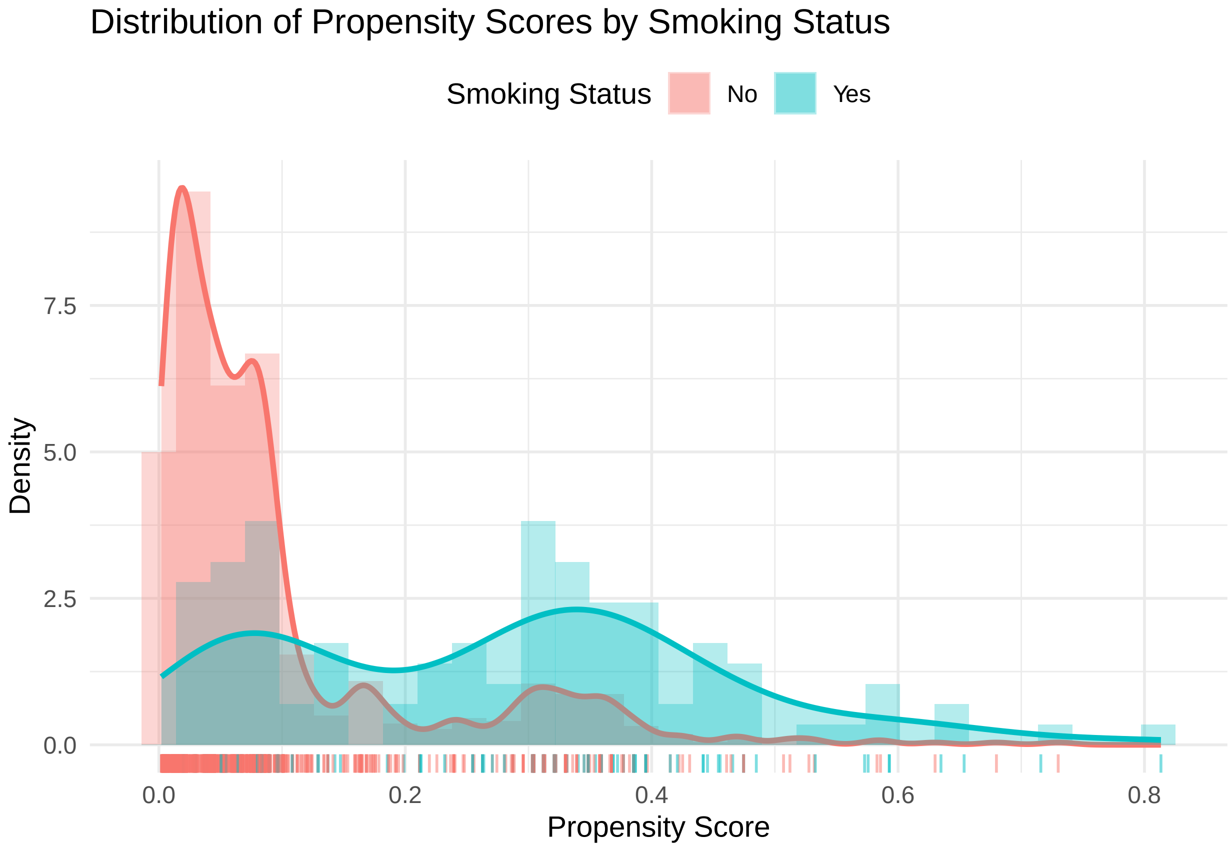
```r
# extract coefficient and 95% CI for smoker
effect <- coef(summary(reg))["smokerY", ]
ci <- confint(reg, "smokerY", level = 0.95)
cat("Estimated ACE:", round(effect[1], 2),
    "\n95% CI:", round(ci[1], 2), "to", round(ci[2], 2), "\n")
```

```
## Estimated ACE: -0.09
## 95% CI: -0.62 to 0.43
```

All else being equal, in the eligible population, the point estimate of ACE suggests that smoking during pregnancy is associated with a reduction of approximately 0.09 weeks in gestational age but 95% CI indicates that this is not statistically significant.

Propensity score weighting approach:
I will fit logisting regression with the covariates (`age`, `education`, `firstep`, `married`, `race`, `welfare`) to estimate the propensity score.

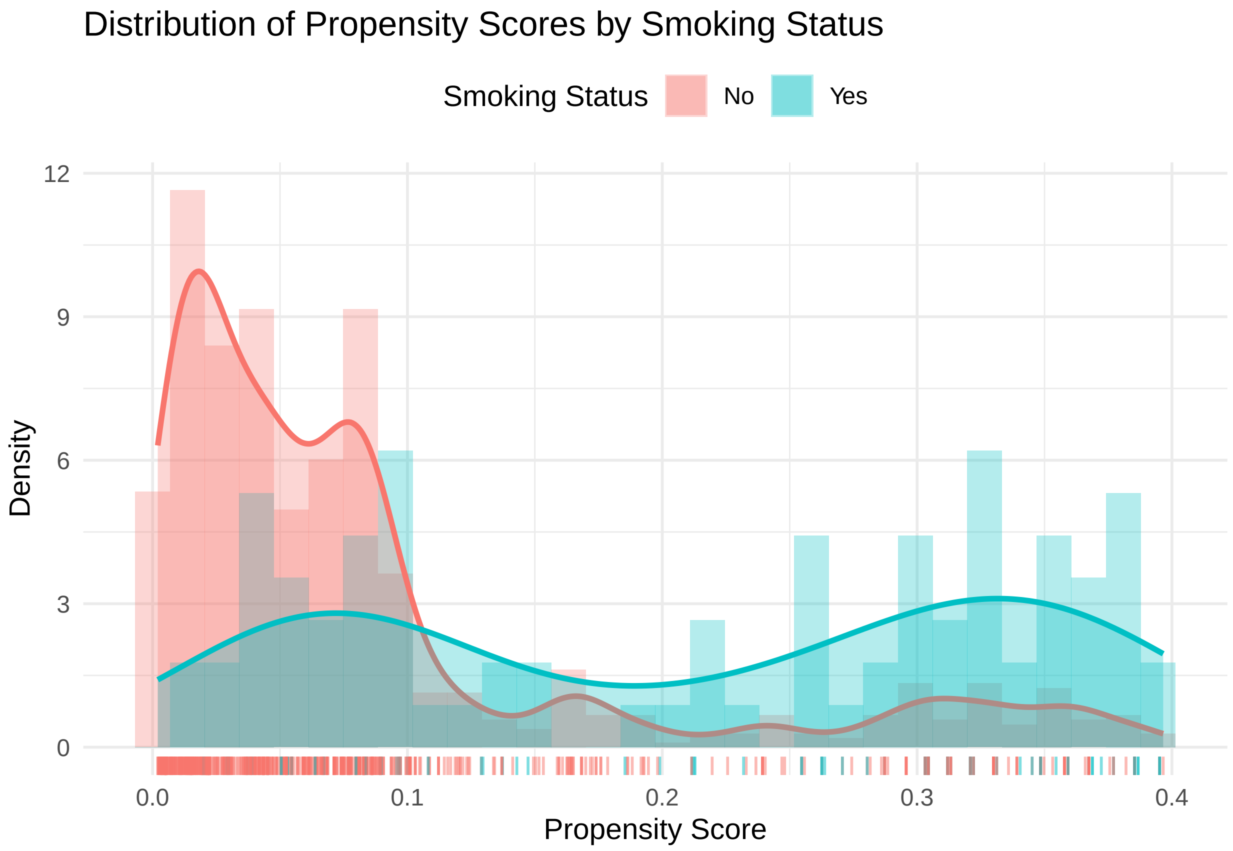
```r
# ps estimation
ps_est <- glm(smoker ~ age + education + firstep + married + race + welfare, data = df_d,  family = bin

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

# visualize propensity scores
ggplot(df_d, aes(x = ps, fill = smoker)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, alpha = 0.3, position = "identity") +  # Hist
  geom_density(aes(color = smoker), alpha = 0.3, size = 1) +  # Density plot
  scale_fill_manual(
    values = c("#F8766D", "#00BFC4"),
    name = "Smoking Status",
    labels = c("No", "Yes"),
    guide = guide_legend(override.aes = list(color = NA))
  ) +
  scale_color_manual(
    values = c("#F8766D", "#00BFC4"),
    guide = "none"
  ) +
  geom_rug(aes(color = smoker), sides = "b", alpha = 0.5) +  # Rug plot
  labs(
    title = "Distribution of Propensity Scores by Smoking Status",
    x = "Propensity Score",
    y = "Density"
  )
```



There is a clear lack of overlap for propensity scores greater than 0.4. Therefore, I will trim the observations with PS >0.4.

```
df_trimmed <- df_d |>
  filter(ps <= 0.4)

# verify the overlap again
ggplot(df_trimmed, aes(x = ps, fill = smoker)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, alpha = 0.3, position = "identity") +  # Hist
```

```r
  geom_density(aes(color = smoker), alpha = 0.3, size = 1) +  # Density plot
  scale_fill_manual(
    values = c("#F8766D", "#00BFC4"),
    name = "Smoking Status",
    labels = c("No", "Yes"),
    guide = guide_legend(override.aes = list(color = NA))
  ) +
  scale_color_manual(
    values = c("#F8766D", "#00BFC4"),
    guide = "none"
  ) +
  geom_rug(aes(color = smoker), sides = "b", alpha = 0.5) +  # Rug plot
  labs(
    title = "Distribution of Propensity Scores by Smoking Status",
    x = "Propensity Score",
    y = "Density"
  )
```



Now, propensity scores largely overlap between the two groups and probabilistic assumption is mostly satisfied.

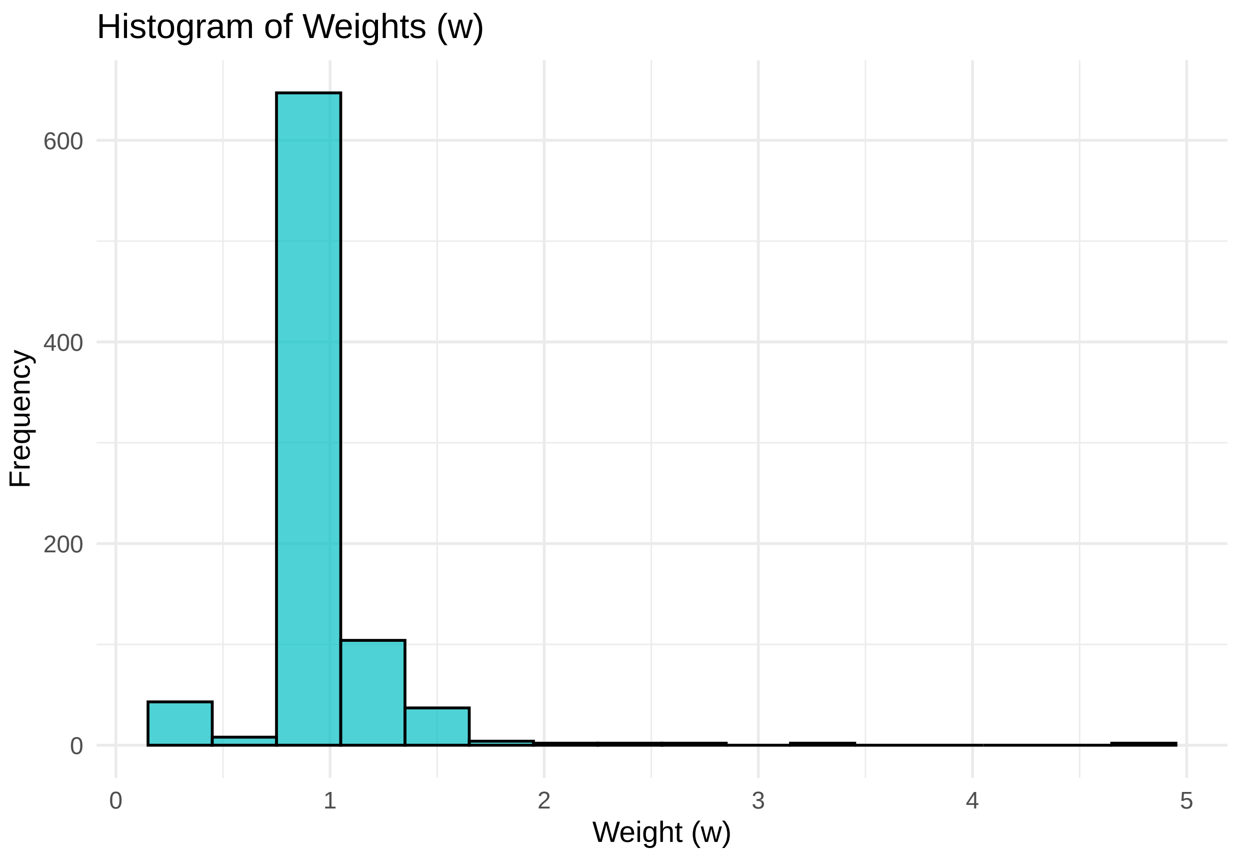
```r
# covariate balance
covariates <- c("age", "education", "firstep", "married", "race", "welfare")
table1 <- CreateTableOne(vars = covariates, strata = "smoker", data = df_trimmed, test = FALSE)
```

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

```
##                         Stratified by smoker
##                          N             Y             SMD
##   n                        770            83
##   age (mean (SD))        23.40 (2.81)  22.33 (2.61)   0.396
##   education (mean (SD)) 13.08 (2.04)  12.07 (1.24)   0.597
##   firstep = 1 (%)          200 (26.0)     27 (32.5)   0.144
##   married = 1 (%)          519 (67.4)     24 (28.9)   0.835
##   race (%)                                            0.514
##      asian                 130 (16.9)      5 ( 6.0)
##      black                  97 (12.6)      6 ( 7.2)
##      hispanic               99 (12.9)      6 ( 7.2)
##      other                   8 ( 1.0)      3 ( 3.6)
##      white                 436 (56.6)     63 (75.9)
##   welfare = 1 (%)           14 ( 1.8)      2 ( 2.4)   0.041
```

We can see that the covariates are significantly imbalanced between the two groups except for `welfare`. Now, I will calculate the stabilized weights ($w_i = \frac{P(T=t)}{P(T=t|X_i)}$, where T is smoker = Y or N):

```r
# calculate the marginal probability of smoking (P(T = Y))
p_treated <- mean(df_trimmed$smoker == "Y")

# calculate stabilized weights
df_trimmed$w_stb <- ifelse(
  df_trimmed$smoker == "Y",
  p_treated / df_trimmed$ps,          # For smokers
  (1 - p_treated) / (1 - df_trimmed$ps)  # For non-smokers
)

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



Finally, I will use bootstrap to estimate the marginal ACE of smoking status on gestational age.

```r
# bootstrapping parameters
boots <- 100
b.holder <- rep(NA, boots)
n <- nrow(df_trimmed)
set.seed(2024)

# marginal probability of smoking (stabilized weights)
p_treated <- mean(df_trimmed$smoker == "Y")

for (i in 1:boots) {
  # resample data with replacement
  S.b <- sample(1:n, size = n, replace = TRUE)
  data.b <- df_trimmed[S.b, ]

  # fit the propensity score model on the bootstrap sample
  ps_model <- glm(smoker ~ age + education + firstep + married + race + welfare,
                  data = data.b, family = binomial(logit))

  # predict propensity scores
  data.b$ps <- predict(ps_model, type = "response")

  # calculate stabilized weights
  data.b$w_stb <- ifelse(
    data.b$smoker == "Y",
    p_treated / data.b$ps,  # For smokers
    (1 - p_treated) / (1 - data.b$ps)  # For non-smokers
  )

  # fit a MSM using weighted regression
  design.b <- svydesign(~1, weights = ~w_stb, data = data.b)
  msm_model <- svyglm(gestation ~ smoker, design = design.b)

  # store the estimated coefficient for smoking
  b.holder[i] <- coef(msm_model)["smokerY"]
}

# calculate the mean and 95% confidence interval
mean_marginal <- mean(b.holder)
ci_marginal <- quantile(b.holder, probs = c(0.025, 0.975))

# display results
list(
  Marginal_Effect = list(
    Estimate = round(mean_marginal, 3),
    CI = round(ci_marginal, 3)
  )
)
```

```
## $Marginal_Effect
## $Marginal_Effect$Estimate
## [1] -0.666
##
## $Marginal_Effect$CI
##   2.5%  97.5%
## -2.663  0.335
```

Given the results above, MACE -0.666 suggests that smoking during pregnancy is associated with a reduction of approximately 0.67 weeks in gestational age but 95% CI indicates that this is not statistically significant.

e

I will used the same trimmed data to examine unadjusted association between smoking status and gestational age as follows:

```
# simple linear regression
reg <- lm(gestation ~ smoker, data = df_trimmed)

summary(reg)
```

```
##
## Call:
## lm(formula = gestation ~ smoker, data = df_trimmed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.987  -0.987   0.013   1.013   6.446
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.98701    0.08399 464.192   <2e-16 ***
## smokerY     -0.43280    0.26925  -1.607    0.108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.331 on 851 degrees of freedom
## Multiple R-squared:  0.003027,   Adjusted R-squared:  0.001855
## F-statistic: 2.584 on 1 and 851 DF,  p-value: 0.1083
```

```r
# extract point estimate and confidence intervals
effect <- coef(summary(reg))["smokerY", "Estimate"]
ci <- confint(reg, "smokerY", level = 0.95)

# results
cat("Unadjusted Point Estimate:", round(effect, 2), "\n")
```

```
## Unadjusted Point Estimate: -0.43
```

```r
cat("95% CI:", round(ci[1], 2), "to", round(ci[2], 2), "\n")
```

```
## 95% CI: -0.96 to 0.1
```

The unadjusted association suggests that smokers have, on average, a gestational age 0.43 weeks shorter compared to non-smokers. However, 95%CI indicates that this is not statistically significant.

All methods in (d) and (e) consistently suggest that smoking during pregnancy is associated with a reduction in gestational age, but the effect is not statistically significant.

f

Initially, the positivity assumption appeared to be violated. I tried to mitigate this by trimming extreme propensity scores and using stabilized weights in the propensity score weighting approach. However, as mentioned in (b) and (c), the presence of unmeasured confounders and potential SUTVA violation in this study may lead to biased causal effect estimates.

g

The population of interest is all pregnant individuals, but the sample at hand is limited to singleton births in King County, WA, in 2001, with many mothers participating in the First Steps program. During the analysis, I excluded some additional individuals to meet causal assumptions. This sampling scheme and analysis method limits generalizability by excluding multiple births and individuals with extreme propensity scores. Therefore, the observed MACE is only valid among the individuals included in the analysis.
Excluding multiple births, which are inherently at higher risk of preterm birth, may underestimate the overall effect of smoking on gestational age. Additionally, as discussed in (c) SUTVA violation, the spillover effect may also underestimate the effect of smoking. Therefore, the observed MACE may be biased towards the null and careful interpretation is required.

2

a

b

c

d

e

f