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:

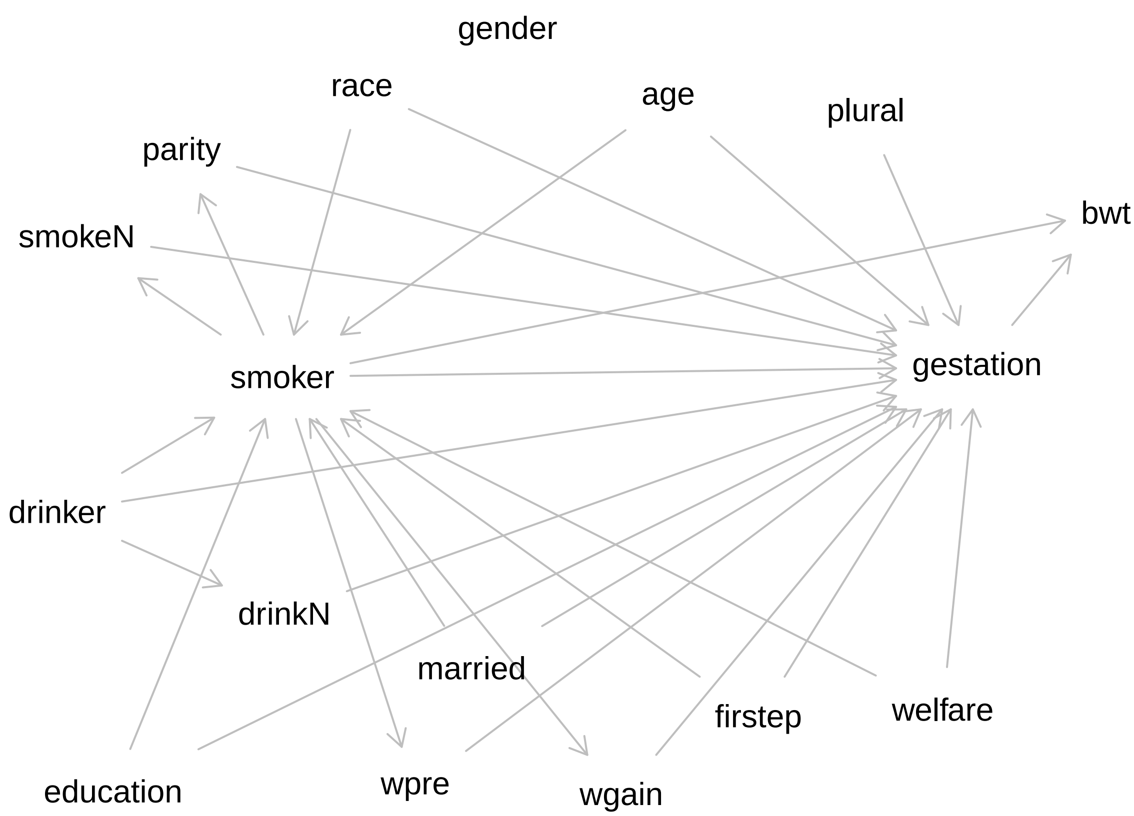
- Consistency: The observed gestational age for each mother corresponds to the actual smoking status during pregnancy (i.e., the recorded smoking status is correct and there is no error in measurement)
- No unmeasured confounding: All confounding factors affecting both smoking and gestational age are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

```r
# 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)
```



`smoker`: Exposure.
`gestation`: Outcome.
`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`: **Potential confounder**. Drinking state affects number of alcoholic drinks per week during pregnancy, and `drinkN` also affects both smoking status and 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.
`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**. Education years are likely to influence smoking status and gestational age.

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`, `drinkN`, `education`, `firstep`, `married`, `race`, `welfare`), and avoid adjusting for mediators (`smokeN`, `parity`, `drinkN`, `wpre`, `wgain`) and colliders (`bwt`).

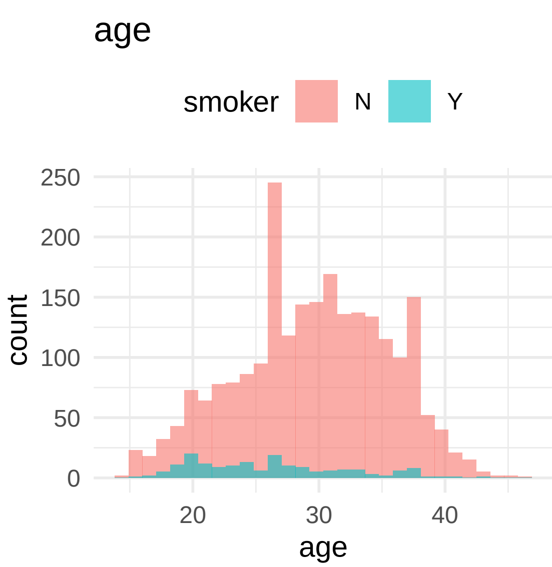
c

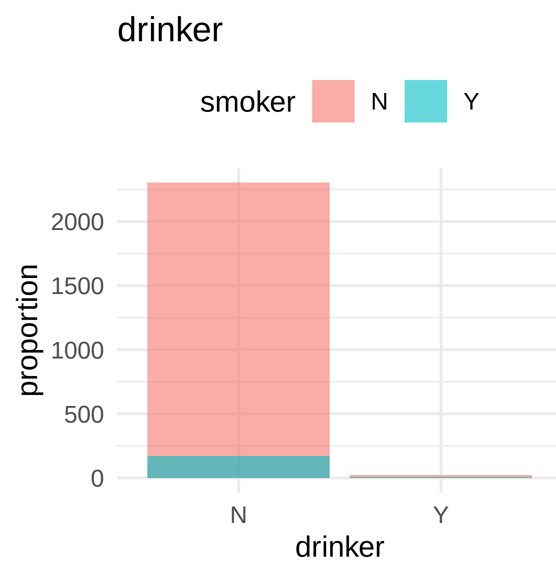
```r
covariates <- c("age", "drinker", "drinkN", "education", "firstep", "married", "race", "welfare")
exposure <- "smoker"
outcome <- "gestation"

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

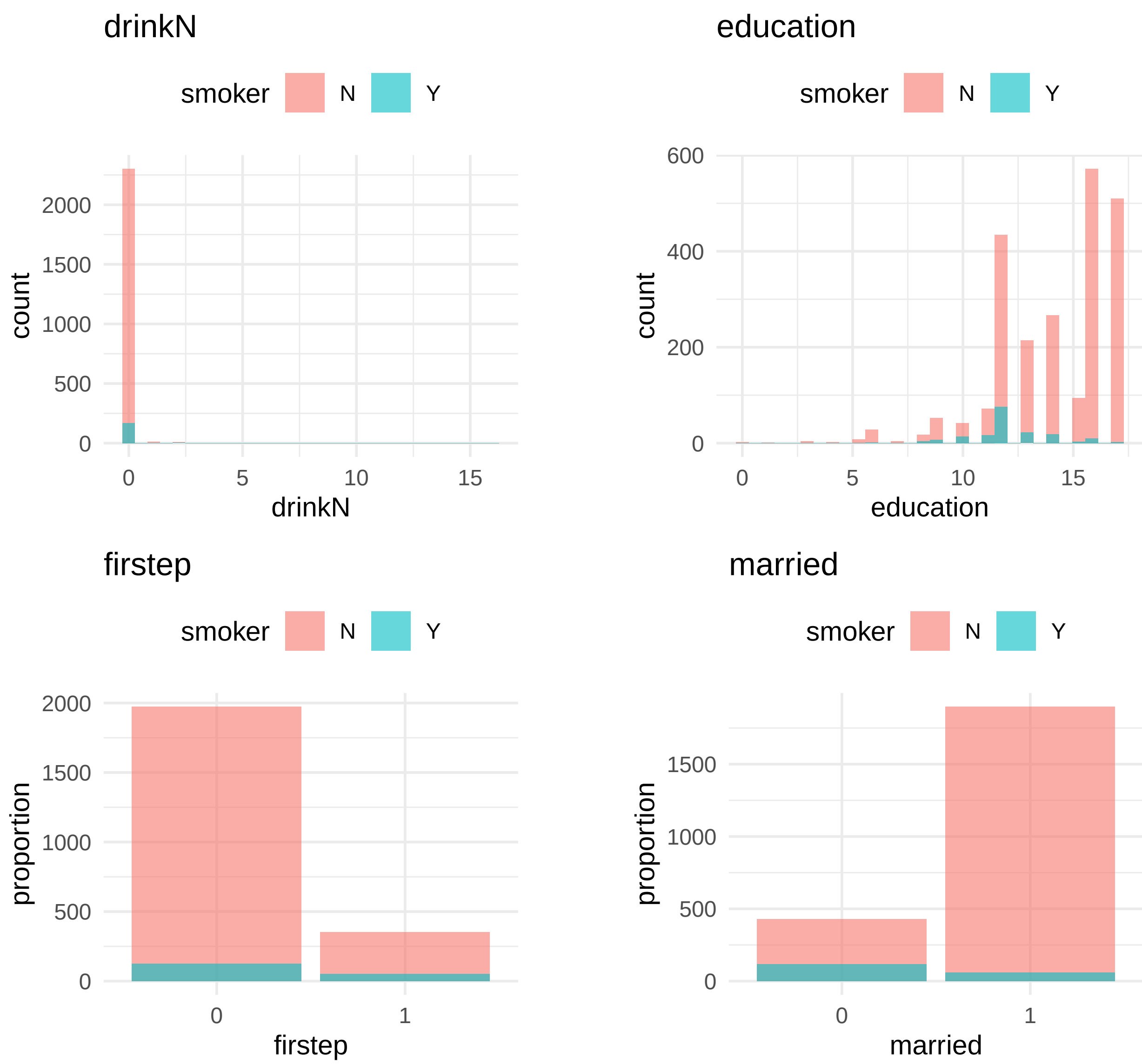
```
##                         Stratified by smoker
##                          N             Y             SMD
##   n                       2325           175
##   age (mean (SD))        29.54 (5.94)  26.16 (6.01)   0.566
##   drinker = Y (%)           23 ( 1.0)      6 ( 3.4)   0.167
##   drinkN (mean (SD))      0.03 (0.48)   0.07 (0.47)   0.095
##   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
```

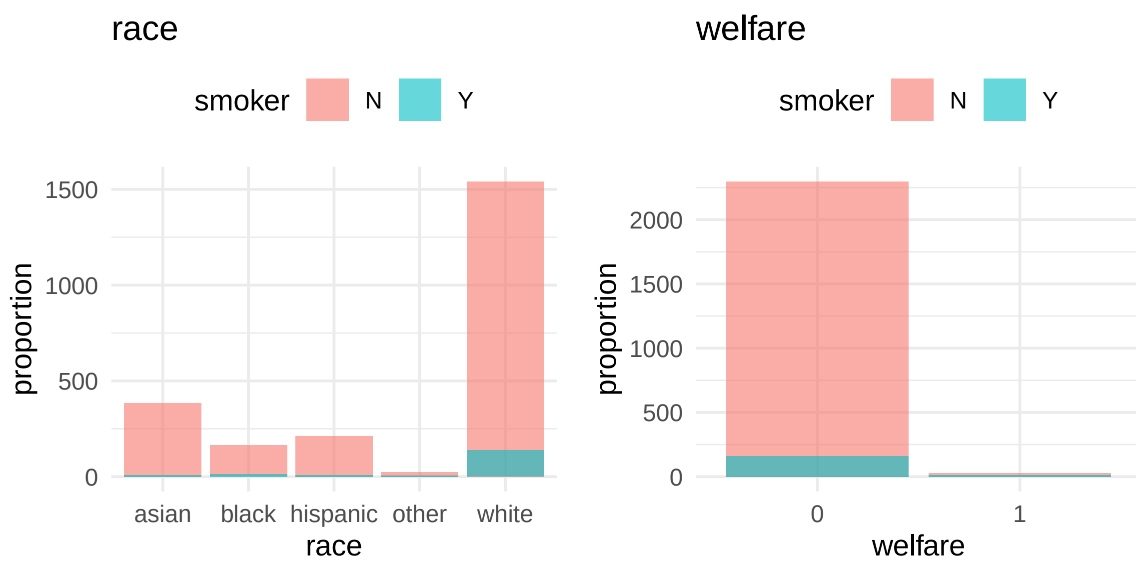
Given SMD >0.1, all the selected covariates edcluding `drinkN` are imbalanced between the two smoking groups, indicating that adjustment of these covariates are required in the later analysis to estimate smoking's causal effect on gestational age.

```
# 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)
}
```






drinkN
smoker
N
Y
count
2000
1500
1000
500
0
0
5
10
15
drinkN
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



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, `drinkN` 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

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

Direct regression adjustment:

```
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
```

```
# 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")
```

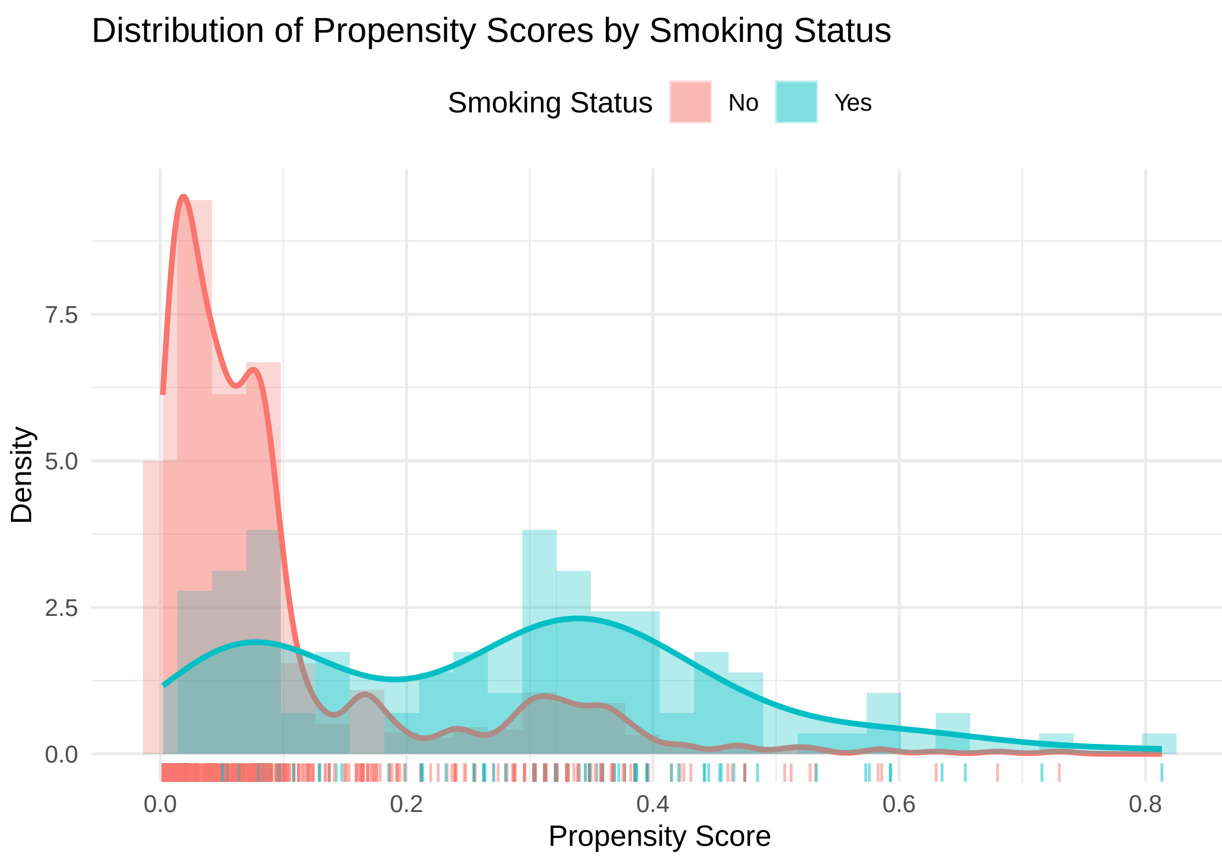
```
## Estimated ACE: -0.09
## 95% CI: -0.62 to 0.43
```

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.

```
# 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(
```

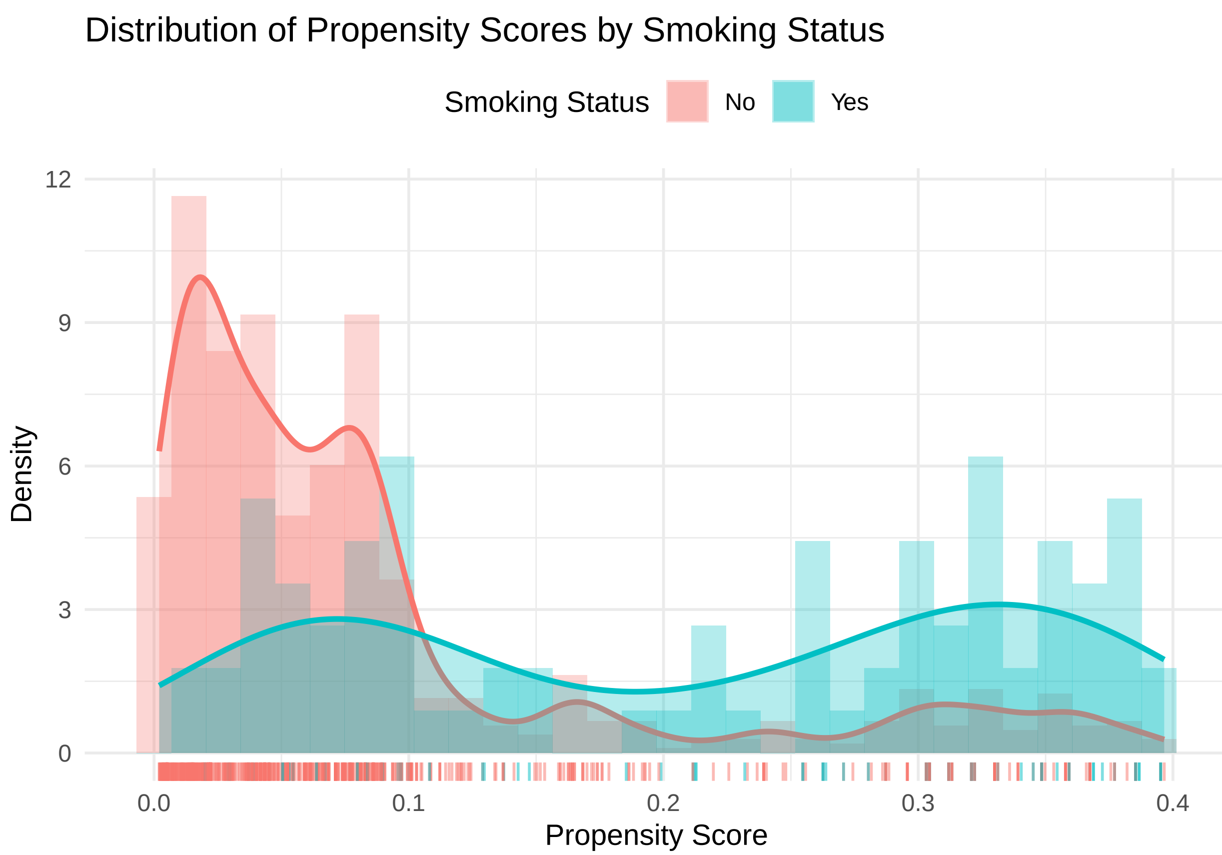
```
    values = c("#F8766D", "#00BFC4"),
    name = "Smoking Status",
    labels = c("No", "Yes"),
    guide = guide_legend(override.aes = list(color = NA))
  ) +
  scale_color_manual(
    values = c("#F8766D", "#00BFC4"),
    guide = "none"
  ) +
  geom_rug(aes(color = smoker), sides = "b", alpha = 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
  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"
  )
```



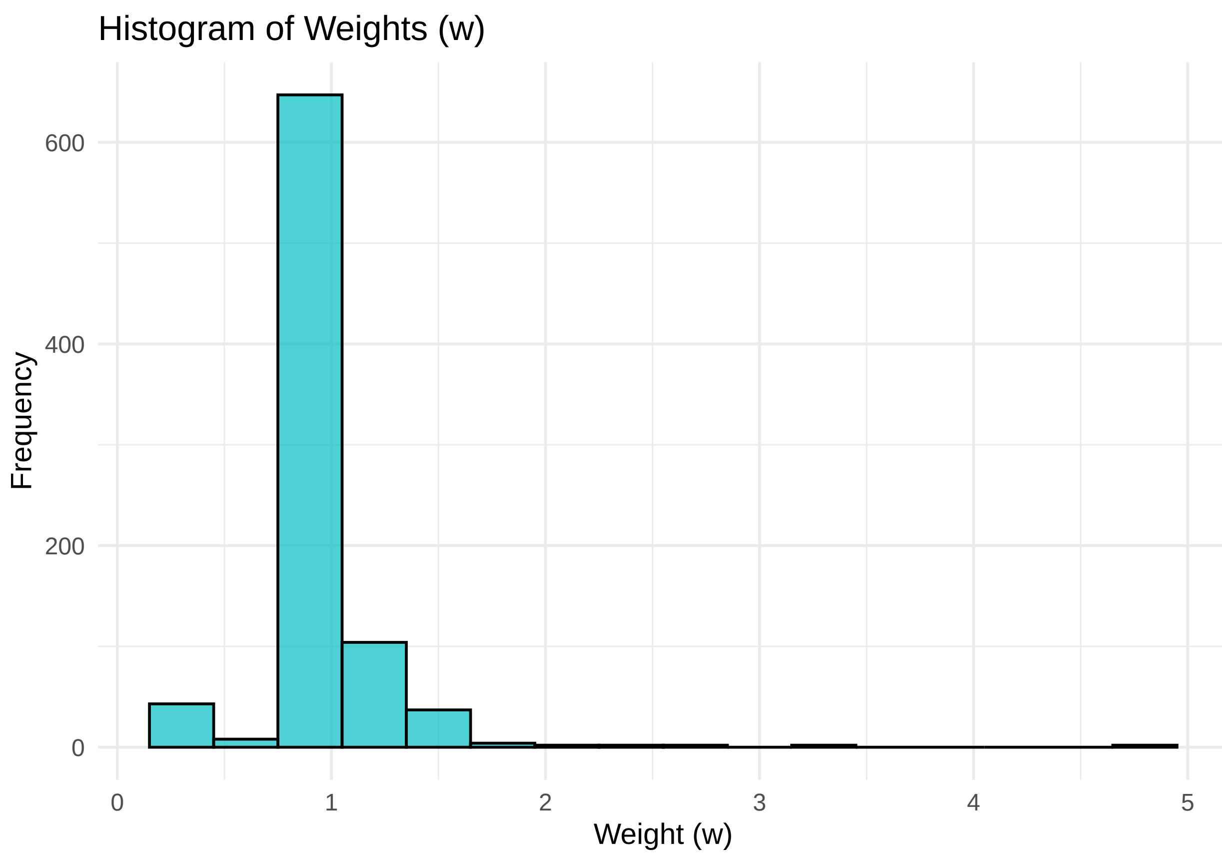
Now, propensity scores largely overlap between the two groups and probabilistic assumption is mostly satisfied.

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

# 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.

```
# 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",
```

```r
    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:

```r
# 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.

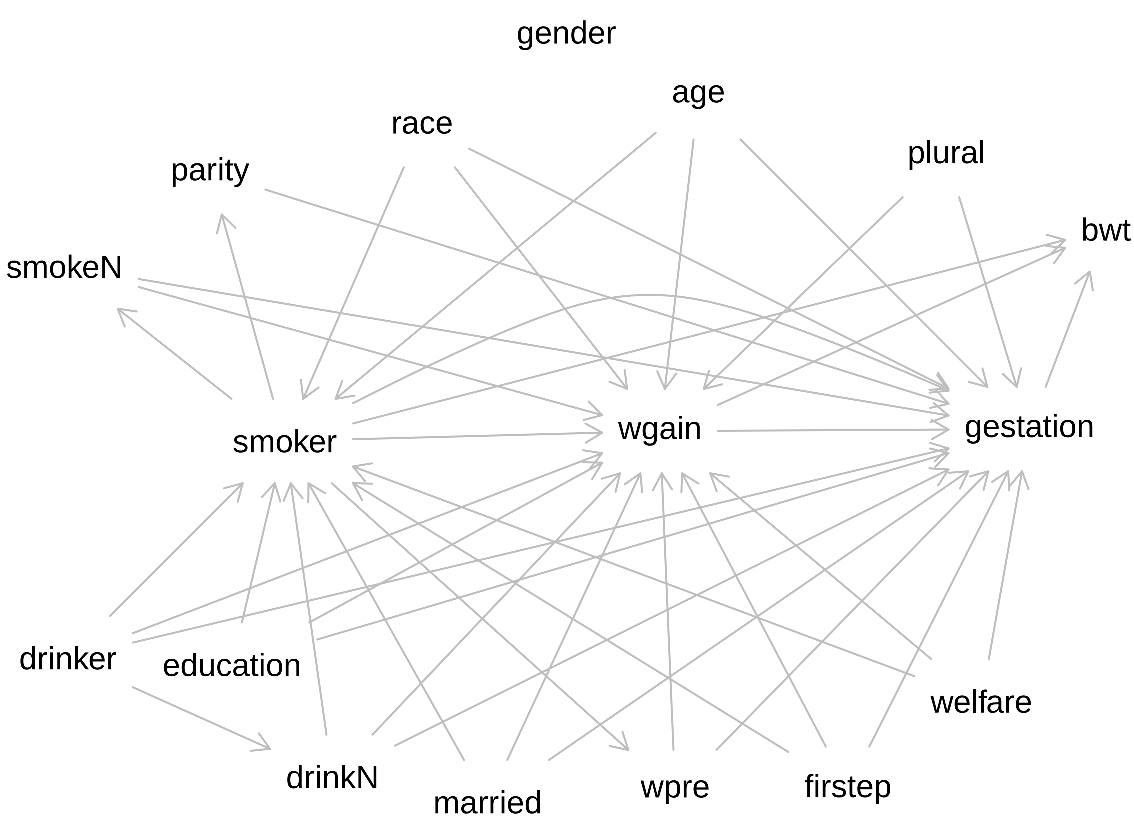
2

In this section, I will investigate whether the effect of smoking on gestational age is mediated by maternal weight gain during pregnancy.

a

```
# DAG
g = dagitty('dag {
age [pos="-0.106,-0.624"]
bwt [pos="0.826,-0.425"]
drinkN [pos="-0.943,0.400"]
drinker [pos="-1.548,0.221"]
education [pos="-1.173,0.231"]
firstep [pos="0.236,0.418"]
gender [pos="-0.408,-0.717"]
gestation [outcome,pos="0.651,-0.124"]
married [pos="-0.588,0.438"]
parity [pos="-1.223,-0.511"]
plural [pos="0.461,-0.537"]
race [pos="-0.738,-0.582"]
smokeN [pos="-1.556,-0.369"]
smoker [exposure,pos="-1.052,-0.106"]
welfare [pos="0.541,0.286"]
wgain [pos="-0.194,-0.121"]
wpre [pos="-0.159,0.423"]
age -> gestation
age -> smoker
age -> wgain
drinkN -> gestation
drinkN -> smoker
drinkN -> wgain
drinker -> drinkN
drinker -> gestation
drinker -> smoker
drinker -> wgain
education -> gestation
education -> smoker
education -> wgain
firstep -> gestation
firstep -> smoker
firstep -> wgain
gestation -> bwt
married -> gestation
married -> smoker
married -> wgain
parity -> gestation
plural -> gestation
plural -> wgain
race -> gestation
race -> smoker
race -> wgain
```

```
smokeN -> gestation
smokeN -> wgain
smoker -> bwt
smoker -> gestation [pos="-0.239,-0.403"]
smoker -> parity
smoker -> smokeN
smoker -> wgain
smoker -> wpre
welfare -> gestation
welfare -> smoker
welfare -> wgain
wgain -> bwt
wgain -> gestation
wpre -> gestation
wpre -> wgain
}
')

plot(g)
```



`smoker`: Exposure.
`gestation`: Outcome.
`wgain`: Potential mediator.
`gender`: Maternal smoking status does not influence the infant's gender, and the infant's gender does not affest gestational age. Gender is not associated with `wgain`, either.
`plural`: **Potential mediator-outcome confounder**. Maternal smoking status does not influence

singleton/twin/triplet pregnancy, and the number of child carriage affect `wgain` and gestational age.
`age`: **Potential confounder for exposure, mediator and outcome**. Depending on the mother's generation, smoking prevalence varies, the magnitude of weight gain changes, and it also influence the gestational age.
`race`: **Potential confounder for exposure, mediator and outcome**. There is a racial difference in smoking prevalence, maternal weight gain and gestational age.
`parity`: Potential mediator. Smokers may be more likely to have multiple parity and parity also affects gestational age.
`married`: **Potential exposure-outcome confounder**. Married people may be more likely to be a non-smoker, and marriage status may impact gestational age.
`bwt`: Potential collider. Smoking status, maternal weight gain and gestational age impacts birth weight.
`smokeN`: **Potential mediator-outcome confounder**. Smoking status impacts number of cigarettes smoked per day during pregnancy, and `smokeN` affects maternal weight gain and gestational age.
`drinkN`: **Potential confounder for exposure, mediator and outcome**. Drinking state affects number of alcoholic drinks per week during pregnancy, and `drinkN` also affects smoking status, maternal weight gain, and gestational age.
`firstep`: **Potential confounder for exposure, mediator and outcome**. The participation influences smoking status, maternal weight gain, and gestational age.
`welfare`: **Potential confounder for exposure, mediator and outcome**. The participation influences both smoking status and gestational age.
`drinker`: **Potential confounder for exposure, mediator and outcome**. Drinkers are more likely to smoke, gain weight during the pregnancy and it affects gestational age.
`wpre`: **Potential mediator-outcome confounder**. Smoking status affects mother's weight in pounds prior to pregnancy, and `wpre` affects both maternal weight gain and gestational age.
`education`: **Potential confounder for exposure, mediator and outcome**. Education years are likely to influence smoking status, maternal weight gain, and gestational age.

b

Given (a), the confounder for `smoker` and `wgain` are: `age`, `race`, `drinkN`, `firstep`, `welfare`, `drinker`, and `education`.
I will estimate the effect of smoking on maternal weight gain using direct regression confounding adjustment using data with inclusion criteria sepcified in 1-(d). Hence, covariates does not include drinking-related variables.

```r
fit1 <- lm(wgain ~ smoker + age + race + firstep + welfare + education, data = df_d)

summary(fit1)
```

```
##
## Call:
## lm(formula = wgain ~ smoker + age + race + firstep + welfare +
##     education, data = df_d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.036  -8.777  -1.100   7.820 108.781
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   52.0866     5.0714  10.271  < 2e-16 ***
## smokerY       -0.1495     1.6150  -0.093 0.926262
## age           -0.5860     0.1993  -2.940 0.003365 **
```

```
## raceblack     -4.5103     1.9693  -2.290 0.022239 *
## racehispanic  -6.8280     2.0061  -3.404 0.000695 ***
## raceother     -1.2498     3.8632  -0.324 0.746380
## racewhite     -1.9376     1.4377  -1.348 0.178097
## firstep1       0.4894     1.1531   0.424 0.671345
## welfare1      -1.8411     3.0476  -0.604 0.545937
## education     -0.1906     0.2873  -0.663 0.507352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.69 on 880 degrees of freedom
## Multiple R-squared:  0.02803,    Adjusted R-squared:  0.01809
## F-statistic:  2.82 on 9 and 880 DF,  p-value: 0.00284
```

```r
# extract coefficient and 95% CI
effect <- coef(summary(fit1))["smokerY", ]
ci <- confint(fit1, "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: -0.15
## 95% CI: -3.32 to 3.02
```

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 0.15 pound in maternal weight but 95% CI indicates that this is not statistically significant.

c

Given (a), the confounder for `wgain` and `gestation` are: `smokeN`, `wpre`, `age`, `race`, `firstep`, `welfare`, and `education` (`plural` has only one type in this study so it's not included in the covariates).
I will estimate the effect of weight gain on gestational age using direct regression confounding adjustment using data with inclusion criteria sepcified in 1-(d).

```r
fit2 <- lm(gestation ~ wgain + smokeN + wpre + age + race + firstep + welfare + education, data = df_d)

summary(fit2)
```

```
##
## Call:
## lm(formula = gestation ~ wgain + smokeN + wpre + age + race +
##     firstep + welfare + education, data = df_d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.629  -1.029   0.250   1.263   6.590
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.438231   0.894505  40.736  < 2e-16 ***
## wgain         0.019794   0.005421   3.651 0.000276 ***
## smokeN       -0.035833   0.026061  -1.375 0.169490
```

```
## wpre          0.003987   0.002150   1.854 0.064008 .
## age           0.019161   0.031940   0.600 0.548718
## raceblack     0.338681   0.315676   1.073 0.283622
## racehispanic  0.271759   0.321975   0.844 0.398878
## raceother    -0.774050   0.612970  -1.263 0.207001
## racewhite    -0.003827   0.231141  -0.017 0.986792
## firstep1      0.118896   0.183276   0.649 0.516685
## welfare1     -1.495712   0.485997  -3.078 0.002151 **
## education     0.059031   0.045743   1.290 0.197219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.333 on 878 degrees of freedom
## Multiple R-squared:  0.03649,    Adjusted R-squared:  0.02442
## F-statistic: 3.023 on 11 and 878 DF,  p-value: 0.000566
```

```r
# extract coefficient and 95% CI
effect <- coef(summary(fit2))["wgain", ]
ci <- confint(fit2, "wgain", level = 0.95)
cat("Estimated ACE:", round(effect[1], 2),
    "\n95% CI:", round(ci[1], 2), "to", round(ci[2], 2), "\n")
```

```
## Estimated ACE: 0.02
## 95% CI: 0.01 to 0.03
```

All else being equal, in the eligible population, the point estimate of ACE suggests that one unit of increase in maternal weight (= 1 pound) is associated with an increase of approximately 0.02 weeks in gestational age. Given 95% CI, this is statistically significant.

d

Given (a), the confounder for `wgain`, `smoker` and `gestation` are:`married`, `smokeN`, `wpre`, `age`, `race`, `firstep`, `welfare`, and `education` I will estimate the effect of weight gain on gestational age using direct regression confounding adjustment using data with inclusion criteria sepcified in 1-(d).

```r
fit3 <- lm(gestation ~ smoker*wgain + married + smokeN + wpre + age + race + firstep + welfare + educati

summary(fit3)
```

```
##
## Call:
## lm(formula = gestation ~ smoker * wgain + married + smokeN +
##     wpre + age + race + firstep + welfare + education, data = df_d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5299  -1.0538   0.2735   1.2188   6.5516
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.5170407  0.9072557  40.250  < 2e-16 ***
## smokerY        0.7074874  0.7232103   0.978 0.328216
```

```
## wgain          0.0220385  0.0058841   3.745 0.000192 ***
## married1       0.4140692  0.1879955   2.203 0.027887 *
## smokeN        -0.0569677  0.0426345  -1.336 0.181835
## wpre           0.0042203  0.0021586   1.955 0.050890 .
## age            0.0004938  0.0333473   0.015 0.988189
## raceblack      0.4096572  0.3168469   1.293 0.196381
## racehispanic   0.2849613  0.3221213   0.885 0.376594
## raceother     -0.7365726  0.6147611  -1.198 0.231185
## racewhite     -0.0317153  0.2316016  -0.137 0.891110
## firstep1       0.1830730  0.1852874   0.988 0.323402
## welfare1      -1.4554417  0.4874534  -2.986 0.002907 **
## education      0.0555719  0.0458604   1.212 0.225931
## smokerY:wgain -0.0098311  0.0148830  -0.661 0.509070
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 875 degrees of freedom
## Multiple R-squared:  0.04254,    Adjusted R-squared:  0.02722
## F-statistic: 2.777 on 14 and 875 DF,  p-value: 0.000468
```

```
# extract coefficient and 95% CI
effect <- coef(summary(fit3))["smokerY", ]
ci <- confint(fit3, "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: 0.71
## 95% CI: -0.71 to 2.13
```

```
effect <- coef(summary(fit3))["wgain", ]
ci <- confint(fit3, "wgain", level = 0.95)
cat("Estimated ACE:", round(effect[1], 2),
    "\n95% CI:", round(ci[1], 2), "to", round(ci[2], 2), "\n")
```

```
## Estimated ACE: 0.02
## 95% CI: 0.01 to 0.03
```

```
effect <- coef(summary(fit3))["smokerY:wgain", ]
ci <- confint(fit3, "smokerY:wgain", level = 0.95)
cat("Estimated ACE:", round(effect[1], 2),
    "\n95% CI:", round(ci[1], 2), "to", round(ci[2], 2), "\n")
```

```
## Estimated ACE: -0.01
## 95% CI: -0.04 to 0.02
```

- All else being equal, in the eligible population, the point estimate of ACE suggests that smoking during pregnancy is associated with an increase of approximately 0.71 weeks in maternal weight but 95% CI indicates that this is not statistically significant
- All else being equal, in the eligible population, the point estimate of ACE suggests that one unit of increase in maternal weight (= 1 pound) is associated with a increase of approximately 0.02 weeks in gestational age. Given 95% CI, this is statistically significant
- All else being equal, in the eligible population, for smokers, the positive effect of weight gain on gestational age is slightly reduced by 0.01 pounds weight gain compared to non-smokers, but 95% CI indicates that this is not statistically significant

e

Covariates: `married`, `smokeN`, `wpre`, `age`, `race`, `firstep`, `welfare`, and `education`

I will estimate causal effects using the following code:

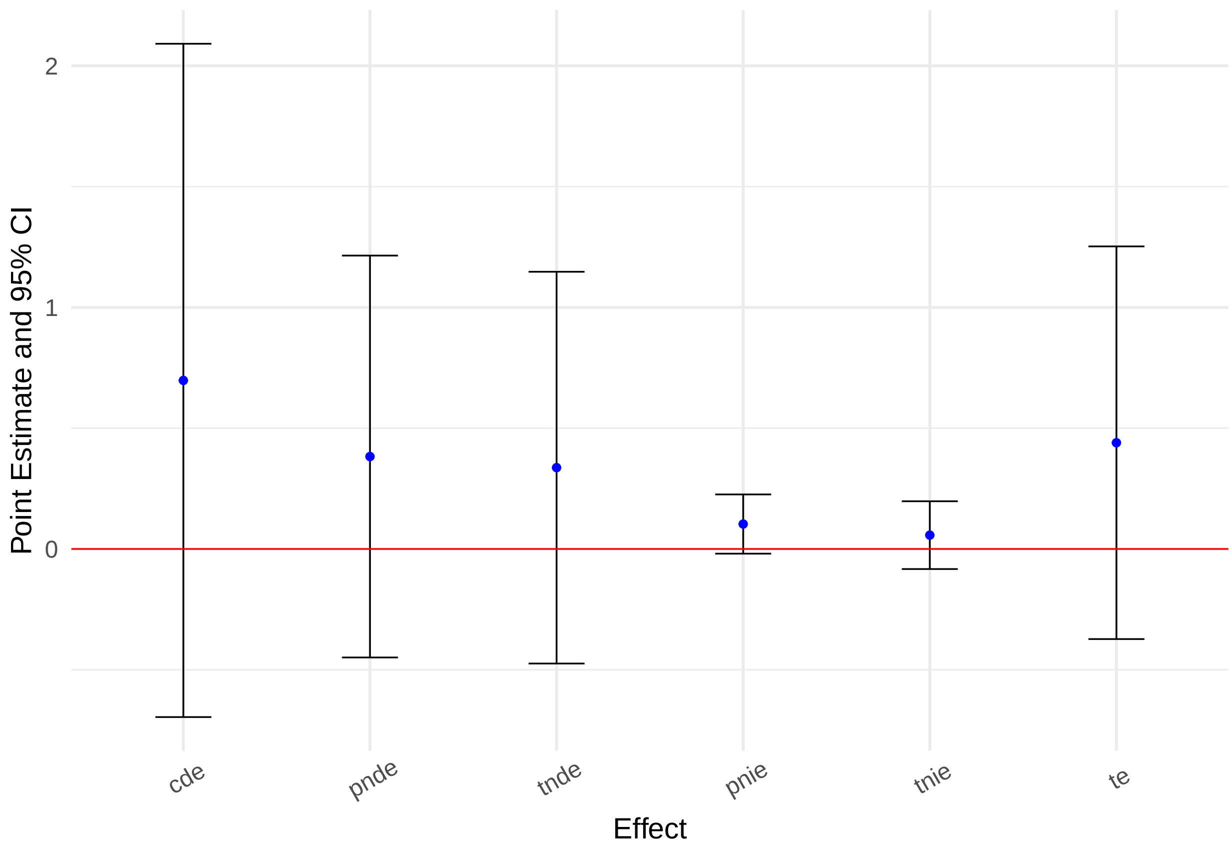
```r
med_result = cmest(
  data = df_d,
  model = "rb", # regression-based approach
  outcome = "gestation",
  exposure = "smoker",
  mediator = "wgain",
  basec = c("married", "smokeN", "wpre", "age", "race", "firstep", "welfare", "education"),
  EMint = TRUE,  # allow exposure-mediator interaction
  mreg = list(type = "linear"),
  yreg = "linear",
  estimation = "paramfunc",
  inference = "delta",
  full = FALSE,
  astar = "N", # reference value
  a = "Y",
  mval = list(1)
)

summary(med_result)
```

```
## Causal Mediation Analysis
##
## # Outcome regression:
##
## Call:
## glm(formula = gestation ~ smoker + wgain + smoker * wgain + married +
##     smokeN + wpre + age + race + firstep + welfare + education,
##     family = gaussian(), data = getCall(x$reg.output$yreg)$data,
##     weights = getCall(x$reg.output$yreg)$weights)
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.5170407  0.9072557  40.250  < 2e-16 ***
## smokerY        0.7074874  0.7232103   0.978 0.328216
## wgain          0.0220385  0.0058841   3.745 0.000192 ***
## married1       0.4140692  0.1879955   2.203 0.027887 *
## smokeN        -0.0569677  0.0426345  -1.336 0.181835
## wpre           0.0042203  0.0021586   1.955 0.050890 .
## age            0.0004938  0.0333473   0.015 0.988189
## raceblack      0.4096572  0.3168469   1.293 0.196381
## racehispanic   0.2849613  0.3221213   0.885 0.376594
## raceother     -0.7365726  0.6147611  -1.198 0.231185
## racewhite     -0.0317153  0.2316016  -0.137 0.891110
## firstep1       0.1830730  0.1852874   0.988 0.323402
## welfare1      -1.4554417  0.4874534  -2.986 0.002907 **
## education      0.0555719  0.0458604   1.212 0.225931
## smokerY:wgain -0.0098311  0.0148830  -0.661 0.509070
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 5.427416)
##
##     Null deviance: 4960  on 889  degrees of freedom
## Residual deviance: 4749  on 875  degrees of freedom
## AIC: 4048
##
## Number of Fisher Scoring iterations: 2
##
##
## # Mediator regressions:
##
## Call:
## glm(formula = wgain ~ smoker + married + smokeN + wpre + age +
##     race + firstep + welfare + education, family = gaussian(),
##     data = getCall(x$reg.output$mreg[[1L]])$data, weights = getCall(x$reg.output$mreg[[1L]])$weights)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.30868    5.26734  10.690  < 2e-16 ***
## smokerY       4.67859    2.54845   1.836  0.06672 .
## married1     -2.67611    1.16186  -2.303  0.02150 *
## smokeN       -0.74841    0.25458  -2.940  0.00337 **
## wpre         -0.05621    0.01321  -4.254 2.33e-05 ***
## age          -0.31364    0.20630  -1.520  0.12879
## raceblack    -4.15846    1.95976  -2.122  0.03412 *
## racehispanic -5.91946    1.98410  -2.983  0.00293 **
## raceother    -0.96625    3.81204  -0.253  0.79996
## racewhite    -0.70464    1.43521  -0.491  0.62358
## firstep1      0.26208    1.14890   0.228  0.81961
## welfare1     -0.88697    3.01951  -0.294  0.76902
## education    -0.29469    0.28421  -1.037  0.30007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 208.7022)
##
##     Null deviance: 195329  on 889  degrees of freedom
## Residual deviance: 183032  on 877  degrees of freedom
## AIC: 7294
##
## Number of Fisher Scoring iterations: 2
##
##
## # Effect decomposition on the mean difference scale via the regression-based approach
##
## Closed-form parameter function estimation with
##  delta method standard errors, confidence intervals and p-values
##
##      Estimate Std.error  95% CIL 95% CIU  P.val
## cde   0.69766   0.71103 -0.69593   2.091 0.3265
## pnde  0.38258   0.42441 -0.44924   1.214 0.3673
## tnde  0.33659   0.41370 -0.47426   1.147 0.4159
```

```
## pnie  0.10311   0.06255 -0.01948   0.226 0.0993 .
## tnie  0.05711   0.07157 -0.08315   0.197 0.4248
## te    0.43970   0.41469 -0.37307   1.252 0.2890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (cde: controlled direct effect; pnde: pure natural direct effect; tnde: total natural direct effect;
##
## Relevant variable values:
## $a
## [1] "Y"
##
## $astar
## [1] "N"
##
## $mval
## $mval[[1]]
## [1] 1
##
##
## $basecval
## $basecval[[1]]
## [1] 0.6101124
##
## $basecval[[2]]
## [1] 0.8865169
##
## $basecval[[3]]
## [1] 148.0056
##
## $basecval[[4]]
## [1] 23.21685
##
## $basecval[[5]]
## [1] 0.11573034 0.11797753 0.01910112 0.59550562
##
## $basecval[[6]]
## [1] 0.2764045
##
## $basecval[[7]]
## [1] 0.02808989
##
## $basecval[[8]]
## [1] 12.87191
```

```
# visualization
ggcmest(med_result) +
ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 30, vjust = 0.8))
```



- Pure natural direct effect (PNDE): 0.383 (95%CI -0.449, 1.214)
- Total natural direct effect (TNDE): 0.337 (95%CI -0.474, 1.147)
- Pure natural indirect effect (PNIE): 0.103 (95%CI -0.0195, 0.226)
- Total natural indirect effect (TNIE): 0.0571 (95%CI -0.0832, 0.197)
- Total effect: 0.440 (95%CI -0.373, 1.252)

f

- Natural direct effect (NDE): In this analysis, the NDE represents how smoking status directly impacts gestational age, independent of its effect through maternal weight gain. When assuming no exposure-mediator interaction, NDE was 0.383 but it was not statistically significant. When accounting for exposure-mediator interaction, NDE was 0.377 but it was not statistically significant
- Natural indirect effect (NIE): In this analysis, the NIE quantifies the mediated pathway, i.e., the effect of smoking status on gestational age that operates through maternal weight gain. When assuming no exposure-mediator interaction, NIE was 0.103 but it was not statistically significant. When accounting for exposure-mediator interaction, NIE was 0.0571 but it was not statistically significant
- Total effect: TE is decomposable into NDE + NIE. In this analysis, the TE was 0.440 but not statistically significant

The following identifiability assumptions are necessary for the causal interpretation:

- $Y_{am} \perp A \mid C$: No unmeasured confounders affect the relationship between exposure A and outcome Y after adjusting for covariates C
  Some unmeasured factors, like income, or residential location, could still confound this relationship.

- $Y_{am} \perp M \mid C, A$: No unmeasured confounders influence the relationship between the mediator M and outcome Y, conditional on exposure and covariates
  Pre-pregnancy health status (e.g., comorbidities) might confound the weight gain-gestational age relationship but were not explicitly adjusted for.
- $M_a \perp A \mid C$: No unmeasured confounders affect the relationship between exposure A and mediator M after adjusting for covariates
  Unmeasured factors like stress or working hours might confound the exposure-mediator relationship.
- $Y_{am} \perp M_{a*} \mid C$: No effect of exposure A that confounds the mediator M and outcome Y relationship
  Smoking may indirectly influence gestational age through other pathway, such as mother's weight in pounds prior to pregnancy.

In summary, there is no statistically significant evidence to support either a direct or mediated effect of smoking status on gestational age. However, the validity of the causal interpretation is weakened by the above identifiability assumption violations.