Assignment 2

Lewis White

2023-03-11

```
# READ IN DATA

smoking <- read_csv(here("SMOKING_EDS241.csv"))
```

Question 1: Application of estimators based on the "treatment ignorability" assumption

The goal is to estimate the causal effect of maternal smoking during pregnancy on infant birth weight using the treatment ignorability assumptions (Lecture 6 & 7). The data are taken from the National Natality Detail Files, and the extract "SMOKING_EDS241.csv"' is a random sample of all births in Pennsylvania during 1989-1991. Each observation is a mother-infant pair.

The outcome and treatment variables are: - birthwgt = birth weight of infant in grams - tobacco = indicator for maternal smoking

The control variables are: mage (mother's age), meduc (mother's education), mblack (=1 if mother black), alcohol (=1 if consumed alcohol during pregnancy), first (=1 if first child), diabete (=1 if mother diabetic), anemia (=1 if mother anemic)

(a) What is the unadjusted mean difference in birth weight of infants with smoking and nonsmoking mothers? Under what assumption does this correspond to the average treatment effect of maternal smoking during pregnancy on infant birth weight? Provide some simple empirical evidence for or against this assumption.

Creating the model

```
# Running a simple linear regression to see the unadjusted
# mean difference in birth weights for smoking vs
# non-smoking mothers

smoking_birthwgt_SLR <- lm(formula = birthwgt ~ tobacco, data = smoking)

se_models <- starprep(smoking_birthwgt_SLR, stat = c("std.error"),
    se_type = "HC2", alpha = 0.05)

stargazer(smoking_birthwgt_SLR, se = se_models, type = "text")
```

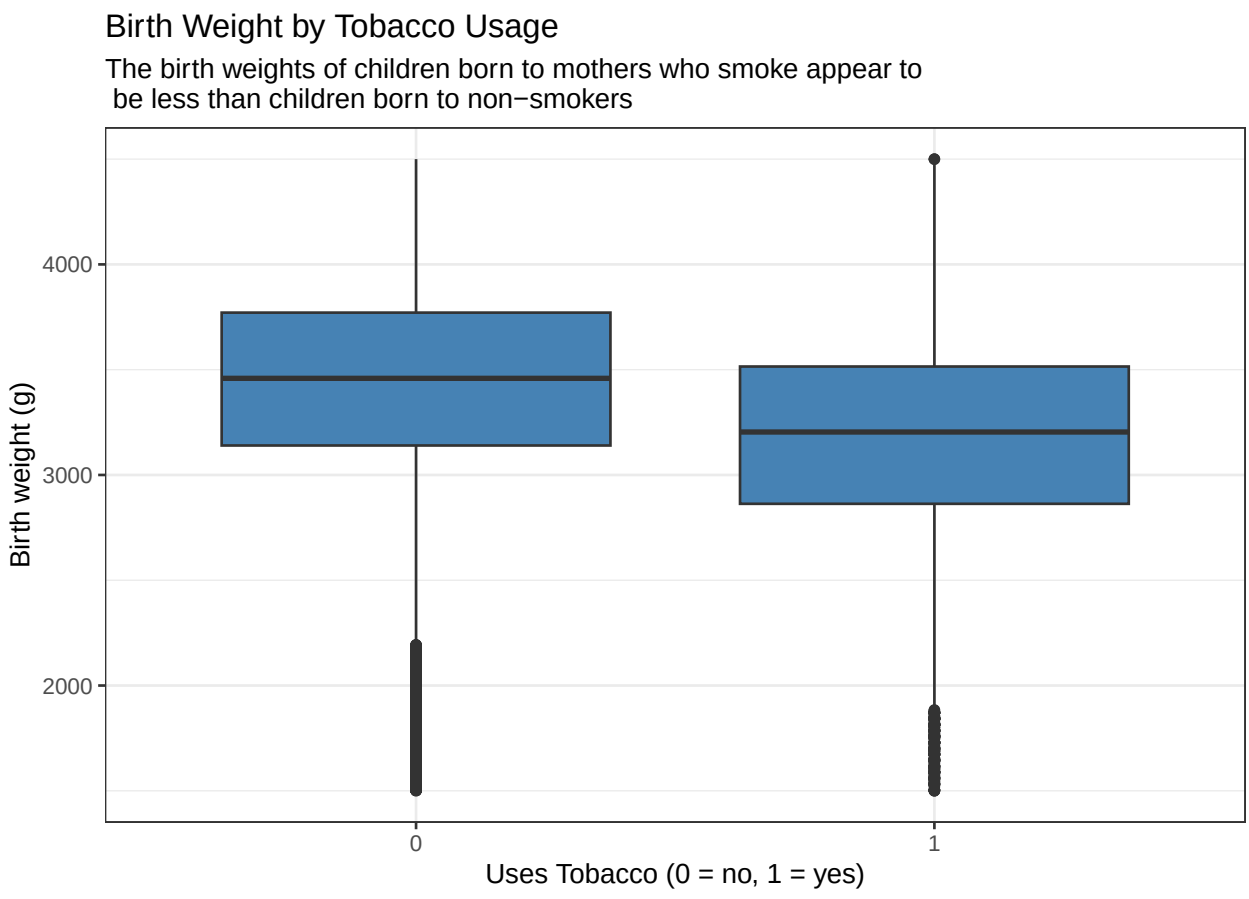
```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                birthwgt
## ------------------------------------------------
## tobacco                      -244.539***
##                               (4.150)
```

```
##
## Constant                      3,430.286***
##                                 (1.781)
##
## ------------------------------------------------
## Observations                     94,173
## R2                               0.037
## Adjusted R2                      0.037
## Residual Std. Error     493.753 (df = 94171)
## F Statistic         3,594.265*** (df = 1; 94171)
## ================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

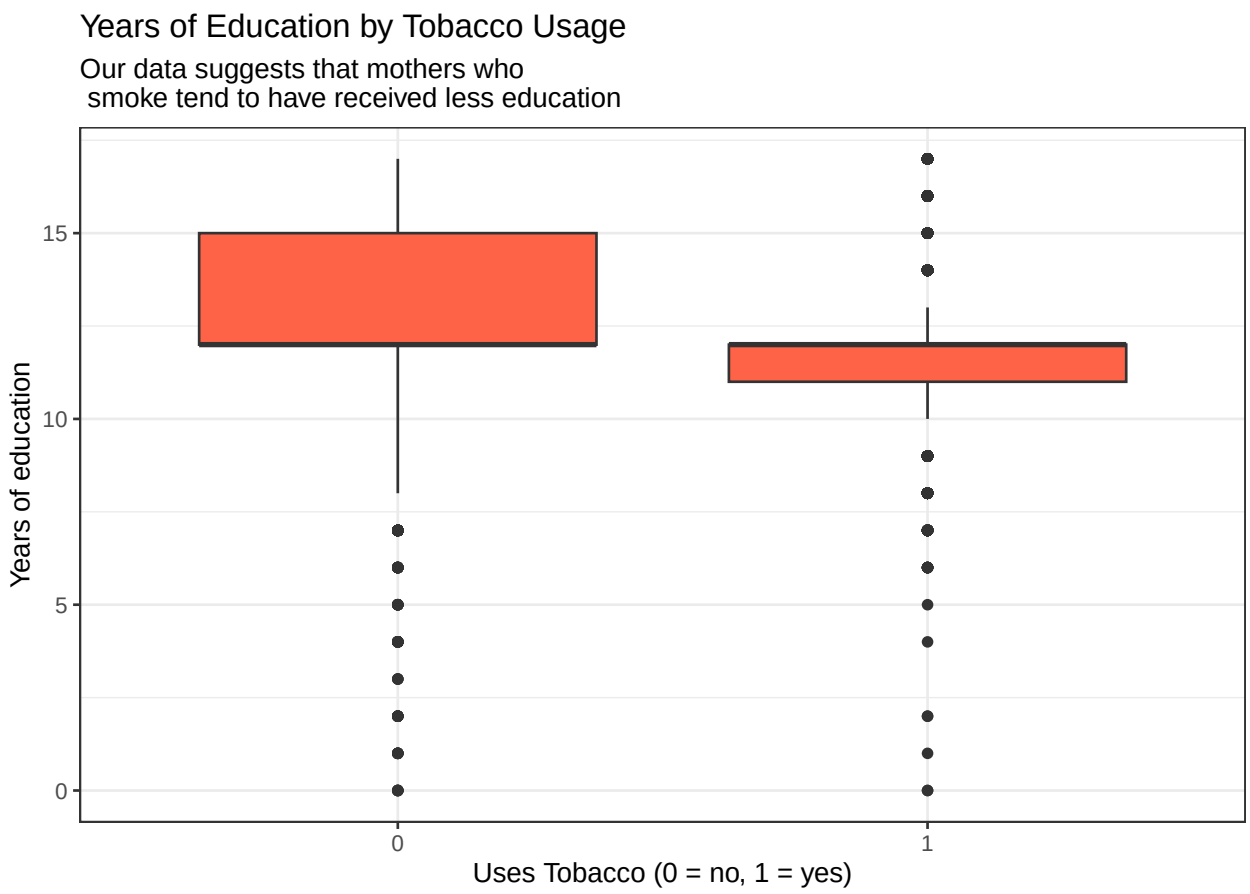
The unadjusted mean difference in birth weights for smoking vs non-smoking mothers is 244.5 grams. This means that on average, moms who smoke have babies that weigh 244.5 grams less than moms who don't smoke.

Checking the assumption

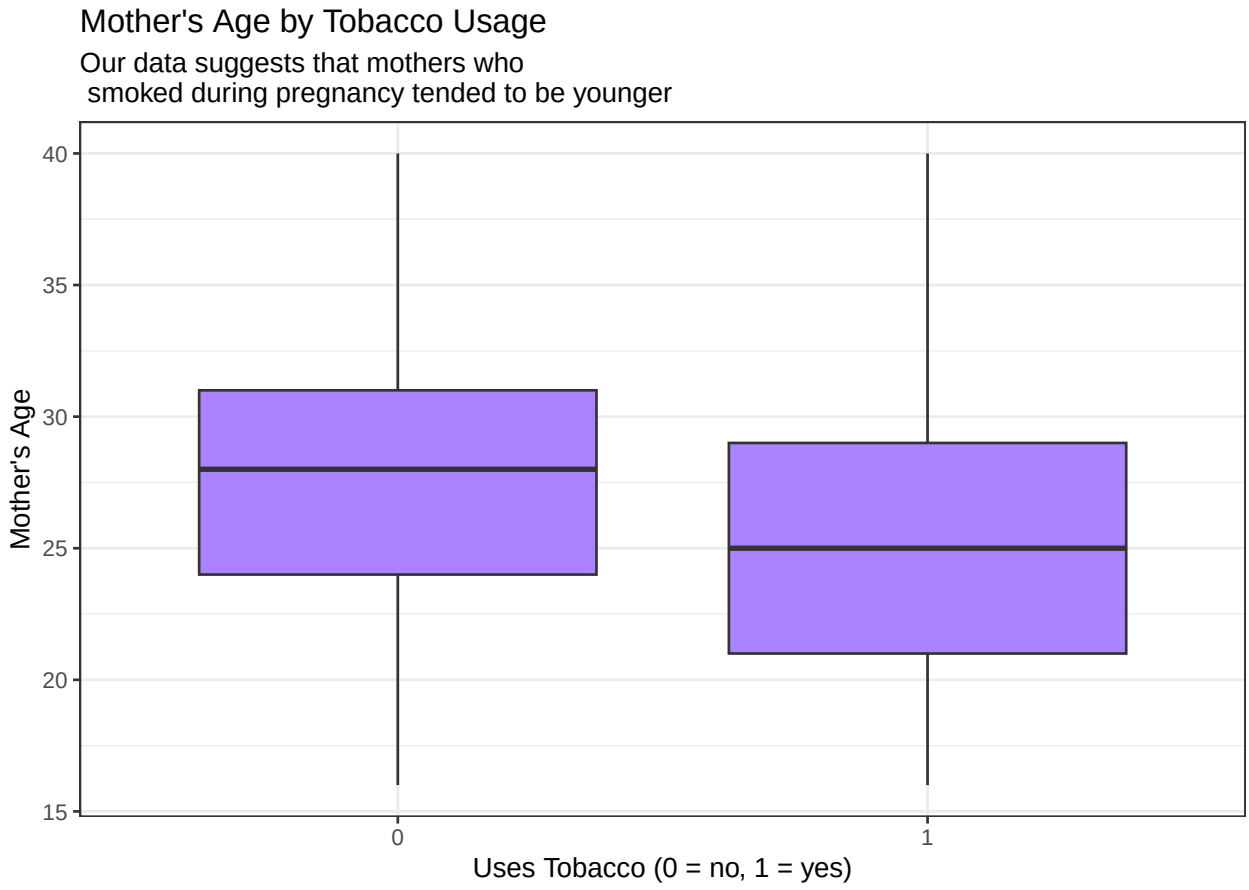
```
# creating a boxplot for smoking and birth weight
ggplot(data = smoking, mapping = aes(x = as.factor(tobacco),
    y = birthwgt)) + geom_boxplot(fill = "steelblue") + labs(x = "Uses Tobacco (0 = no, 1 = yes)",
    y = "Birth weight (g)", title = "Birth Weight by Tobacco Usage",
    subtitle = "The birth weights of children born to mothers who smoke appear to \n be less than childr
    theme_bw()
```



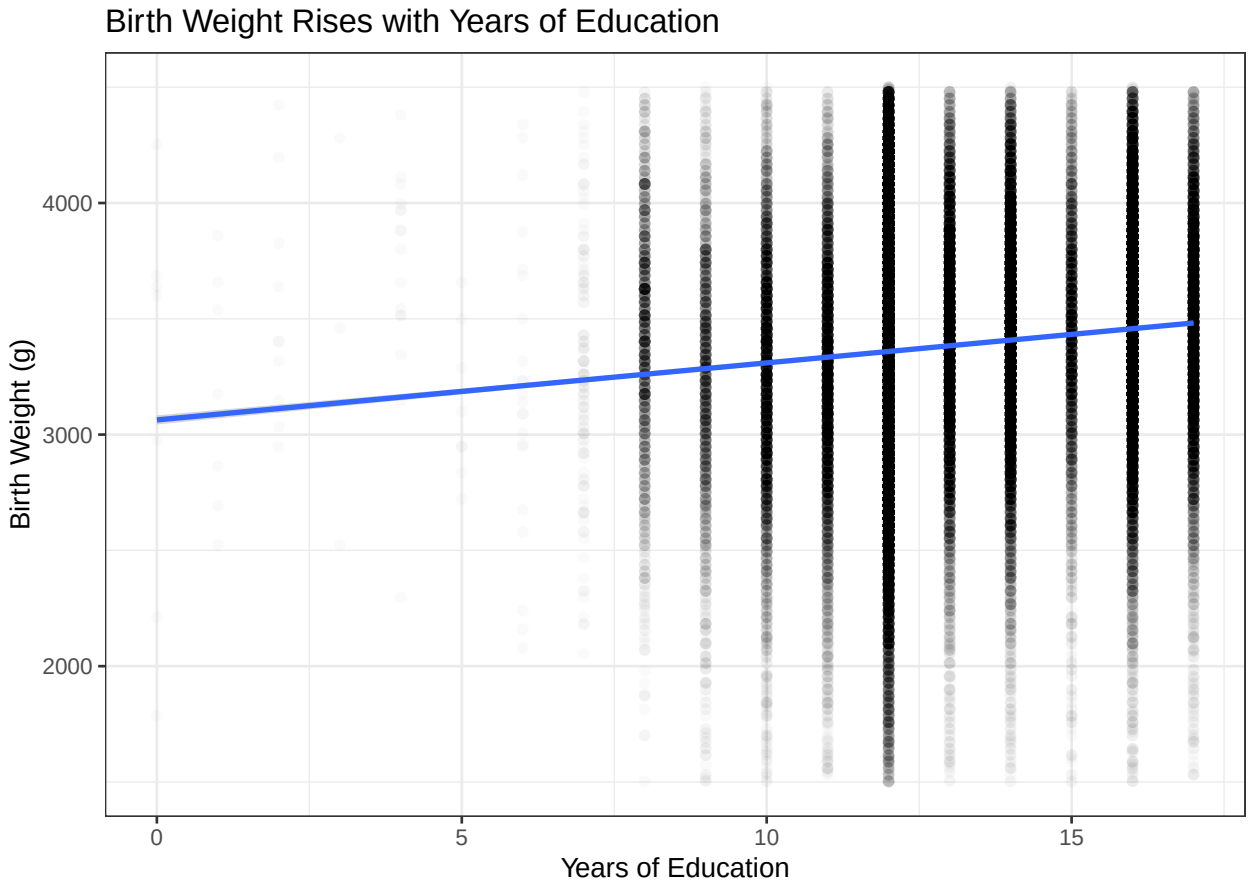
```
# creating a boxplot for smoking and years of education
ggplot(data = smoking, mapping = aes(x = as.factor(tobacco),
    y = meduc)) + geom_boxplot(fill = "tomato1") + labs(x = "Uses Tobacco (0 = no, 1 = yes)",
    y = "Years of education", title = "Years of Education by Tobacco Usage",
    subtitle = "Our data suggests that mothers who \n smoke tend to have received less education") +
    theme_bw()
```



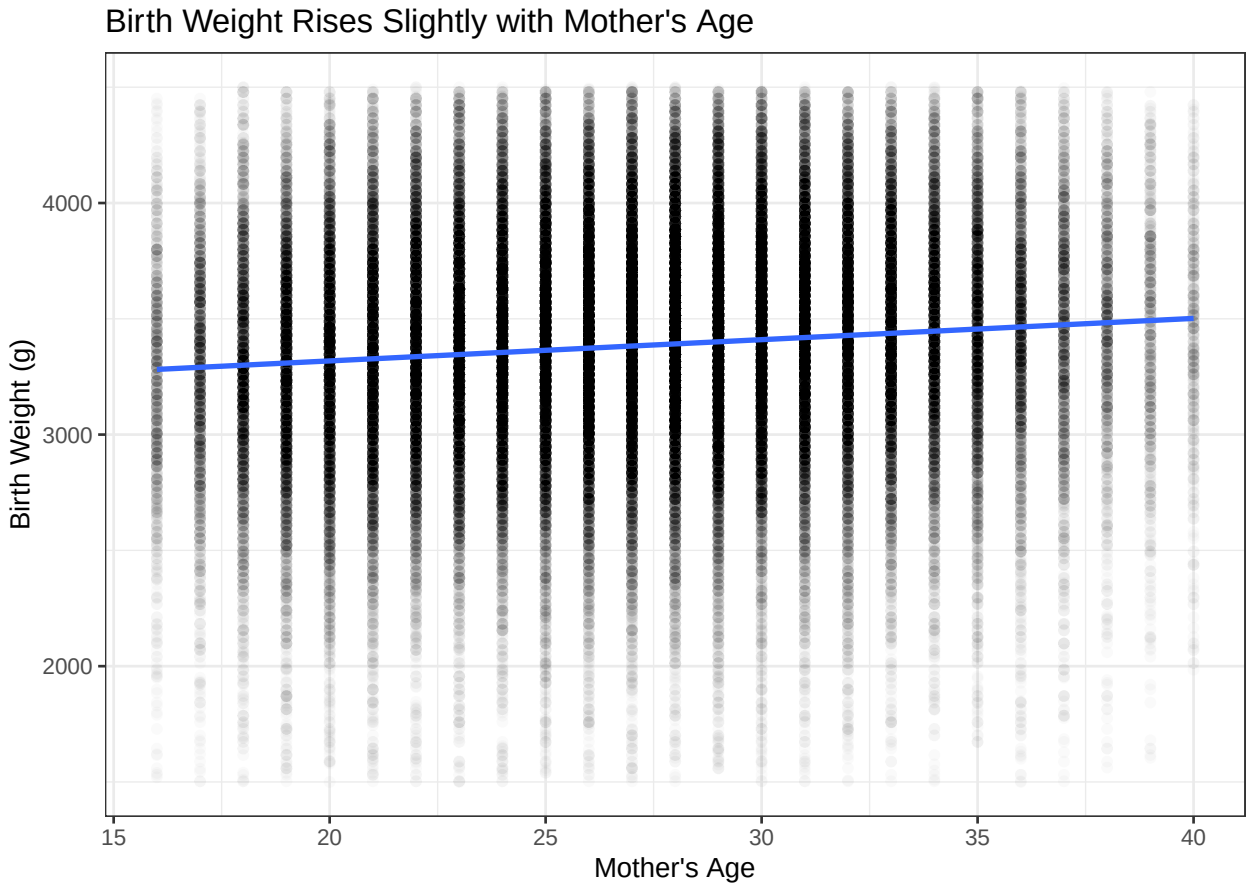

```
# creating a boxplot for smoking and mother's age
ggplot(data = smoking, mapping = aes(x = as.factor(tobacco),
    y = mage)) + geom_boxplot(fill = "mediumpurple1") + labs(x = "Uses Tobacco (0 = no, 1 = yes)",
    y = "Mother's Age", title = "Mother's Age by Tobacco Usage",
    subtitle = "Our data suggests that mothers who \n smoked during pregnancy tended to be younger") +
    theme_bw()
```




```
# creating a graph for years of education and birth weight
ggplot(data = smoking, mapping = aes(x = meduc, y = birthwgt)) +
    geom_point(alpha = 0.02) + geom_smooth(method = "lm") + labs(x = "Years of Education",
    y = "Birth Weight (g)", title = "Birth Weight Rises with Years of Education") +
    theme_bw()
```




```
# creating a graph for mother's age and birth weight
ggplot(data = smoking, mapping = aes(x = mage, y = birthwgt)) +
    geom_point(alpha = 0.02) + geom_smooth(method = "lm") + labs(x = "Mother's Age",
    y = "Birth Weight (g)", title = "Birth Weight Rises Slightly with Mother's Age") +
    theme_bw()
```



The unadjusted mean difference in birth weight would correspond to the average treatment effect (ATE) of maternal smoking during if the treatment ignorability assumption is satisfied. In other words, if we assume that there are no other differences between smoking and non-smoking mothers that could possibly affect infant birth weight, then the unadjusted mean difference in birth weight would correspond to the ATE of smoking during pregnancy on birth weight. This would also imply that there is no omitted variable bias in the model.

Looking at box plots above, it is clear that there are other predictors that are correlated with smoking activity. Additionally, these variables also appear to be somewhat related to the outcome variable, which indicates that the treatment ignorability assumption is likely not satisfied here.

(b) Assume that maternal smoking is randomly assigned conditional on the observable covariates listed above. Estimate the effect of maternal smoking on birth weight using an OLS regression with linear controls for the covariates. Report the estimated coefficient on tobacco and its standard error.

```
# running a linear regression with birth weight as the
# outcome variable and all columns as predictors

smoking_birthwgt_MLR <- lm(formula = birthwgt ~ ., data = smoking)

se_models <- starprep(smoking_birthwgt_MLR, stat = c("std.error"),
    se_type = "HC2", alpha = 0.05)

stargazer(smoking_birthwgt_MLR, se = se_models, type = "text")
```

```
##
```

```
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                birthwgt
## ------------------------------------------------
## anemia                          -4.796
##                                (17.874)
##
## diabete                         73.228***
##                                 (13.235)
##
## tobacco                        -228.073***
##                                  (4.277)
##
## alcohol                        -77.350***
##                                 (14.039)
##
## mblack                         -240.030***
##                                  (5.348)
##
## first                          -96.944***
##                                  (3.488)
##
## mage                             -0.694*
##                                  (0.368)
##
## meduc                           11.688***
##                                  (0.862)
##
## Constant                      3,362.258***
##                                  (12.076)
##
## ------------------------------------------------
## Observations                     94,173
## R2                                0.072
## Adjusted R2                       0.072
## Residual Std. Error    484.733 (df = 94164)
## F Statistic         909.176*** (df = 8; 94164)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

With the new model, the coefficient for smoking is -228.0731 grams with a standard error of 4.277 grams. While holding everything else constant, our model predicts that mothers who smoke during pregnancy will give birth to children that are 228 grams lighter than mothers who don't smoke during pregnancy.

The coefficient is not quite as strong as before (-244.5), while the standard errors are similar to the previous model (4.150)

(c) Use the exact matching estimator to estimate the effect of maternal smoking on birth weight.

For simplicity, consider the following covariates in your matching estimator: create a 0-1 indicator for mother's age (=1 if mage>=34), and a 0-1 indicator for mother's education (1 if meduc>=16), mother's race (mblack), and alcohol consumption indicator (alcohol). These 4 covariates will create 16 cells. Report the estimated average treatment effect of smoking on birthweight using the exact matching estimator and its linear regression analogue.

```
# create indicators for mothers age and years of education
smoking_indicators <- smoking %>%
    mutate(age_over_34 = ifelse(mage >= 34, 1, 0), edu_over_16 = ifelse(meduc >=
        16, 1, 0)) %>%
    dplyr::select(-c(mage, meduc, anemia, diabete, first))

# Create a new variable that combines the values of the 4
# covariates of interest
smoking_indicators$factor_var <- paste0(smoking_indicators$mblack,
    smoking_indicators$alcohol, smoking_indicators$age_over_34,
    smoking_indicators$edu_over_16)

# Convert the new variable to a factor with 16 levels
smoking_indicators$factor_var <- factor(smoking_indicators$factor_var,
    levels = unique(smoking_indicators$factor_var))
```

```
#multivariate matching estimator. create a table with the exact matches for smokers vs non-smokers and

match_table <- smoking_indicators %>%
  group_by(factor_var, tobacco) %>%
  summarise(n_obs = n(),
            Y_mean= mean(birthwgt, na.rm = T))%>% #Calculate number of observations and Y mean by X by
  ungroup()%>%
  mutate(total_obs = sum(n_obs))%>% #Calculate total number of observations
  group_by(tobacco)%>%
  mutate(total_obs_tobacco = sum(n_obs))%>% #Calculate total number of observations by treatment cells
  group_by(factor_var)%>%
  mutate(Y_diff = lead(Y_mean)-Y_mean,
         W_ATE = sum(n_obs)/total_obs,
         W_ATT = lead(n_obs)/lead(total_obs_tobacco))%>% #Calculate difference in outcome and ATE and A
  ungroup()%>%
  mutate(ATE=sum(W_ATE*Y_diff, na.rm= T),
         ATT=sum(W_ATT*Y_diff, na.rm= T))%>% #Calculate ATE and ATT
  mutate_if(is.numeric, round, 2) #Round data

ATE <- unique(match_table$ATE)
ATE
```

```
## [1] -225.37
```

```
ATT <- unique(match_table$ATT)
ATT
```

```
## [1] -227.71
```

According to the exact match methods, the average treatment effect of maternal smoking on infant birth weight is -225.37 grams. In other words, while controlling for exact matches, our model predicts that mothers who smoke during pregnancy will give birth to infants that are 225 grams lighter than mothers who don't smoke during pregnancy.

```
# MULTIVARIATE MATCHING AS REGRESSION ESTIMATOR
reg_ate <- lm(formula = birthwgt ~ tobacco + factor_var, data = smoking_indicators)

se_models = starprep(reg_ate, stat = c("std.error"), se_type = "HC2",
    alpha = 0.05)
```

```
stargazer(reg_ate, se = se_models, type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                birthwgt
## ------------------------------------------------
## tobacco                       -226.245***
##                                (4.220)
##
## factor_var0000                -37.809***
##                                (4.535)
##
## factor_var0010                -27.450***
##                                (7.641)
##
## factor_var1010                -289.496***
##                                (24.372)
##
## factor_var0011                  3.016
##                                (8.118)
##
## factor_var1000                -279.648***
##                                (6.731)
##
## factor_var1001                -158.585***
##                                (19.275)
##
## factor_var0101                 50.702
##                                (38.568)
##
## factor_var1101                -257.007**
##                                (127.408)
##
## factor_var0100                -100.933***
##                                (20.801)
##
## factor_var0110                -140.662***
##                                (45.299)
##
## factor_var1011                -183.997***
##                                (38.696)
##
## factor_var0111                 -11.072
##                                (55.358)
##
## factor_var1100                -421.816***
##                                (30.141)
##
## factor_var1111                 -223.560
##                                (198.931)
##
## factor_var1110                -481.671***
```

```
##                                (79.522)
##
## Constant                     3,483.682***
##                                (3.982)
##
## ------------------------------------------------
## Observations                    94,173
## R2                               0.063
## Adjusted R2                      0.063
## Residual Std. Error    487.101 (df = 94156)
## F Statistic         393.603*** (df = 16; 94156)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

My coefficient using exact matching in a linear model is -226.245, which is very similar to the ATE I obtained above.

```
# I did the following at first, but some of the predictors
# didn't have coefficients.

# I believe factor_var0111, factor_var1100, factor_var1111
# and factor_var1110 were removed due to collinearity, but
# that shouldn't impact the coefficient for tobacco

reg_ate <- lm(formula = birthwgt ~ tobacco + factor_var, data = smoking_indicators)

se_models = starprep(reg_ate, stat = c("std.error"), se_type = "HC2",
    alpha = 0.05)

stargazer(reg_ate, se = se_models, type = "text")
```