Econometrics PS3 q3

Yuhuan Huang

2025-05-09

```
library("haven")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("tidyr")
library("psych")
library("skimr")
library("ggplot2")
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
data <- read_dta("ps3_2025.dta")
```

```
#Q1
last_15 <- tail(colnames(data), 15)
print(last_15)
```

```
##  [1] "cardiac"  "lung"     "diabetes" "herpes"   "chyper"   "phyper"
##  [7] "pre4000"  "preterm"  "tobacco"  "cigar"    "cigar6"   "alcohol"
## [13] "drink"    "drink5"   "wgain"
```

```r
missing_values <- list(
  cardiac = c(8,9),
  lung = c(8,9),
  diabetes = c(8,9),
  herpes = c(8,9),
  chyper = c(8,9),
  phyper = c(8,9),
  pre4000 = c(8,9),
  preterm = c(8,9),
  tobacco = 9,
  cigar = 99,
  cigar6 = 6,
  alcohol = 5,
  drink = 99,
  drink5 = 5,
  wgain = 99
)


clean_data <- data %>%
  mutate(across(
    all_of(names(missing_values)),
    ~ ifelse(.x %in% missing_values[[cur_column()]], NA, .x)
  ))


clean_data <- clean_data %>% drop_na()

describe(clean_data)
```

```
##           vars     n    mean     sd median trimmed    mad    min     max   range
## rectype      1 92789    1.26   0.44      1    1.20   0.00   1.00    2.00    1.00
## pldel3       2 92789    1.02   0.13      1    1.00   0.00   1.00    2.00    1.00
## birattnd     3 92789    1.20   0.56      1    1.05   0.00   1.00    5.00    4.00
## cntocpop     4 92789    1.44   1.14      2    1.43   1.48   0.00    3.00    3.00
## stresfip     5 92789   41.74   2.15     42   42.00   0.00   0.00   55.00   55.00
## dmage        6 92789   27.77   5.70     28   27.77   5.93  12.00   49.00   37.00
## ormoth       7 92789    0.09   0.52      0    0.00   0.00   0.00    5.00    5.00
## mrace3       8 92789    1.26   0.65      1    1.07   0.00   1.00    3.00    2.00
## dmeduc       9 92789   13.21   2.27     12   13.24   1.48   0.00   17.00   17.00
## dmar        10 92789    1.25   0.43      1    1.19   0.00   1.00    2.00    1.00
## adequacy    11 92789    1.30   0.55      1    1.19   0.00   1.00    3.00    2.00
## nlbnl       12 92789    0.97   1.15      1    0.78   1.48   0.00   12.00   12.00
## dlivord     13 92789    1.99   1.17      2    1.79   1.48   1.00   14.00   13.00
## dtotord     14 92789    2.42   1.51      2    2.18   1.48   1.00   24.00   23.00
## totord9     15 92789    2.41   1.45      2    2.18   1.48   1.00    8.00    7.00
## monpre      16 92789    2.50   1.32      2    2.31   1.48   0.00    9.00    9.00
## nprevist    17 92789   11.16   3.52     12   11.21   2.97   0.00   49.00   49.00
## disllb      18 92789  350.52 362.32     73  339.26  83.03   0.00  777.00  777.00
## isllb10     19 92789    3.32   3.19      4    3.05   5.93   0.00    9.00    9.00
## dfage       20 92789   30.07   6.41     30   29.90   5.93  13.00   78.00   65.00
## orfath      21 92789    0.10   0.53      0    0.00   0.00   0.00    5.00    5.00
## dfeduc      22 92789   13.28   2.33     12   13.28   1.48   0.00   17.00   17.00
```

```
##              vars  n   mean    sd median trimmed   mad    min     max   range
## birmon         23 92789   6.48  3.39      7    6.48  4.45   1.00   12.00   11.00
## weekday        24 92789   4.05  1.88      4    4.06  2.97   1.00    7.00    6.00
## dgestat        25 92789  39.15  2.44     39   39.33  1.48  17.00   47.00   30.00
## csex           26 92789   1.49  0.50      1    1.48  0.00   1.00    2.00    1.00
## dbrwt          27 92789 3359.95 588.63   3398 3384.23 510.01 217.25 6067.00 5849.75
## dplural        28 92789   1.03  0.18      1    1.00  0.00   1.00    4.00    3.00
## omaps          29 92789   8.12  1.26      8    8.36  1.48   0.00   10.01   10.01
## fmaps          30 92789   9.00  0.70      9    9.04  0.00   0.00   10.00   10.00
## clingest       31 92789  39.11  2.05     40   39.38  1.48  17.00   44.00   27.00
## delmeth5       32 92789   1.55  1.01      1    1.33  0.00   1.00    5.00    4.00
## anemia         33 92789   1.99  0.10      2    2.00  0.00   1.00    2.00    1.00
## cardiac        34 92789   1.99  0.08      2    2.00  0.00   1.00    2.00    1.00
## lung           35 92789   1.99  0.08      2    2.00  0.00   1.00    2.00    1.00
## diabetes       36 92789   1.97  0.16      2    2.00  0.00   1.00    2.00    1.00
## herpes         37 92789   1.99  0.08      2    2.00  0.00   1.00    2.00    1.00
## chyper         38 92789   1.99  0.09      2    2.00  0.00   1.00    2.00    1.00
## phyper         39 92789   1.97  0.17      2    2.00  0.00   1.00    2.00    1.00
## pre4000        40 92789   1.99  0.12      2    2.00  0.00   1.00    2.00    1.00
## preterm        41 92789   1.99  0.12      2    2.00  0.00   1.00    2.00    1.00
## tobacco        42 92789   1.84  0.37      2    1.93  0.00   1.00    2.00    1.00
## cigar          43 92789   1.91  5.29      0    0.41  0.00   0.00   98.00   98.00
## cigar6         44 92789   0.35  0.86      0    0.10  0.00   0.00    5.00    5.00
## alcohol        45 92789   1.99  0.10      2    2.00  0.00   1.00    2.00    1.00
## drink          46 92789   0.03  0.64      0    0.00  0.00   0.00   91.00   91.00
## drink5         47 92789   0.02  0.23      0    0.00  0.00   0.00    4.00    4.00
## wgain          48 92789  30.38 11.89     30   29.97 10.38   0.00   98.00   98.00
##              skew kurtosis   se
## rectype      1.08    -0.82 0.00
## pldel3       7.23    50.31 0.00
## birattnd     3.41    13.55 0.00
## cntocpop    -0.05    -1.42 0.00
## stresfip   -10.94   151.58 0.01
## dmage        0.00    -0.50 0.02
## ormoth       6.73    49.93 0.00
## mrace3       2.22     3.03 0.00
## dmeduc      -0.09     0.12 0.01
## dmar         1.15    -0.67 0.00
## adequacy     1.68     1.88 0.00
## nlbnl        1.90     6.40 0.00
## dlivord      1.93     6.66 0.00
## dtotord      1.75     5.54 0.00
## totord9      1.33     1.92 0.00
## monpre       1.75     3.99 0.00
## nprevist     0.40     4.62 0.01
## disllb       0.32    -1.88 1.19
## isllb10      0.29    -1.32 0.01
## dfage        0.39     0.59 0.02
## orfath       6.57    47.52 0.00
## dfeduc      -0.05     0.17 0.01
## birmon       0.00    -1.17 0.01
## weekday     -0.03    -1.14 0.01
## dgestat     -1.67     8.68 0.01
## csex         0.05    -2.00 0.00
## dbrwt       -0.72     2.33 1.93
```

```
## dplural    6.67    50.54 0.00
## omaps     -2.76    10.20 0.00
## fmaps     -4.70    44.51 0.00
## clingest  -2.91    15.48 0.01
## delmeth5   1.51     0.71 0.00
## anemia    -9.87    95.32 0.00
## cardiac  -12.06   143.47 0.00
## lung     -11.61   132.88 0.00
## diabetes  -5.97    33.67 0.00
## herpes   -12.70   159.23 0.00
## chyper   -11.48   129.87 0.00
## phyper    -5.47    27.88 0.00
## pre4000   -8.05    62.84 0.00
## preterm   -8.14    64.31 0.00
## tobacco   -1.86     1.45 0.00
## cigar      3.44    15.04 0.02
## cigar6     2.42     4.74 0.00
## alcohol  -10.11   100.21 0.00
## drink     75.38  8747.75 0.00
## drink5    13.63   201.83 0.00
## wgain      0.51     1.36 0.04
```

```
#Q2

data1 <- clean_data %>%
  mutate(nonsmoking = if_else((tobacco == 2 & cigar6 == 0),1,0))

data1 <- data1 %>%
  mutate(smoking = if_else((nonsmoking == 0),1,0))

grouping_data2 <- data1 %>%
  group_by(smoking) %>%
  summarise(
    mean_apgar1 = mean(omaps),
    mean_apgar5 = mean(fmaps),
    mean_birthweight = mean(dbrwt)
  )


data2_diff <- grouping_data2 %>%
  summarise(
    apgar1_diff = diff(mean_apgar1),
    apgar5_diff = diff(mean_apgar5),
    birthweight_diff = diff(mean_birthweight)
  )

print(grouping_data2)
```

```
## # A tibble: 2 x 4
##   smoking mean_apgar1 mean_apgar5 mean_birthweight
##     <dbl>       <dbl>       <dbl>            <dbl>
## 1       0        8.12        9.01            3412.
## 2       1        8.10        8.97            3087.
```

```r
print(data2_diff)
```

```
## # A tibble: 1 x 3
##   apgar1_diff apgar5_diff birthweight_diff
##         <dbl>       <dbl>            <dbl>
## 1     -0.0172     -0.0394            -324.
```

```r
#data1 %>%
#  group_by(smoking) %>%
#  skim()

predetermined_variables <- c("stresfip","ormoth","mrace3","orfath","birmon","weekday",
                             "csex","dplural","anemia","cardiac","diabetes","herpes"
                             ,"chyper","phyper","pre4000","preterm")

print(predetermined_variables)
```

```
##  [1] "stresfip" "ormoth"   "mrace3"   "orfath"   "birmon"   "weekday"
##  [7] "csex"     "dplural"  "anemia"   "cardiac"  "diabetes" "herpes"
## [13] "chyper"   "phyper"   "pre4000"  "preterm"
```

```r
formular_str0 <- paste("dbrwt ~ smoking + ",paste(predetermined_variables, collapse="+"))
formular_obj0 <- as.formula(formular_str0)

model <- lm(formular_obj0 , data = data1)
summary(model)
```

```
##
## Call:
## lm(formula = formular_obj0, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3318.7  -299.4    17.4   335.4  2832.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4182.5703   101.2764  41.299  < 2e-16 ***
## smoking     -328.3412     4.8166 -68.169  < 2e-16 ***
## stresfip       1.5598     0.8186   1.905  0.05673 .
## ormoth       -34.0082     4.4541  -7.635 2.27e-14 ***
## mrace3      -131.3214     2.6966 -48.700  < 2e-16 ***
## orfath       -25.0806     4.3758  -5.732 9.98e-09 ***
## birmon        -1.4090     0.5188  -2.716  0.00661 **
## weekday       -0.6027     0.9357  -0.644  0.51951
## csex        -124.0941     3.5221 -35.233  < 2e-16 ***
## dplural     -894.1438    10.0434 -89.028  < 2e-16 ***
## anemia        20.7495    17.7642   1.168  0.24279
## cardiac       39.3873    21.5297   1.829  0.06734 .
## diabetes    -116.3330    11.1199 -10.462  < 2e-16 ***
## herpes       -36.2097    22.6320  -1.600  0.10962
## chyper       239.1731    20.5597  11.633  < 2e-16 ***
```

```
## phyper       200.9766    10.2665  19.576  < 2e-16 ***
## pre4000     -496.0495    14.6267 -33.914  < 2e-16 ***
## preterm      375.3339    14.7839  25.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 536.2 on 92771 degrees of freedom
## Multiple R-squared:  0.1704, Adjusted R-squared:  0.1703
## F-statistic:  1121 on 17 and 92771 DF,  p-value: < 2.2e-16
```

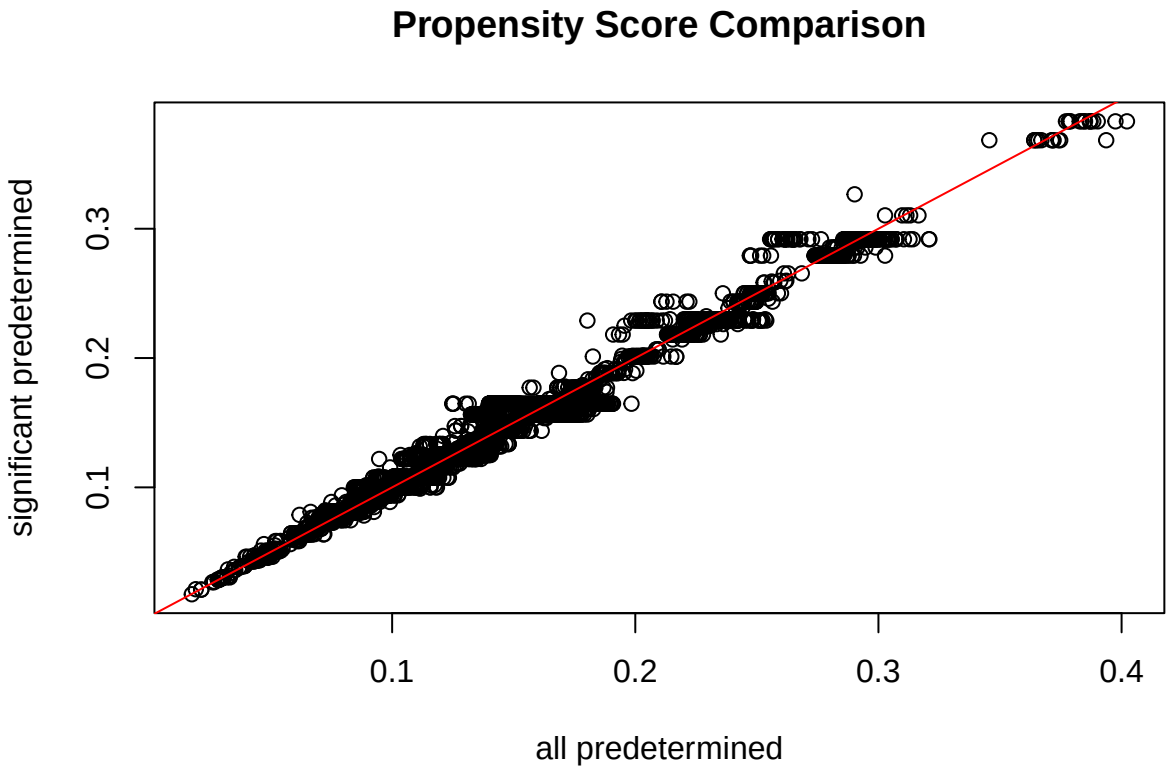
```r
#Q3

formula_str <- paste("smoking ~", paste(predetermined_variables, collapse="+"))
formula_obj <- as.formula(formula_str)

#logit regression
logit_model <- glm(formula_obj, data = data1, family = binomial())
summary(logit_model)
```

```
##
## Call:
## glm(formula = formula_obj, family = binomial(), data = data1)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.2974638  0.5525892  -5.967 2.41e-09 ***
## stresfip     0.0437932  0.0059754   7.329 2.32e-13 ***
## ormoth      -0.1798075  0.0268464  -6.698 2.12e-11 ***
## mrace3      -0.0313170  0.0138872  -2.255   0.0241 *
## orfath       0.0264586  0.0236169   1.120   0.2626
## birmon      -0.0004963  0.0026478  -0.187   0.8513
## weekday     -0.0081692  0.0047700  -1.713   0.0868 .
## csex        -0.0066725  0.0179770  -0.371   0.7105
## dplural     -0.2454643  0.0567684  -4.324 1.53e-05 ***
## anemia      -0.4106429  0.0810633  -5.066 4.07e-07 ***
## cardiac      0.1409225  0.1142135   1.234   0.2173
## diabetes    -0.0434160  0.0565909  -0.767   0.4430
## herpes      -0.0997803  0.1118386  -0.892   0.3723
## chyper       0.1630214  0.1095801   1.488   0.1368
## phyper       0.4938301  0.0620781   7.955 1.79e-15 ***
## pre4000      0.5747389  0.0913262   6.293 3.11e-10 ***
## preterm     -0.7402960  0.0619075 -11.958  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 81479  on 92788  degrees of freedom
## Residual deviance: 81052  on 92772  degrees of freedom
## AIC: 81086
##
## Number of Fisher Scoring iterations: 4
```

```r
logit_vals <- predict(logit_model, type = "link")
data1$pscore <- predict(logit_model, type = "response")
#print(data1$pscore)

sig_predetermined_vars <- c("stresfip","ormoth","mrace3","dplural","anemia",
                            "phyper","pre4000","preterm")

formula_str2 <- paste("smoking ~", paste(sig_predetermined_vars, collapse="+"))
formula_obj2 <- as.formula(formula_str2)

logit_model2 <- glm(formula_obj2, data = data1, family = binomial())
summary(logit_model2)
```

```
##
## Call:
## glm(formula = formula_obj2, family = binomial(), data = data1)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.023686   0.392164  -7.710 1.26e-14 ***
## stresfip     0.043790   0.005976   7.327 2.35e-13 ***
## ormoth      -0.160954   0.020913  -7.696 1.40e-14 ***
## mrace3      -0.031110   0.013883  -2.241    0.025 *
## dplural     -0.246767   0.056759  -4.348 1.38e-05 ***
## anemia      -0.409576   0.081020  -5.055 4.30e-07 ***
## phyper       0.492781   0.062035   7.944 1.96e-15 ***
## pre4000      0.573556   0.091279   6.284 3.31e-10 ***
## preterm     -0.737435   0.061865 -11.920  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 81479  on 92788  degrees of freedom
## Residual deviance: 81061  on 92780  degrees of freedom
## AIC: 81079
##
## Number of Fisher Scoring iterations: 4
```

```r
logit_vals2 <- predict(logit_model2, type = "link")
data1$pscore_sig <- predict(logit_model2, type = "response")
#print(data1$pscore_sig)

#compare the two propensity scores
plot(data1$pscore, data1$pscore_sig,
     xlab = "all predetermined", ylab = "significant predetermined", main = "Propensity Score Comparison
abline(0, 1, col = "red")
```




```
#include propensity score as a covariate
model_covariate <- lm(dbrwt ~ smoking + pscore_sig, data = data1)
summary(model_covariate)
```

```
##
## Call:
## lm(formula = dbrwt ~ smoking + pscore_sig, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3275.8  -311.1    23.9   355.7  2650.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3291.388     12.517 262.956   <2e-16 ***
## smoking     -327.879      5.177 -63.339   <2e-16 ***
## pscore_sig   757.360     77.689   9.749   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 576.2 on 92786 degrees of freedom
## Multiple R-squared:  0.04174,    Adjusted R-squared:  0.04172
## F-statistic:  2021 on 2 and 92786 DF,  p-value: < 2.2e-16
```

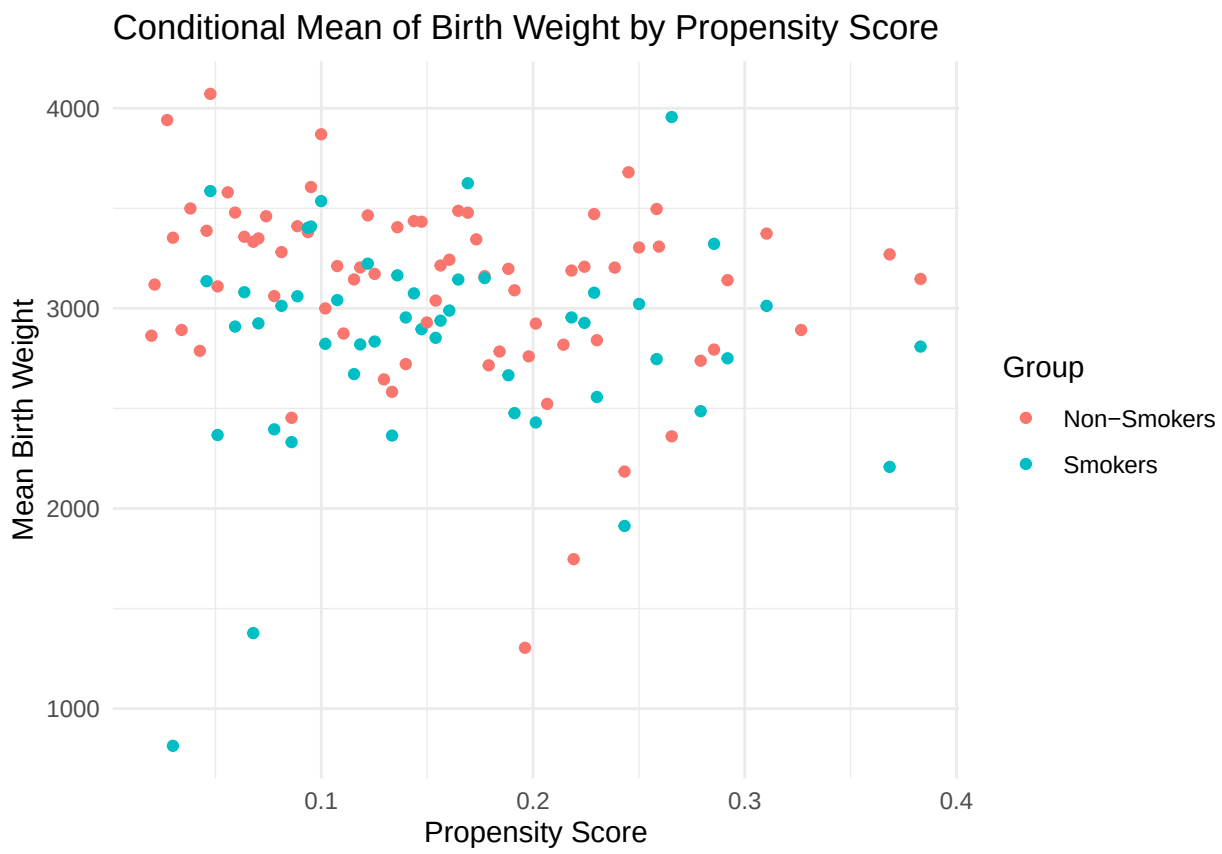
```r
#weighting with propensity score
sum_smoking <- 0
sum_nonsmkoing <- 0
for (i in 1:nrow(data1)){
  p <- data1$pscore_sig[i]
  y <- data1$dbrwt[i]
  d <- data1$smoking[i]

  if (is.na(p) || is.na(y) || is.na(d) || p == 0 || p == 1) {
    next
  }

  if(d==1){
    sum_smoking <- sum_smoking + y/p
  }else if(d==0){
    sum_nonsmkoing <- sum_nonsmkoing + y/(1-p)
  }
}

weighted_ATE = 1/(nrow(data1)) * (sum_smoking - sum_nonsmkoing)
print(weighted_ATE)
```

```
##        17
## -321.7174
```

```r
#blocking on propensity score

data1 <- data1 %>%
  mutate(bin = cut(pscore_sig, breaks = 100, include.lowest = TRUE, labels = FALSE))

data_smokers <- data1 %>%
  filter(smoking == 1) %>%
  group_by(bin) %>%
  summarize(pscore_mean = mean(pscore_sig), y1 = mean(dbrwt), .groups = "drop")

data_nonsmokers <- data1 %>%
  filter(smoking == 0) %>%
  group_by(bin) %>%
  summarize(pscore_mean = mean(pscore_sig), y0 = mean(dbrwt), .groups = "drop")

bin_estimates <- data_nonsmokers %>%
  left_join(data_smokers, by = "bin", suffix = c("_0", "_1")) %>%
  mutate(tau_b = y1 - y0)

bin_weights <- data1 %>%
  group_by(bin) %>%
  summarize(n_bin = n(), .groups = "drop") %>%
  mutate(weight = n_bin / sum(n_bin))

bin_estimates <- bin_estimates %>%
  left_join(bin_weights, by = "bin") %>%
  mutate(w_tau_b = weight * tau_b)
```

```r
tau_sb <- sum(bin_estimates$w_tau_b, na.rm = TRUE)
#print(tau_sb)


plot_data <- bin_estimates %>%
  select(bin, pscore_mean = pscore_mean_0, y0, y1) %>%
  pivot_longer(cols = c(y0, y1), names_to = "group", values_to = "y_mean") %>%
  mutate(group = ifelse(group == "y0", "Non-Smokers", "Smokers"))

ggplot(plot_data, aes(x = pscore_mean, y = y_mean, color = group)) +
  geom_point()+
  labs(
    x = "Propensity Score",
    y = "Mean Birth Weight",
    title = "Conditional Mean of Birth Weight by Propensity Score",
    color = "Group"
  ) +
  theme_minimal()
```

```
## Warning: Removed 23 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```r
# Q5
data3 <- data1
```

```r
data3 <- data3 %>%
  mutate(lowbw = if_else(dbrwt < 2500, 1, 0))

data3 <- data3 %>%
  mutate(bin = cut(pscore_sig, breaks = 100, include.lowest = TRUE, labels = FALSE))

data_smokers3 <- data3 %>%
  filter(smoking == 1) %>%
  group_by(bin) %>%
  summarize(pscore_mean = mean(pscore_sig),
            y1 = mean(lowbw),
            .groups = "drop")

data_nonsmokers3 <- data3 %>%
  filter(smoking == 0) %>%
  group_by(bin) %>%
  summarize(pscore_mean = mean(pscore_sig),
            y0 = mean(lowbw),
            .groups = "drop")

bin_estimates3 <- data_nonsmokers3 %>%
  left_join(data_smokers3, by = "bin", suffix = c("_0", "_1")) %>%
  mutate(tau_b = y1 - y0)


bin_weights3 <- data3 %>%
  group_by(bin) %>%
  summarize(n_bin = n(), .groups = "drop") %>%
  mutate(weight = n_bin / sum(n_bin))

bin_estimates3 <- bin_estimates3 %>%
  left_join(bin_weights3, by = "bin") %>%
  mutate(w_tau_b = weight * tau_b)

tau_sb3 <- sum(bin_estimates3$w_tau_b, na.rm = TRUE)
#print(tau_sb3)


plot_data3 <- bin_estimates3 %>%
  select(bin, pscore_mean = pscore_mean_0, y0, y1) %>%
  pivot_longer(cols = c(y0, y1), names_to = "group", values_to = "y_mean") %>%
  mutate(group = ifelse(group == "y0", "Non-Smokers", "Smokers"))

ggplot(plot_data3, aes(x = pscore_mean, y = y_mean, color = group)) +
  geom_point() +
  labs(
    x = "Estimated Propensity Score",
    y = "Proportion of Low Birth Weight",
    title = "Conditional Probability of Low Birth Weight by Propensity Score",
    color = "Group"
  ) +
  theme_minimal()
```

```
## Warning: Removed 23 rows containing missing values or values outside the scale range
## (`geom_point()`).
```