HW5

Daniel Koslovsky

2024-11-20

##This submission is my work alone and complies with the 31202 integrity policy. Add you initials to indicate your agreement: DK

#HW5

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tinytex)

# Question 2
pset5_simulated_data <- read.csv("pset5_simulated_data.csv")

# a The people who gain from this treatment are the people
# who receive the treatment and do not contract the
# disease.
marginal_distributions <- pset5_simulated_data |>
    group_by(treatment) |>
    summarize(disease = mean(disease)) |>
    mutate(no_disease = 1 - disease)

print(marginal_distributions)
```

```
## # A tibble: 2 x 3
##   treatment disease no_disease
##       <int>   <dbl>      <dbl>
## 1         0  0.175       0.825
## 2         1  0.0506      0.949
```

The bounds for the fraction of people who benefit from the treatment are [.124, .175]

#b The people who lose from the treatment are the people who received the treatment and still contracted the disease. The bounds for the fraction of people who lose from the treatment are [0, .0506]

```r
# Question 3
library(wooldridge)
data("bwght2")

# a Checking to see if it is reasonable to assume that the
# observations missing cigs are missing at random
missing_cig <- bwght2 |>
    filter(is.na(cigs)) |>
    summarise(mean_age = mean(mage), sd_age = sd(mage), mean_bwght = mean(bwght),
        sd_bwght = sd(bwght))
print(missing_cig)
```

```
##   mean_age   sd_age mean_bwght sd_bwght
## 1 29.95455 5.314661   3274.573 657.2657
```

```r
no_missing_cig <- bwght2 |>
    filter(!is.na(cigs)) |>
    summarise(mean_age = mean(mage), sd_age = sd(mage), mean_bwght = mean(bwght),
        sd_bwght = sd(bwght))
print(no_missing_cig)
```

```
##   mean_age   sd_age mean_bwght sd_bwght
## 1 29.53252 4.734722   3409.206 570.2629
```

```r
# Dropping all missing observations for cigs as missing at
# random assumption seems reasonable since mean and
# standard deviation of age and bwght variables seem
# similar for observations with and without missing cigs
# observations.
bwght2 <- bwght2 |>
    filter(!is.na(cigs))

bwght2_smoker <- bwght2 |>
    mutate(smoker = case_when(cigs > 0 ~ 1, cigs == 0 ~ 0))

summary(select(bwght2_smoker, mage, mblck, bwght, smoker))
```

```
##       mage           mblck             bwght          smoker
##  Min.   :16.00   Min.   :0.00000   Min.   : 360   Min.   :0.00000
##  1st Qu.:26.00   1st Qu.:0.00000   1st Qu.:3081   1st Qu.:0.00000
##  Median :29.00   Median :0.00000   Median :3430   Median :0.00000
##  Mean   :29.53   Mean   :0.05691   Mean   :3409   Mean   :0.08537
##  3rd Qu.:33.00   3rd Qu.:0.00000   3rd Qu.:3771   3rd Qu.:0.00000
##  Max.   :44.00   Max.   :1.00000   Max.   :5204   Max.   :1.00000
```

```r
# b
model_no_controls <- lm(bwght ~ smoker, data = bwght2_smoker)
model_age <- lm(bwght ~ smoker + mage, data = bwght2_smoker)
model_age_squared <- lm(bwght ~ smoker + mage + magesq, data = bwght2_smoker)
fsm_model <- lm(bwght ~ smoker + as.factor(mage), data = bwght2_smoker)

summary(model_no_controls)
```

```
##
## Call:
## lm(formula = bwght ~ smoker, data = bwght2_smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3067.41  -327.41    12.59   362.59  1776.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3427.41      14.29 239.767  < 2e-16 ***
## smoker       -213.24      48.93  -4.358 1.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 567.3 on 1720 degrees of freedom
## Multiple R-squared:  0.01092,    Adjusted R-squared:  0.01035
## F-statistic:    19 on 1 and 1720 DF,  p-value: 1.387e-05
```

```
summary(model_age)
```

```
##
## Call:
## lm(formula = bwght ~ smoker + mage, data = bwght2_smoker)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3079.8  -325.6     7.6   357.5  1775.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3317.863     87.086  38.099  < 2e-16 ***
## smoker      -208.148     49.079  -4.241 2.34e-05 ***
## mage           3.695      2.897   1.275    0.202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 567.2 on 1719 degrees of freedom
## Multiple R-squared:  0.01186,    Adjusted R-squared:  0.01071
## F-statistic: 10.31 on 2 and 1719 DF,  p-value: 3.524e-05
```

```
summary(model_age_squared)
```

```
##
## Call:
## lm(formula = bwght ~ smoker + mage + magesq, data = bwght2_smoker)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3092.6  -326.1    13.0   351.4  1749.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 2301.5348   390.3192   5.897 4.46e-09 ***
## smoker      -199.1605    49.1072  -4.056 5.22e-05 ***
## mage          73.9028    26.4447   2.795  0.00525 **
## magesq        -1.1825     0.4427  -2.671  0.00764 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.2 on 1718 degrees of freedom
## Multiple R-squared:  0.01594,    Adjusted R-squared:  0.01423
## F-statistic: 9.279 on 3 and 1718 DF,  p-value: 4.361e-06
```

```
summary(fsm_model)
```

```
##
## Call:
## lm(formula = bwght ~ smoker + as.factor(mage), data = bwght2_smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3091.29  -323.43    12.78   354.77  1740.47
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3743.00     327.94  11.414  < 2e-16 ***
## smoker              -200.09      49.68  -4.028 5.89e-05 ***
## as.factor(mage)17   -874.64     464.08  -1.885   0.0596 .
## as.factor(mage)18   -354.68     374.05  -0.948   0.3431
## as.factor(mage)19   -737.07     374.21  -1.970   0.0490 *
## as.factor(mage)20   -455.16     349.77  -1.301   0.1933
## as.factor(mage)21   -298.08     343.58  -0.868   0.3858
## as.factor(mage)22   -424.59     339.51  -1.251   0.2113
## as.factor(mage)23   -438.75     335.60  -1.307   0.1913
## as.factor(mage)24   -292.79     336.03  -0.871   0.3837
## as.factor(mage)25   -337.95     333.70  -1.013   0.3113
## as.factor(mage)26   -314.54     332.36  -0.946   0.3441
## as.factor(mage)27   -284.22     332.53  -0.855   0.3928
## as.factor(mage)28   -296.97     331.03  -0.897   0.3698
## as.factor(mage)29   -327.77     330.90  -0.991   0.3221
## as.factor(mage)30   -279.47     331.31  -0.844   0.3990
## as.factor(mage)31   -301.94     331.58  -0.911   0.3626
## as.factor(mage)32   -267.75     331.72  -0.807   0.4197
## as.factor(mage)33   -291.71     332.79  -0.877   0.3809
## as.factor(mage)34   -255.80     334.58  -0.765   0.4447
## as.factor(mage)35   -343.35     334.90  -1.025   0.3054
## as.factor(mage)36   -279.87     337.86  -0.828   0.4076
## as.factor(mage)37   -302.40     338.97  -0.892   0.3725
## as.factor(mage)38   -403.56     339.73  -1.188   0.2350
## as.factor(mage)39   -250.33     350.62  -0.714   0.4754
## as.factor(mage)40   -324.52     350.62  -0.926   0.3548
## as.factor(mage)41   -671.20     414.82  -1.618   0.1058
## as.factor(mage)42   -565.64     464.08  -1.219   0.2231
## as.factor(mage)43   -488.00     518.52  -0.941   0.3468
## as.factor(mage)44   -591.00     463.78  -1.274   0.2027
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 568 on 1692 degrees of freedom
## Multiple R-squared:  0.02458,    Adjusted R-squared:  0.007866
## F-statistic: 1.471 on 29 and 1692 DF,  p-value: 0.05112
```

```
# c
bwght_weighted <- bwght2_smoker |>
    group_by(mage) |>
    mutate(p_hat = sum(smoker)/n(), w_ATE = (smoker * (1/p_hat) +
        (1 - smoker) * (1/(1 - p_hat))))

matching_no_controls <- lm(bwght ~ smoker, data = bwght_weighted,
    weight = w_ATE)
matching_age <- lm(bwght ~ smoker + mage, data = bwght_weighted,
    weight = w_ATE)
matching_age_squared <- lm(bwght ~ smoker + mage + magesq, data = bwght_weighted,
    weight = w_ATE)
fsm_matching <- lm(bwght ~ smoker + as.factor(mage), data = bwght_weighted,
    weight = w_ATE)

summary(matching_no_controls)
```

```
##
## Call:
## lm(formula = bwght ~ smoker, data = bwght_weighted, weights = w_ATE)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -6190.1  -368.0    10.3   391.9  3405.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3430.00      18.81 182.349  < 2e-16 ***
## smoker       -192.52      26.60  -7.237 6.96e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 768.2 on 1666 degrees of freedom
##   (54 observations deleted due to missingness)
## Multiple R-squared:  0.03048,    Adjusted R-squared:  0.0299
## F-statistic: 52.38 on 1 and 1666 DF,  p-value: 6.957e-13
```

```
summary(matching_age)
```

```
##
## Call:
## lm(formula = bwght ~ smoker + mage, data = bwght_weighted, weights = w_ATE)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -6271.1  -367.7    10.3   390.2  3433.3
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3344.357     89.107  37.532  < 2e-16 ***
## smoker      -192.523     26.602  -7.237 6.96e-13 ***
## mage           2.926      2.975   0.983    0.326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 768.2 on 1665 degrees of freedom
##   (54 observations deleted due to missingness)
## Multiple R-squared:  0.03104,    Adjusted R-squared:  0.02988
## F-statistic: 26.67 on 2 and 1665 DF,  p-value: 3.964e-12
```

```
summary(matching_age_squared)
```

```
##
## Call:
## lm(formula = bwght ~ smoker + mage + magesq, data = bwght_weighted,
##     weights = w_ATE)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -6221.9  -358.4    12.4   392.3  3446.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2703.4188   417.7814   6.471 1.28e-10 ***
## smoker      -192.5229    26.5901  -7.240 6.81e-13 ***
## mage          47.8237    28.7471   1.664   0.0964 .
## magesq        -0.7679     0.4890  -1.570   0.1165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 767.9 on 1664 degrees of freedom
##   (54 observations deleted due to missingness)
## Multiple R-squared:  0.03248,    Adjusted R-squared:  0.03073
## F-statistic: 18.62 on 3 and 1664 DF,  p-value: 7.015e-12
```

```
summary(fsm_matching)
```

```
##
## Call:
## lm(formula = bwght ~ smoker + as.factor(mage), data = bwght_weighted,
##     weights = w_ATE)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -5277.9  -376.8    21.0   404.5  3423.2
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2976.26     310.07   9.599  < 2e-16 ***
## smoker             -192.52      26.28  -7.327 3.68e-13 ***
```

```
## as.factor(mage)18   339.75     353.22   0.962   0.3363
## as.factor(mage)19   -68.74     353.22  -0.195   0.8457
## as.factor(mage)20   109.39     330.24   0.331   0.7405
## as.factor(mage)21   436.95     324.44   1.347   0.1782
## as.factor(mage)22   431.98     320.67   1.347   0.1781
## as.factor(mage)23   397.47     316.97   1.254   0.2100
## as.factor(mage)24   522.03     317.32   1.645   0.1001
## as.factor(mage)25   546.69     315.21   1.734   0.0830 .
## as.factor(mage)26   547.64     313.95   1.744   0.0813 .
## as.factor(mage)27   578.86     314.11   1.843   0.0655 .
## as.factor(mage)28   431.30     312.67   1.379   0.1679
## as.factor(mage)29   392.96     312.58   1.257   0.2089
## as.factor(mage)30   449.33     312.96   1.436   0.1513
## as.factor(mage)31   421.37     313.19   1.345   0.1787
## as.factor(mage)32   502.37     313.35   1.603   0.1091
## as.factor(mage)33   467.17     314.36   1.486   0.1374
## as.factor(mage)34   445.46     316.01   1.410   0.1588
## as.factor(mage)35   191.51     316.36   0.605   0.5450
## as.factor(mage)36   497.08     319.14   1.558   0.1195
## as.factor(mage)37   638.53     320.18   1.994   0.0463 *
## as.factor(mage)39   635.39     331.18   1.919   0.0552 .
## as.factor(mage)40   514.09     331.18   1.552   0.1208
## as.factor(mage)42   176.00     438.11   0.402   0.6879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 758.8 on 1643 degrees of freedom
##   (54 observations deleted due to missingness)
## Multiple R-squared:  0.0671, Adjusted R-squared:  0.05347
## F-statistic: 4.924 on 24 and 1643 DF,  p-value: 7.849e-14
```

The coefficients for smoking in the OLS estimates are larger than for the matching estimates and they change more depending on the specification. The standard errors for smoking the OLS estimares are larger than for the matching estimates.

```r
# d create list of ages in treatment group
age_treatment_list <- bwght2_smoker |>
    filter(smoker == 1) |>
    distinct(mage)

print(age_treatment_list)
```

```
##    mage
## 1    24
## 2    33
## 3    34
## 4    37
## 5    18
## 6    30
## 7    29
## 8    19
## 9    28
## 10   31
```

```
## 11   20
## 12   27
## 13   23
## 14   22
## 15   26
## 16   21
## 17   25
## 18   32
## 19   40
## 20   35
## 21   39
## 22   36
## 23   42
## 24   17
```

```r
# create list of ages in control group
age_control_list <- bwght2_smoker |>
    filter(smoker == 0) |>
    distinct(mage)

print(age_control_list)
```

```
##    mage
## 1    26
## 2    33
## 3    28
## 4    23
## 5    27
## 6    41
## 7    32
## 8    16
## 9    25
## 10   29
## 11   31
## 12   30
## 13   24
## 14   22
## 15   19
## 16   36
## 17   35
## 18   38
## 19   37
## 20   34
## 21   21
## 22   20
## 23   39
## 24   44
## 25   43
## 26   40
## 27   42
## 28   17
## 29   18
```

```r
# compare ages in treatment and control group
control_only_ages <- setdiff(age_control_list, age_treatment_list)
print(control_only_ages)
```

```
##   mage
## 1   41
## 2   16
## 3   38
## 4   44
## 5   43
```

```r
treatment_only_ages <- setdiff(age_treatment_list, age_control_list)
print(treatment_only_ages)
```

```
## [1] mage
## <0 rows> (or 0-length row.names)
```

The conditional support assumption fails. While every value of age in the treatment group has a matching value in the control group, there is no matching observation in the treatment group for the ages 16, 38, 41, 43, and 44 in the control group.

For the conditional independence assumption to be met in this case, we would need to assume that conditional on age, whether or not a baby's mother is a smoker is random. This also implies that the birth weight for the babyies whose mother's did not smoke are equal to the counterfactual birth weight that the babies whose mother's do smoke would have had if their mother did not smoke.

```r
# Question 4
pset5_moderna <- read.csv("pset5_moderna.csv")

# a
covid_incidence <- pset5_moderna |>
    group_by(treat) |>
    summarize(covid_incidence = mean(covid))

print(covid_incidence)
```

```
## # A tibble: 2 x 2
##   treat covid_incidence
##   <int>           <dbl>
## 1     0         0.0191
## 2     1         0.00134
```

```r
infection_rate_reduction <- covid_incidence[[1, 2]]/covid_incidence[[2,
    2]]
print(infection_rate_reduction)
```

```
## [1] 14.21926
```

```r
vaccine_eff <- (covid_incidence[[2, 2]] - covid_incidence[[1,
    2]])/covid_incidence[[1, 2]]
print(vaccine_eff)
```

```
## [1] -0.9296729
```

#b The infection rate in the treatment group was over 14 times smaller than in the control group. The efficacy of the vaccine in the experimental setting was -0.9297, meaning that the estimated effect of the vaccine on the infection rate is to reduce it by 92.97%.

#c The change in the infection rate during the beginning of 2021 as compared to the infection rate during the period of the experiment reduces the external validity of the experiment. The increased infection rate indicates that something in the disease environment has changed. If that change were, for example, a new variant that the vaccine is less effective against, then the estimated efficacy of the trial would be greater than the true efficacy of the vaccine. An example of this can be seen in the CDC data, which showed the vaccines as significantly less effective in August 2021 compared to in the trial, likely because of the Delta variant.