Assignment 4

Grant Jackson

2024-10-02

Theory Questions:

Question 1:

i: I would subset the data to have only 30 year old individuals, and then calculate the average outcomes for the treated group (Di = 1) and untreated group (Di = 0) from the subset. The difference between the average outcomes is the conditional average treatment effect for 30 year old individuals.

ii: After conducting the same process described above for each age group (25-55 years old), I would calculate a weighted average of these conditional effects, with the weights being proportional to the number of individuals in each age group in the sample.

Question 2:

i: The conditional independence assumption likely will not hold in this case. While conditioning on a college degree could control for motivation, there are unobservable factors among individuals with college degrees that affect whether they take an online course or not. An individual's current employment status or job market conditions could affect both their decision to take an online coding class and potential outcomes (returns).

ii: The conditional independence assumption likely will not hold in this case. An unobservable factor could influence where new apartments are built, such as neihborhoods with lower crime rates, which may not be fully represented by neihborhood income, could attract new apartments to be built and also independently affect home prices.

iii: The conditional independence assumption likely is also violated in this case. Matching individuals on average weekly drinking may not control for all factors that home effect. Individuals who smoke may be more likely to have a poor diet or not exercise frequent, which affects later life health outcomes. This is not fully captured by alcohol consumption and would lead to biased estimates.

Coding Exercise:

Loading packages and data:

Question 1:

```r
# Creating dummy variable that is 1 if education is 12 years or more (HS Grad)
df$m_hs_degree <- df$medu >= 12

# No matching; constructing a pre-match matchit object
m.out0 <- matchit(
  mbsmoke ~ mage + m_hs_degree + mrace + fbaby,
  data = df,
  method = NULL,
  distance = "glm"
)

# Checking balance prior to matching
summary(m.out0)
```

Checking Initial Imbalance:

```
##
## Call:
## matchit(formula = mbsmoke ~ mage + m_hs_degree + mrace + fbaby,
##     data = df, method = NULL, distance = "glm")
##
## Summary of Balance for All Data:
##                  Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                0.2210        0.1781          0.4523     1.5694
## mage                   25.1667       26.8105         -0.3101     0.8818
## m_hs_degreeTRUE         0.6840        0.8621         -0.3830          .
## mrace                   0.8090        0.8478         -0.0986          .
## fbaby                   0.3715        0.4531         -0.1689          .
##                  eCDF Mean eCDF Max
## distance            0.1190   0.2338
## mage                0.0510   0.1593
## m_hs_degreeTRUE     0.1781   0.1781
## mrace               0.0388   0.0388
## fbaby               0.0816   0.0816
##
## Sample Sizes:
##           Control Treated
## All          3778     864
## Matched      3778     864
## Unmatched       0       0
## Discarded       0       0
```

```
# 1:1 NN PS matching w/o replacement
m.out1 <- matchit(
  mbsmoke ~ mage + m_hs_degree + mrace + fbaby,
  data = df,
  method = "nearest",
  distance = "glm"
)

# Checking balance after matching
summary(
  m.out1,
  un = FALSE
)
```

Matching:

```
##
## Call:
## matchit(formula = mbsmoke ~ mage + m_hs_degree + mrace + fbaby,
##     data = df, method = "nearest", distance = "glm")
##
## Summary of Balance for Matched Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance               0.2210        0.2210          0.0006     1.0010
## mage                  25.1667       25.1539          0.0024     0.9907
## m_hs_degreeTRUE        0.6840        0.6898         -0.0124          .
## mrace                  0.8090        0.8079          0.0029          .
## fbaby                  0.3715        0.3634          0.0168          .
##                 eCDF Mean eCDF Max Std. Pair Dist.
## distance           0.0002   0.0046          0.0009
## mage               0.0012   0.0046          0.0600
## m_hs_degreeTRUE    0.0058   0.0058          0.0174
## mrace              0.0012   0.0012          0.0501
## fbaby              0.0081   0.0081          0.0503
##
## Sample Sizes:
##           Control Treated
## All          3778     864
## Matched       864     864
## Unmatched    2914       0
## Discarded       0       0
```

Interpretation: The initial imbalance results shows noticeable differences between mothers who smoked during pregnancy (treated group) and mothers who didn't smoke during pregnancy (untreated group). They also reveal strong evidence that smoking mothers are on average younger, less likely to complete highschool, while a mothers race or first born don't have a strong difference with smoking or not. The post matching balance results make all the standardized mean differences very close to zero, which effectively helps reduce bias in estimating the effect of smoking during pregnancy on birth weight.

Question 2:

```
# Nearest neighbor match for mage and medu exactly on mmarried, mhisp, and alcohol
m.out3 <- matchit(
  mbsmoke ~ medu + mage,
  data = df,
  exact = ~ mmarried + mhisp + alcohol,
  method = "nearest", distance = "mahalanobis"
)
```

```
## Warning: Fewer control units than treated units in some `exact` strata; not all
## treated units will get a match.
```

```
# Getting matched dataset on matchit function results
matched_df <- match.data(m.out3)

# Estimating difference-in-means estimator
feols(
  bweight ~ i(mbsmoke),
  data = matched_df,
  vcov = "HC1"
)
```

```
## OLS estimation, Dep. Var.: bweight
## Observations: 1,680
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 3393.387    20.5175 165.39017 < 2.2e-16 ***
## mbsmoke::1  -249.296    28.1079  -8.86925 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 575.7   Adj. R2: 0.044211
```

```
# sometimes people do this
# feols(
#   bweight ~ i(mbsmoke) + medu + mage, data = matched_df, vcov = "HC1"
# )
```

Interpretation: The coefficient of -249.296 is the estimated average treatment effect of smoking during pregnancy on birth weight. In other words, through matching mothers who smoked during pregnancy with similar mothers who didn't smoke based on several covariates, its estimated that smoking during pregnancy reduces their birth weight by on average 249.3 grams. This is statistically significant with a p-value less than 2.2e-16.

Question 3:

```
# Regression adjustment
reg_adj <- lm(
  bweight ~ mbsmoke + mmarried + mhisp + alcohol + mage + medu,
  data = df
)

# Displaying the results
summary(reg_adj)
```

```
##
## Call:
## lm(formula = bweight ~ mbsmoke + mmarried + mhisp + alcohol +
##     mage + medu, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3114.24  -307.02    23.84   345.82  2002.42
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3148.072     52.210  60.296  < 2e-16 ***
## mbsmoke     -214.053     22.316  -9.592  < 2e-16 ***
## mmarried     159.238     21.081   7.554 5.07e-14 ***
## mhisp         32.464     46.376   0.700   0.4840
## alcohol      -61.942     47.518  -1.304   0.1924
## mage           2.385      1.727   1.381   0.1674
## medu           6.282      3.724   1.687   0.0916 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 562.8 on 4635 degrees of freedom
## Multiple R-squared:  0.05574,    Adjusted R-squared:  0.05452
## F-statistic:  45.6 on 6 and 4635 DF,  p-value: < 2.2e-16
```

Question 4:

```r
# Estimate propensity scores and weights
w_out <- weightit(
  mbsmoke ~ mmarried + mhisp + alcohol + mage + medu,
  data = df,
  method = "ps",
  estimand = "ATE"
)

# Checking the balance after weighting
bal_tab <- bal.tab(w_out)
print(bal_tab)
```

Estimating propensity score and checking balance:

```
## Balance Measures
##                 Type Diff.Adj
## prop.score Distance   0.0346
## mmarried     Binary  -0.0151
## mhisp        Binary  -0.0061
## alcohol      Binary   0.0012
## mage        Contin.  -0.0944
## medu        Contin.  -0.1499
##
## Effective sample sizes
##            Control Treated
## Unadjusted 3778.    864.
## Adjusted   3599.59  672.25
```

```r
# Estimate the IPTW estimator
iptw_est <- feols(
  bweight ~ mbsmoke,
  data = df,
  weights = w_out$weights,
  vcov = "HC1"
)
summary(iptw_est)
```

Estimating the IPTW estimator:

```
## OLS estimation, Dep. Var.: bweight
## Observations: 4,642
## Weights: w_out$weights
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 3399.572    9.81662 346.30786 < 2.2e-16 ***
## mbsmoke     -231.049   24.23525  -9.53361 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 802.0   Adj. R2: 0.03885
```

Interpretation: The weighting generally improved balance across the covariates. This process reduces the effective sample size, especially in the treated group. The small difference in propensity score distance indicates the weighting has successfully balanced the overall probability of treatment assignment between groups.