# DS-UA 201: Problem Set 4

Vanessa (Ziwei) Xu

December 8, 2020

This problem set is due at **11:59 pm on Tuesday, December 8th**. The data are on the course website.

Please upload your solutions as a .pdf file saved as `Yourlastname_Yourfirstinitial_pset4.pdf`). In addition, an electronic copy of your .Rmd file (saved as `Yourlastname_Yourfirstinitial_pset4.Rmd`) must be submitted to the course website at the same time. We should be able to run your code without error messages. Please note on your problem set if you collaborated with another student and, if so, whom. In order to receive credit, homework submissions must be substantially started and all work must be shown. Late assignments will not be accepted.

## Problem 1

Despite heated political and media rhetoric, there are few causal estimates of the effect of expanded healthcare insurance on healthcare outcomes. One landmark study, the Oregon Health Insurance Experiment, covered new ground by utilizing a randomized control trial implemented by the state government of Oregon. To allocate a limited number of eligible coverage slots for the state's Medicaid expansion, about 30,000 low-income, uninsured adults (out of about 90,000 wait-list applicants) were randomly selected by lottery to be allowed to apply for Medicaid coverage. Researchers collected observable measures of health (blood pressure, cholesterol, and blood sugar levels), as well as hospital visitation and healthcare expenses for 6,387 selected adults and 5,842 not selected adults.

For this problem, you will need the `OHIE.dta` file. The code below will load this dataset

```
ohie <- read_dta("OHIE.dta")
ohie
```

```
## # A tibble: 20,745 x 59
##    weight_total_inp tab1_gender_inp tab2dia_dx_post~ tab2hbp_dx_post~
##               <dbl>       <dbl+lbl>        <dbl+lbl>        <dbl+lbl>
##  1            1.15       1 [Female]           0 [No]           0 [No]
##  2            0.897      0 [Male]             0 [No]           1 [Yes]
##  3            0         NA                   NA               NA
##  4            1          1 [Female]           0 [No]           0 [No]
##  5            1.21       0 [Male]             0 [No]           0 [No]
##  6            1          0 [Male]             0 [No]           0 [No]
##  7            0         NA                   NA               NA
##  8            1.00       1 [Female]           0 [No]           0 [No]
##  9            1.20       0 [Male]             0 [No]           0 [No]
## 10            1          0 [Male]             0 [No]           0 [No]
## # ... with 20,735 more rows, and 55 more variables:
```

```
## #   tab2chl_dx_post_lottery <dbl+lbl>, tab2dep_dx_post_lottery <dbl+lbl>,
## #   tab3_pcs8_score <dbl>, tab3_mcs8_score <dbl>,
## #   tab5_usual_clinic_inp <dbl+lbl>, tab5_needmet_med_inp <dbl+lbl>,
## #   tab5_chl_chk_inp <dbl+lbl>, tab5_pap_chk_inp <dbl+lbl>,
## #   tab5_fobt_chk_inp <dbl+lbl>, tab5_col_chk_inp <dbl+lbl>,
## #   tab5_psa_chk_inp <dbl+lbl>, tab5_did_flu_inp <dbl+lbl>,
## #   tab2cvd_risk_point <dbl>, tab4_catastrophic_exp_inp <dbl+lbl>,
## #   tab4_owe_inp <dbl+lbl>, tab4_borrow_inp <dbl+lbl>,
## #   tab1_hispanic_inp <dbl+lbl>, tab1_race_white_inp <dbl+lbl>,
## #   tab1_race_black_inp <dbl+lbl>, tab1_race_nwother_inp <dbl+lbl>,
## #   tab2a1c_inp <dbl>, tab2hdl_inp <dbl>, tab2chl_inp <dbl>,
## #   tab2bp_sar_inp <dbl>, tab2bp_dar_inp <dbl>, tab5_rx_num_mod_inp <dbl>,
## #   tab2hbp_diure_med_inp <dbl+lbl>, tab2antihyperlip_med_inp <dbl+lbl>,
## #   tab2diabetes_med_inp <dbl+lbl>, tab2antidep_med_inp <dbl+lbl>,
## #   household_id <dbl>, treatment <dbl+lbl>, ohp_all_ever_admin <dbl+lbl>,
## #   tab1_age_19_34_inp <dbl>, tab1_age_35_49_inp <dbl>,
## #   tab1_age_50_64_inp <dbl>, tab1_itvw_english_inp <dbl>,
## #   tab3_pain_low_inp <dbl>, tab3_health_change_noworse <dbl+lbl>,
## #   tab5_obese <dbl>, tab2bp_hyper <dbl>, tab2a1c_dia <dbl>, tab2chl_h <dbl>,
## #   tab2hdl_low <dbl>, tab2phqtot_high <dbl>, tab5_med_qual_bin_inp <dbl+lbl>,
## #   tab5_smk_curr_bin_inp <dbl+lbl>, tab3_poshappiness_bin_inp <dbl+lbl>,
## #   tab5_mam50_chk_inp <dbl>, tab5_doc_num_mod_inp <dbl>,
## #   tab5_ed_num_mod_inp <dbl>, tab5_surg_num_mod_inp <dbl>,
## #   tab5_hosp_num_mod_inp_2 <dbl>, tab4_any_oop_inp <dbl>,
## #   tab4_tr_tot_spend_inp <dbl>
```

The variables you will need are:

`treatment` - Selected in the lottery

`ohp_all_ever_admin` - Ever enrolled in Medicaid from matched notification date to September 30, 2009 (actually "took" the treatment)

`tab2bp_hyper` - Outcome: Binary indicator for elevated blood pressure (defined a systolic pressure of 140mm Hg or more and a diastolic pressure of 90mm Hg or more)

`tab2phqtot_high` - Outcome: Binary indicator for a positive screening result for depression (defined as a score of 10 or higher on the Patient Health Questionaire - 8)

`tab4_catastrophic_exp_inp` - Outcome: Indicator for catastrophic medical expenditure (total out-of-pocket medical expenses $\geq 30\%$ of household income)

`tab5_needmet_med_inp` - Outcome: Participant feels that they received all needed medical care in past 12 months (binary indicator)

## Question A

Estimate the intent-to-treat effects of assignment to treatment (being eligible to apply) on each of the four outcomes (elevated blood pressure, depression, catastrophic medical expenditure, and whether respondents had their health care needs met). Provide 95% confidence intervals for each estimate and interpret your results.

---

```r
# filter out n/a results
ohie_1 <- subset(ohie, tab2bp_hyper == 1 | tab2bp_hyper == 0)
# fit the regression model
lm_robust(tab2bp_hyper ~ treatment, data = ohie_1)
```

```
##                 Estimate  Std. Error    t value      Pr(>|t|)    CI Lower
## (Intercept)  0.159106529 0.004795019 33.1816279 3.214924e-231  0.14970753
## treatment   -0.001600248 0.006620705 -0.2417035  8.090140e-01 -0.01457788
##                CI Upper    DF
## (Intercept) 0.16850553 12186
## treatment   0.01137739 12186
```

```r
# filter out n/a results
ohie_2 <- subset(ohie, tab2phqtot_high == 1 | tab2phqtot_high == 0)
# fit the regression model
lm_robust(tab2phqtot_high ~ treatment, data = ohie_2)
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)    CI Lower
## (Intercept)  0.30366672 0.006033841 50.327265 0.000000e+00  0.29183944
## treatment   -0.03493247 0.008206692 -4.256583 2.091319e-05 -0.05101889
##                 CI Upper    DF
## (Intercept)  0.31549401 12159
## treatment   -0.01884605 12159
```

```r
# filter out n/a results
ohie_4 <- subset(ohie, tab4_catastrophic_exp_inp == 1 | tab4_catastrophic_exp_inp == 0)
# fit the regression model
lm_robust(tab4_catastrophic_exp_inp ~ treatment, data = ohie_4)
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)    CI Lower
## (Intercept)  0.05382436 0.003003077 17.923072 6.769195e-71  0.04793784
## treatment   -0.01526897 0.003879414 -3.935896 8.336278e-05 -0.02287326
##                  CI Upper    DF
## (Intercept)  0.059710889 11793
## treatment   -0.007664678 11793
```

```r
# filter out n/a results
ohie_5 <- subset(ohie, tab5_needmet_med_inp == 1 | tab5_needmet_med_inp == 0)
# fit the regression model
lm_robust(tab5_needmet_med_inp ~ treatment, data = ohie_5)
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)   CI Lower   CI Upper
## (Intercept) 0.61240576 0.006378049 96.017723 0.000000e+00 0.59990377 0.62490774
## treatment   0.03445945 0.008745816  3.940107 8.189875e-05 0.01731626 0.05160263
##                DF
## (Intercept) 12214
## treatment   12214
```

The estimate of intent-to-treat effects of assignment to treatment on elevated blood pressure is -0.002; and the 95% confidence intervals is [-0.015, 0.011]. We would fail to reject the null of no treatment effect.

The estimate of intent-to-treat effects of assignment to treatment on depression is -0.035; and the 95% confidence intervals is [-0.051, -0.019]. We would reject the null of no treatment effect.

The estimate of intent-to-treat effects of assignment to treatment on catastrophic medical expenditure is -0.015; and the 95% confidence intervals is [-0.023, -0.008]. We would reject the null of no treatment effect.

The estimate of intent-to-treat effects of assignment to treatment on whether respondents had their health care needs met is 0.034; and the 95% confidence intervals is [0.017, 0.052]. We would reject the null of no treatment effect.

---

## Question B

Suppose that researchers actually wanted to estimate the effect of Medicaid enrollment on each of the four outcomes. Suppose they first used a naive regression of each of the the outcomes on the indicator of Medicaid enrollment. Report 95% confidence intervals for each of your estimates and interpret your results. Why might these be biased estimates for the causal effect of Medicaid enrollment?

---

```
lm_robust(tab2bp_hyper ~ ohp_all_ever_admin, data = ohie_1)
```

```
##                       Estimate  Std. Error   t value   Pr(>|t|)    CI Lower
## (Intercept)         0.16344605 0.003965969 41.212132 0.00000000  0.15567213
## ohp_all_ever_admin -0.01805395 0.007162470 -2.520632 0.01172708 -0.03209353
##                       CI Upper    DF
## (Intercept)         0.171219984 12186
## ohp_all_ever_admin -0.004014375 12186
```

```
lm_robust(tab2phqtot_high ~ ohp_all_ever_admin, data = ohie_2)
```

```
##                       Estimate  Std. Error   t value     Pr(>|t|)   CI Lower
## (Intercept)         0.27129192 0.004773484 56.833108 0.000000e+00 0.26193513
## ohp_all_ever_admin 0.04931657 0.009237089  5.338974 9.515626e-08 0.03121041
##                       CI Upper    DF
## (Intercept)         0.28064871 12159
## ohp_all_ever_admin 0.06742274 12159
```

```
lm_robust(tab4_catastrophic_exp_inp ~ ohp_all_ever_admin, data = ohie_4)
```

```
##                       Estimate  Std. Error  t value     Pr(>|t|)    CI Lower
## (Intercept)         0.04893693 0.002351357 20.81221 1.658488e-94  0.04432788
## ohp_all_ever_admin -0.01072603 0.004051915 -2.64715 8.128113e-03 -0.01866845
##                       CI Upper    DF
## (Intercept)         0.053545976 11793
## ohp_all_ever_admin -0.002783606 11793
```

```
lm_robust(tab5_needmet_med_inp ~ ohp_all_ever_admin, data = ohie_5)
```

```
##                       Estimate  Std. Error    t value     Pr(>|t|)   CI Lower
## (Intercept)        0.61281433 0.005219932 117.398900 0.000000e+00 0.60258244
## ohp_all_ever_admin 0.06126608 0.009482118   6.461223 1.078051e-10 0.04267963
##                       CI Upper    DF
## (Intercept)        0.62304622 12214
## ohp_all_ever_admin 0.07985253 12214
```

The estimate of the effect of Medicaid enrollment on elevated blood pressure is -0.018; and the 95% confidence intervals is [-0.032, -0.004]. We would reject the null of no treatment effect.

The estimate of the effect of Medicaid enrollment on depression is 0.049; and the 95% confidence intervals is [0.031, 0.067]. We would reject the null of no treatment effect.

The estimate of the effect of Medicaid enrollment on catastrophic medical expenditure is -0.011; and the 95% confidence intervals is [-0.019, -0.003]. We would reject the null of no treatment effect.

The estimate of the effect of Medicaid enrollment on whether respondents had their health care needs met is 0.061; and the 95% confidence intervals is [0.043, 0.080]. We would reject the null of no treatment effect.

However, these are biased estimates because randomization is broken by non-compliance.

---

## Question C

Suppose we were to use assignment to treatment as an instrument for actually receiving Medicaid coverage.

Consider that not everyone who was selected to apply for Medicaid actually ended up applying and receiving coverage. Likewise, some applicants who were not selected to receive the treatment nevertheless were eventually covered. What were the compliance rates (the level of Medicaid enrollment) for subjects who were selected and subjects who were not selected? Use a "first stage" regression to estimate the effect of being selected on Medicaid enrollment to estimate the compliance rates. Is the instrument of assignment-to-treatment a strong instrument for actual Medicaid enrollment?

---

```r
lm_robust(ohp_all_ever_admin ~ treatment, data = ohie)
```

```
##              Estimate  Std. Error  t value Pr(>|t|)  CI Lower  CI Upper    DF
## (Intercept) 0.1454545 0.003467304 41.95033        0 0.1386584 0.1522507 20743
## treatment   0.2363811 0.005891465 40.12264        0 0.2248334 0.2479288 20743
```

The instrument of assignment-to-treatment is a very strong instrument for actual Medicaid enrollment.

---

## Question D

Discuss whether the exclusion restriction holds in this design.

---

```r
cov(ohie_1$tab2bp_hyper, ohie_1$treatment)
```

```
## [1] -0.0003992859
```

```r
cov(ohie_2$tab2phqtot_high, ohie_2$treatment)
```

```
## [1] -0.008716422
```

```r
cov(ohie_4$tab4_catastrophic_exp_inp, ohie_4$treatment)
```

```
## [1] -0.003810733
```

```r
cov(ohie_5$tab5_needmet_med_inp, ohie_5$treatment)
```

```
## [1] 0.008598482
```

Exclusion restriction holds such that the instrument affects the outcome Y only through the channel X but not directly. The covariance are close to 0 but not exactly. I would argue that exclusion restriction holds at some degree since there are finite number of observations but in theory they do no hold.

---

## Question E

Now estimate the effect of Medicaid enrollment on each of the four outcomes using two-stage least squares estimators. Report 95% confidence intervals for your estimates and interpret your results. Compare the estimates to those you obtained in Question C.

---

```r
iv_robust(tab2bp_hyper ~ ohp_all_ever_admin | treatment, data = ohie_1)
```

```
##                       Estimate  Std. Error    t value     Pr(>|t|)    CI Lower
## (Intercept)         0.160076358 0.008180946 19.5669739 5.713710e-84  0.14404041
## ohp_all_ever_admin -0.006299556 0.026059156 -0.2417406 8.089852e-01 -0.05737964
##                      CI Upper    DF
## (Intercept)        0.17611231 12186
## ohp_all_ever_admin 0.04478052 12186
```

```r
iv_robust(tab2phqtot_high ~ ohp_all_ever_admin | treatment, data = ohie_2)
```

```
##                       Estimate Std. Error   t value      Pr(>|t|)   CI Lower
## (Intercept)         0.3248452 0.01039309 31.255885 2.383439e-206  0.3044731
## ohp_all_ever_admin -0.1376126 0.03286133 -4.187675  2.838127e-05 -0.2020260
##                      CI Upper    DF
## (Intercept)         0.34521729 12159
## ohp_all_ever_admin -0.07319914 12159
```

```r
iv_robust(tab4_catastrophic_exp_inp ~ ohp_all_ever_admin | treatment, data = ohie_4)
```

```
##                       Estimate  Std. Error   t value     Pr(>|t|)    CI Lower
## (Intercept)         0.06314350 0.005080412 12.428816 3.029925e-35  0.05318506
## ohp_all_ever_admin -0.06036067 0.015425260 -3.913105 9.162856e-05 -0.09059672
##                       CI Upper    DF
## (Intercept)         0.07310195 11793
## ohp_all_ever_admin -0.03012461 11793
```

```r
iv_robust(tab5_needmet_med_inp ~ ohp_all_ever_admin | treatment, data = ohie_5)
```

```
##                      Estimate Std. Error   t value    Pr(>|t|)   CI Lower
## (Intercept)         0.5915172 0.01086399 54.447502 0.00000e+00 0.57022203
## ohp_all_ever_admin 0.1354509 0.03441917  3.935333 8.35417e-05 0.06798387
##                      CI Upper    DF
## (Intercept)         0.6128123 12214
## ohp_all_ever_admin 0.2029179 12214
```

The estimate of the effect of Medicaid enrollment on elevated blood pressure is -0.006; and the 95% confidence intervals is [-0.057, 0.045]. We would fail to reject the null of no treatment effect.

The estimate of the effect of Medicaid enrollment on depression is -0.138; and the 95% confidence intervals is [-0.202, -0.073]. We would reject the null of no treatment effect.

The estimate of the effect of Medicaid enrollment on catastrophic medical expenditure is -0.060; and the 95% confidence intervals is [-0.091, -0.030]. We would reject the null of no treatment effect.

The estimate of the effect of Medicaid enrollment on whether respondents had their health care needs met is 0.135; and the 95% confidence intervals is [0.068, 0.203]. We would reject the null of no treatment effect.

Compared to the results from Question C, the estimates are significantly lower. ******

### Question F

What additional assumptions do you have to make in order to interpret your estimates from Question E as Average Treatment Effects for the entire sample?

---

Apart from exclusion restriction, other assumptions needed in order to interpret estimates from Question E as Average Treatment Effects for the entire sample are randomization of instrument, first-stage relationship, and monotonicity. Z has to have an effect on D, but it should only go in one direction at the individual level. Z should also be independent of both sets of potential outcomes as well.

---

## Problem 2

In this problem, we will return to the Card and Krueger (1994) minimum wage study, which used the change in minimum wage laws in New Jersey relative to neighboring Pennsylvania in order to estimate the effect of minimum wage increases on employment via a difference-in-differences design.The full citation to the paper is

Card, David, and Alan B. Krueger. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." The American Economic Review 84.4 (1994): 772-793.

This problem will look at the effect of the minimum wage increases on observed wages paid. You will be using the `minwage.csv` dataset for this problem. Below is the code to load the data into R

```
# Load the data for Card and Krueger (1994)
minwage <- read_csv("minwage.csv")
```

Card and Krueger conducted a survey of fast food restaurants in New Jersey and eastern Pennsylvania in two waves. The first wave was conducted in February/March 1992, about a month prior to the minimum wage increase in New Jersey. The second wave was conducted in November/December of 1992, about 8 months after the minimum wage increase in New Jersey. Each survey wave asked a similar set of questions about the characteristics of the restaurant (employment, starting wages, etc..). Prior to the April, 1992 increase, the minimum wage in both Pennsylvania and New Jersey was the federal minimum of $4.25 per hour. After April 1992, New Jersey's minimum wage was raised by state law to $5.05 per hour while Pennsylvania's remained at $4.25.

The relevant variables you may need are

- `STATE`: Treatment indicator: 1 if New Jersey, 0 if Pennsylvania
- `CHAIN`: Fast food franchise (categorical): 1=Burger King; 2=KFC; 3=Roy Rogers; 4=Wendy's
- `EMPFT`: Number of full-time employees (first survey wave)
- `WAGE_ST`: Starting wage (dollars per hour) (first survey wave)
- `PSODA`: Price of a medium soda (first survey wave)
- `PFRY`: Price of small fries (first survey wave)
- `PENTREE`: Price of an entree (first survey wave)
- `EMPFT2`: Number of full-time employees (second survey wave)
- `WAGE_ST2`: Starting wage (dollars per hour) (second survey wave)
- `PSODA2`: Price of a medium soda (second survey wave)
- `PFRY2`: Price of small fries (second survey wave)
- `PENTREE2`: Price of an entree (second survey wave)

## Question A

Using the difference-in-differences estimator, estimate the effect of New Jersey's minimum wage increase on starting wages (in dollars per hour). Provide a 95% confidence interval and interpret your results

---

```r
# drop unites with n/a results
minwage2 <- subset(minwage, !is.na(WAGE_ST) & !is.na(WAGE_ST2) & !is.na(EMPFT) & !is.na(EMPFT2))

# Create a variable for change in starting wages
minwage2$CHANGE <- minwage2$WAGE_ST2 - minwage2$WAGE_ST

# get the DiD estimate
lm_robust(CHANGE ~ STATE, data = minwage2)
```

```
##               Estimate Std. Error    t value     Pr(>|t|)   CI Lower   CI Upper
## (Intercept) -0.0380597 0.04707944 -0.8084144 4.193919e-01 -0.1306485 0.05452908
## STATE        0.5095717 0.05107896  9.9761575 7.854316e-21  0.4091173 0.61002616
##               DF
## (Intercept) 356
## STATE       356
```

The estimate of the effect of New Jersey's minimum wage increase on starting wage is 0.510 dollars per hour; and the 95% confidence intervals is [0.409, 0.610]. We would reject the null of no treatment effect.

---

## Question B

Suppose we were just interested in estimating the effect of the minimum wage increase only on those restaurants in New Jersey that were paying minimum wage as their starting wage prior to the increase. Using a difference-in-differences estimator and including all Pennsylvania restaurants as the control group, estimate the effect of the minimum wage increase on starting wages in New Jersey restaurants that were paying the minimum wage ($4.25 per hour) as their starting wage prior to the law. Provide a 95% confidence interval and compare your result to the result in Question 1. Discuss why the two might differ.

---

```r
# drop all observations where restaurants in New Jersey weren't paying minimum wage as their starting w
minwage3 <- minwage2[!(minwage2$WAGE_ST != 4.25 & minwage2$STATE == 1),]
minwage3$CHANGE <- minwage3$WAGE_ST2 - minwage3$WAGE_ST

# get the DiD estimate
lm_robust(CHANGE ~ STATE, data = minwage3)
```

```
##               Estimate Std. Error    t value     Pr(>|t|)   CI Lower   CI Upper
## (Intercept) -0.0380597 0.04707944 -0.8084144 4.200676e-01 -0.1310459 0.05492653
## STATE        0.8461242 0.04725470 17.9056109 7.292633e-40  0.7527918 0.93945660
##               DF
## (Intercept) 158
## STATE       158
```

The estimate of the effect is 0.846 dollars per hour; and the 95% confidence intervals is [0.753, 0.939]. We would reject the null of no treatment effect. The estimate is larger than from Question A probably because there is more room for starting wage to be raised for the restaurants in New Jersey that were paying minimum wage as their starting wage prior to the increase. And we also exclude restaurants that were already paying above the new minimum wage as well.

---

## Question C

Now estimate the effect of the minimum wage increase on starting wages in those restaurants in New Jersey that were minimally affected by the law (those restaurants already paying $5.00 per hour or above as their starting wage in the pre-treatment period). Provide a 95% confidence interval and compare your result to the findings from Questions 1 and 2. Discuss what this result tells us about the validity of the difference-in-differences identification strategy used by Card and Krueger (Hint: What sort of diagnostic strategy is this?)

---

```
# drop all observations where restaurants in New Jersey weren't minimally affected by the law
minwage4 <- minwage2[!(minwage2$WAGE_ST < 5 & minwage2$STATE == 1),]
minwage4$CHANGE <- minwage4$WAGE_ST2 - minwage4$WAGE_ST

# get the DiD estimate
lm_robust(CHANGE ~ STATE, data = minwage4)
```

```
##                Estimate Std. Error    t value  Pr(>|t|)    CI Lower   CI Upper
## (Intercept) -0.03805970 0.04707944 -0.8084144 0.4203066 -0.13118749 0.05506809
## STATE        0.03402985 0.05415328  0.6283987 0.5308286 -0.07309069 0.14115040
##               DF
## (Intercept) 132
## STATE       132
```

The estimate of the effect is 0.034 dollars per hour; and the 95% confidence intervals is [-0.073, 0.141]. We would fail to reject the null of no treatment effect. The estimate is a lot smaller than from both Question A and B. This is because there is no incentive for those restaurants to raise their minimum wage anymore since it's already abiding the law.

---

## Question D

Estimate the average effect of the New Jersey minimum wage increase on the average price of a regular meal (the combination of an entree, small fries and a medium soda) using a difference-in-differences design. For the purposes of this problem, drop any observations with missing data. Report a 95% confidence interval and interpret your results.

---

```
# drop observations with missing data
minwage6 <- subset(minwage, !is.na(WAGE_ST) & !is.na(WAGE_ST2) & !is.na(EMPFT)
                   & !is.na(EMPFT2) & !is.na(PSODA) & !is.na(PSODA2) & !is.na(PFRY)
                   & !is.na(PENTREE) & !is.na(PENTREE2) & !is.na(PFRY2))

# Create a variable for the combination of a regular meal
minwage6$meal <- minwage6$PFRY2 + minwage6$PENTREE2 + minwage6$PSODA2 - minwage6$PENTREE - minwage6$PSOD

# get the DiD estimate
lm_robust(meal ~ STATE, data = minwage6)
```

```
##               Estimate Std. Error    t value   Pr(>|t|)    CI Lower  CI Upper
## (Intercept) -0.0350000 0.04114873 -0.8505731 0.39563860 -0.11595430 0.0459543
## STATE        0.1107634 0.04700821  2.3562556 0.01905903  0.01828135 0.2032454
##               DF
## (Intercept) 322
## STATE       322
```

The estimate of the average effect of the New Jersey minimum wage increase on the average price of a regular meal is 0.111 dollars; and the 95% confidence intervals is [0.018, 0.203]. We would reject the null of no treatment effect. This shows us that by raising the minimum wage, the average price of a meal was also impacted and increased by 0.11 dollars.

---

## Question E

We will now return to the effect of the minimum wage law on full time employment, the main focus of the Card/Krueger paper.

We are concerned about potential violations of the parallel trends assumption. One way of addressing this is to adjust directly for pre-treatment covariates that differ between Pennsylvania and New Jersey and may account for differential trends. We may, for example, be concerned that some restaurant chains are overrepresented in the New Jersey sample relative to Pennsylvania and may exhibit differential trends over time.

Matching is one approach to addressing this problem. Use exact matching to adjust for the type of chain and estimate the ATT of the minimum wage increase on the change in full time employment from wave 1 to wave 2. Report a 95% asymptotic confidence interval using the Abadie-Imbens standard error. Compare your results to the unadjusted difference-in-differences estimate. Lastly, conduct a balance test for your matching solution and discuss the pre-matching imbalance between New Jersey/Pennsylvania and whether matching was successful in reducing that imbalance.

Hint: The `exact=T` option in the `Match()` function in the `Matching` package will use all valid exact matches for each treated unit.

---

```r
# before matching / unadjusted
lm_robust(CHANGE ~ STATE, data = minwage2)
```

```
##               Estimate Std. Error    t value     Pr(>|t|)   CI Lower   CI Upper
## (Intercept) -0.0380597 0.04707944 -0.8084144 4.193919e-01 -0.1306485 0.05452908
## STATE        0.5095717 0.05107896  9.9761575 7.854316e-21  0.4091173 0.61002616
##               DF
## (Intercept) 356
## STATE       356
```

```r
exact_match <- Match(Y = minwage2$CHANGE, Tr = minwage2$STATE,
                     X = minwage2[, c('CHAIN')],
                     M = 1, exact = T, ties = T, estimand = 'ATT', Weight = 2)
summary(exact_match)
```

```
##
## Estimate...  0.51916
## AI SE......  0.04751
## T-stat.....  10.927
## p.val......  < 2.22e-16
##
## Original number of observations..............  358
## Original number of treated obs...............  291
## Matched number of observations...............  291
## Matched number of observations  (unweighted).  5788
##
## Number of obs dropped by 'exact' or 'caliper'  0
```

```
# 95% confidence interval
exact_match_CI <- c(exact_match$est - qnorm(.975)*exact_match$se, exact_match$est + qnorm(.975)*exact_ma
exact_match_CI
```

```
## [1] 0.4260376 0.6122741
```

```
# use the MatchBaance function to diagnose the balance
balance_match <- MatchBalance(STATE ~ CHAIN, data = minwage2, match.out = exact_match)
```

```
##
## ***** (V1) CHAIN *****
##                        Before Matching        After Matching
## mean treatment........     2.0859             2.0859
## mean control.........     2.0896             2.0859
## std mean diff.........   -0.34197                  0
##
## mean raw eQQ diff.....     0.1194                  0
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  0
##
## mean eCDF diff........   0.029505                  0
## med  eCDF diff........   0.030415                  0
## max  eCDF diff........   0.057188                  0
##
## var ratio (Tr/Co).....    0.83657                  1
## T-test p-value........    0.98135                  1
## KS Bootstrap p-value..       0.65                  1
## KS Naive p-value......    0.99417                  1
## KS Statistic..........   0.057188        2.7105e-20
```

Using exact matching, we estimate an ATT of the minimum wage increase of 0.519, with a standard error of 0.048. This corresponds to a 95% CI of [0.426, 0.612]. We would reject the null of no treatment effect. Compared to before matching / unadjusted data, the estimate is larger and there is a bigger effect. Using a balance test, we can see that the standard mean difference has dropped to 0 after matching. I would argue it was pretty successful.

---