TA Session 4

TA's 2020

May 3rd, 2020

1.Matching

1.1 Motivation to use matching methods

Previously, the generic framework to estimate treatment effects is to specify a parametric functional form:

$$Y_i = \beta_0 + \beta_1 \times D_i + \alpha \times X_i + u_i$$

To some extent, we are implicitly assuming for a (linear) functional form, and the treatment effects are constant(β_1) across the population of interest. Neither of these two assumptions hold water in most cases. However, with matching methods, we can group our observations based on whether they share exactly same or similar characteristics (X's). **Nonparametric estimates can be made, (e.g taking the mean difference)**, and **the functional form is irrelevant**. In other words, matching methods are more flexible, and are able to capture heterogeneous treatment effects.

In many ways, matching is a very intuitive estimator: Compare two identical people, treat one and leave the other untreated. The difference must be the the treatment effect. "There are two important assumptions to make, to make the matching method a theoretically feasible approach to estimating treatment effects:

Common Support Assumption (CSA): For all the possible X's, we should be able to observe both the treated and untreated. In other words, for any given X^0, when we have a sufficiently large sample, we should be able to see both the treated and untreated. Under this circumstance, we can observe both the treated and untreated in a given cell, and then the treatment effects in that cell can be estimated.

$$0 < Pr(D_i|X = x^0) < 1, \forall x^0$$

Conditional Independence Assumption (CIA): the potential outcomes of a given individual are orthogonal to the treatment, conditional on all the X's. In other words, for any given X, whether being treated will not affect an individual's potential outcome. Only with this assumption made, can we safely say the observed outcomes of the matched counterparts are good estimates of the selve's counterfacts.

$$(Y_{1,i}, Y_{0,i}) \perp D_i|X$$

Here we present an example of how matching works and how to check if matching succeeds in R:

```
# Matching example
library(MatchIt) # This is the package commonly used to do data preprocessing in R
set.seed(1234)
card$college <- ifelse(card$educ>12, 1, 0)
card_cov <- c('lwage', 'age', 'IQ')
# We want to estimate the ATE of going to college on log(wage)

summary(lm(lwage ~ college + age + IQ, data = card))
```

```
##
## Call:
## lm(formula = lwage ~ college + age + IQ, data = card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55816 -0.23207  0.02166  0.25287  1.58066
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.4207949  0.1015750  43.522  < 2e-16 ***
## college     0.0570413  0.0189787   3.006  0.00268 **
## age         0.0463887  0.0028240  16.427  < 2e-16 ***
## IQ          0.0055095  0.0006055   9.099  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3836 on 2057 degrees of freedom
##   (949 observations deleted due to missingness)
## Multiple R-squared:  0.1583, Adjusted R-squared:  0.157
## F-statistic: 128.9 on 3 and 2057 DF,  p-value: < 2.2e-16
```

```
# First, we check the covariate and outcome means in the two groups
card %>%
  group_by(college) %>%
  select(one_of(card_cov)) %>%
  summarise_all(funs(mean(., na.rm = T)))
```

```
## # A tibble: 2 x 4
##   college lwage   age    IQ
##     <dbl> <dbl> <dbl> <dbl>
## 1       0  6.16  28.2  94.6
## 2       1  6.36  28.0 108.
```

The background variables are seemingly different between the two groups.

```
# Second, we do use the exact method to match the data
card_nomissed <- card %>%  # MatchIt does not allow missing values
  select(lwage, college, one_of(card_cov)) %>% na.omit()

mod_match <- matchit(college ~ age + IQ, #formula: treat ~ x1+x2+...(background variables)
                     method = "exact", # set the matching method you want to use
                     data = card_nomissed)
df_m <- match.data(mod_match)
# Do the balance test on the two groups
df_m %>%
  group_by(college) %>%
  select(one_of(card_cov)) %>%
  summarise_all(funs(mean(., na.rm = T)))
```

```
## Adding missing grouping variables: `college`
```

```
## # A tibble: 2 x 4
```

```
##   college lwage   age    IQ
##     <dbl> <dbl> <dbl> <dbl>
## 1       0  6.30  28.2  98.9
## 2       1  6.36  28.0 103.
```

```
# To do the balance more formally using regressions
summary(lm(age~college, data = df_m), type='text')
```

```
##
## Call:
## lm(formula = age ~ college, data = df_m)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.169 -2.169 -1.012  1.988  5.988
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.1691     0.1169 240.924   <2e-16 ***
## college      -0.1569     0.1587  -0.989    0.323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.914 on 1357 degrees of freedom
## Multiple R-squared:  0.00072,    Adjusted R-squared:  -1.639e-05
## F-statistic: 0.9777 on 1 and 1357 DF,  p-value: 0.3229
```

```
summary(lm(IQ~college, data = df_m), type='text')
```

```
##
## Call:
## lm(formula = IQ ~ college, data = df_m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.473  -6.473   0.527   7.527  29.137
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.8631     0.4433 223.008  < 2e-16 ***
## college       4.6098     0.6016   7.663 3.45e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.05 on 1357 degrees of freedom
## Multiple R-squared:  0.04148,    Adjusted R-squared:  0.04077
## F-statistic: 58.72 on 1 and 1357 DF,  p-value: 3.447e-14
```

We can see that the matching did well on age, but not so well on IQ. This balance check helps us see if we succeed on doing matching. Ideally, we should not be able to reject the null hypothesis of mean difference for each covariate.

You can go here to go over the entire process of doing matching estimation in R.

1.2 Exact Matching

Please refer to Slide7_page18. The procedures of exact matching unfolds as:

- a) Divide data into cells uniquely defined by the covariates
- b) For each value of $X = x$ (each cell), calculate $\bar{Y}_T$ and $\bar{Y}_U$
- c) Calculate $\bar{Y}_T - \bar{Y}_U$ for each $X = x$
- d) Estimate τ^{ATE} as a weighted average of the results in step c).

Remember to use different weights when estimating ATE versus ATT versus ATN:

- When estimating ATE, the weight should be the counts of observations in a cell over the total counts of observations.

$$\hat{\tau}^{ATE} = \sum_{k=1}^{N} \frac{N_k}{N} \hat{\tau}_k$$

- When estimating ATT, the weight should be the counts of treated in a cell over the total counts of treated individuals.

$$\hat{\tau}^{ATT} = \sum_{k=1}^{N_T} \frac{N_{k,T}}{N_T} \hat{\tau}_k$$

- When estimating ATN, the weight should be the counts of untreated in a cell over the total counts of untreated individuals.

$$\hat{\tau}^{ATN} = \sum_{k=1}^{N_U} \frac{N_{k,U}}{N_U} \hat{\tau}_k$$

Notice that the cell estimates remain unchanged. What distinguished these three estimates are **weights**. However, we should always caution ourselves how big are the cell sizes. The cell size could be so small that there are no enough available observations to make the cell estimation. The method could be jeopardized when the data is marked with super high dimensionality, or when the variables are continuous.

1.2.1 Exact matching using ddply

The function `ddply` from the `library (plyr)` splits up the dataset into cells uniquely defined by the specified covariates. This function is useful to perform exact matching when we cannot use the `Matchit` package.

```
library(plyr) #necessary to load the plyr package after the dplyr package to use ddply

df <- card %>%
  ddply(.(IQ, age, college), #splitting the dataset up into cells uniquely defined
  summarise,
  avg = mean(lwage)) #make a new summary statistic variable
head(df)
```

```
##   IQ age college      avg
## 1 50  28       0 6.437752
## 2 51  27       0 6.109248
## 3 53  27       0 5.874931
## 4 54  27       1 6.437752
## 5 54  29       0 6.396930
## 6 55  29       0 6.437752
```

```
# Please check how ddply works by tying ?ddply
```

1.3 Other matching estimators:

The principle of all matching methods is: for a given data point, apply a criterion to determine the "nearest" counterparts, then use the counterparts' outcomes as the counterfacts to make the estimation. Here is a menu of some popular matching criteria:

- Propensity Score Matching: estimate the propensity score first using a logistic regression
- Bandwidth Matching (Slide 7, pp29-30)
- K Nearest Neighbors (Slide 7, pp27-28)
- Kernel Estimator: making estimates at a point (the kernel) by taking the weighted average of its surroundings. The weight can vary based on what density functions you employ. Some commonly used density functions include Epanechnikov Density and Gaussian Density function.

2. Instrumental Variable

2.1 Endogeneity Problem

When there is endogeneity problem (the following cases) in the regression model, the model will have $cov(x, u) \neq 0$, which means we will have bias in estimating the treatment effect.

- Omitted Variable Bias
- Measurement error on Y
- Reverse Causality (Simultaneity) For example we are trying to measure the effect of education on earnings in the following equation

$$log(wage) = \beta_0 + \beta_1 education + \beta_2 experience + \beta_3 (experience)^2 + u$$

Since *selection on observables* requires strong assumptions and we are not able to observe and measure everything that influences wage, there is OVB in the model and $cov(education, u) \neq 0$. It is not possible to have an consistent estimate for the effect of education.

We can use IV estimator to get the quasi-random variation in treatment variable.

2.2 Math explanation for IV

For a simple model

$$Y_i = \alpha + \tau D_i + \beta X_i + \epsilon_i$$

We separate D_i into two parts $D_i = B_i \epsilon_i + C_i$ with $cov(C_i, \epsilon_i) = 0$. Then we can rewrite the model as

$$Y_i = \alpha + \tau C_i + \beta X_i + (1 + \tau B_i)\epsilon_i$$

In reality, we do not observe the components of D_i, so the best thing we can do is to use an IV. The idea behind IV is to find a variable Z which is correlated with C, the exogenous part of D_i, and is uncorrelated with ϵ.

2.3 Requirements for a valid IV

As we mentioned above, it is not possible to get the treatment effect of education (educ) on earnings(log(wage)), therefore, we use an IV of *whether someone grew up near a 4-year college (nearc4)* to get the IV estimate. How do we know if nearc4 is a good IV? We need it to satisfy the following restrictions:

Exclusion restriction $cov(Z_i, \epsilon_i) = 0$

- IV is *not* correlated with ϵ. Because we want to recover the quasi-random part of variation in D_i, which means we don't want our IV to be correlated with the outcome variable to get the endogeneity problem.
- By satisfying exclusion restriction, IV does not directly affect Y_i, it only affects Y_i through D_i.
- This is fundamentally *untestable*!! Because ϵ is *not observable*!

In the education example, the first restriction that *whether someone grew up near a 4-year college (nearc4)* needs to satisfy to be a valid IV is that it is not directly correlated with earnings, but it can affect earnings through education. This is *not testable*. But we would assume whether someone grew up near a 4-year college can't directly affect someone's earning. It only makes sense when the environment you grow up affects earnings through educational attainment.

Instrument Condition (Relevance) $cov(Z_i, D_i) \neq 0$

- IV must be *correlated* with the treatment.
- If the correlation is too weak, it might not be a strong IV.
- If the IV and D_i is too closely correlated, the exclusion restriction might fail.

In the education example, the first restriction that *whether someone grew up near a 4-year college (nearc4)* needs to satisfy to be a valid IV is that it needs to be correlated with *education (educ)*. Usually we use F-test in First Stage to test for this assumption, we have an example for this in 2.5.1, here let's check the covariance first.

```
# Get the covariance of educ and nearc4
cov(card$educ, card$nearc4)
```

```
## [1] 0.179836
```

And we can see from the output above that education (educ) and whether someone grew up near a 4-year college (nearc4) are positively correlated.

2.4 Two Stage Least Squares (2SLS) and Reduced form

The IV estimator (effect of treatment) is actually

$$\hat{\tau}^{IV} = \frac{cov(Z_i, Y_i)}{cov(Z_i, D_i)}$$

2.4.1 First stage

- Regress endogenous D_i on all exogenous variables: $D_i = \gamma Z_i + \beta X_i + \eta_i$
- And store predicted value $\hat{D}_i$. It checks instrument condition.
- $\hat{\gamma}$ is the *effect of instrument on the treatment.*

2.4.2 Second stage

- Regress outcome Y_i on predicted $\hat{D}_i$ and other Xs: $Y_i = \tau \hat{D}_i + \delta X_i + \epsilon_i$
- $\hat{\tau}$ in this equation is our *IV estimate. The effect of our treatment on outcomes.*
- Attention: the standard error is *wrong* here! (Use canned routine).

2.4.3 Reduced form

- Regress Y_i on instrument Z_i: $Y_i = \alpha + \theta Z_i + \pi Xi + \eta_i$
- This does not recover $\hat{\tau}^{IV}$. But $\hat{\theta}$ tells you the *effect of instrument on outcome.*

$$\hat{\tau}^{IV} = \frac{\hat{\theta}}{\hat{\gamma}}$$

The IV estimate is just the *effect of the instrument on outcome, weighted by how much the instrument moves treatment.*

2.5 Education and Earning example to get IV estimate

If we just run the original regression

$$log(wage) = \beta_0 + \beta_1 education + \beta_2 experience + \beta_3 (experience)^2 + u$$

, we will get a biased estimate because of endogeneity problem.

```
biased_reg <- lm(lwage~educ+exper+expersq, data = card)
biased_estimate <- biased_reg$coefficients[[2]]
cat("The biased effect of education on earning in the original model is",biased_estimate)
```

```
## The biased effect of education on earning in the original model is 0.0931707
```

```
summary(biased_reg)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + expersq, data = card)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9375 -0.2592  0.0248  0.2682  1.4694
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.4685405  0.0686899  65.054  < 2e-16 ***
## educ         0.0931707  0.0035802  26.024  < 2e-16 ***
```

```
## exper        0.0897828  0.0070636  12.711  < 2e-16 ***
## expersq     -0.0024859  0.0003377  -7.361 2.35e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3982 on 3006 degrees of freedom
## Multiple R-squared:  0.1958, Adjusted R-squared:  0.195
## F-statistic:   244 on 3 and 3006 DF,  p-value: < 2.2e-16
```

2.5.1 First stage

In this step, we can run the following model to estimate the *effect of instrument on treatment*, which is the effect of whether someone grew up near a 4-year college (nearc4) on education. In this step, we also *save the predicted value of education* in this regression to use it in the second stage.

$$education = \gamma_0 + \gamma_1 near\ college + \gamma_2 experience + \gamma_3 (experience)^2 + u$$

```
# First stage
first <- lm(educ~nearc4+exper+expersq, data = card)
predict_educ <- fitted(first)
gamma <- first$coefficients[[2]]
cat("The effect of instrument on treatment is",gamma)
```

```
## The effect of instrument on treatment is 0.6002325
```

```
summary(first)
```

```
##
## Call:
## lm(formula = educ ~ nearc4 + exper + expersq, data = card)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4736 -1.4874 -0.1374  1.3529  5.6282
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.573446   0.170170  97.393  < 2e-16 ***
## nearc4       0.600232   0.078804   7.617 3.46e-14 ***
## exper       -0.422514   0.034817 -12.135  < 2e-16 ***
## expersq      0.000235   0.001704   0.138     0.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.009 on 3006 degrees of freedom
## Multiple R-squared:  0.4372, Adjusted R-squared:  0.4367
## F-statistic: 778.4 on 3 and 3006 DF,  p-value: < 2.2e-16
```

This step also shows that whether someone grew up near a 4-year college (nearc4) is positively correlated with education. And by checking the F-statistic 778.4, it is larger than 20, by rule of thumb, we know the instrument condition is satisfied.

2.5.2 Second Stage

In this step, we we can run the following model to estimate the *effect of our treatment on outcomes*, which is the effect of exogenous part of education on earnings, also the IV estimator.

$$log(wage) = \beta_0 + \beta_1 predicted\_education + \beta_2 experience + \beta_3 (experience)^2 + u$$

```
# Second Stage
second <- lm(lwage~predict_educ+exper+expersq, data = card)
tau_IV <- second$coefficients[[2]]
cat("The effect of treatment on outcome is", tau_IV)
```

```
## The effect of treatment on outcome is 0.2587155
```

```
summary(second)
```

```
##
## Call:
## lm(formula = lwage ~ predict_educ + exper + expersq, data = card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75292 -0.28332  0.02355  0.28756  1.45414
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6539851  0.4842956   3.415 0.000646 ***
## predict_educ  0.2587155  0.0284116   9.106  < 2e-16 ***
## exper         0.1596791  0.0141659  11.272  < 2e-16 ***
## expersq      -0.0024875  0.0003688  -6.745 1.82e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4348 on 3006 degrees of freedom
## Multiple R-squared:  0.04109,    Adjusted R-squared:  0.04013
## F-statistic: 42.94 on 3 and 3006 DF,  p-value: < 2.2e-16
```

Therefore, we can see that the IV estimate is very different from what we got in the original regression, which is negatively biased.

2.5.3 Reduced form

What we can also do besides the 2SLS is the reduced form regression, which is estimating the *effect of instrument on outcomes*, which is the effect of whether someone grew up near a 4-year college (nearc4) on earnings.

$$log(wage) = \theta_0 + \theta_1 near\ college + \theta_2 experience + \theta_3 (experience)^2 + u$$

```
# Reduced form
reduced_form <- lm(lwage~nearc4+exper+expersq, data = card)
theta <- reduced_form$coefficients[[2]]
cat("The effect of instrument on outcome is", theta)
```

```
## The effect of instrument on outcome is 0.1552895
```

```r
summary(reduced_form)
```

```
##
## Call:
## lm(formula = lwage ~ nearc4 + exper + expersq, data = card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75292 -0.28332  0.02355  0.28756  1.45414
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9417926  0.0368257 161.349  < 2e-16 ***
## nearc4       0.1552895  0.0170535   9.106  < 2e-16 ***
## exper        0.0503681  0.0075346   6.685 2.74e-11 ***
## expersq     -0.0024267  0.0003688  -6.579 5.55e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4348 on 3006 degrees of freedom
## Multiple R-squared:  0.04109,    Adjusted R-squared:  0.04013
## F-statistic: 42.94 on 3 and 3006 DF,  p-value: < 2.2e-16
```

We can see that whether someone grew up near a 4-year college (nearc4) is positively correlated with the earnings. *But this is not the IV estimator*. If we use the coefficient of first stage and reduced form, it is also possible to get the same IV estimator:

```r
theta/gamma
```

```
## [1] 0.2587155
```

2.5.4 The AER package

```r
library(AER)

iv <- ivreg(lwage ~ educ + exper + expersq | nearc4 + exper + expersq, data = card)
print(summary(iv))
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + expersq | nearc4 + exper +
##     expersq, data = card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41878 -0.33574  0.01427  0.34507  1.99293
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.6539851  0.5801706   2.851  0.00439 **
## educ         0.2587155  0.0340361   7.601 3.90e-14 ***
```

```
## exper        0.1596791  0.0169702   9.409  < 2e-16 ***
## expersq     -0.0024875  0.0004418  -5.631 1.96e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5209 on 3006 degrees of freedom
## Multiple R-Squared: -0.3762, Adjusted R-squared: -0.3775
## Wald test: 29.92 on 3 and 3006 DF,  p-value: < 2.2e-16
```