SOSC 5340 Tutorial Three

Matching, FE, DID, and Causal Forest

Yabin YIN
HKUST

April, 2020

Set working directory to the current directory

Remark: Need to save current R file before using *getActiveDocumentContext*

R Packages

R packages for matching estimator:

- *Matching*: https://cran.r-project.org/web/packages/Matching/
- *MatchIt*: https://cran.r-project.org/web/packages/MatchIt/index.html
- Read the *reference manual* and *vignettes*.
- Sekhon, J.S. Multivariate and propensity score matching software with automated balance optimization: the matching package for **R**. *Journal of Statistical Software*, 42(7): 1-52, 2011.

R packages for FE estimator:

- *plm*: https://cran.r-project.org/web/packages/plm/index.html
  - provides various estimators for linear models for panel data
  - can adjust standard errors
  - can perform various tests
  - can implement IV estimation
- *lfe*: https://cran.r-project.org/web/packages/lfe/index.html
  - linear models with multiple group fixed effects
  - deals with many levels of "fixed effect"
  - allows for multi-way clustering s.e.
  - can implement IV estimation
- *fixest*: https://cran.r-project.org/web/packages/fixest/index.html
  - fast for models with multiple fixed-effects
  - panel GLM, MLE, and non-linear MLE
- *pglm*: https://cran.r-project.org/web/packages/pglm/index.html
- Read the *reference manual* and *vignettes*.

R packages for Diff-in-Diffs estimator:

DID estimation can be done by the **lm()** function or functions from other packages.

DID is a common stratefy for natural experiments. New:

- Andrew Goodman-Bacon. 2018. Difference-in-Differences with Variation in Treatment Timing. (https://www.nber.org/papers/w25018)
- Anton Strezhnev. 2018. Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs. (https://www.antonstrezhnev.com/research)

R packages for causal forest:

- *grf*: https://cran.r-project.org/web/packages/grf/grf.pdf

Matching

We will use `Matching` package to match treatment and control group based on several methods.

We use data from **Dehejia and Wahba (1999 JASA)** as an example. This paper studied the effect of a job training (National Support Work) on the income of its participants. The job training is a random experiment, with 185 obs in the treatment group and 260 in the control group.

- *age*: age;
- *educ*: years of schooling;
- *black*: black or not;
- *hisp*: hispanic or not;
- *married*: married or not;
- *nodegr*: have high school diploma or not;
- *re74*, *re75*, *re78*: real earnings in 1974, 1975 and 1978, respectively;
- *u74*, *u75*: unemployed or not in 1974 and 1975, respectively;
- *treat*: participant of job training or not.

```
## library packages
library(Matching)
```

```
## Loading required package: MASS
```

```
## ##
## ##  Matching (Version 4.9-7, Build Date: 2020-02-05)
## ##  See http://sekhon.berkeley.edu/matching for additional documentation.
## ##  Please cite software as:
## ##   Jasjeet S. Sekhon. 2011. ``Multivariate and Propensity Score Matching
## ##   Software with Automated Balance Optimization: The Matching package for R.''
## ##   Journal of Statistical Software, 42(7): 1-52.
## ##
```

```
data('lalonde') ## Dehejia and Wahba (1999 JASA)

## data processing
Y <- lalonde$re78 ## Y is the dependent variable, income in 1978 (re78)
```

```r
Tr <- lalonde$treat ## Tr is an indicator of whether in the treatment group

## estimate the propensity scores using the glm() function
glm.ps <- glm(Tr ~ age + educ + black + hisp + married + nodegr + re74 + re75,
            family = binomial,
            data = lalonde)
```

Then, we will use `Match` function in `Matching` package to match. type *?Match* to see help document:

- **Y** is a vector containing the outcome of interest;
- **Tr** is a vector indicating the observations which are in the treatment regime and those which are not;
- **X** is a matrix containing the variables we wish to match on. This matrix may contain the actual observed covariates or the propensity score or a combination of both;
- **estimand** is a character string for the estimand. The default estimand is "ATT";
- **M** is a scalar for the number of matches which should be found. The default is one-to-one matching;
- **caliper** is the distance which is acceptable for any match. Observations which are outside of the caliper are dropped. For example, caliper=.25 means that all matches not equal to or within .25 standard deviations of each covariate in X are dropped;
- **replace** denotes whether matching should be done with replacement, by default is TURE. if `replace=F`, the order of matches generally matters. Matches will be found in the same order as the data are sorted. Matching without replacement will generally increase bias.

```r
## one-to-one matching with replacement, match on educ and marital status, ATT
match1 <- Match(Y=Y, Tr=Tr, X=lalonde[,c('educ', 'married')], replace = T)
summary(match1)
```

```
##
## Estimate...  1740.5
## AI SE......  738.67
## T-stat.....  2.3562
## p.val......  0.018461
##
## Original number of observations..............  445
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  5838
```

```r
## one-to-one matching without replacement, match on propensity score, ATT
match2 <- Match(Y = Y, Tr = Tr, X = glm.ps$fitted, replace = F)
summary(match2)
```

```
##
## Estimate...  2080.9
## SE.........  639.75
## T-stat.....  3.2527
## p.val......  0.0011431
##
## Original number of observations..............  445
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  185
```

```r
# one-to-one matching with replacement, match on propensity score, ATE
match3 <- Match(Y = Y, Tr = Tr, X = glm.ps$fitted, estimand = "ATE", replace = T)
summary(match3)
```

```
##
## Estimate...  2088.1
## AI SE......  726.19
## T-stat.....  2.8755
## p.val......  0.0040341
##
## Original number of observations..............  445
## Original number of treated obs...............  185
## Matched number of observations...............  445
## Matched number of observations  (unweighted).  725
```

```r
# one-to-multiple matching with replacement, match on propensity score, ATT
match4 <- Match(Y=Y, Tr = Tr, X = glm.ps$fitted, M=2, caliper = 0.25,
                replace = T)
summary(match4)
```

```
##
## Estimate...  2546.5
## AI SE......  753.12
## T-stat.....  3.3812
## p.val......  0.00072162
##
## Original number of observations..............  445
## Original number of treated obs...............  185
## Matched number of observations...............  181
## Matched number of observations  (unweighted).  475
##
## Caliper (SDs)........................................   0.25
## Number of obs dropped by 'exact' or 'caliper'  4
```

```r
# the following two are equivalent
m1 = Match(Y = Y, Tr = Tr, X = glm.ps$fitted)
m1 = Match(Y = Y, Tr = Tr, X = glm.ps$fitted, estimand = "ATT",
          M = 1, replace = TRUE)
```

Use `MatchBalance()` from `Matching` to examine how well the matching procedure did in producing balance. If the balance results printed by `MatchBalance` are not good enough, one would go back and change either the propensity score model or some parameter of how the matching is done.

```r
## Tests for Univariate Balance
MatchBalance(Tr ~ nodegr, match.out = match1, nboots = 1000, data = lalonde)
```

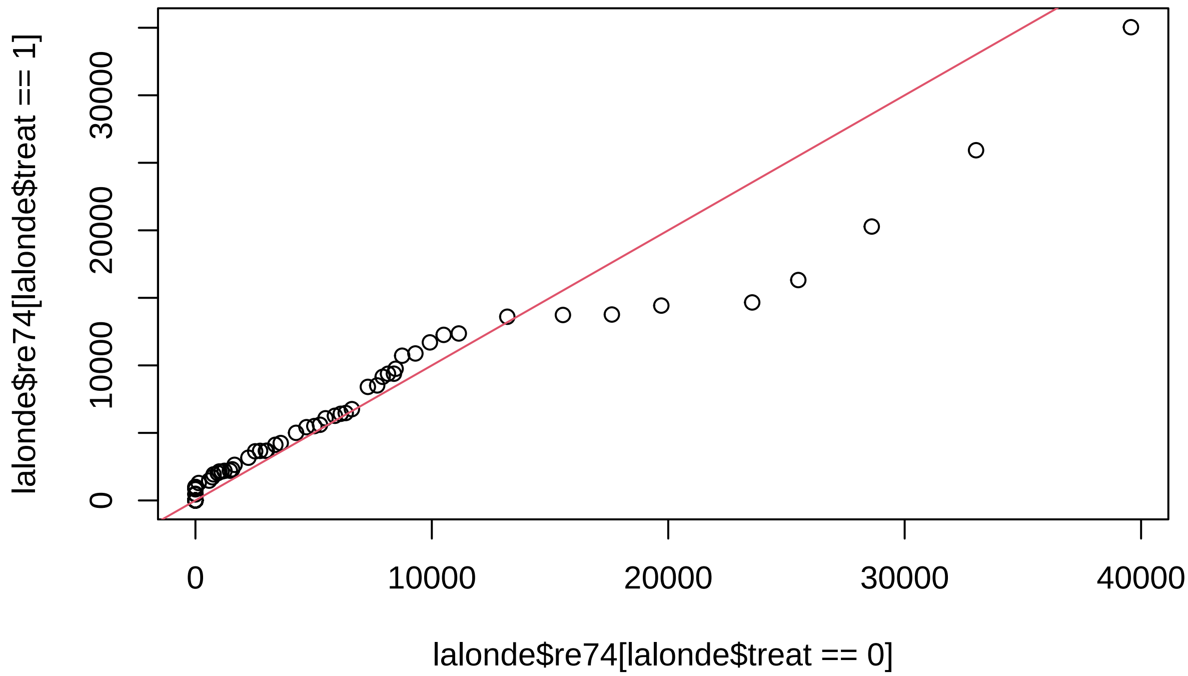
```
##
## ***** (V1) nodegr *****
##                        Before Matching        After Matching
## mean treatment........    0.70811            0.70811
## mean control..........    0.83462            0.70811
```

```
## std mean diff.........    -27.751                 0
##
## mean raw eQQ diff.....    0.12432                 0
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....          1                 0
##
## mean eCDF diff........   0.063254                 0
## med  eCDF diff........   0.063254                 0
## max  eCDF diff........    0.12651                 0
##
## var ratio (Tr/Co).....     1.4998                 1
## T-test p-value........  0.0020368                 1
```

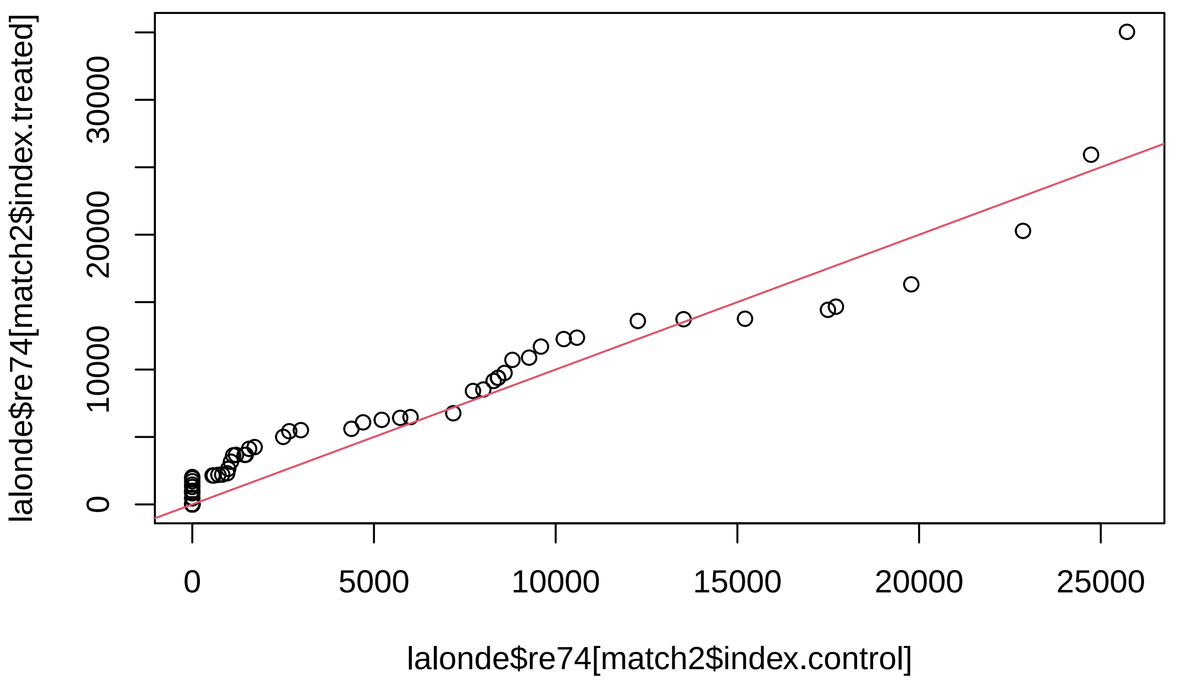
```r
MatchBalance(Tr ~ re74, match.out = match2, nboots = 1000, data = lalonde)
```

```
##
## ***** (V1) re74 *****
##                        Before Matching        After Matching
## mean treatment........     2095.6              2095.6
## mean control..........       2107              1744.9
## std mean diff.........   -0.23437              7.1753
##
## mean raw eQQ diff.....     487.98              502.29
## med  raw eQQ diff.....          0                   0
## max  raw eQQ diff.....       8413              9319.2
##
## mean eCDF diff........   0.019223            0.031411
## med  eCDF diff........     0.0158            0.021622
## max  eCDF diff........   0.047089            0.081081
##
## var ratio (Tr/Co).....     0.7381              1.1218
## T-test p-value........    0.98186             0.45906
## KS Bootstrap p-value..      0.559               0.186
## KS Naive p-value......    0.97023             0.57731
## KS Statistic..........   0.047089            0.081081
```

```r
## plot: before matching
qqplot(lalonde$re74[lalonde$treat==0], lalonde$re74[lalonde$treat==1])
abline(coef = c(0, 1), col = 2)
```




```r
## plot: after matching
qqplot(lalonde$re74[match2$index.control], lalonde$re74[match2$index.treated])
abline(coef = c(0, 1), col = 2)
```



Tests for Multivariate Balance

```r
## propensity score model proposed by Dehejia and Wahba (1999)
dw.pscore <- glm(Tr ~ age + I(age^2) + educ + I(educ^2) + black + hisp +
                   married + nodegr + re74 + I(re74^2) + re75 +
                   I(re75^2) + u74 + u75,
```

```r
                 family = binomial, data = lalonde)
# estimate the ATT
dw.rr <- Match(Y = Y, Tr = Tr, X = dw.pscore$fitted)
summary(dw.rr)
```

```
##
## Estimate...  2153.3
## AI SE......  825.4
## T-stat.....  2.6088
## p.val......  0.0090858
##
## Original number of observations..............  445
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  346
```

```r
# ## Tests for Multivariate Balance
MatchBalance(Tr ~ age + I(age^2) + educ + I(educ^2) + black + hisp +
               married + nodegr + re74 + I(re74^2) + re75 + I(re75^2) + u74 + u75 +
               I(re74 * re75) + I(age * nodegr) + I(educ * re74) + I(educ * re75),
             data = lalonde, match.out = dw.rr, nboots = 1000)
```

```
##
## ***** (V1) age *****
##                        Before Matching        After Matching
## mean treatment........     25.816             25.816
## mean control..........     25.054             25.006
## std mean diff.........     10.655             11.317
##
## mean raw eQQ diff.....    0.94054            0.41618
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          7                  9
##
## mean eCDF diff........   0.025364           0.010597
## med  eCDF diff........   0.022193          0.0086705
## max  eCDF diff........   0.065177           0.049133
##
## var ratio (Tr/Co).....     1.0278             1.0662
## T-test p-value........    0.26594            0.23472
## KS Bootstrap p-value..      0.528               0.55
## KS Naive p-value......     0.7481            0.79781
## KS Statistic..........   0.065177           0.049133
##
##
## ***** (V2) I(age^2) *****
##                        Before Matching        After Matching
## mean treatment........     717.39             717.39
## mean control..........     677.32             673.08
## std mean diff.........     9.2937             10.275
##
## mean raw eQQ diff.....     56.076             28.948
## med  raw eQQ diff.....         43                  0
## max  raw eQQ diff.....        721                909
```

```
##
## mean eCDF diff........   0.025364           0.010597
## med  eCDF diff........   0.022193          0.0086705
## max  eCDF diff........   0.065177           0.049133
##
## var ratio (Tr/Co).....     1.0115            0.91516
## T-test p-value........    0.33337            0.31819
## KS Bootstrap p-value..      0.528               0.55
## KS Naive p-value......     0.7481            0.79781
## KS Statistic..........   0.065177           0.049133
##
##
## ***** (V3) educ *****
##                        Before Matching        After Matching
## mean treatment........     10.346             10.346
## mean control..........     10.088              10.48
## std mean diff.........     12.806            -6.6749
##
## mean raw eQQ diff.....    0.40541            0.16185
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          2                  2
##
## mean eCDF diff........   0.028698           0.011561
## med  eCDF diff........   0.012682          0.0086705
## max  eCDF diff........    0.12651           0.052023
##
## var ratio (Tr/Co).....     1.5513             1.1917
## T-test p-value........    0.15017            0.45021
## KS Bootstrap p-value..       0.02              0.341
## KS Naive p-value......   0.062873            0.73726
## KS Statistic..........    0.12651           0.052023
##
##
## ***** (V4) I(educ^2) *****
##                        Before Matching        After Matching
## mean treatment........     111.06             111.06
## mean control..........     104.37             113.21
## std mean diff.........     17.012             -5.466
##
## mean raw eQQ diff.....     8.7189             3.1098
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....         60                 60
##
## mean eCDF diff........   0.028698           0.011561
## med  eCDF diff........   0.012682          0.0086705
## max  eCDF diff........    0.12651           0.052023
##
## var ratio (Tr/Co).....     1.6625             1.2716
## T-test p-value........   0.053676            0.51046
## KS Bootstrap p-value..       0.02              0.341
## KS Naive p-value......   0.062873            0.73726
## KS Statistic..........    0.12651           0.052023
##
##
```

```
## ***** (V5) black *****
##                        Before Matching        After Matching
## mean treatment........    0.84324           0.84324
## mean control..........    0.82692           0.85946
## std mean diff.........     4.4767           -4.4482
##
## mean raw eQQ diff.....   0.016216          0.0086705
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........  0.0081601          0.0043353
## med  eCDF diff........  0.0081601          0.0043353
## max  eCDF diff........    0.01632          0.0086705
##
## var ratio (Tr/Co).....    0.92503             1.0943
## T-test p-value........    0.64736            0.57783
##
##
## ***** (V6) hisp *****
##                        Before Matching        After Matching
## mean treatment........   0.059459           0.059459
## mean control..........    0.10769           0.048649
## std mean diff.........    -20.341             4.5591
##
## mean raw eQQ diff.....   0.048649          0.0057803
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........   0.024116          0.0028902
## med  eCDF diff........   0.024116          0.0028902
## max  eCDF diff........   0.048233          0.0057803
##
## var ratio (Tr/Co).....    0.58288             1.2083
## T-test p-value........   0.064043            0.41443
##
##
## ***** (V7) married *****
##                        Before Matching        After Matching
## mean treatment........    0.18919            0.18919
## mean control..........    0.15385            0.16667
## std mean diff.........     8.9995              5.735
##
## mean raw eQQ diff.....   0.037838           0.017341
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........   0.017672          0.0086705
## med  eCDF diff........   0.017672          0.0086705
## max  eCDF diff........   0.035343           0.017341
##
## var ratio (Tr/Co).....     1.1802             1.1045
## T-test p-value........    0.33425            0.46741
##
##
```

```
## ***** (V8) nodegr *****
##                       Before Matching        After Matching
## mean treatment........    0.70811           0.70811
## mean control..........    0.83462           0.69189
## std mean diff.........    -27.751            3.5572
##
## mean raw eQQ diff.....    0.12432          0.014451
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....          1                 1
##
## mean eCDF diff........   0.063254         0.0072254
## med  eCDF diff........   0.063254         0.0072254
## max  eCDF diff........    0.12651          0.014451
##
## var ratio (Tr/Co).....     1.4998           0.96957
## T-test p-value........  0.0020368           0.49161
##
##
## ***** (V9) re74 *****
##                       Before Matching        After Matching
## mean treatment........     2095.6            2095.6
## mean control..........       2107            1624.3
## std mean diff.........   -0.23437            9.6439
##
## mean raw eQQ diff.....     487.98            467.33
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....       8413             12410
##
## mean eCDF diff........   0.019223          0.019782
## med  eCDF diff........     0.0158          0.018786
## max  eCDF diff........   0.047089          0.046243
##
## var ratio (Tr/Co).....     0.7381            2.2663
## T-test p-value........    0.98186           0.22745
## KS Bootstrap p-value..      0.584             0.233
## KS Naive p-value......    0.97023            0.8532
## KS Statistic..........   0.047089          0.046243
##
##
## ***** (V10) I(re74^2) *****
##                       Before Matching        After Matching
## mean treatment........   28141434          28141434
## mean control..........   36667413          13117852
## std mean diff.........    -7.4721            13.167
##
## mean raw eQQ diff.....   13311731          10899373
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....  365146387         616156569
##
## mean eCDF diff........   0.019223          0.019782
## med  eCDF diff........     0.0158          0.018786
## max  eCDF diff........   0.047089          0.046243
##
## var ratio (Tr/Co).....    0.50382            7.9006
```

```
## T-test p-value........    0.51322           0.08604
## KS Bootstrap p-value..      0.584             0.233
## KS Naive p-value......    0.97023            0.8532
## KS Statistic..........   0.047089          0.046243
##
##
## ***** (V11) re75 *****
##                        Before Matching        After Matching
## mean treatment........     1532.1            1532.1
## mean control..........     1266.9            1297.6
## std mean diff.........     8.2363            7.2827
##
## mean raw eQQ diff.....     367.61            211.42
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....     2110.2            8195.6
##
## mean eCDF diff........   0.050834          0.023047
## med  eCDF diff........   0.061954          0.023121
## max  eCDF diff........    0.10748          0.057803
##
## var ratio (Tr/Co).....     1.0763            1.4291
## T-test p-value........    0.38527           0.33324
## KS Bootstrap p-value..      0.058             0.171
## KS Naive p-value......    0.16449           0.60988
## KS Statistic..........    0.10748          0.057803
##
##
## ***** (V12) I(re75^2) *****
##                        Before Matching        After Matching
## mean treatment........   12654753          12654753
## mean control..........   11196530           8896263
## std mean diff.........     2.6024            6.7076
##
## mean raw eQQ diff.....    2840830           2887443
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....  101657197         344942969
##
## mean eCDF diff........   0.050834          0.023047
## med  eCDF diff........   0.061954          0.023121
## max  eCDF diff........    0.10748          0.057803
##
## var ratio (Tr/Co).....     1.4609             3.559
## T-test p-value........    0.77178           0.37741
## KS Bootstrap p-value..      0.058             0.171
## KS Naive p-value......    0.16449           0.60988
## KS Statistic..........    0.10748          0.057803
##
##
## ***** (V13) u74 *****
##                        Before Matching        After Matching
## mean treatment........    0.70811           0.70811
## mean control..........       0.75           0.68458
## std mean diff.........    -9.1895            5.1608
##
```

```
## mean raw eQQ diff.....   0.037838          0.017341
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....          1                 1
##
## mean eCDF diff........   0.020946         0.0086705
## med  eCDF diff........   0.020946         0.0086705
## max  eCDF diff........   0.041892          0.017341
##
## var ratio (Tr/Co).....     1.1041           0.95721
## T-test p-value........    0.33033           0.52298
##
##
## ***** (V14) u75 *****
##                        Before Matching        After Matching
## mean treatment........        0.6               0.6
## mean control..........    0.68462           0.62072
## std mean diff.........    -17.225           -4.2182
##
## mean raw eQQ diff.....   0.081081          0.031792
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....          1                 1
##
## mean eCDF diff........   0.042308          0.015896
## med  eCDF diff........   0.042308          0.015896
## max  eCDF diff........   0.084615          0.031792
##
## var ratio (Tr/Co).....     1.1133            1.0194
## T-test p-value........   0.068031           0.46507
##
##
## ***** (V15) I(re74 * re75) *****
##                        Before Matching        After Matching
## mean treatment........   13118591          13118591
## mean control..........   14530303           8958064
## std mean diff.........    -2.7799            8.1928
##
## mean raw eQQ diff.....    3278733           3085879
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....  188160151         211819713
##
## mean eCDF diff........   0.022723          0.014519
## med  eCDF diff........   0.014449          0.014451
## max  eCDF diff........   0.061019          0.037572
##
## var ratio (Tr/Co).....    0.69439            2.7882
## T-test p-value........    0.79058           0.30299
## KS Bootstrap p-value..       0.31             0.385
## KS Naive p-value......    0.81575           0.96754
## KS Statistic..........   0.061019          0.037572
##
##
## ***** (V16) I(age * nodegr) *****
##                        Before Matching        After Matching
## mean treatment........     17.968            17.968
```

```
## mean control..........     20.608            17.294
## std mean diff.........    -20.144            5.1366
##
## mean raw eQQ diff.....     2.7189           0.60405
## med  raw eQQ diff.....          1                 0
## max  raw eQQ diff.....         18                17
##
## mean eCDF diff........   0.020386         0.0090105
## med  eCDF diff........  0.0061331         0.0072254
## max  eCDF diff........    0.12651          0.037572
##
## var ratio (Tr/Co).....     1.3301           0.98044
## T-test p-value........   0.027633           0.48453
## KS Bootstrap p-value..      0.027              0.83
## KS Naive p-value......   0.062873           0.96754
## KS Statistic..........    0.12651          0.037572
##
##
## ***** (V17) I(educ * re74) *****
##                        Before Matching        After Matching
## mean treatment........      22899             22899
## mean control..........      21067             17069
## std mean diff.........      3.191            10.157
##
## mean raw eQQ diff.....     4775.1            5443.8
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....     173996            267977
##
## mean eCDF diff........   0.018141          0.016409
## med  eCDF diff........   0.015281          0.014451
## max  eCDF diff........    0.04553          0.049133
##
## var ratio (Tr/Co).....     1.1152            2.9191
## T-test p-value........    0.73471           0.18059
## KS Bootstrap p-value..      0.619             0.195
## KS Naive p-value......    0.97849           0.79781
## KS Statistic..........    0.04553          0.049133
##
##
## ***** (V18) I(educ * re75) *****
##                        Before Matching        After Matching
## mean treatment........      15881             15881
## mean control..........      12981             13051
## std mean diff.........     8.5349            8.3267
##
## mean raw eQQ diff.....     3760.4            2235.4
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....      46244            124045
##
## mean eCDF diff........   0.050006          0.022441
## med  eCDF diff........   0.064293          0.020231
## max  eCDF diff........     0.1052          0.057803
##
## var ratio (Tr/Co).....     1.1901            1.6746
```

```
## T-test p-value........    0.35903            0.25369
## KS Bootstrap p-value..      0.067              0.177
## KS Naive p-value......    0.18269            0.60988
## KS Statistic..........     0.1052           0.057803
##
##
## Before Matching Minimum p.value: 0.0020368
## Variable Name(s): nodegr  Number(s): 8
##
## After Matching Minimum p.value: 0.08604
## Variable Name(s): I(re74^2)  Number(s): 10
```

Note: Sometimes matching even gives you a worse result, you may find the variable *re74* is the case

Recover the Matched Dataset

```
## recover datasets
treated.data <- lalonde[dw.rr$index.treated, ]
control.data <- lalonde[dw.rr$index.control, ]
matched.data <- rbind(treated.data, control.data)

## extract variables
Y2 <- dw.rr$mdata$Y # the outcome vector of matched dataset
Tr2 <- dw.rr$mdata$Tr # the treatment indicator of matched dataset
X2 <- dw.rr$mdata$X # The X matrix contains matched pairs.
```

Fixed Effect

Let's use `plm`, `lfe` and `fixest` to fit fixed effect model.

Empirical example: *Aghion, Van Reenen, and Zingales (2013 AER)*

@Aghion2013Innovation studied the relationship between institutional ownership and innovation. We replicate column 1 of Table 1 of this paper (see page 283).

```
## library packages
library(plm)
library(lfe)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'lfe'
```

```
## The following object is masked from 'package:plm':
##
##     sargan
```

```
library(fixest)
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'lmtest'

## The following object is masked from 'package:lfe':
##
##     waldtest
```

```r
## load the data: from the "sandwich" package (Aghion, Van Reenen, and Zingales, 2013 AER)
data("InstInnovation")

## Least Square Dummy Variable (LSDV)
### with firm dummies and time dummies (Fixed effects as a dummy variable model)
fe_lsdv <- lm(log(cites+1)~institutions+log(I(capital/employment)+1)+log(sales+1)
              +factor(industry)+factor(year),
              data = InstInnovation)
se_lsdv <- coeftest(fe_lsdv, vcov. = vcovCL(fe_lsdv, cluster = ~industry+year))[,2]
```

```
## Warning in sqrt(diag(se)): NaNs produced
```

When fitting a fixed effect model on panel data, `plm()` is preferred than LSDV. - **effect**: 'individual', 'time', 'twoways', or 'nested'; - **model**: 'pooling'(pooled OLS), 'within'(fixed effect), 'between'(between), 'random'(random effects), 'fd'(first differences).

```r
## Fixed effect using `plm`
### transform to panel data
InstInnovation_p <- pdata.frame(InstInnovation, index = c("company", "year"), drop.index = TRUE)
### note: index identifies id and time

## Within estimator: one-way (time) FE + industry FE(Fixed effects as deviation from means)
fe_within <- plm(log(cites+1)~institutions+log(I(capital/employment)+1)+
                   log(sales+1)+factor(industry),
                 effect = "time",
                 model = "within",
                 data = InstInnovation_p)
se_within <- coeftest(fe_within,
                      vcov. = vcovHC(fe_within, cluster = "group"))[,2]

## First-difference: with industry dummies (Fixed effects as difference in time)
fe_fd <- plm(log(cites+1)~institutions+log(I(capital/employment)+1)+
               log(sales+1)+factor(industry),
             effect = "individual",
             model = "fd",
             data = InstInnovation_p)
```

```r
se_fd <- coeftest(fe_fd,
                  vcov. = vcovHC(fe_fd, cluster ="group"))[,2]

# show the results
library(texreg)
```

```
## Version:  1.37.5
## Date:     2020-06-17
## Author:   Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
```

```r
screenreg(list(fe_lsdv, fe_within, fe_fd),
          se = list(se_lsdv, se_within, se_fd),
          custom.model.names = c("ln(Cites) LSDV", "ln(Cites) Within", "ln(Cites) FD"),
          custom.coef.names = c("Share of institutions", "ln(K/L)", "ln(Sales)"),
          omit.coef = c("(Intercept)|(industry)|(company)|(year)"),
          stars = c(0.01, 0.05, 0.1),
          digits = 4)
```

```
##
## =====================================================================
##                          ln(Cites) LSDV  ln(Cites) Within  ln(Cites) FD
## ---------------------------------------------------------------------
## Share of institutions     0.0060 ***      0.0060 ***        0.0018
##                          (0.0010)        (0.0010)          (0.0015)
## ln(K/L)                   0.4304 ***      0.4304 ***        0.2614 **
##                          (0.0391)        (0.0391)          (0.1025)
## ln(Sales)                 0.6123 ***      0.6123 ***        0.1439 *
##                          (0.0138)        (0.0138)          (0.0765)
## ---------------------------------------------------------------------
## R^2                       0.5753          0.5020            0.0020
## Adj. R^2                  0.5650          0.4900            0.0015
## Num. obs.              6208            6208              5405
## =====================================================================
## *** p < 0.01; ** p < 0.05; * p < 0.1
```

Note: First difference gives us very different results from within fixed effect. It's because FD and FE have different assumptions and FD usually generate missing values. Generally, we prefer results from FE and use FD as a robustness check.

Now, let's fit a twoway fixed effect model, fixing at company level and year level.

```r
## LSDV (with company dummies and year dummies)
fe_lsdv2 <- lm(log(cites+1)~institutions+log(I(capital/employment)+1)+log(sales+1)
               +factor(company)+factor(year),
               data = InstInnovation)
se_lsdv2 <- coeftest(fe_lsdv2,
                     vcov. = vcovCL(fe_lsdv2, cluster = ~company+year))[,2]
```

```
## Warning in sqrt(diag(se)): NaNs produced
```

```r
## twoway fixed effect
fe_within2 <- plm(log(cites+1)~institutions+log(I(capital/employment)+1)+log(sales+1),
                  effect = "twoways",
                  model = "within",
                  data = InstInnovation_p)
se_within2 <- coeftest(fe_within2,
                       vcov. = vcovHC(fe_within2, cluster = 'group'))[,2]
```

Alternative packages: `lfe` and `fixest`, more efficient with large panels, and clustered and robust standard errors are handled more elegantly compared to `plm`

```r
## the felm() function from the lfe package
fe_1 <- felm(log(cites+1)~institutions+log(I(capital/employment)+1)+log(sales+1) # Y and Xs
             | company + year # fixed effects
             | 0 # IVs
             | company+year, # clusters
             data = InstInnovation)

## compare the results
screenreg(list(fe_lsdv2, fe_within2, fe_1),
          se = list(se_lsdv2, se_within2,
                    summary(fe_1)$coefficients[,2]),
          custom.model.names = c("LSDV Firm+Year",
                                 "Within Firm+Year(plm)",
                                 "Within Firm+Year(felm)"),
          custom.coef.names = c("Share of institutions", "ln(K/L)", "ln(Sales)"),
          omit.coef = c("(Intercept)|(industry)|(company)|(year)"),
          stars = c(0.01, 0.05, 0.1),
          digits = 4)
```

```
##
## =======================================================================================
##                           LSDV Firm+Year  Within Firm+Year(plm)  Within Firm+Year(felm)
## ---------------------------------------------------------------------------------------
## Share of institutions     0.0020          0.0020                 0.0020
##                          (0.0014)        (0.0014)               (0.0026)
## ln(K/L)                   0.0390          0.0390                 0.0390
##                          (0.0751)        (0.0751)               (0.1407)
## ln(Sales)                 0.0726          0.0726                 0.0726
##                          (0.0479)        (0.0479)               (0.1011)
## ---------------------------------------------------------------------------------------
## R^2                       0.8040          0.0009
## Adj. R^2                  0.7745         -0.1497
## Num. obs.              6208            6208                   6208
## R^2 (full model)                                                 0.8040
## R^2 (proj model)                                                 0.0009
## Adj. R^2 (full model)                                            0.7745
## Adj. R^2 (proj model)                                           -0.1497
## Num. groups: company                                           803
## Num. groups: year                                                9
## =======================================================================================
## *** p < 0.01; ** p < 0.05; * p < 0.1
```

```r
# the feols() function from the fixest package
fe_2 <- feols(I(log(cites+1))~institutions+
                log(I(capital/employment)+1)+log(sales+1) # Y and Xs
              |company+year, # fixed effects
              data = InstInnovation)
summary(fe_2, cluster=~company+year)
```

```
## OLS estimation, Dep. Var.: I(log(cites + 1))
## Observations: 6,208
## Fixed-effects: company: 803,  year: 9
## Standard-errors: Two-way (company & year)
##                                Estimate Std. Error  t value Pr(>|t|)
## institutions                   0.001998   0.002585 0.772616 0.461963
## log(I(capital/employment) + 1) 0.039018   0.142341 0.274116 0.790940
## log(sales + 1)                 0.072647   0.102412 0.709362 0.498249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.0053     Adj. R2: 0.774509
##                Within R2: 8.885e-4
```

Diff in Diffs

DID estimation can be done by the **lm()** function or functions from other packages.

Empirical example: **Card and Krueger (1994 AER)**, this paper examines the effect of minimum wage increase on the employment:

- **fte**: full time-equivalent employees
- **nj**: =1 if New Jersey (first d: location difference)
- **d**: =1 if after NJ mini wage increases (second d: time difference)

```r
## library packages
library(foreign)
## load data: Card and Krueger (1994 AER)
minwage <- read.dta("njmin3.dta")

# regression
did <- lm(fte~nj*d, data = minwage)
summary(did)
```

```
##
## Call:
## lm(formula = fte ~ nj * d, data = minwage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.166  -6.439  -1.027   4.473  64.561
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.331      1.072  21.767   <2e-16 ***
```

```
## nj             -2.892      1.194  -2.423   0.0156 *
## d              -2.166      1.516  -1.429   0.1535
## nj:d            2.754      1.688   1.631   0.1033
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.406 on 790 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.007401,   Adjusted R-squared:  0.003632
## F-statistic: 1.964 on 3 and 790 DF,  p-value: 0.118
```

Causal Forest (Advanced)

Basically, causal forest predicts the counterfactual, then we will get an estimation of individual level treatment effect $\tau_i = Y_i^1 - Y_i^0$ (see lecture 7 slides).

We use `grf` package to fit it. data used here is from *Dehejia and Wahba (1999 JASA)*.

```
## library packages and load data
library(grf)

## split data into training and test sets
set.seed(333)
train <- sample(1:nrow(lalonde), round(nrow(lalonde) * .5))
trainset <- lalonde[train, ]
testset <- lalonde[-train, ]
```

Now let's fit the causal forest using `causal_forest()` function from `grf` package. The `causal_forest()` has 3 primary inputs:

- **X** is a matrix of the covariates which we are using to predict heterogeneity in treatment effects;
- **Y** is a vector of the outcome of interest;
- **W** is the treatment assignment.

The crucial thing here is that all of these must be numeric, which means that we need to dummy code the factor variables.

```
X = as.matrix(trainset[, -c(9, 12)])
Y = trainset$re78
W = as.numeric(trainset$treat)

## fit a causal forest
cf <- causal_forest(X = X, Y = Y, W = W, num.trees = 5000, seed = 333)
```

Estimate CATE and CATT using `average_treatment_effect()` function

```
# Estimate the conditional average treatment effect on the full sample (CATE).
average_treatment_effect(cf, target.sample = "all")
```

```
## estimate  std.err
## 195.5849 817.9638
```

```r
# Estimate the conditional average treatment effect on the treated sample (CATT).
average_treatment_effect(cf, target.sample = "treated")
```

```
## estimate  std.err
## 334.5393 857.0773
```

Predict on test set

```r
preds <- predict(object = cf,
                 newdata = as.matrix(testset[, -c(9, 12)]),
                 estimate.variance = TRUE) # tell grf to include variance estimates

## assign the predictions (the estimated treatment effects) to the test data frame so that we can use t
testset$preds <- preds$predictions
testset$se <- sqrt(preds$variance.estimates)
```

We would also like to know the nature of the heterogeneity: What variables are useful for targeting based on treatment effects?

The `grf` package also has a `variable_importance()` function to realize it.

```r
## variable importance
library(dplyr)
```

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

## The following objects are masked from 'package:plm':
##
##     between, lag, lead

## The following object is masked from 'package:MASS':
##
##     select

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

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

```r
cf %>%
  variable_importance() %>%
  as.data.frame() %>%
  mutate(variable = colnames(cf$X.orig)) %>%
  arrange(desc(V1))
```

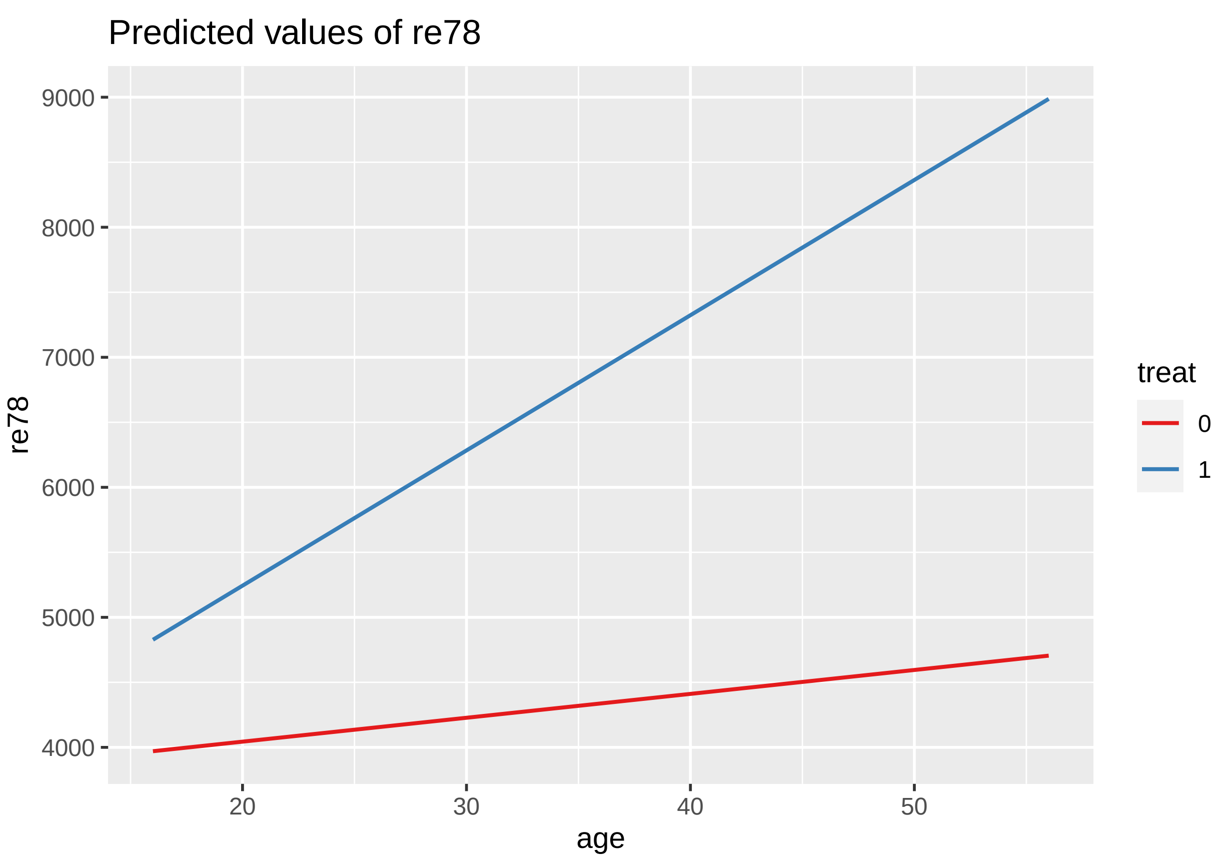
```
##               V1 variable
## 1  0.392963160      age
## 2  0.209172409     educ
## 3  0.136536896     re75
## 4  0.118915821     re74
## 5  0.049440238      u74
## 6  0.048663194   nodegr
## 7  0.023248646      u75
## 8  0.019443364  married
## 9  0.001616271    black
## 10 0.000000000     hisp
```

plot individual level treatment effect on covariates

```
library(ggplot2)
library(sjPlot)
```

```
## Install package "strengejacke" from GitHub (`devtools::install_github("strengejacke/strengejacke")`)
```

```
## traditional linear interaction
lm_interaction <- lm(re78 ~ age*treat+.-re78, data = lalonde)
plot_model(lm_interaction, type = "int", ci.lvl = NA)
```




```
## individual treatment effect
trainset$age2 <- cut(trainset$age, breaks = c(0, 20, 25, 30, 35, 40, 45, Inf),
```

```
                      right = F, labels = c(1:7))

ate <- data.frame()
for (i in 1:7) {
  df <- as.data.frame(t(average_treatment_effect(cf, target.sample = "all",
                                              subset=trainset$age2==i)))
  ate <- rbind(ate, df)
}
ate$age <- c(20, 25, 30, 35, 40, 45, 50)

ate %>% ggplot() +
  geom_line(aes(x = age, y = estimate)) +
  geom_line(aes(x = age, y = estimate+1.96*std.err), linetype='dashed')+
  geom_line(aes(x = age, y = estimate-1.96*std.err), linetype='dashed')+
  labs(x='age', y='CATE')+
  theme_light()
```