Exercises: Week March 30

Econometrics Prof. Conlon

Ulrich Atz

2021-03-30

This weeks packages

```
library(tidyverse)
library(sampleSelection)
library(MatchIt)
```

Selection Example

The code for this example can be found at: https://cran.r-project.org/web/packages/sampleSelection/vignettes/selection.pdf

1. In this case we are going to work backwards. I will give you the code that esimates the selection model, and you will write down the equations (with estimated coefficients) and explain what is the selection problem, and how is it addressed here.

The data is described as follows:

> The Mroz87 data frame contains data about 753 married women. These data are collected within the "Panel Study of Income Dynamics" (PSID). Of the 753 observations, the first 428 are for women with positive hours worked in 1975, while the remaining 325 observations are for women who did not work for pay in 1975.

We are interested in a regression model that explains wage as dependent variable with education as presumably the key variable of interest. Labor force participation, however, is conditional on a woman's situation; here modeled as depending on age, family income, number of kids, and education. Age and experience for example are correlated. The average effect of experience on wage therefore suffers likely from a selection bias.

First step regression:

$$P(\widehat{\text{lfp} = 1}) = \Phi[-4.16 + 0.19(\text{age}) - 0.002(\text{age}^2) + 0.00005(\text{faminc}) - 0.45(\text{kids}_{\text{TRUE}}) + 0.1(\text{educ})]$$

Second step regression:

$$\widehat{\text{wage}} = -0.97 + 0.02(\text{exper}) + 0.0001(\text{exper}^2) + 0.42(\text{educ}) + 0.44(\text{city}) - 1.1(\text{inv\_mills})$$

The second method estimates the parameters via maximum likelihood, i.e in one step. We can find the log-likelihood function as equation (12) in the vignette.

```
greeneTS <- selection( lfp ~ age + I( age^2 ) + faminc + kids + educ,
    wage ~ exper + I( exper^2 ) + educ + city,
    data = Mroz87, method = "2step" )
summary(greeneTS)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 753 observations (325 censored and 428 observed)
## 14 free parameters (df = 740)
## Probit selection equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.157e+00  1.402e+00  -2.965 0.003127 **
## age          1.854e-01  6.597e-02   2.810 0.005078 **
## I(age^2)    -2.426e-03  7.735e-04  -3.136 0.001780 **
## faminc       4.580e-06  4.206e-06   1.089 0.276544
## kidsTRUE    -4.490e-01  1.309e-01  -3.430 0.000638 ***
## educ         9.818e-02  2.298e-02   4.272 2.19e-05 ***
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.9712003  2.0593505  -0.472    0.637
## exper        0.0210610  0.0624646   0.337    0.736
## I(exper^2)   0.0001371  0.0018782   0.073    0.942
## educ         0.4170174  0.1002497   4.160 3.56e-05 ***
## city         0.4438379  0.3158984   1.405    0.160
## Multiple R-Squared:0.1264,   Adjusted R-Squared:0.116
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio   -1.098      1.266  -0.867    0.386
## sigma            3.200         NA      NA       NA
## rho             -0.343         NA      NA       NA
## --------------------------------------------
```

```r
Mroz87$yhat <- predict(greeneTS)
```

```r
greeneML <- selection( lfp ~ age + I( age^2 ) + faminc + kids + educ,
    wage ~ exper + I( exper^2 ) + educ + city, data = Mroz87,
    maxMethod = "BHHH", iterlim = 500 )
summary(greeneML)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## Maximum Likelihood estimation
## BHHH maximisation, 62 iterations
## Return code 8: successive function values within relative tolerance limit (reltol)
## Log-Likelihood: -1581.259
## 753 observations (325 censored and 428 observed)
## 13 free parameters (df = 740)
## Probit selection equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.120e+00  1.410e+00  -2.921  0.00359 **
## age          1.840e-01  6.584e-02   2.795  0.00532 **
## I(age^2)    -2.409e-03  7.735e-04  -3.115  0.00191 **
## faminc       5.676e-06  3.890e-06   1.459  0.14493
## kidsTRUE    -4.507e-01  1.367e-01  -3.298  0.00102 **
## educ         9.533e-02  2.400e-02   3.973  7.8e-05 ***
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.9537242  1.6745690  -1.167    0.244
## exper        0.0284295  0.0753989   0.377    0.706
```

```
## I(exper^2)  -0.0001151  0.0023339  -0.049    0.961
## educ         0.4562471  0.0959626   4.754 2.39e-06 ***
## city         0.4451424  0.4255420   1.046    0.296
##    Error terms:
##       Estimate Std. Error t value Pr(>|t|)
## sigma  3.10350    0.08368  37.088   <2e-16 ***
## rho   -0.13328    0.22296  -0.598     0.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

2. Explain the difference between the two-step and MLE estimates above. How does the procedure differ? Which do you prefer and why?

The 2-step procedure first estimates a probit model and uses the results to construct the inverse Mill's ration, which then gets used as an additional variable in the second step, the OLS model.

ML ought to be theoretically most efficient. However, as the vignette points out, "the two-step solution allows certain generalisations more easily than ML, and is more robust in certain circumstances." For example, the optimization algorithm may not converge.

3. Now compare these results to a naive OLS regression of just the outcome (wages) that does not account for the selection effects from labor force participation. How do the coefficients in the outcome equation change?

Experience matters a lot more in the model, e.g. the coefficients are statistically significant. The coefficient on education appears to not change much.

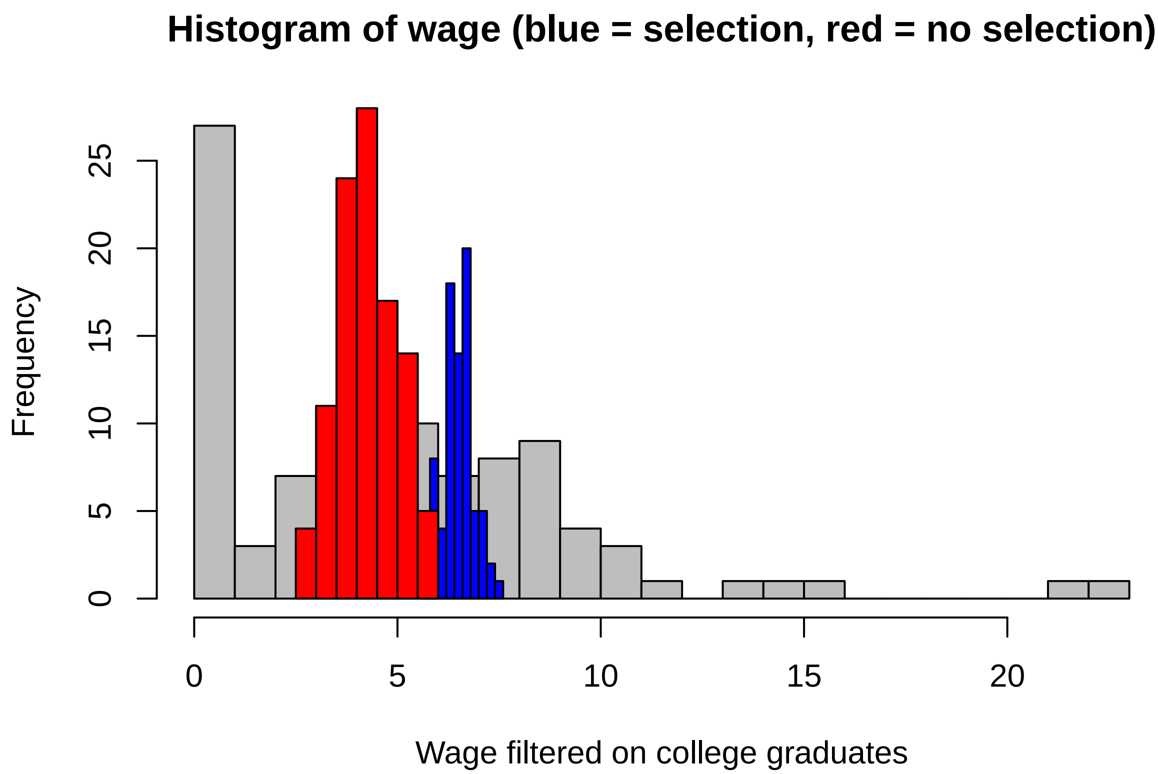
```r
naive <- lm(wage ~ exper + I( exper^2 ) + educ + city, data = Mroz87)
Mroz87$yhat_naive <- predict(naive)

summary(naive)
```

```
##
## Call:
## lm(formula = wage ~ exper + I(exper^2) + educ + city, data = Mroz87)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0863 -1.7436 -0.4114  1.1285 23.7700
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.183011   0.613924  -6.814 1.96e-11 ***
## exper        0.187911   0.039026   4.815 1.78e-06 ***
## I(exper^2)  -0.003277   0.001259  -2.602  0.00944 **
## educ         0.414801   0.048553   8.543  < 2e-16 ***
## city         0.072734   0.229465   0.317  0.75135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.975 on 748 degrees of freedom
## Multiple R-squared:  0.1621, Adjusted R-squared:  0.1576
## F-statistic: 36.17 on 4 and 748 DF,  p-value: < 2.2e-16
```

4. Plot the distribution of observed wages and predicted wages for college graduates (education >= 16) for the model with and without selection for labor force participation.

The model with selection

```
hist(Mroz87[Mroz87$educ >= 16, "wage"], col = 'gray', breaks = 30,
     main = "Histogram of wage (blue = selection, red = no selection)",
     xlab = "Wage filtered on college graduates")
hist(Mroz87[Mroz87$educ >= 16, "yhat"], col = 'blue', add = T)
hist(Mroz87[Mroz87$educ >= 16, "yhat_naive"], col = 'red', add = T)
```



Matching

Following the vignette at: https://cran.r-project.org/web/packages/MatchIt/vignettes/MatchIt.html#assessing-the-quality-of-matches

1. Discuss the balance table using the following unadjusted sample

The balance table gives the standardized mean differences (exception: cobalt binary variables). The covariates are clearly not balanced at conventional levels.

```
m.out0 <- matchit(treat ~ age + educ + race + married +
                     nodegree + re74 + re75, data = lalonde,
                 method = NULL, distance = "glm")
summary(m.out0)
```

```
##
## Call:
## matchit(formula = treat ~ age + educ + race + married + nodegree +
##     re74 + re75, data = lalonde, method = NULL, distance = "glm")
##
## Summary of Balance for All Data:
```

```
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance          0.5774        0.1822          1.7941     0.9211    0.3774
## age              25.8162       28.0303         -0.3094     0.4400    0.0813
## educ             10.3459       10.2354          0.0550     0.4959    0.0347
## raceblack         0.8432        0.2028          1.7615          .    0.6404
## racehispan        0.0595        0.1422         -0.3498          .    0.0827
## racewhite         0.0973        0.6550         -1.8819          .    0.5577
## married           0.1892        0.5128         -0.8263          .    0.3236
## nodegree          0.7081        0.5967          0.2450          .    0.1114
## re74           2095.5737     5619.2365         -0.7211     0.5181    0.2248
## re75           1532.0553     2466.4844         -0.2903     0.9563    0.1342
##            eCDF Max
## distance     0.6444
## age          0.1577
## educ         0.1114
## raceblack    0.6404
## racehispan   0.0827
## racewhite    0.5577
## married      0.3236
## nodegree     0.1114
## re74         0.4470
## re75         0.2876
##
##
## Sample Sizes:
##           Control Treated
## All           429     185
## Matched       429     185
## Unmatched       0       0
## Discarded       0       0
```

```
cobalt::bal.tab(m.out0, thresholds = c(m = .05))
```

```
## Call
##  matchit(formula = treat ~ age + educ + race + married + nodegree +
##     re74 + re75, data = lalonde, method = NULL, distance = "glm")
##
## Balance Measures
##                  Type Diff.Un      M.Threshold.Un
## distance     Distance  1.7941
## age           Contin. -0.3094 Not Balanced, >0.05
## educ          Contin.  0.0550 Not Balanced, >0.05
## race_black     Binary  0.6404 Not Balanced, >0.05
## race_hispan    Binary -0.0827 Not Balanced, >0.05
## race_white     Binary -0.5577 Not Balanced, >0.05
## married        Binary -0.3236 Not Balanced, >0.05
## nodegree       Binary  0.1114 Not Balanced, >0.05
## re74          Contin. -0.7211 Not Balanced, >0.05
## re75          Contin. -0.2903 Not Balanced, >0.05
##
## Balance tally for mean differences
##                     count
## Balanced, <0.05         0
## Not Balanced, >0.05     9
##
```

```
## Variable with the greatest mean difference
##  Variable Diff.Un      M.Threshold.Un
##      re74 -0.7211 Not Balanced, >0.05
##
## Sample sizes
##     Control Treated
## All     429     185
```

2. Perform 4 nearest neighbor matching using the Mahalbanois distance and the above covariates for real earnings in 1978. Give me your best estimate of the ATE and ATT of the treatment status.

```
nn4_att <- matchit(treat ~ age + educ + race + married + nodegree + re74 + re75,
               data = lalonde,
               method ="nearest", distance = "mahalanobis", ratio = 4,
               estimand = "ATT")
```

```
## Warning: Not all treated units will get 4 matches.
```

```
d_nn4_att <-  match.data(nn4_att)

att <- lm(educ ~ treat + age + race + married + nodegree + re74 + re75,
          data = d_nn4_att, weights = weights)

#
# lmtest::coeftest(ate, vcov. = sandwich::vcovHC) %>%
#   broom::tidy() %>% filter(term == "treat") %>% mutate(term = "ATE"),
lmtest::coeftest(att, vcov. = sandwich::vcovCL, cluster = ~subclass) %>%
  broom::tidy() %>% filter(term == "treat") %>% mutate(term = "ATT")
```

```
## # A tibble: 1 x 5
##   term  estimate std.error statistic p.value
##   <chr>    <dbl>     <dbl>     <dbl>   <dbl>
## 1 ATT      0.467     0.159      2.94 0.00338
```

3. Is the ATE greater or less than the ATT, explain why this is a sensible outcome and what this implies for the ATUT.