selection-Apr8.R

danny

2020-04-08

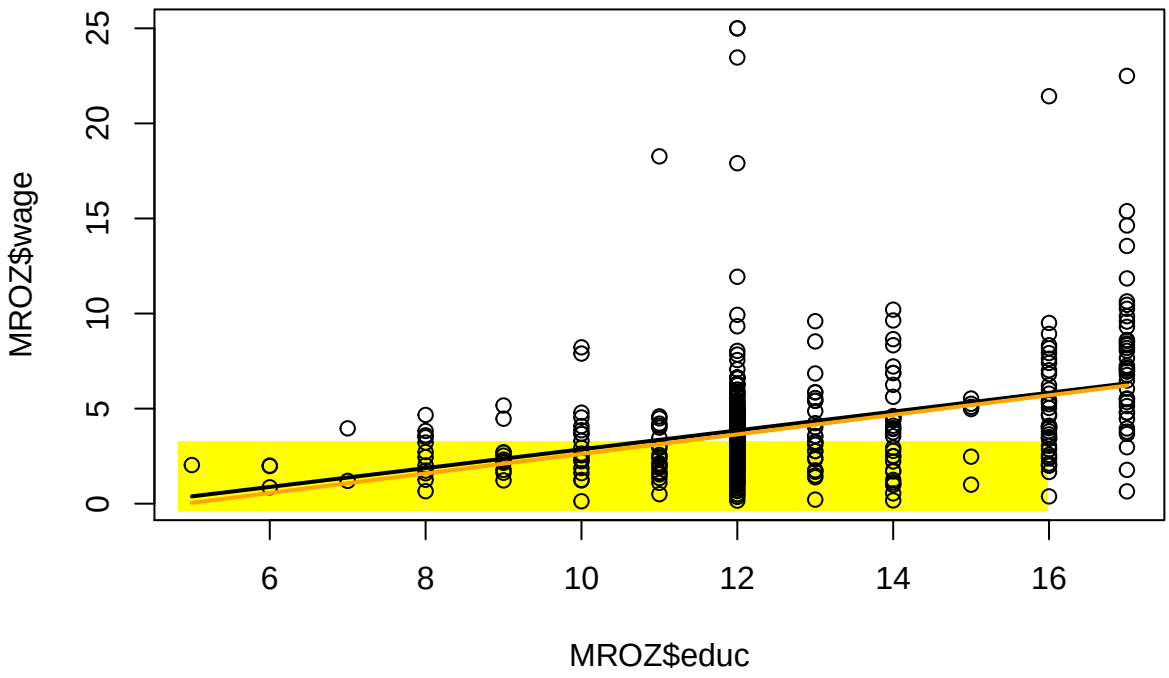
```
# Author: Gordon Burtch and Gautam Ray
# Course: MSBA 6440
# Session: Selection and Measurement Error
# Topic: Selection Model Example
# Lecture 9


suppressWarnings(suppressPackageStartupMessages({
library(sampleSelection)
library(stargazer)
}))

MROZ <-read.csv("MROZ.csv")

MROZ$kids <- (MROZ$kidslt6 + MROZ$kidsge6)

# Female labor supply (lfp = labour force participation)

## Outcome equations without correcting for selection
# I() means "as-is" -- do calculation in parentheses then use as variable

## Comparison of linear regression and selection model

outcome1 <- lm(wage ~ educ, data = MROZ)
summary(outcome1)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = MROZ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6797 -1.6658 -0.4556  0.8794 21.1487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.09237    0.84829  -2.467    0.014 *
## educ         0.49531    0.06595   7.511 3.49e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.114 on 426 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.1169, Adjusted R-squared:  0.1149
## F-statistic: 56.41 on 1 and 426 DF,  p-value: 3.486e-13
```

```r
selection1 <- selection(selection = lfp ~ age + I(age^2) + faminc + kidslt6 + educ,
                        outcome = wage ~ educ,
                        data = MROZ, method = "2step")
summary(selection1)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 753 observations (325 censored and 428 observed)
## 11 free parameters (df = 743)
## Probit selection equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.399e-01  1.514e+00  -0.092    0.926
## age         -1.174e-02  6.876e-02  -0.171    0.864
## I(age^2)    -2.567e-04  7.808e-04  -0.329    0.742
## faminc       3.233e-06  4.297e-06   0.752    0.452
## kidslt6     -8.531e-01  1.144e-01  -7.457 2.47e-13 ***
## educ         1.166e-01  2.365e-02   4.931 1.01e-06 ***
## Outcome equation:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.52489    1.30609  -1.933   0.0536 .
## educ         0.51403    0.07869   6.532  1.2e-10 ***
## Multiple R-Squared:0.1173,   Adjusted R-Squared:0.1132
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio   0.3149     0.7235   0.435    0.663
## sigma           3.1151         NA      NA       NA
## rho             0.1011         NA      NA       NA
## --------------------------------------------
```

```r
plot(MROZ$wage ~ MROZ$educ)
curve(outcome1$coeff[1] + outcome1$coeff[2]*x, col="black", lwd="2", add=TRUE)
curve(selection1$coeff[7] + selection1$coeff[8]*x, col="orange", lwd="2", add=TRUE)
```




```
## A more complete model comparison

outcome2 <- lm(wage ~ exper + I( exper^2 ) + educ + city, data = MROZ)
summary(outcome2)
```

```
##
## Call:
## lm(formula = wage ~ exper + I(exper^2) + educ + city, data = MROZ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6021 -1.6012 -0.4787  0.8950 21.2762
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.5609920  0.9288390  -2.757  0.00608 **
## exper        0.0324982  0.0615864   0.528  0.59800
## I(exper^2)  -0.0002602  0.0018378  -0.142  0.88747
## educ         0.4809623  0.0668679   7.193 2.91e-12 ***
## city         0.4492741  0.3177735   1.414  0.15815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.111 on 423 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.1248, Adjusted R-squared:  0.1165
```

```
## F-statistic: 15.08 on 4 and 423 DF,  p-value: 1.569e-11
```

```
## Correcting for selection

selection.twostep2 <- selection(selection = lfp ~ age + I(age^2) + faminc + kidslt6 + educ,
                                outcome = wage ~ exper + I(exper^2) + educ + city,
                                data = MROZ, method = "2step")
summary(selection.twostep2)
```

```
## --------------------------------------------
## 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) -1.399e-01  1.514e+00  -0.092    0.926
## age         -1.174e-02  6.876e-02  -0.171    0.864
## I(age^2)    -2.567e-04  7.808e-04  -0.329    0.742
## faminc       3.233e-06  4.297e-06   0.752    0.452
## kidslt6     -8.531e-01  1.144e-01  -7.457 2.48e-13 ***
## educ         1.166e-01  2.365e-02   4.931 1.01e-06 ***
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.7413454  1.3679742  -2.004   0.0454 *
## exper        0.0334859  0.0614715   0.545   0.5861
## I(exper^2)  -0.0003096  0.0018477  -0.168   0.8670
## educ         0.4887549  0.0795133   6.147 1.29e-09 ***
## city         0.4467138  0.3162288   1.413   0.1582
## Multiple R-Squared:0.1248,   Adjusted R-Squared:0.1145
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio  0.13220    0.73970   0.179    0.858
## sigma          3.09469         NA      NA       NA
## rho            0.04272         NA      NA       NA
## --------------------------------------------
```

```
selection.mle <- selection(selection = lfp ~ age + I(age^2) + faminc + kids + educ,
                           outcome = wage ~ exper + I(exper^2) + educ + city,
                           data = MROZ, method = "mle")
summary(selection.mle)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 3 iterations
## Return code 2: successive function values within tolerance limit
## Log-Likelihood: -1579.498
## 753 observations (325 censored and 428 observed)
## 13 free parameters (df = 740)
## Probit selection equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.709e+00  1.399e+00  -2.652 0.008183 **
```

```
## age          1.649e-01  6.484e-02   2.543 0.011182 *
## I(age^2)    -2.189e-03  7.541e-04  -2.903 0.003808 **
## faminc       4.581e-06  4.525e-06   1.012 0.311667
## kids        -1.507e-01  3.830e-02  -3.935  9.1e-05 ***
## educ         9.061e-02  2.341e-02   3.870 0.000118 ***
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.2332665  1.3302676  -1.679   0.0936 .
## exper        0.0291691  0.0620275   0.470   0.6383
## I(exper^2)  -0.0001513  0.0018553  -0.082   0.9350
## educ         0.4679380  0.0766012   6.109 1.62e-09 ***
## city         0.4467800  0.3160013   1.414   0.1578
##    Error terms:
##       Estimate Std. Error t value Pr(>|t|)
## sigma  3.09755    0.10907  28.400   <2e-16 ***
## rho   -0.07081    0.20547  -0.345     0.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

```
## Heckman model selection "by hand" ##

seleqn1 <- glm(lfp ~ age + I(age^2) + faminc + kidslt6 + educ, family=binomial(link="probit"),
               data=MROZ)
summary(seleqn1)
```

```
##
## Call:
## glm(formula = lfp ~ age + I(age^2) + faminc + kidslt6 + educ,
##     family = binomial(link = "probit"), data = MROZ)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0359  -1.1386   0.6860   0.9789   2.1831
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.399e-01  1.507e+00  -0.093    0.926
## age         -1.174e-02  6.852e-02  -0.171    0.864
## I(age^2)    -2.567e-04  7.784e-04  -0.330    0.742
## faminc       3.233e-06  4.353e-06   0.743    0.458
## kidslt6     -8.531e-01  1.149e-01  -7.425 1.13e-13 ***
## educ         1.166e-01  2.367e-02   4.926 8.38e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  931.42  on 747  degrees of freedom
## AIC: 943.42
##
## Number of Fisher Scoring iterations: 4
```

```r
## Calculate inverse Mills ratio by hand ##

MROZ$IMR <- dnorm(seleqn1$linear.predictors)/pnorm(seleqn1$linear.predictors)

## Outcome equation correcting for selection ##

outeqn1 <- lm(wage ~ exper + I(exper^2) + educ + city + IMR, data=MROZ, subset=(lfp==1))
summary(outeqn1)
```

```
##
## Call:
## lm(formula = wage ~ exper + I(exper^2) + educ + city + IMR, data = MROZ,
##     subset = (lfp == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6074 -1.6048 -0.4736  0.8876 21.2940
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.7413490  1.3773664  -1.990   0.0472 *
## exper        0.0334859  0.0619076   0.541   0.5889
## I(exper^2)  -0.0003096  0.0018608  -0.166   0.8679
## educ         0.4887551  0.0800561   6.105 2.33e-09 ***
## city         0.4467137  0.3184647   1.403   0.1614
## IMR          0.1322070  0.7448157   0.178   0.8592
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.115 on 422 degrees of freedom
## Multiple R-squared:  0.1248, Adjusted R-squared:  0.1145
## F-statistic: 12.04 on 5 and 422 DF,  p-value: 6.495e-11
```

```r
## compare to selection package -- coefficients right, se's wrong

summary(selection.twostep2)
```

```
## --------------------------------------------
## 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) -1.399e-01  1.514e+00  -0.092    0.926
## age         -1.174e-02  6.876e-02  -0.171    0.864
## I(age^2)    -2.567e-04  7.808e-04  -0.329    0.742
## faminc       3.233e-06  4.297e-06   0.752    0.452
## kidslt6     -8.531e-01  1.144e-01  -7.457 2.48e-13 ***
## educ         1.166e-01  2.365e-02   4.931 1.01e-06 ***
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.7413454  1.3679742  -2.004   0.0454 *
```

```
## exper        0.0334859  0.0614715   0.545   0.5861
## I(exper^2)  -0.0003096  0.0018477  -0.168   0.8670
## educ         0.4887549  0.0795133   6.147 1.29e-09 ***
## city         0.4467138  0.3162288   1.413   0.1582
## Multiple R-Squared:0.1248,   Adjusted R-Squared:0.1145
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio  0.13220    0.73970   0.179    0.858
## sigma          3.09469         NA      NA       NA
## rho            0.04272         NA      NA       NA
## -----------------------------------------------
```

```
stargazer(outeqn1,selection.twostep2,type="text",title="Heckman Two-step vs.Heckman by Hand",
          column.labels = c("Heckman By Hand","Heckman Command"))
```

```
##
## Heckman Two-step vs.Heckman by Hand
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                         wage
##                                  OLS              selection
##                            Heckman By Hand     Heckman Command
##                                  (1)                 (2)
## ---------------------------------------------------------------
## exper                            0.033               0.033
##                                 (0.062)             (0.061)
##
## I(exper2)                       -0.0003             -0.0003
##                                 (0.002)             (0.002)
##
## educ                           0.489***            0.489***
##                                 (0.080)             (0.080)
##
## city                             0.447               0.447
##                                 (0.318)             (0.316)
##
## IMR                              0.132
##                                 (0.745)
##
## Constant                        -2.741**            -2.741**
##                                 (1.377)             (1.368)
##
## ---------------------------------------------------------------
## Observations                      428                 753
## R2                               0.125
## Adjusted R2                      0.114
## rho                                                  0.043
## Inverse Mills Ratio                                0.132 (0.740)
## Residual Std. Error    3.115 (df = 422)
## F Statistic         12.040*** (df = 5; 422)
## ===============================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
```