Week 4: Endogeneity, Training Exercises

Coursera/Erasmus U., Econometric Methods and Applications

Anthony Nguyen

Training Exercise 4.5

Notes:

- This exercise uses the datafile `TrainExer45`and requires a computer.
- The dataset `TrainExer45` is available on the website.

Questions

In this exercise we reconsider the example of lecture 4.5. In this lecture we related the Grade Point Average (GPA) of learners in an engineering MOOC to the participation in a preparatory course. The dataset contains the following variables:

- GPA: Grade Point Average in Engineerinc MOOC
- Gender: 0/1 dummy for gender (1: male, 0: female)
- Participation: 0/1 dummy for participation in a preparatory mathematics course (1: did participate, 0: did not participate)
- Email: 0/1 dummy for receiving an email invitation to take the preparatory course (1: received invitation, 0: did not receive invitation)

(a) Redo the OLS estimation of the coefficients in a model that explains GPA using a constant, gender and prepartory course participation. Also calculate the standard errors and t-values. Confirm that you obtain the same results as mentioned in the lecture.

(b) Use the email dummy as an instrument to perform 2SLS estimation. First do the first-stage regression:

$$\text{Participation} = \gamma_1 + \gamma_2 \text{Gender} + \gamma_3 \text{Email} + \eta.$$

Next calculate the predicted values according to this regression and perform OLS on the model:

$$\text{GPA} = \beta_1 + \beta_2 \text{Gender} + \beta_3 \widehat{\text{Participation}} + \varepsilon.$$

Confirm that the parameter estimates that you obtain are the same as reported in the lecture.

(c) Obtain the standard errors that correspond to the final regression in the previous part. These do **not** match with the standard errors reported in the lecture! Why are the standard errors from part (b) wrong?

(d) Calculate the ratio between the standard errors in part (b) and those reported in the lecture. Why is the obtained ratio the same for all parameters? Explain how we can also obtain this ratio using different residual series.

Answers

(a) Redo the OLS estimation of the coefficients in a model that explains GPA using a constant, gender and prepartory course participation. Also calculate the standard errors and t-values. Confirm that you obtain the same results as mentioned in the lecture.

```
mod1 <- lm(GPA~GENDER+PARTICIPATION, data = TrainExer45)

summary(mod1)
```

```
##
## Call:
## lm(formula = GPA ~ GENDER + PARTICIPATION, data = TrainExer45)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00981 -0.48166  0.01022  0.47071  1.80200
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.77111    0.03397 169.874  < 2e-16 ***
## GENDER        -0.21376    0.04431  -4.824 1.63e-06 ***
## PARTICIPATION  0.82437    0.04686  17.593  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6985 on 997 degrees of freedom
## Multiple R-squared:  0.2441, Adjusted R-squared:  0.2426
## F-statistic:   161 on 2 and 997 DF,  p-value: < 2.2e-16
```

(b) Use the email dummy as an instrument to perform 2SLS estimation.

First do the first-stage regression:

```
# Perform first stage regression on PARTICIPATION by instruments
mod2 <- lm(PARTICIPATION~GENDER+EMAIL, data = TrainExer45)

# Save fitted values
PAR_FIT <- predict(mod2)

#display regression results
summary(mod2)
```

```
##
## Call:
## lm(formula = PARTICIPATION ~ GENDER + EMAIL, data = TrainExer45)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5625 -0.1496 -0.1011  0.4375  0.8989
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.10112    0.02290   4.415 1.12e-05 ***
## GENDER       0.04846    0.02690   1.801   0.0719 .
## EMAIL        0.41290    0.02690  15.349  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 997 degrees of freedom
## Multiple R-squared:  0.1956, Adjusted R-squared:  0.194
## F-statistic: 121.2 on 2 and 997 DF,  p-value: < 2.2e-16
```

Next calculate the predicted values according to this regression and perform OLS on the model:

```
mod3 <- lm(GPA~GENDER+PAR_FIT, data = TrainExer45)

summary(mod3)
```

```
##
## Call:
## lm(formula = GPA ~ GENDER + PAR_FIT, data = TrainExer45)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07155 -0.57838 -0.05513  0.59162  2.23636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.94787    0.05117 116.239  < 2e-16 ***
## GENDER      -0.17276    0.05121  -3.373 0.000771 ***
## PAR_FIT      0.24050    0.12246   1.964 0.049819 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7981 on 997 degrees of freedom
## Multiple R-squared:  0.01325,    Adjusted R-squared:  0.01127
## F-statistic: 6.694 on 2 and 997 DF,  p-value: 0.001294
```

(c) Obtain the standard errors that correspond to the final regression in the previous part. These do not match with the standard errors reported in the lecture! Why are the standard errors from part (b) wrong?

The standard errors from our 2SLS in part (b) are:

0.0511694, 0.0512149, 0.1224594

These differ from the lecture, because here, the wrong estimate was used to calculate the variance of epsilon. In other words, we *should* use the residuals from the regression on PARTICIPATION and *not* the residuals from the fitted values of PARTICIPATION (PAR_FIT) from the first stage of the 2SLS.

The ratio of the Standard errors from part (b) and the standard errors from the lecture are:

$$\frac{\text{SE part (b)}}{\text{SE lecture}} = 1.063$$

So, the standard errors obtained in the 2SLS are about 6% too high. The correct estimate of the variance of epsilon is based on the residuals obtained using *actual* participation, not *predicted* participation. Therefore, the ratio 1.063 is equal to:

$$\sqrt{\frac{\hat{\sigma^2}_{correct}}{\hat{\sigma^2}_{incorrect}}}$$