Test Exercise 4

Jason

25/01/2018

Questions

A challenging and very relevant economic problem is the measurement of the returns to schooling. In this question we will use the following variables on 3010 US men:

- *logw*: log wage
- *educ*: number of years of schooling
- *age*: age of the individual in years
- *exper*: working experience in years
- *smsa*: dummy indicating whether the individual lived in a metropolitan area
- *south*: dummy indicating whether the individual lived in the south
- *nearc*: dummy indicating whether the individual lived near a 4-year college
- *dadeduc*: education of the individual's father (in years)
- *momeduc*: education of the individual's mother (in years)

This data is a selection of the data used by D. Card (1995)1

Question 1

Use OLS to estimate the parameters of the model

$$logw = 1 + 2educ + 3exper + 4exper2 + 5smsa + 6south + .$$

Give an interpretation to the estimated 2 coefficient

```
wageData$experSquared <- (wageData$exper)^2

olsModel <- lm(logw ~ educ + exper + experSquared + smsa + south, data = wageData)
olsModel$coefficients
```

```
##  (Intercept)         educ        exper experSquared         smsa
##  4.611014446  0.081579706  0.083835685 -0.002202115  0.150800573
##        south
## -0.175176080
```

The coefficient of 2 which is 0.084, indicates that for every additional year of schooling the individual has, the value of log wage is expected to increase by 0.084.

Question 2

OLS may be inconsistent in this case as educ and exper may be endogenous. Give a reason why this may be the case. Also indicate whether the estimate in part (a) is still useful.

Variables *educ* and *exper* are likely to be endogenous as they may be influence by other factors such as cost of education and motivation to study/work. There may also be a correlation between *educ* and *exper* such that longer education time means less experience in the workforce.

As such, if the variables are not exogenous, they cause OLS to be inconsistent.

Question 3

Give a motivation why *age* and age^2 can be used as instruments for *exper* and $exper^2$.

age and age^2 can be used as instruments as they are both exogneous, have strong correlation with exper and exper2 and have no correlation with ϵ

Question 4

Run the first-stage regression for educ for the two-stage least squares estimation of the parameters in the model above when age, age2, nearc, dadeduc, and momeduc are used as additional instruments. What do you conclude about the suitability of these instruments for schooling?

```
wageData$ageSquared <- (wageData$age)^2
model1Stage <- lm(educ ~ age + ageSquared + nearc + daded + momed, data = wageData)
```

Based on the t-statistic of variables *age*, age^2, *nearc*, *daded* and *momed*, they are all correlated with the endogenous variable *educ* with *daded* and *momed* the having the strongest correlation. This makes sense since parents that have strong educational backgrounds will more than likely have a strong influence on the education of their child. As such, they are all suitable to be used as instruments.

Question 5

Estimate the parameters of the model for log wage using two-stage least squares where you correct for the endogeneity of education and experience. Compare your result to the estimate in part (a).

```
model2SLS <- ivreg(logw ~ educ + exper + experSquared + smsa + south | age + ageSquared + nearc + daded
model2SLS$coefficients
```

```
##  (Intercept)         educ        exper experSquared         smsa
##  4.416903899  0.099842919  0.072866858 -0.001639293  0.134937031
##        south
## -0.158986861
```

Comparing this model to (a), variables *educ*, *exper*, *smsa* still have a positive effect on the log wage value with the positive effect of *educ* slightly stronger then the previous model. All other variables still have a similar negative effect on log wage.

Question 6

Perform the Sargan test for validity of the instruments. What is your conclusion?

```
summary(model2SLS, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = logw ~ educ + exper + experSquared + smsa + south |
##     age + ageSquared + nearc + daded + momed + smsa + south,
##     data = wageData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7494 -0.2360  0.0266  0.2498  1.3468
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4169039  0.1154208  38.268  < 2e-16 ***
## educ          0.0998429  0.0065738  15.188  < 2e-16 ***
## exper         0.0728669  0.0167134   4.360 1.35e-05 ***
## experSquared -0.0016393  0.0008381  -1.956   0.0506 .
## smsa          0.1349370  0.0167695   8.047 1.21e-15 ***
## south        -0.1589869  0.0156854 -10.136  < 2e-16 ***
##
## Diagnostic tests:
##                                  df1  df2 statistic p-value
## Weak instruments (educ)            5 3002   145.511 < 2e-16 ***
## Weak instruments (exper)           5 3002  1257.258 < 2e-16 ***
## Weak instruments (experSquared)    5 3002  1098.430 < 2e-16 ***
## Wu-Hausman                         2 3002     5.709 0.00335 **
## Sargan                             2   NA     3.702 0.15705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3844 on 3004 degrees of freedom
## Multiple R-Squared: 0.2512,  Adjusted R-squared: 0.2499
## Wald test: 175.9 on 5 and 3004 DF,  p-value: < 2.2e-16
```

```
qchisq(0.05, df=5, lower.tail = FALSE)
```

```
## [1] 11.0705
```

The Sargan test statistic which is 3.702 is lower than the 5% of chi square distribution with 5 degrees of freedom which is 11.07. As such we cannot reject the null hypothesis, and therefore the instruments are valid.

Appendix

```
summary(olsModel)
```

```
##
## Call:
## lm(formula = logw ~ educ + exper + experSquared + smsa + south,
##     data = wageData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71487 -0.22987  0.02268  0.24898  1.38552
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6110144  0.0678950  67.914  < 2e-16 ***
## educ          0.0815797  0.0034990  23.315  < 2e-16 ***
## exper         0.0838357  0.0067735  12.377  < 2e-16 ***
## experSquared -0.0022021  0.0003238  -6.800 1.26e-11 ***
## smsa          0.1508006  0.0158360   9.523  < 2e-16 ***
## south        -0.1751761  0.0146486 -11.959  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3813 on 3004 degrees of freedom
## Multiple R-squared:  0.2632, Adjusted R-squared:  0.2619
## F-statistic: 214.6 on 5 and 3004 DF,  p-value: < 2.2e-16
```

```
summary(model1Stage)
```

```
##
## Call:
## lm(formula = educ ~ age + ageSquared + nearc + daded + momed,
##     data = wageData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4573  -1.4968  -0.2734   1.6843   7.5636
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.923273   4.010502  -1.477 0.139796
## age          0.992550   0.281060   3.531 0.000419 ***
## ageSquared  -0.017075   0.004878  -3.500 0.000472 ***
## nearc        0.528751   0.092698   5.704 1.28e-08 ***
## daded        0.202048   0.015665  12.898  < 2e-16 ***
## momed        0.248379   0.017036  14.580  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.346 on 3004 degrees of freedom
## Multiple R-squared:  0.233,  Adjusted R-squared:  0.2317
## F-statistic: 182.5 on 5 and 3004 DF,  p-value: < 2.2e-16
```

```
summary(model2SLS)
```

```
##
## Call:
## ivreg(formula = logw ~ educ + exper + experSquared + smsa + south |
##     age + ageSquared + nearc + daded + momed + smsa + south,
##     data = wageData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7494 -0.2360  0.0266  0.2498  1.3468
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4169039  0.1154208  38.268  < 2e-16 ***
## educ          0.0998429  0.0065738  15.188  < 2e-16 ***
## exper         0.0728669  0.0167134   4.360 1.35e-05 ***
## experSquared -0.0016393  0.0008381  -1.956   0.0506 .
## smsa          0.1349370  0.0167695   8.047 1.21e-15 ***
## south        -0.1589869  0.0156854 -10.136  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3844 on 3004 degrees of freedom
```

```
## Multiple R-Squared: 0.2512,  Adjusted R-squared: 0.2499
## Wald test: 175.9 on 5 and 3004 DF,  p-value: < 2.2e-16
```