Week4Assignment

Pradeepta Das

16 November 2020

Let's see the data

```
head(data)
```

```
##       logw educ age exper smsa south nearc daded momed
## 1 6.306275    7  29    16    1     0     0  9.94 10.25
## 2 6.175867   12  27     9    1     0     0  8.00  8.00
## 3 6.580639   12  34    16    1     0     0 14.00 12.00
## 4 5.521461   11  27    10    1     0     1 11.00 12.00
## 5 6.591674   12  34    16    1     0     1  8.00  7.00
## 6 6.214608   12  26     8    1     0     1  9.00 12.00
```

```
summary(data)
```

```
##       logw            educ            age            exper         smsa
##  Min.   :4.605   Min.   : 1.00   Min.   :24.00   Min.   : 0.000   0: 864
##  1st Qu.:5.977   1st Qu.:12.00   1st Qu.:25.00   1st Qu.: 6.000   1:2146
##  Median :6.287   Median :13.00   Median :28.00   Median : 8.000
##  Mean   :6.262   Mean   :13.26   Mean   :28.12   Mean   : 8.856
##  3rd Qu.:6.564   3rd Qu.:16.00   3rd Qu.:31.00   3rd Qu.:11.000
##  Max.   :7.785   Max.   :18.00   Max.   :34.00   Max.   :23.000
##  south     nearc        daded            momed
##  0:1795   0: 957   Min.   : 0.000   Min.   : 0.00
##  1:1215   1:2053   1st Qu.: 8.000   1st Qu.: 9.00
##                    Median : 9.940   Median :11.00
##                    Mean   : 9.989   Mean   :10.34
##                    3rd Qu.:12.000   3rd Qu.:12.00
##                    Max.   :18.000   Max.   :18.00
```

(a) Use OLS to estimate the parameters of the model and Give an interpretation to the estimated β_2 coefficient.

```
data$exper_sq <- (data$exper)^2
ols_model <- lm(logw~educ+exper+exper_sq+smsa+south, data=data)
summary(ols_model)
```

```
##
## Call:
## lm(formula = logw ~ educ + exper + exper_sq + smsa + south, data = data)
##
## 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 ***
## exper_sq    -0.0022021  0.0003238  -6.800 1.26e-11 ***
## smsa1        0.1508006  0.0158360   9.523  < 2e-16 ***
## south1      -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
```

the coefficient for education is β_2 which is +ve. This indecates that log wage is positively correlated to education. Therefore, with each additional year of schooling the wage increases by about exp(0.082), or by 1.085 or ~8.5%.

(b) 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.

Endogeneous means the explanatory variables are stochastic and are correlated to the residuals. These might also occue due to measurement errors. In this case the OLS doesn't properly estimate β. (as, for n -> inf the OLS estimator would converge to wrong β / diverge!)

It is possible the wage, experience and education variables to be affected by some other variable (i.e. ability, social class, family support, etc.) in a way, such as, a higher ability to lead to a higher wage, longer education and less experience (due to long education) and vice versa.

In this case, these variables would be endogenous and the OLS estimates would be biased and inconsistent, therefore not useful anymore.

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

Age is obviously exogenous as it cannot be influenced by the people, and it is also obviously related to experience as younger people cannot have a very long experience.

So it's a good instrument for the experience variable. And the same applies for their squared values.

(d) Run the first-stage regression for educ for the two-stage least squares estimation of the parameters in the model above when age, age^2, nearc, dadeduc, and momeduc are used as additional instruments. What do you conclude about the suitability of these instruments for schooling?

```r
data$age_sq <- data$age^2
edu_model <- lm(formula = educ ~ age + age_sq + smsa + south + nearc + daded + momed, data = data)
summary(edu_model)
```

```
##
## Call:
## lm(formula = educ ~ age + age_sq + smsa + south + nearc + daded +
##     momed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2777  -1.5450  -0.2224   1.6957   7.2250
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.652354   3.976343  -1.421 0.155277
## age          0.989610   0.278714   3.551 0.000390 ***
## age_sq      -0.017019   0.004838  -3.518 0.000441 ***
## smsa1        0.529566   0.101504   5.217 1.94e-07 ***
## south1      -0.424851   0.091037  -4.667 3.19e-06 ***
## nearc1       0.264554   0.099085   2.670 0.007626 ** 
## daded        0.190443   0.015611  12.199  < 2e-16 ***
## momed        0.234515   0.017028  13.773  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.326 on 3002 degrees of freedom
## Multiple R-squared:  0.2466, Adjusted R-squared:  0.2448 
## F-statistic: 140.4 on 7 and 3002 DF,  p-value: < 2.2e-16
```

The additional instruments (age, age^2, nearc, daded, and momed) are significantly correlated with the education. This is especially true about the later two (daded and momed) due to their high t-statistics, which makes perfect sense as highly educated parents are more likely to support and promote their children education as well.

So, the instrument variables and the endogenous variable educ are significantly related.

```r
data$educ_f <- edu_model$fitted.values
summary(edu_model$fitted.values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    7.87   12.53   13.39   13.26   14.14   17.10
```

```r
summary(data$educ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   12.00   13.00   13.26   16.00   18.00
```

Similaly, for expr:

```r
exper_model <- lm(formula = exper ~ age + age_sq + smsa + south + nearc + daded + momed, data = data)
summary(exper_model)
```

```
## 
## Call:
## lm(formula = exper ~ age + age_sq + smsa + south + nearc + daded + 
##     momed, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.2250 -1.6957  0.2224  1.5450 11.2777 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.347646   3.976343  -0.087 0.930337    
## age          0.010390   0.278714   0.037 0.970266    
## age_sq       0.017019   0.004838   3.518 0.000441 ***
## smsa1       -0.529566   0.101504  -5.217 1.94e-07 ***
## south1       0.424851   0.091037   4.667 3.19e-06 ***
## nearc1      -0.264554   0.099085  -2.670 0.007626 ** 
## daded       -0.190443   0.015611 -12.199  < 2e-16 ***
## momed       -0.234515   0.017028 -13.773  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 2.326 on 3002 degrees of freedom
## Multiple R-squared:  0.6853, Adjusted R-squared:  0.6845
## F-statistic: 933.7 on 7 and 3002 DF,  p-value: < 2.2e-16
```

```r
data$exper_f <- exper_model$fitted.values
summary(exper_model$fitted.values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.730   6.122   8.283   8.856  11.390  18.505
```

```r
exper_sq_model <- lm(formula = exper_sq ~ age + age_sq + smsa + south + nearc + daded + momed, data = da
summary(exper_sq_model)
```

```
##
## Call:
## lm(formula = exper_sq ~ age + age_sq + smsa + south + nearc +
##     daded + momed, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -164.28  -27.39   -0.20   23.05  380.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 681.3828    84.8457   8.031 1.38e-15 ***
## age         -54.0654     5.9471  -9.091  < 2e-16 ***
## age_sq        1.2799     0.1032  12.399  < 2e-16 ***
## smsa1       -11.8031     2.1659  -5.450 5.46e-08 ***
## south1       10.6147     1.9425   5.464 5.02e-08 ***
## nearc1       -5.7804     2.1142  -2.734  0.00629 **
## daded        -3.3142     0.3331  -9.949  < 2e-16 ***
## momed        -4.7333     0.3633 -13.028  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.64 on 3002 degrees of freedom
## Multiple R-squared:  0.6567, Adjusted R-squared:  0.6559
## F-statistic: 820.4 on 7 and 3002 DF,  p-value: < 2.2e-16
```

```r
data$exper_sq_f <- exper_sq_model$fitted.values
summary(exper_sq_model$fitted.values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -32.77   41.24   78.90   95.58  141.21  300.67
```

```r
summary(data)
```

```
##       logw            educ            age            exper         smsa
##  Min.   :4.605   Min.   : 1.00   Min.   :24.00   Min.   : 0.000   0: 864
##  1st Qu.:5.977   1st Qu.:12.00   1st Qu.:25.00   1st Qu.: 6.000   1:2146
##  Median :6.287   Median :13.00   Median :28.00   Median : 8.000
##  Mean   :6.262   Mean   :13.26   Mean   :28.12   Mean   : 8.856
##  3rd Qu.:6.564   3rd Qu.:16.00   3rd Qu.:31.00   3rd Qu.:11.000
##  Max.   :7.785   Max.   :18.00   Max.   :34.00   Max.   :23.000
##  south     nearc        daded            momed            exper_sq
##  0:1795   0: 957   Min.   : 0.000   Min.   : 0.00   Min.   :  0.00
##  1:1215   1:2053   1st Qu.: 8.000   1st Qu.: 9.00   1st Qu.: 36.00
##                     Median : 9.940   Median :11.00   Median : 64.00  
##                     Mean   : 9.989   Mean   :10.34   Mean   : 95.58  
##                     3rd Qu.:12.000   3rd Qu.:12.00   3rd Qu.:121.00  
##                     Max.   :18.000   Max.   :18.00   Max.   :529.00  
##      age_sq           educ_f          exper_f          exper_sq_f    
##  Min.   : 576.0   Min.   : 7.87   Min.   : 1.730   Min.   :-32.77  
##  1st Qu.: 625.0   1st Qu.:12.53   1st Qu.: 6.122   1st Qu.: 41.24  
##  Median : 784.0   Median :13.39   Median : 8.283   Median : 78.90  
##  Mean   : 800.5   Mean   :13.26   Mean   : 8.856   Mean   : 95.58  
##  3rd Qu.: 961.0   3rd Qu.:14.14   3rd Qu.:11.390   3rd Qu.:141.21  
##  Max.   :1156.0   Max.   :17.10   Max.   :18.505   Max.   :300.67  
```

(e) 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).

```
fit <- ivreg(formula = logw ~ educ + exper + exper_sq + smsa + south |  age + age_sq + smsa + south + ne

summary(fit)
```

```
## 
## Call:
## ivreg(formula = logw ~ educ + exper + exper_sq + smsa + south | 
##     age + age_sq + smsa + south + nearc + daded + momed, data = data)
## 
## 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 ***
## exper_sq    -0.0016393  0.0008381  -1.956   0.0506 .  
## smsa1        0.1349370  0.0167695   8.047 1.21e-15 ***
## south1      -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
```

This can also be obtained by using the fitted variables we obtained in the last step:

```
sls2_model <- lm(formula = logw ~ educ_f + exper_f + exper_sq_f + smsa + south, data = data)
summary(sls2_model)
```

```
## 
## Call:
## lm(formula = logw ~ educ_f + exper_f + exper_sq_f + smsa + south, 
##     data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.67797 -0.23820  0.01715  0.26700  1.46756
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.4169039  0.1178606  37.476  < 2e-16 ***
## educ_f       0.0998429  0.0067128  14.874  < 2e-16 ***
## exper_f      0.0728669  0.0170667   4.270 2.02e-05 ***
## exper_sq_f  -0.0016393  0.0008559  -1.915   0.0555 .
## smsa1        0.1349370  0.0171240   7.880 4.54e-15 ***
## south1      -0.1589869  0.0160170  -9.926  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3925 on 3004 degrees of freedom
## Multiple R-squared:  0.2192, Adjusted R-squared:  0.2179
## F-statistic: 168.6 on 5 and 3004 DF,  p-value: < 2.2e-16
```

```r
coefs2SLS <- matrix(summary(sls2_model)$coefficients[,1])
```

We can see that both models look a bit similar, and that both education and experience still have a positive effect while the squared experience still has a negative effect to logw.

The 2SLS education estimated effect size of about 10% is a bit larger than the OLS estimation of about 8.2%, while the 2SLS experience estimated effect size of about 7.3% is a bit smaller than the OLS estimation of about 8.4%. And both 2SLS and OLS estimated a (small) negative 0.2% effect size for the squared experience variable.

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

```r
data$final_residual <- residuals(sls2_model)

fres <-lm(formula = final_residual ~ smsa + south + age + age_sq + nearc + daded + momed, data = data)
summary(fres)
```

```
##
## Call:
## lm(formula = final_residual ~ smsa + south + age + age_sq + nearc +
##     daded + momed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68021 -0.23801  0.01513  0.26883  1.46398
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1258215  0.6707128   0.188    0.851
## smsa1       -0.0033465  0.0171213  -0.195    0.845
## south1       0.0022260  0.0153557   0.145    0.885
## age         -0.0093315  0.0470122  -0.198    0.843
## age_sq       0.0001591  0.0008160   0.195    0.845
## nearc1       0.0135079  0.0167132   0.808    0.419
## daded       -0.0041052  0.0026333  -1.559    0.119
## momed        0.0041134  0.0028721   1.432    0.152
##
## Residual standard error: 0.3924 on 3002 degrees of freedom
## Multiple R-squared:  0.00118,    Adjusted R-squared:  -0.001149
## F-statistic: 0.5065 on 7 and 3002 DF,  p-value: 0.8303
```

R2: very low, with only 0.1% of logwage residuals explained.

```
sargan.tstat = nrow(data) * summary(fres)$r.squared
sargan.tstat
```

```
## [1] 3.55069
```

Critical value:

```
qchisq(0.95, df = 8-6, lower.tail = TRUE) #8 variables and 6 instruments
```

```
## [1] 5.991465
```

3.55 is smaller than 5.99 we do not reject the null hypothesis. So the instruments seem to be valid. so the instruments are not related with errors of the linear model called on logwage and are not omitted variables in the model, so they qualify correctly as instruments.

#Hausman Test: with p-value < 0.01 rejects the null hypothesis, so educ, exper and exper2 are endogenous, as expected, #that is they are related to ϵ, the model's errors.