# Test4_YM

May 13, 2018

```
In [2]: import pandas as pd
        import numpy as np
        import statsmodels.api as sm
        from patsy import dmatrices, dmatrix
```

@author: Yiming Cai

### 0.0.1 Question (a)

Use OLS to estimate the parameters of the model

$$logw = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 smsa + \beta_6 south + \epsilon$$

Give an interpretation to the estimated 2 coefficient.

```
In [2]: df = pd.read_excel("Test4_data.xls")
        df.head()
```

Out[2]:

| | logw | educ | age | exper | smsa | south | nearc | daded | momed |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 6.306275 | 7 | 29 | 16 | 1 | 0 | 0 | 9.94 | 10.25 |
| 1 | 6.175867 | 12 | 27 | 9 | 1 | 0 | 0 | 8.00 | 8.00 |
| 2 | 6.580639 | 12 | 34 | 16 | 1 | 0 | 0 | 14.00 | 12.00 |
| 3 | 5.521461 | 11 | 27 | 10 | 1 | 0 | 1 | 11.00 | 12.00 |
| 4 | 6.591674 | 12 | 34 | 16 | 1 | 0 | 1 | 8.00 | 7.00 |

```
In [3]: y, X = dmatrices("logw ~ educ + exper + np.square(exper) + smsa + south", df)
```

The OLS estimation result is given as follows:

```
In [4]: mod = sm.OLS(y, X).fit()
        print (mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   logw   R-squared:                       0.263
Model:                            OLS   Adj. R-squared:                  0.262
Method:                 Least Squares   F-statistic:                     214.6
Date:                Sun, 13 May 2018   Prob (F-statistic):         3.70e-196
Time:                        21:48:56   Log-Likelihood:                -1365.6
```

```
No. Observations:                3010   AIC:                             2743.
Df Residuals:                    3004   BIC:                             2779.
Df Model:                           5
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           4.6110      0.068     67.914      0.000       4.478       4.744
educ                0.0816      0.003     23.315      0.000       0.075       0.088
exper               0.0838      0.007     12.377      0.000       0.071       0.097
np.square(exper)   -0.0022      0.000     -6.800      0.000      -0.003      -0.002
smsa                0.1508      0.016      9.523      0.000       0.120       0.182
south              -0.1752      0.015    -11.959      0.000      -0.204      -0.146
==============================================================================
Omnibus:                       52.759   Durbin-Watson:                   1.853
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.537
Skew:                          -0.261   Prob(JB):                     2.63e-14
Kurtosis:                       3.476   Cond. No.                     1.26e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

### 0.0.2 Answer (a)

Other things being equal, taking one year education, the expected log of wage (logw) would increase by 0.086. Or put it another way, because

$$log(\frac{w2}{w1}) = 0.0816 \Rightarrow \frac{w2}{w1} = e^{0.0816} = 1.085 \Rightarrow w2 = (1 + 8.5\%)w1 \Rightarrow$$

one additional year of education is associated with 8.5% increase on expected wage level.

### 0.0.3 Question (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.

### 0.0.4 Answer (b)

educ and exper might be endogenous due to **ommited variables**. For example, individual's characteristics are likely to influence individual's educ and exper. Individual with higher intellectual ability and motivation would likely to obtain higher education, i.e., more number of years of schooling (educ). Also, people with hard-working ethics tend to have more working experience. All these characteristics are likely to positively influence wage level but not included in the model. Therefore, **educ** and **exper** are endogenouse, resulting estimate in part(a) being inconsistent.

### 0.0.5 Question (c)

Give a motivation why **age** and **age2** can be used as instruments for exper and exper2.

### 0.0.6 Answer(c)

Older people tend to have longer working experience, nevertheless, the wage is unlikely to influenced by age itself. Therefore, **age** and **age^{2}** is likely to be correlated with **exper** and **exper^{2}** but uncorrelated with error term ($\epsilon$), which suffice them to be instruments for exper and exper2.

### 0.0.7 Question (d)

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?

```
In [5]: y2, X2 = dmatrices("educ ~ age + np.square(age) + nearc + daded + momed", df)
```

```
In [6]: first_stage_mod = sm.OLS(y2, X2).fit()
        print (first_stage_mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   educ   R-squared:                       0.233
Model:                            OLS   Adj. R-squared:                  0.232
Method:                 Least Squares   F-statistic:                     182.5
Date:                Sun, 13 May 2018   Prob (F-statistic):          4.51e-170
Time:                        21:48:56   Log-Likelihood:                -6835.1
No. Observations:                3010   AIC:                         1.368e+04
Df Residuals:                    3004   BIC:                         1.372e+04
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept         -5.9233      4.011     -1.477      0.140     -13.787       1.940
age                0.9926      0.281      3.531      0.000       0.441       1.544
np.square(age)    -0.0171      0.005     -3.500      0.000      -0.027      -0.008
nearc              0.5288      0.093      5.704      0.000       0.347       0.711
daded              0.2020      0.016     12.898      0.000       0.171       0.233
momed              0.2484      0.017     14.580      0.000       0.215       0.282
==============================================================================
Omnibus:                       21.480   Durbin-Watson:                   1.778
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.916
Skew:                          -0.070   Prob(JB):                     3.19e-07
```

```
Kurtosis:                       3.468   Cond. No.                     7.72e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.72e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The above result suggests:
There are enough instruments.
The p-values suggest instruments are correlated with educ.
Therefore, these instruments are suitable for schooling. However, the validity of these instruments require following Sargon test.

#### 0.0.8 Question (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).

> As suggested by Quesiton (b) and Question (c). $age, age2, nearc, dadeduc, andmomeduc$ can be used as instruments for $educ$, and $age$ and $age^2$ would be instruments for $expr$ and $expr^2$ respectively.

```
In [7]: y3, X3 = dmatrices("exper ~  age ", df )
        expr_stage1_mod = sm.OLS(y3, X3).fit()
        print (expr_stage1_mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  exper   R-squared:                       0.582
Model:                            OLS   Adj. R-squared:                  0.582
Method:                 Least Squares   F-statistic:                     4193.
Date:                Sun, 13 May 2018   Prob (F-statistic):               0.00
Time:                        21:48:57   Log-Likelihood:                -7234.2
No. Observations:                3010   AIC:                         1.447e+04
Df Residuals:                    3008   BIC:                         1.448e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -19.4732      0.440    -44.236      0.000     -20.336     -18.610
age            1.0075      0.016     64.754      0.000       0.977       1.038
==============================================================================
Omnibus:                       33.319   Durbin-Watson:                   1.593
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               35.593
Skew:                           0.227   Prob(JB):                     1.87e-08
Kurtosis:                       3.279   Cond. No.                         256.
```

```
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
In [8]: y4, X4 = dmatrices("np.square(exper) ~ np.square(age)", df )
        expr2_stage1_mod = sm.OLS(y4, X4).fit()
        print (expr2_stage1_mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       np.square(exper)   R-squared:                       0.553
Model:                            OLS   Adj. R-squared:                  0.553
Method:                 Least Squares   F-statistic:                     3723.
Date:                Sun, 13 May 2018   Prob (F-statistic):               0.00
Time:                        21:48:57   Log-Likelihood:                -16417.
No. Observations:                3010   AIC:                         3.284e+04
Df Residuals:                    3008   BIC:                         3.285e+04
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept       -183.1527      4.683    -39.110      0.000    -192.335    -173.970
np.square(age)     0.3482      0.006     61.017      0.000       0.337       0.359
==============================================================================
Omnibus:                      741.263   Durbin-Watson:                   1.629
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2955.288
Skew:                           1.158   Prob(JB):                         0.00
Kurtosis:                       7.267   Cond. No.                     3.73e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.73e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Calculated the predicted values for $educ, exper, exper^2$

```
In [9]: educ_explained = first_stage_mod.predict(X2)
```

```
In [10]: exper_explained = expr_stage1_mod.predict(X3)
```

```
In [11]: exper2_explained = expr2_stage1_mod.predict(X4)
```

Next, use the predicted values as variables the second stage OLS

```
In [12]: df_2sls = df[["logw","smsa", "south"] ].copy()
         df_2sls["educ_explained"] = educ_explained
         df_2sls["exper_explained"] = exper_explained
         df_2sls["exper2_explained"] = exper2_explained

In [13]: df_2sls.head()

Out[13]:        logw  smsa  south  educ_explained  exper_explained  exper2_explained
         0  6.306275     1      0       13.054551         9.743111        109.662945
         1  6.175867     1      0       12.031066         7.728195         70.667282
         2  6.580639     1      0       13.893543        14.780402        219.338246
         3  5.521461     1      0       14.159474         7.728195         70.667282
         4  6.591674     1      0       11.968116        14.780402        219.338246

In [14]: y, X_stage2 = dmatrices("logw ~ educ_explained + exper_explained+ exper2_explained+ sm

In [16]: stage2_mod = sm.OLS(y, X_stage2).fit()
         print (stage2_mod.summary())

                            OLS Regression Results
==============================================================================
Dep. Variable:                   logw   R-squared:                       0.219
Model:                            OLS   Adj. R-squared:                  0.218
Method:                 Least Squares   F-statistic:                     168.6
Date:                Sun, 13 May 2018   Prob (F-statistic):          2.00e-158
Time:                        21:49:08   Log-Likelihood:                -1452.9
No. Observations:                3010   AIC:                             2918.
Df Residuals:                    3004   BIC:                             2954.
Df Model:                           5
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            4.8025      0.197     24.322      0.000       4.415       5.190
educ_explained       0.0543      0.006      9.243      0.000       0.043       0.066
exper_explained      0.1246      0.047      2.648      0.008       0.032       0.217
exper2_explained    -0.0042      0.002     -1.797      0.072      -0.009       0.000
smsa                 0.1646      0.016     10.064      0.000       0.133       0.197
south               -0.1862      0.015    -12.211      0.000      -0.216      -0.156
==============================================================================
Omnibus:                       58.465   Durbin-Watson:                   1.836
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               70.093
Skew:                          -0.276   Prob(JB):                     6.02e-16
Kurtosis:                       3.504   Cond. No.                     3.26e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.26e+03. This might indicate that there are
```

```
strong multicollinearity or other numerical problems.
```

below is the estimate from (a), in comparison, the impact of education on wage becomes less but the effect of experience grows. The non-linear term exper^2 remains negative but almost doubles the effect.

```
In [17]: print (mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   logw   R-squared:                       0.263
Model:                            OLS   Adj. R-squared:                  0.262
Method:                 Least Squares   F-statistic:                     214.6
Date:                Sun, 13 May 2018   Prob (F-statistic):          3.70e-196
Time:                        21:49:09   Log-Likelihood:                -1365.6
No. Observations:                3010   AIC:                             2743.
Df Residuals:                    3004   BIC:                             2779.
Df Model:                           5
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            4.6110      0.068     67.914      0.000       4.478       4.744
educ                 0.0816      0.003     23.315      0.000       0.075       0.088
exper                0.0838      0.007     12.377      0.000       0.071       0.097
np.square(exper)    -0.0022      0.000     -6.800      0.000      -0.003      -0.002
smsa                 0.1508      0.016      9.523      0.000       0.120       0.182
south               -0.1752      0.015    -11.959      0.000      -0.204      -0.146
==============================================================================
Omnibus:                       52.759   Durbin-Watson:                   1.853
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.537
Skew:                          -0.261   Prob(JB):                     2.63e-14
Kurtosis:                       3.476   Cond. No.                     1.26e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

### 0.0.9 Question (f)

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

1 . Calculate the residuals using formula \$e_{2SLS} = y - Xb_{2SLS} \$

```
In [18]: e_2sls = df.logw.values  - stage2_mod.predict(X)
```

2 . Regress $e_{2SLS}$ on $Z$, where $Z$ is (constant, age, age2, nearc, dadeduc, momeduc, smsa, south)

```
In [19]: z = dmatrix("age + np.square(age) + nearc + daded + momed + smsa + south",
                     data= df ,return_type= "dataframe")
```

```
In [20]: z_mod = sm.OLS(e_2sls, z).fit()
         print (z_mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.022
Model:                            OLS   Adj. R-squared:                  0.020
Method:                 Least Squares   F-statistic:                     9.757
Date:                Sun, 13 May 2018   Prob (F-statistic):           4.53e-12
Time:                        21:49:10   Log-Likelihood:                -1404.7
No. Observations:                3010   AIC:                             2825.
Df Residuals:                    3002   BIC:                             2873.
Df Model:                           7
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          1.2780      0.660      1.935      0.053      -0.017       2.573
age               -0.1058      0.046     -2.285      0.022      -0.197      -0.015
np.square(age)     0.0018      0.001      2.292      0.022       0.000       0.003
nearc              0.0242      0.016      1.469      0.142      -0.008       0.056
daded              0.0058      0.003      2.230      0.026       0.001       0.011
momed              0.0146      0.003      5.158      0.000       0.009       0.020
smsa              -0.0122      0.017     -0.726      0.468      -0.045       0.021
south              0.0158      0.015      1.044      0.296      -0.014       0.045
==============================================================================
Omnibus:                       53.838   Durbin-Watson:                   1.843
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               63.790
Skew:                          -0.265   Prob(JB):                     1.41e-14
Kurtosis:                       3.478   Cond. No.                     7.72e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.72e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

3 . calculate $nR^2$, and $nR^2 \sim \chi^2(m-k)$, where m = 8, k = 6

```
In [21]: n = 3010
         R_2 = 0.022
         n * R_2
```

Out[21]: 66.22

since the critical value for $\chi^2(2)$ at 5% confidence level is 5.99 and 66.22 > 5.99, therefore, we reject the null hypothesis and correlation Z and /epsilon is 0.

Conclusion: the instrument variables are actually not valid, further refinements are required.