Test4_YM

May 14, 2018

```
In [40]: 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:                Mon, 14 May 2018   Prob (F-statistic):          3.70e-196
Time:                        20:28:05   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 (ϵ), 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 [21]: df.head()
```

Out[21]:

	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 [22]: y2, X2 = dmatrices("educ ~ age + np.square(age) + nearc + daded + momed + smsa + south
```

```
In [23]: first_stage_mod = sm.OLS(y2, X2).fit()
         print (first_stage_mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   educ   R-squared:                       0.247
Model:                            OLS   Adj. R-squared:                  0.245
Method:                 Least Squares   F-statistic:                     140.4
Date:                Mon, 14 May 2018   Prob (F-statistic):          2.14e-179
Time:                        20:28:44   Log-Likelihood:                -6808.2
No. Observations:                3010   AIC:                         1.363e+04
Df Residuals:                    3002   BIC:                         1.368e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -5.6524      3.976     -1.421      0.155     -13.449       2.144
```

```
age                0.9896      0.279      3.551      0.000       0.443       1.536
np.square(age)    -0.0170      0.005     -3.518      0.000      -0.027      -0.008
nearc              0.2646      0.099      2.670      0.008       0.070       0.459
daded              0.1904      0.016     12.199      0.000       0.160       0.221
momed              0.2345      0.017     13.773      0.000       0.201       0.268
smsa               0.5296      0.102      5.217      0.000       0.331       0.729
south             -0.4249      0.091     -4.667      0.000      -0.603      -0.246
==============================================================================
Omnibus:                       13.809   Durbin-Watson:                   1.796
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               17.748
Skew:                          -0.053   Prob(JB):                     0.000140
Kurtosis:                       3.361   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*, *age*2, *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 [24]: y3, X3 = dmatrices("exper ~ age + np.square(age) + nearc + daded + momed + smsa + sou
         expr_stage1_mod = sm.OLS(y3, X3).fit()
         print (expr_stage1_mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  exper   R-squared:                       0.685
Model:                            OLS   Adj. R-squared:                  0.685
Method:                 Least Squares   F-statistic:                     933.7
Date:                Mon, 14 May 2018   Prob (F-statistic):               0.00
Time:                        20:29:12   Log-Likelihood:                -6808.2
No. Observations:                3010   AIC:                         1.363e+04
Df Residuals:                    3002   BIC:                         1.368e+04
Df Model:                           7
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -0.3476      3.976     -0.087      0.930      -8.144       7.449
age             0.0104      0.279      0.037      0.970      -0.536       0.557
np.square(age)  0.0170      0.005      3.518      0.000       0.008       0.027
nearc          -0.2646      0.099     -2.670      0.008      -0.459      -0.070
daded          -0.1904      0.016    -12.199      0.000      -0.221      -0.160
momed          -0.2345      0.017    -13.773      0.000      -0.268      -0.201
smsa           -0.5296      0.102     -5.217      0.000      -0.729      -0.331
south           0.4249      0.091      4.667      0.000       0.246       0.603
==============================================================================
Omnibus:                       13.809   Durbin-Watson:                   1.796
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               17.748
Skew:                           0.053   Prob(JB):                     0.000140
Kurtosis:                       3.361   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.
```

```
In [25]: y4, X4 = dmatrices("np.square(exper) ~ age + np.square(age) + nearc + daded + momed +
         expr2_stage1_mod = sm.OLS(y4, X4).fit()
         print (expr2_stage1_mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       np.square(exper)   R-squared:                       0.657
Model:                            OLS   Adj. R-squared:                  0.656
Method:                 Least Squares   F-statistic:                     820.4
Date:                Mon, 14 May 2018   Prob (F-statistic):               0.00
Time:                        20:29:23   Log-Likelihood:                -16020.
No. Observations:                3010   AIC:                         3.206e+04
Df Residuals:                    3002   BIC:                         3.210e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    681.3828     84.846      8.031      0.000     515.021     847.744
age          -54.0654      5.947     -9.091      0.000     -65.726     -42.405
np.square(age)   1.2799      0.103     12.399      0.000       1.077       1.482
nearc         -5.7804      2.114     -2.734      0.006      -9.926      -1.635
daded         -3.3142      0.333     -9.949      0.000      -3.967      -2.661
```

```
momed          -4.7333      0.363    -13.028      0.000      -5.446      -4.021
smsa          -11.8031      2.166     -5.450      0.000     -16.050      -7.556
south          10.6147      1.943      5.464      0.000       6.806      14.423
==============================================================================
Omnibus:                      658.664   Durbin-Watson:                   1.823
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3018.668
Skew:                           0.981   Prob(JB):                         0.00
Kurtosis:                       7.496   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.
```

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

```
In [26]: educ_explained = first_stage_mod.predict(X2)

In [27]: exper_explained = expr_stage1_mod.predict(X3)

In [28]: exper2_explained = expr2_stage1_mod.predict(X4)
```

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

```
In [29]: 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 [30]: df_2sls.head()
```

Out[30]:

	logw	smsa	south	educ_explained	exper_explained	exper2_explained
0	6.306275	1	0	13.559710	9.440290	96.593284
1	6.175867	1	0	12.589499	8.410501	78.458225
2	6.580639	1	0	14.330376	13.669624	207.685943
3	5.521461	1	0	14.363443	6.636557	43.802227
4	6.591674	1	0	12.279697	15.720303	245.456999

```
In [31]: y, X_stage2 = dmatrices("logw ~ educ_explained + exper_explained+ exper2_explained+ sm

In [32]: 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:                Mon, 14 May 2018   Prob (F-statistic):          1.84e-158
Time:                        20:29:46   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            4.4169      0.118     37.476      0.000       4.186       4.648
educ_explained       0.0998      0.007     14.874      0.000       0.087       0.113
exper_explained      0.0729      0.017      4.270      0.000       0.039       0.106
exper2_explained    -0.0016      0.001     -1.915      0.056      -0.003    3.88e-05
smsa                 0.1349      0.017      7.880      0.000       0.101       0.169
south               -0.1590      0.016     -9.926      0.000      -0.190      -0.128
==============================================================================
Omnibus:                       58.101   Durbin-Watson:                   1.836
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               69.727
Skew:                          -0.274   Prob(JB):                     7.23e-16
Kurtosis:                       3.505   Cond. No.                     1.96e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.96e+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 [34]: 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:                Mon, 14 May 2018   Prob (F-statistic):          3.70e-196
Time:                        20:42:57   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 [35]: 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 [36]: z = dmatrix("age + np.square(age) + nearc + daded + momed + smsa + south",
                 data= df ,return_type= "dataframe")

In [37]: z_mod = sm.OLS(e_2sls, z).fit()
         print (z_mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.5282
Date:                Mon, 14 May 2018   Prob (F-statistic):              0.814
Time:                        20:43:06   Log-Likelihood:                -1388.1
No. Observations:                3010   AIC:                             2792.
Df Residuals:                    3002   BIC:                             2840.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
Intercept          0.1258      0.657      0.192      0.848      -1.162       1.414
age               -0.0093      0.046     -0.203      0.839      -0.100       0.081
np.square(age)     0.0002      0.001      0.199      0.842      -0.001       0.002
nearc              0.0135      0.016      0.825      0.409      -0.019       0.046
daded             -0.0041      0.003     -1.592      0.111      -0.009       0.001
momed              0.0041      0.003      1.462      0.144      -0.001       0.010
smsa              -0.0033      0.017     -0.200      0.842      -0.036       0.030
south              0.0022      0.015      0.148      0.882      -0.027       0.032
==============================================================================
Omnibus:                       54.658   Durbin-Watson:                   1.864
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               65.671
Skew:                          -0.263   Prob(JB):                     5.49e-15
Kurtosis:                       3.498   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 [39]: n = 3010
         R_2 = z_mod.rsquared
         n * R_2

Out[39]: 3.7023886431634678
```

since the critical value for $\chi^2(2)$ at 5% confidence level is 5.99 and $3.7 < 5.99$, 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.