Test Exercise 4

April 24, 2018

```
In [1]: import numpy as np
        import pandas as pd
        import statsmodels.api as sm
        import matplotlib.pyplot as plt
        import seaborn as sns
        sns.set_style('whitegrid')
        %matplotlib inline
```

```
/anaconda/envs/research/lib/python3.5/site-packages/statsmodels/compat/pandas.py:56: FutureWar
  from pandas.core import datetools
```

```
In [2]: df = pd.read_excel('TestExer4_Wage-round1.xls')
```

```
In [3]: df.head()
```

Out[3]:

	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

0.1 Part (a)

The coefficient for `educ` in the OLS estimate is 0.0816. This means that when education increases by 1 year `logw` increases by 0.082.

```
In [4]: df['exper2'] = df['exper']**2
```

```
In [5]: X = df[['educ', 'exper', 'exper2', 'smsa', 'south']]
        X = sm.add_constant(X)
        y = df['logw']

        model = sm.OLS(y,X)
        result = model.fit()

        print(result.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:                Tue, 24 Apr 2018   Prob (F-statistic):          3.70e-196
Time:                        03:11:48   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]
------------------------------------------------------------------------------
const          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
exper2        -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.
```

0.2 Part (b)

`Intelligence`, `Efficiency` of a person may be factors that could make `educ` and `exper` endogenous.

In this case OLS is very useful as it is inconsistent so the estimate in Part(a) is ot useful.

0.3 Part (c)

`age` is related to `exper` as older people usually have a lot of experience. So, `age` and `age2` would be highly correlated with `exper` and `exper2`.

0.4 Part (d)

0.4.1 First Stage Regression

All the instruments have high correlation with educ as evidenced by their p-values. As the endogenous variable and the instrument variables have high correlation, they are suitable instruments for schooling.

```
In [6]: df['age2'] = df['age']**2
```

```
In [7]: y1 = df['educ']
        X1 = df[['smsa', 'south', 'age', 'age2', 'nearc', 'daded', 'momed']]
        X1 = sm.add_constant(X1)

        model1 = sm.OLS(y1,X1)
        res1 = model1.fit()

        print(res1.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:                Tue, 24 Apr 2018   Prob (F-statistic):          2.14e-179
Time:                        03:11:49   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]
------------------------------------------------------------------------------
const         -5.6524      3.976     -1.421      0.155     -13.449       2.144
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
age            0.9896      0.279      3.551      0.000       0.443       1.536
age2          -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
==============================================================================
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 [8]: y2 = df['exper']
        X2 = df[['smsa', 'south', 'age', 'age2', 'nearc', 'daded', 'momed']]
        X2 = sm.add_constant(X2)

        model2 = sm.OLS(y2,X2)
        res2 = model2.fit()

        print(res2.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:                Tue, 24 Apr 2018   Prob (F-statistic):               0.00
Time:                        03:11:49   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]
------------------------------------------------------------------------------
const         -0.3476      3.976     -0.087      0.930      -8.144       7.449
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
age            0.0104      0.279      0.037      0.970      -0.536       0.557
age2           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
==============================================================================
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 [9]: y3 = df['exper2']
        X3 = df[['smsa', 'south', 'age', 'age2', 'nearc', 'daded', 'momed']]
        X3 = sm.add_constant(X3)

        model3 = sm.OLS(y3,X3)
        res3 = model3.fit()

        print(res3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  exper2   R-squared:                       0.657
Model:                            OLS   Adj. R-squared:                  0.656
Method:                 Least Squares   F-statistic:                     820.4
Date:                Tue, 24 Apr 2018   Prob (F-statistic):               0.00
Time:                        03:11:49   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]
------------------------------------------------------------------------------
const        681.3828     84.846      8.031      0.000     515.021     847.744
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
age          -54.0654      5.947     -9.091      0.000     -65.726     -42.405
age2           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
==============================================================================
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.
```

```
In [10]: df['pred_educ'] = res1.predict(X1)
         df['pred_exper'] = res2.predict(X2)
         df['pred_exper2'] = res3.predict(X3)

In [11]: y4 = df['logw']
```