Notes:
- See website for how to submit your answers and how feedback is organized
- This exercise uses the datafile TestExer4 Wage and requires a computer.
- The dataset TestExer4 Wage is available on the website.

Goals and skills being used:
- Obtain insight in consequences of endogeneity
- Practice with identifying causes of endogeneity
- Practice with identifying valid instruments
- Obtain hands-on experience with applying 2SLS and the Sargan test

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)

(a) Use OLS to estimate the parameters of the model

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

Give an interpretation to the estimated β2 coefficient.

In [1]:

```
%matplotlib inline
import sys
sys.path.append('/Users/CJ/Documents/bitbucket/xforex_v1/xforex_v3')
import pandas as pd
import matplotlib.pyplot as plt
from datetime import datetime
from xforex.BackTesting.econometrics_tools import Econometrics_Tool
import numpy as np

dat = pd.read_csv(
        '/Users/CJ/Documents/bitbucket/xforex_v1/xforex_v3/training/econometrics/we
dat.describe()
```

Out[1]:

ogw	educ	age	exper	smsa	south	nearc
010.000000	3010.000000	3010.000000	3010.000000	3010.000000	3010.000000	3010.0000
.261832	13.263455	28.119601	8.856146	0.712957	0.403654	0.682060
.443798	2.676913	3.137004	4.141672	0.452457	0.490711	0.465753
.605170	1.000000	24.000000	0.000000	0.000000	0.000000	0.000000
.976985	12.000000	25.000000	6.000000	0.000000	0.000000	0.000000
.286928	13.000000	28.000000	8.000000	1.000000	0.000000	1.000000
.563503	16.000000	31.000000	11.000000	1.000000	1.000000	1.000000
.784889	18.000000	34.000000	23.000000	1.000000	1.000000	1.000000

In [2]:

```
dat['exper2'] = dat['exper']**2
model_ols = Econometrics_Tool().linear_fit(dat[['educ', 'exper','exper2','smsa','so
                                             dat['logw'])
model_ols.summary()
```

Out[2]:

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:	Wed, 14 Sep 2016	**Prob (F-statistic):**	3.70e-196
Time:	14:46:05	**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|**	**[95.0% Conf. Int.]**
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

```
**ans(a):**
β2 means every 1 year schooling will increase 8.5% wage (exp(β2)-1) if other
conditions are the same
```

```
(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.
**ans(b):**
For example, job certification may increase both  educ and exper increase also
increase the wage.
```

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

ans(b):
higher age doesn't always indicate high wage but often indicate high working years.

(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 [19]:

```
dat['age2'] = dat['age']**2
model_stage1 = Econometrics_Tool()\
.linear_fit(dat[['age', 'age2','daded','momed','exper','exper2','smsa','south']], \
                                            dat['educ'])
model_stage1.summary()
```

Out[19]:

OLS Regression Results

Dep. Variable:	educ	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	4.635e+28
Date:	Wed, 14 Sep 2016	**Prob (F-statistic):**	0.00
Time:	15:19:40	**Log-Likelihood:**	83184.
No. Observations:	3010	**AIC:**	-1.664e+05
Df Residuals:	3001	**BIC:**	-1.663e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	-6.0000	4.39e-13	-1.37e+13	0.000	-6.000 -6.000
age	1.0000	3.13e-14	3.19e+13	0.000	1.000 1.000
age2	-2.81e-16	5.45e-16	-0.516	0.606	-1.35e-15 7.87e-16
daded	1.28e-15	1.66e-15	0.772	0.440	-1.97e-15 4.53e-15
momed	-3.886e-16	1.82e-15	-0.214	0.831	-3.96e-15 3.18e-15
exper	-1.0000	4.75e-15	-2.11e+14	0.000	-1.000 -1.000
exper2	-5.117e-17	2.22e-16	-0.230	0.818	-4.87e-16 3.85e-16
smsa	-2.04e-15	1e-14	-0.203	0.839	-2.17e-14 1.76e-14
south	1.804e-15	9.37e-15	0.192	0.847	-1.66e-14 2.02e-14

Omnibus:	579.232	**Durbin-Watson:**	0.159
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	257.054
Skew:	0.549	**Prob(JB):**	1.52e-56
Kurtosis:	2.081	**Cond. No.**	8.30e+04

```
(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).
```

In [14]:

```
dat['edu-fit'] = model_stage1.fittedvalues
model_stage2 = Econometrics_Tool().linear_fit(dat[['edu-fit', 'exper','exper2','sms
                                                  dat['logw'])
model_stage2.summary()
```

Out[14]:

OLS Regression Results

Dep. Variable:	logw	**R-squared:**	0.159
Model:	OLS	**Adj. R-squared:**	0.158
Method:	Least Squares	**F-statistic:**	113.8
Date:	Wed, 14 Sep 2016	**Prob (F-statistic):**	1.93e-110
Time:	15:12:39	**Log-Likelihood:**	-1564.1
No. Observations:	3010	**AIC:**	3140.
Df Residuals:	3004	**BIC:**	3176.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	5.0762	0.096	53.031	0.000	4.889 5.264
edu-fit	0.0649	0.006	10.256	0.000	0.052 0.077
exper	0.0542	0.007	7.663	0.000	0.040 0.068
exper2	-0.0021	0.000	-6.043	0.000	-0.003 -0.001
smsa	0.1748	0.017	10.356	0.000	0.142 0.208
south	-0.1979	0.016	-12.584	0.000	-0.229 -0.167

Omnibus:	44.161	**Durbin-Watson:**	1.814
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	53.476
Skew:	-0.224	**Prob(JB):**	2.44e-12
Kurtosis:	3.476	**Cond. No.**	1.66e+03

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

In [15]:

```
model_sargan = Econometrics_Tool().linear_fit(dat[['age', 'age2','daded','momed']],
                                              model_stage2.resid)
model_sargan.summary()
```

Out[15]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.046
Model:	OLS	**Adj. R-squared:**	0.045
Method:	Least Squares	**F-statistic:**	36.08
Date:	Wed, 14 Sep 2016	**Prob (F-statistic):**	1.72e-29
Time:	15:16:41	**Log-Likelihood:**	-1493.5
No. Observations:	3010	**AIC:**	2997.
Df Residuals:	3005	**BIC:**	3027.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
const	-0.4151	0.680	-0.611	0.541	-1.748 0.917
age	0.0015	0.048	0.032	0.974	-0.092 0.095
age2	0.0004	0.001	0.544	0.587	-0.001 0.002
daded	-0.0036	0.003	-1.348	0.178	-0.009 0.002
momed	0.0046	0.003	1.587	0.113	-0.001 0.010

Omnibus:	49.899	**Durbin-Watson:**	1.828
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	59.161
Skew:	-0.251	**Prob(JB):**	1.42e-13
Kurtosis:	3.469	**Cond. No.**	7.72e+04

In [20]:

```
import scipy.stats as stats
sargan_stat = model_sargan.nobs * model_sargan.rsquared
print sargan_stat

degree_freedom = model_stage1.df_model + 1 - (model_stage2.df_model +1)
crit = stats.chi2.ppf(q = 0.95, # Find the critical value for 95% confidence*
                      df = degree_freedom)   # *

print("Critical value")
print(crit)

p_value = 1 - stats.chi2.cdf(sargan_stat,  # Find the p-value
                             df=degree_freedom)
print("P value")
print(p_value)
```

```
137.931588214
Critical value
7.81472790325
P value
0.0
```

ans(e): significate at 5% level
reject H0. -> instruments are not valid