Quantitative methods in finance - multiple regressions and tests

Eric Vansteenberghe

September 28, 2017

Contents

Link to the source codes and the data sets used in this lecture

1 Introduction

This exercise is largely inspired by the book **Econometrie des fondements a la modelisation** from Stephen Bazen and Mareva Sabatier published in 2007.

2 Data

We focus on the a salary equation as in Schooling, Experience and Earnings of Mincer 1974 work. We find a sample with 534 observation here. We import that data set as usual:

```
import pandas as pd
import os
os.chdir('/Users/skimeur/Google␣Drive/empirical_finance')

df=pd.read_excel('R_data/wagesmicrodata.xls',sheetname='Data',header=0,index_col=0)
#Get rid of the first row
df=df.ix[1:]
```

We change some of the variables:

```
#Male is 1
df['SEX']=1-df['SEX']
#Wage in log
```

```
import numpy as np
for i in range(0,len(df)):
    df.iloc[i]['WAGE']=np.log(df.iloc[i]['WAGE'])
#Experience squared
df['EXPERIENCE2']=df['EXPERIENCE']**2
df=df.astype(float)
```

3 Descriptive statistics

From our data set, we are interesting in the descriptive statistics:

```
df.mean()
df.std()
df[df.SEX==1].mean()
df[df.SEX==0].mean()
df[df.SEX==1].std()
df[df.SEX==0].std()
```

We find as in the book:

Variable	**Mean**	**Standard Deviation**
Wage	2.06	0.53
Sex	0.54	0.5
Education	13.01	2.62
Experience	17.82	12.38
Wage (men)	2.17	0.53
Wage(women)	1.93	0.49

4 Wage determinants estimation

We want to test the following regression:

$$\log(\text{Wage}_i) = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{education}_i + \beta_3 \text{experience}_i + \beta_4 \text{experience}_i^2 + \epsilon_i \tag{1}$$

We apply an ordinary least square regression:

```
import statsmodels.formula.api as smf
model = smf.ols('WAGE ~ SEX+EDUCATION+EXPERIENCE+EXPERIENCE2',data=df).fit()
print(model.summary())
```

We obtain the following results:

	coef	std err
Intercept	0.3437	0.122
SEX	0.2570	0.039
EDUCATION	0.0913	0.008
EXPERIENCE	0.0361	0.005
EXPERIENCE2	-0.0005	0.000

And a R^2 close to 30%. That is to say that our model explains 30% of the variance of the wages.

4.0.1 Interpreting the coefficient R2 of a multiple regression

As seen in previous exercise, the R^2 can be interpreted as the proportion of the total variation of Y explained by our multiple regression.

$$R^2 = \frac{\sum \hat{Y}_i^2}{\sum Y_i^2}$$

Adding explanatory variables would increase the numerator of this R^2, we compute an adjusted R^2:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

with n the number of observations and k the number of explanatory variables.

4.1 Fisher test

The Fisher test is testing H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against H_1: there is at least one coefficient different from 0.

The Fisher statistics is:

$$F_{k-1,n-k} = \frac{R^2}{1-R^2}\frac{n-k}{k-1}$$

We obtain the following statistics: F-statistic: 75
We can compute it "manually" and find the same result:

R2=0.297
#Number of estimated parameters (or rather dimension)
k=5
#Number of observations
n=**len**(df)
F=(R2/(k−1))/((1−R2)/(n−k))

If we look in a Fisher table, for 4 degree of liberty with $\alpha = 5\%$, we have 4 degree of liberty, we find a threshold around 2.4. Our calcultaed F-statistic is higher than this threshold hence we reject H_0 and our model has some explanatory power.

4.2 Test of significance of parameter estimates

As the variance of $\hat{\beta}_i$ are unknown, with use a t-distribution with $n-k$ degrees of freedom to test. The value we test is:

$$t = \frac{\hat{\beta}_i - 0}{\sigma_{\hat{\beta}_i}}$$

For our regression, we find that all t-student statistics is higher than 1.96 which is the threshold for an infinite degree of liberty with $\alpha = 5\%$.
We can thus say that the sex, education and experience have significative effect on the observed wages. Wan can for example express the impact of the years of experience of a worker on his salary as, its marginal effect:

$$\frac{\delta \log \text{WAGE}}{\delta \text{EXPERIENCE}} = \beta_3$$

model.tvalues
#If we only want the t−value for the variable EDUCATION
model.tvalues[2]

4.3 How much do men earn more than women according to our model

According to our model, if a woman earn a wage w, a man will earn (ceteris paribus): $\exp\left(\log(w)+\beta_1\right) = w\exp(0.257) = 1.29 \times w$
According to our model, a man earns around 30% more than a woman, ceteris paribus.

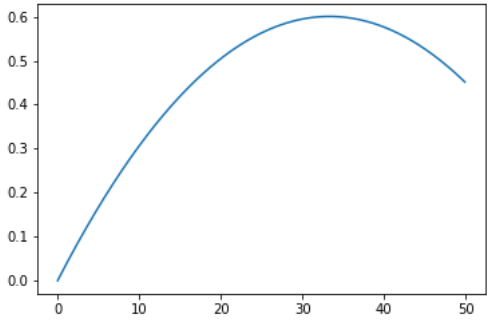
4.4 Non-linear effect of experience

Our t-student test statistics confirm that β_3 and β_4 are significant. β_4 being negative means that the higher the level of a worker's experience, additional experience compare to his peers will have less impact.
We can plot:

$$\log\text{Wage difference} = \beta_3\text{experience} + \beta_4\text{experience}^2 \tag{2}$$

```
beta3=model.params[3]
beta4=model.params[4]
import matplotlib.pyplot as plt
x = np.arange(0,50,0.1)
plt.plot(x,beta3*x+beta4*x**2,'-')
plt.xlabel=('wage difference')
plt.ylabel=('years of experience')
plt.savefig('experience.png')
plt.show()
#We can also plot how the impact of experience decrease with the years of experience
plt.plot(x,beta3+2*beta4*x,'-')
```

So we plot the wage difference according to the number of years of experience:



5 Chow test - importance of belonging to a union

We want to determine if our coefficients $beta_j$ are the same if workers are member of a union or not.
We do two additional regression on the union members and non-union members.

```
#Regression on non−union workers
model0 = smf.ols('WAGE ~ SEX+EDUCATION+EXPERIENCE+EXPERIENCE2',data=df[df.UNION==0]).fit()
print(model0.summary())
#Regression on union workers
model1 = smf.ols('WAGE ~ SEX+EDUCATION+EXPERIENCE+EXPERIENCE2',data=df[df.UNION==1]).fit()
print(model1.summary())
```

We can compute manually the Chow test statistics:

$$C = \frac{\frac{SCR-(SCR_0+SCR_1)}{k}}{\frac{SCR_0+SCR_1}{n-2k}}$$

Where SCR is the sum of residuals of a given regression.

```
numerator=(((model.resid**2).sum()−((model0.resid**2).sum()+(model1.resid**2).sum()))/k)
denominator=(((model0.resid**2).sum()+(model1.resid**2).sum())/(n−2*k))
```

C=numerator/denominator

The null hypothesis H_0 is that β are the same in both samples. We have to look into a Fisher table to accept or reject this hypothesis $F_{k,n-2k}$.

We find a Chow statistics of 5.94 which is above the threshohld around 2.4 at a risk of 5%. Therefore we can reject the hypothesis of homogeneity of coefficients. Therefore the variables sex, education and experience do not have the same effect whether the worker is a union member of not.

6 Is the model well specified? - Ramsey test

We can execute a Ramsey test.

1. we test the model $y_i = \beta x_i + \epsilon$, we obtain $\hat{y}_i$,then compute $\hat{y}_i^2$
2. we test the model $y_i = \alpha x_i + \delta \hat{y}_i^2 + \mu_i$
3. we test the significativity of the coefficient δ with a t-student test, if we find that $\hat{\delta} = 0$ then our model is well specified

```
beta0=model.params[0]
beta1=model.params[1]
beta2=model.params[2]
beta3=model.params[3]
beta4=model.params[4]

df['hatWAGE']=(beta0+beta1*df['SEX']+beta2*df['EDUCATION']+beta3*df['EXPERIENCE']+beta4*df['EXPERIENCE2'])**2

modelramsey=smf.ols('WAGE ~ SEX+EDUCATION+EXPERIENCE+EXPERIENCE2+hatWAGE',data=df).fit()
print(modelramsey.summary())
```

We find a t-student statistic for $\hat{\delta}$ of 0.056 which is below the theoretical threshold of 1.95 hence we do not reject the hypothesis H_0: $\hat{\delta} = 0$ and our initial model is well specified.

7 Heteroskedasticity - White test

As it would be hard to know from which variable would be the cause of an heteroscedasticity, i.e. it would be hard to find that one of the explanatory variables, z_i would make that $var(\epsilon_i) = \sigma^2 f(z_i)$. Hence we resort to the White test:

1. we estimate by ols the model $y_i = \beta x_i + \epsilon_i$
2. we regress by ols $\hat{\epsilon}_i^2$ on the explanatory variables x_i, the squared explanatory variables x_i^2 and the products of explanatory variables $x_i x_j$
3. we compare the $N \times R^2$ against the threshold $\chi^2(p)$ with p the number of explanatory variables in the second regression

With the smf.ols() function, we can use : or *: ":" adds a new column to the design matrix with the product of the other two columns. âĂIJ*âĂİ will also include the individual columns that were multiplied together.

```
df['EXPERIENCE4']=df['EXPERIENCE2']**2
df['EDUCATION2']=df['EDUCATION']**2
modelwhite=smf.ols('WAGE ~ SEX+EDUCATION+EXPERIENCE+EXPERIENCE2+EDUCATION2+EXPERIENCE4+SEX:EDUCATION+SEX:EXPERIENCE+SEX:EXPERIENCE
print(modelwhite.summary())
R2w=0.321
W=n*R2w
```

We find a test statistic which is $W = 171.41$, which follow a Chi-2 with $\frac{k(k+1)}{2}$ degrees of freedom and which is above the statistical threshold at 5%, 23.35 hence we reject the null hypothesis H_0 of absence of heteroscedasticity.
We need to apply a correction of White in order to have non-biased estimators. Because so far, the coefficients found are acceptable but not the computed standard deviations. We would need to use the corrected formula to compute the standard deviations.
Has explained in the manual here, it is already implemented for you in python, you just need to use HC0_se instead of resid:

model.HC0_se

We get the White robust standard errors:

	coef	std err (White)
Intercept	0.3437	0.122
SEX	0.2570	0.039
EDUCATION	0.0913	0.008
EXPERIENCE	0.0361	0.005
EXPERIENCE2	-0.0005	0.000

So in fact, the standard errors python was giving us were already to robust ones.

8 Endogeneity - Hausman test

In our case, it is not easy to propose an instrument variable in our database. Would you propose one?

9 Should we add RACE in our regression model?

If we try to add RACE as an explanatory variable:

modelbis = smf.ols('WAGE ~ SEX+EDUCATION+EXPERIENCE+EXPERIENCE2+RACE',data=df).fit()
print(modelbis.summary())

We find a t-student coefficient below the threshold of 1.96 hence we do not reject the null hypothesis that the coefficient for RACE in our model is 0. Hence we could say that this variable doesn't bring much to our model.
Make a similar exercise with the marital status.

10 Klein test - multi-colinearity

The Klein test is simple to implement, we test for the explanatory variables whether: $R^2 < | \, corr(x_i, x_j) \, |$

df[['SEX','EDUCATION','EXPERIENCE']].corr()

Looking at the results, we see that the correlation coefficient between education and experience is higher than the R^2 of our model:

	SEX	EDUCATION	EXPERIENCE
SEX	1.000	-0.002	-0.075
EDUCATION	-0.002	1.000	-0.353
EXPERIENCE	-0.075	-0.353	1.000

There is a colinearity between education and experience, but its level 35% can be considered as non problematic.

11 Conclusion

The selected and estimated model is:

$$\log(\hat{\text{Wage}}_i) = 0.34 \quad + \quad 0.257 \quad \text{sex}_i + \quad 0.091 \quad \text{education}_i + \quad 0.036 \quad \text{experience}_i - \quad 0.0005 \quad \text{experience}_i^2$$

and explains 30% of the wage variations.
It faces an heteroscedasticity bias, but our coefficients are still significant.
It could face endogeneity bias, but we could not find an instrument variable to perform an Hausman test.
It faces some mutlicolinearity issues but we can accept this, considering the level of colinearity.