Quantitative methods in finance - avoiding spurious regression with Python and financial time series

Eric Vansteenberghe

September 30, 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 Importing financial data from Yahoo Finance

We start by loading some relevant packages for Python:

import pandas_datareader.data as web
import datetime
import pandas as pd
import numpy as np

Then we define the start and end date of our import as in Chapter 5 section 1 of the aforementioned book:

start = datetime.datetime(1995, 1, 16)
end = datetime.datetime(1999, 2, 26)

Then we download the 'Close' value for the CAC 40 and the Dow Jones indexes:

dow= web.DataReader("^DJI", 'yahoo', start, end)['Close']
cac40=web.DataReader("^FCHI", 'yahoo', start, end)['Close']

Next we create a DataFrame from the two downloaded series, in logarithm, note that we need to transpose the DataFrame (.T):

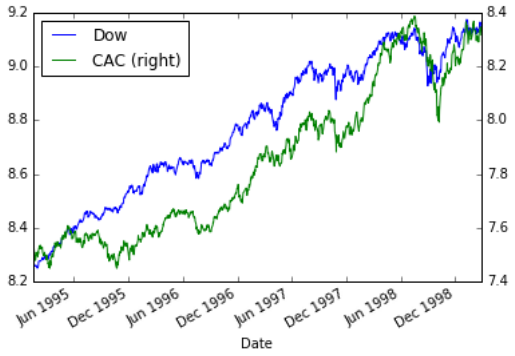
df=pd.DataFrame([np.log(dow),np.log(cac40)]).T
df.columns=['Dow','CAC']

We want to fill in the blank (not a number) by a moving average over seven days:

df = df.fillna(pd.rolling_mean(df, 7, min_periods=1).shift(−3))

We can now plot the times series:

df.plot(secondary_y='CAC')



3 A first spurious regression

Let's do a spurious regression of the CAC 40 on the Dow Jones over the period 16-01-1995 and 26-02-1999:

$$\text{LCAC}_t = \alpha + \beta \text{LDow}_t + u_t \tag{1}$$

With an ols method, first we import the package, then we fit the model, finally we print the summary of the OLS regression:

```
import statsmodels.formula.api as smf
model = smf.ols('CAC ~ Dow',data=df).fit()
print(model.summary())
```

We find a R^2 of 89% and:

$$\underset{}{\text{LCAC}_t} = \underset{(0.096)}{-1.150} + \underset{(0.011)}{1.025} \quad \text{LDow}_t \tag{2}$$

The first clue for our regression to be spurious, is that the Durbin-Watson test statistic is close to 0: 0.019
That is typical fro time series.

4 Our series were non-stationary! Our regression could be spurious.

We use the augmented Dickey-Fuller test (or unit root test).
In its simplest form, the Dickey-Fuller test the null hypothesis H_0: $\rho - 1 = 0$ with a Student test over the equation:

$$\Delta y_t = (\rho - 1) y_{t-1} + \epsilon_t \tag{3}$$

We perform the augmented Dickey-Fuller test after importing the relevant package:

```
import statsmodels.tsa.stattools as st
st.adfuller(df['CAC'].dropna())
st.adfuller(df['Dow'].dropna())
```

For the CAC 40 index, we find a test statistic of -0.2 and for the Dow of -1.34.
We conclude that both series are non-stationary and that the first regression we did might be spurious.

We perform a manual computation of the aD-F test on the Dow, using the following equation:

$$\Delta \text{LDow}_t = \underset{(x)}{\alpha} + \underset{(y)}{\beta}\,\text{LDow}_{t-1} + \underset{(z)}{\gamma}\,\Delta \text{LDow}_{t-1} \tag{4}$$

```
manual=pd.DataFrame([df['Dow'].diff(1),df['Dow'].shift(1),df['Dow'].diff(1).shift(1)]).T
manual.columns=['Y','X1','X2']
model2 = smf.ols('Y ~ X1 + X2',data=manual).fit()
print(model2.summary())
```

As suggested in the book, the statistic that we can compute is $\frac{\beta}{y} = \frac{-0.0015}{0.001} = -1.5$ which is closed to the value found above, considering the possible rounding errors.
We compare this value to the critical value at 5%, given by:

$$\text{Critical value} = -2.862 - \frac{2.738}{T} - \frac{8.36}{T^2}$$

As we are above that critical value, we cannot reject H_0 and our series are non-stationary.

4.1 Augmented Dickey-Fuller test

Exercise:
We can augment the previous test with the regression:

$$\Delta Y_t = \alpha + \sum_j \beta_j \Delta Y_{t-j} + \theta t + \gamma Y_{t-1} + u_t \tag{5}$$

And perform an F-test for the null hypothesis: $H_0 : \theta = \gamma = 0$, if it is accepted, there is a unit root.

5 Are the CAC 40 and the Dow Jones I(1)?

To test if the CAC 40 and the Dow Jones indexes are integrated of order one, we apply an augmented Dickey-Fuller test on their first differences:

```
st.adfuller(df['CAC'].diff(1).dropna())
st.adfuller(df['Dow'].diff(1).dropna())
```

We find the aD-F test statistics for:

- CAC 40 of -13.38
- Dow of -20.3

Those values are both below the critical value of -2.864 hence both series ΔLCAC_t and ΔLDow_t are stationary.
As in the previous section, we can compute the aD-F test statistic manually:

```
manual=pd.DataFrame([df['Dow'].diff(1).diff(1),df['Dow'].diff(1).shift(1),df['Dow'].diff(1).diff(1).shift(1)]).T
manual.columns=['Y','X1','X2']
model3 = smf.ols('Y ~ X1 + X2',data=manual).fit()
print(model3.summary())
```

The statistic that we can compute manually is $\frac{-1.0278}{0.043} = -23.9$ which is closed to the value found above, considering the possible rounding errors.

6 Non-spurious linear regression

Finally we can perform a non-spurious linear regression for the equation:

$$\Delta\text{LCAC}_t = \alpha + \beta\Delta\text{LDow}_t + u_t \tag{6}$$

```
model4= smf.ols('CAC ~ Dow',data=df.diff(1)).fit()
print(model4.summary())
```

We find a R^2 of 13% and:

$$\Delta\text{LCAC}_t = \underset{(0)}{0.0003} + \underset{(0.037)}{0.4819} \quad \Delta\text{LDow}_t \tag{7}$$

As $\log(\text{CAC } 40_t) - \log(\text{CAC } 40_{t-1}) \simeq$ Daily Returns, we can conclude that over the studied time period, the daily returns of the Dow explain 13% of the daily returns of the CAC 40.

6.1 Gauss-Markov hypotheses

Underlying our linear regression, we assume the Gauss-Markov hypotheses for our Ordinary Least Square estimate to be unbiased and efficient:

- $E(u_t) = 0$
- Homoskedasticity: the conditional variance of the error term is constant over time: $Var(u_t) = \sigma^2$
- Error terms are uncorrelated: $cov(u_i, u_j) = 0 \forall i \neq j$
- Our explanatory variable is deterministic (not random): $cov(\Delta\text{LDow}_t, u_t) = E(\Delta\text{LDow}_t \times u_t) - E(\Delta\text{LDow}_t) \times E(u_t) = \Delta\text{LDow}_t E(u_t) - \Delta\text{LDow}_t E(u_t) = 0$

Exercise: check is those hypotheses are respected with our data set.
If those hypotheses are respected, the OLS estimator is then BLUE (best linear unbiased estimator).
Exercise: you can manually compute the OLS estimator:

The programme is:

$$\min_{\alpha,\beta} \sum_t \left(\Delta\text{LCAC}_t - \alpha - \beta\Delta\text{LDow}_t\right)^2$$

With no constant:

$$\hat{\beta} = \frac{\sum_t \Delta\text{LDow}_t \times \Delta\text{LCAC}_t}{\sum_t \Delta\text{LDow}_t^2}$$

With a constant α:

$$\hat{\beta} = \frac{T\sum_t \Delta\text{LDow}_t \times \Delta\text{LCAC}_t - \sum_t Delta\text{LDow}_t \sum_t \Delta\text{LCAC}_t}{T\sum_t \Delta\text{LDow}_t^2 - \left(\sum_t \Delta\text{LDow}_t\right)^2}$$

7 Are the CAC 40 and the Dow Jones cointegrated?

The CAC 40 and the Dow Jones indexes are very unlikely to be cointegrated as in terms of level we do not expect to see a long term relationship between both index as the level would rather be linked to the expected earnings of the companies, which would differ in France and in the United-States of America.
The cointegration test is simply, the null hypothesis H_0 is no cointegration.

st.coint(df['CAC'],df['Dow'])

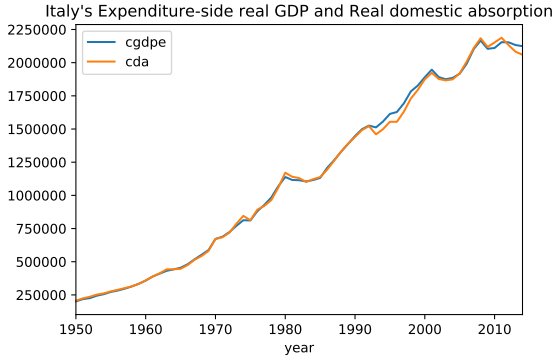
The test statistics is -2.66 which is above the critical value at 5% of -3.345. Hence those time series are not cointegrated.
We could also "manually" perform an augmented Dickey-Fuller test on the residuals of the equation 1:

st.adfuller(model.resid)

7.1 Cointegrated variables

We use data from the Penn World Table of Italy's:

- cda: Real domestic absorption, (real consumption plus investment), at current PPPs (in mil. 2011USD)
- cgdpe: Expenditure-side real GDP at current PPPs (in mil. 2011USD)



In other term, there is no unit root in the error terms of the regression:

$$Y_t = \alpha + \beta X_t + \epsilon_t$$

Exercise: you can build an error-correction model by OLS:

$$\Delta Y_t = \gamma_0 + \gamma_1 \Delta X_T + \gamma_3 \epsilon_{t-1} + \nu_t$$

with

$$\epsilon_t = Y_t - \hat{\alpha} - \hat{\beta} X_t$$

8 Final model candidate

As in the book, it is proposed to use a autoregressive distributed lag model:

$$\Delta \text{LCAC}_t = \alpha_1 + \alpha_2 \Delta \text{LDow}_t + \alpha_3 \Delta \text{LDow}_{t-1} + \alpha_4 \Delta \text{LCAC}_{t-1} + u_t \tag{8}$$

```
final=pd.DataFrame([df['CAC'].diff(1),df['Dow'].diff(1),df['Dow'].diff(1).shift(1),df['CAC'].diff(1).shift(1)]).T
final.columns=['Y','X1','X2','X3']
modelf = smf.ols('Y ~ X1 + X2 + X3',data=final).fit()
print(modelf.summary())
```

We find that the instantaneous rise of 1% of the Dow result in an instantaneous rise of the CAC 40 of 0.49% and the final effect o 0.8% as $\frac{0.49+0.42}{1-(-0.12)}$.

8.1 Breusch Godfrey - are the residual correlated at any order?

The Breusch Godfrey test as pointed out in the book seems to indicate that we have second order autocorrelation. The null hypothesis, H_0 is that there is no autocorrelation of the residuals at any order.

8.1.1 Manually in python

We can manually implement the Breusch adn Godfrey test:

1. we regress the model 8
2. we use the residuals from the above regression and test: $\hat{u}_t = \alpha_1 + \alpha_2 \hat{u}_{t-1} + \alpha_3 \hat{u}_{t-2} + \alpha_4 \Delta \text{LDow}_t + \alpha_5 \Delta \text{LDow}_{t-1} + \alpha_6 \Delta \text{LCAC}_{t-1} + v_t$
3. we test the statistic $LM = nR^2$ against the threshold $\chi^2_{0.05}(2) = 5.99$

```
breuschgodfreyf=pd.DataFrame([modelf.resid,modelf.resid.shift(1),modelf.resid.shift(2),df['Dow'].diff(1),df['Dow'].diff(1).shift(1),df['CAC'].diff(1).shift(1)]).T
breuschgodfreyf.columns=['u','u1','u2','X1','X2','X3']
breuschgodfreyf=breuschgodfreyf.dropna(axis=0)
modelbgf = smf.ols('u ~ u1 + u2 + X1 + X2 + X3',data=breuschgodfreyf).fit()
print(modelbgf.summary())
LM=(len(df)-2)*modelbgf.rsquared
LM
```

We find a statistics of 2.64 which is below the threshold thus we do not reject the null hypothesis.

8.1.2 With the python function

We use the python function smsdia.acorr_breusch_godfrey, as an output, you get:

- Lagrange multiplier test statistic
- p-value for Lagrange multiplier test
- fstatistic for F test
- pvalue for F test

9 ARCH model of the CAC 40

Frist we use the Robert Engle method to test the presence of an ARCH process. From the regression 8 we did, we use the residuals and regress:

$$\hat{u}_t^2 = \alpha + \beta \hat{u}_{t-1}^2 + \nu_t \tag{9}$$

Then we use the R^2 of that regression and the length of the time series (here 1062 days) and compare it to the threshold 3.94.

```
df_ARCH=pd.DataFrame([modelf.resid**2,modelf.resid.shift(1)**2]).T
df_ARCH.columns=['Y','X']
modelARCH=smf.ols('Y ~ X',data=df_ARCH).fit()
```

```
modelARCH.summary()
R2=modelARCH.rsquared
T=len(df_ARCH)
ARCH=T*R2
```

We find a $R^2 = 17\%$ and we have $T = 1062$ the length of our time series. Hence our test statistic is above the threshold χ^2_1: $17.7 > 3.94$, therefore the null hypothesis H_0: $\beta = 0$ is rejected and we are facing an ARCH process.

10 Durbin-Watson test

The errors of a regression model are supposed to be uncorrelated. This can break down with time series. The errors can be serially correlated. Hence an errors associated with one observation will then carry over to future observations. This is relevant namely with net present value model where future dividends impact the current stock price.
The serial correlation of the errors can be written as: $u_t = \rho u_{t-1} + \theta_t$
The Durbin-Watson test the null hypothesis H_0: $\rho = 0$.

We assume a positive correlation if $DW < d_L$, H_0 is accepted (no correlation) if $d_U < DW < 4 - D_Y$ a negative correlation if $DW > 4 - D_L$, and the test is inconclusive otherwise.

Note that the Durbin-Watson test cannot be used if the regression equation contains a lagged dependent variable.
We can apply this test to our equation 6:

```
#Durbin−Watson test
import statsmodels
statsmodels.stats.stattools.durbin_watson(model4.resid)
```

We find a value of $DW = 2.25$ which could tend to show that there is no serial correlation ($\rho \neq 0$) as $DW = 2(1 - \hat{\rho})$, but in fact, looking at the Durbin-Watson table for n observations and k explanatory variables at the 5% level of significance, we find that $d_u \simeq 1.65$ and $d_l \simeq 1.69$, $DW > d_u$ hence we do not reject the null hypothesis, and there is negative serial correlation present in our initial model 6.
Exercise:
Estimate ρ:

$$Y_t = \alpha(1-\rho) + \rho Y_{t-1} + \beta X_t - \beta\rho X_{t-1} + \nu_t$$

Then re-estimate the model on the transformed variables:

$$Y_t - \hat{\rho} Y_{t-1} = \alpha(1-\hat{\rho}) + \beta X_t - \beta\hat{\rho} X_{t-1} + u_t - \hat{\rho} u_{t-1}$$

Exercise to go further:
With a lagged endogenous variable as in our final model 8, you could use the statistics:

$$h = \left(1 - \frac{DW}{2}\right)\sqrt{\frac{T}{1 - TVar(\hat{\alpha}_4)}}$$

statsmodels.stats.stattools.durbin_watson(modelf.resid)
h=(1−statsmodels.stats.stattools.durbin_watson(modelf.resid)/2)∗np.sqrt(**len**(df)/(1−**len**(df)∗0.029∗∗2))

With $Var(\hat{\alpha}_4)$ the square of the standard error observed in the summary of the regression 8. Can you suppose h to be normally distributed with unit variance and the normal distribution table be used?

11 Granger causality

So far we have been working on identifying the sensitivity of the CAC 40 changes to the Dow Jones changes. We have not been trying to identify if the Dow Jones changes are causing the CAC 40's or vice-versa.

A stationary variable X Granger-causes another stationary variable Y if: $E(Y|Y_{t-k}, X_{t-k}) \neq E(Y|Y_{t-k})$. That is: the history of variable X brings information to the prediction of variable Y. In fact, Granger-causality means that from the historical observations, one variable precedes the other.

11.1 Illustration with the bivariate direct Granger test

We are considering two variables and use the following regression model:

$$Y_t = \sum_{j=1}^{k} \alpha_j Y_{t-j} + \sum_{i=1}^{q} \beta_i X_{t-i} + D_t + \epsilon_t \tag{10}$$

with D_t the deterministic and ϵ_t the errors.

We can first perform an F-test to test for the null hypothesis: H_0: $\forall j, \alpha_j = 0$ and $\forall i, \beta_i = 0$, against H_1: there is at least one coefficient different from 0.
The F-statistic of our unrestricted model is:

$$F = \frac{(R^2/(k+q-1))}{(1-R^2)/(n-k-q)}$$

with $k+q$ the number of estimated parameters and n the number of observations. If this is above the critical value (from the Fisher table, for $F_{k+q-1,n-k}$ with $\alpha = 5\%$, we reject H_0 but before we can conclude whether X is Granger-causing Y, we need to perform a Wald test as follow:

We use the Wald test with the null hypothesis H_0: $\forall i, \beta_i = 0$. The restricted model is:

$$Y_t = \sum_{j=1}^{k} \alpha_j Y_{t-j} + D_t + \epsilon_t \tag{11}$$

There are q restrictions, n is the length of our time series, and k the lag. It can be demonstrated that the statistics of the test can be written as (with SSE the sum of squared residuals:

$$F - stat = \frac{R_U^2 - R_R^2}{1 - R_U^2} \frac{n-k}{q} = \frac{SSE_R - SSE_U}{SSE_U} \frac{n-k}{q} \tag{12}$$

if this exceed the tabular value of $F_{q,n-k-q}$, then the null hypothesis is rejected.

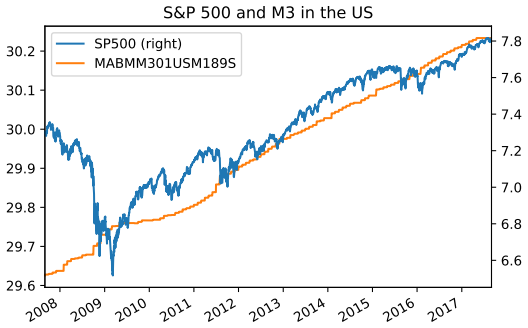
11.2 Sims and Modified Sims test

Exercise: implement the Sims and Modified Sims tests.

12 Error Correction Model

For the following, we use the code VECM_vansteenberghe.py

We download from FRED the M3 for the United States and the S&P 500 index, taking their log.



We perform augmented Dickey-Fuller test, both series are $I(1)$ and we also do a cointegration test: we reject the null hypothesis H_0 that both series are not cointegrated.
We first estimate the long run relationship from the spurious model:

$$Y_t = \alpha + \beta X_t + \epsilon_t$$

and we estimate the disequilibrium errors or the cointegrating residuals:

$$\epsilon_t = Y_t - \hat{\alpha} - \hat{\beta} X_t$$

We then compute the error correction model were all variable are now stationary:

$$\Delta Y_t = \gamma_0 + \gamma_1 \Delta X_t + \gamma_2 \epsilon_{t-1} + \nu_t$$

The results with these data are note very convincing, we will rather use again the Penn World Table of Italy's:

- cda: Real domestic absorption, (real consumption plus investment), at current PPPs (in mil. 2011USD)
- cgdpe: Expenditure-side real GDP at current PPPs (in mil. 2011USD)

We can say that our model explains 89.6% of the variance of the observed changes. And that **for a** 1% **increase in Italy's expenditure-side real GDP there is a** 1.09% **increase in Italy's real domestic absorption** ($\gamma_1 = 1.09$ and is significantly different than 0 with a t-value of 20.93 well above 1.96). Remember that we are using here log and:

Applying a Taylor expansion of $\ln(x)$ when x is close to 1:

$$\ln(x) = \ln(1) + \frac{x-1}{1} - \frac{(x-1)^2}{2} + \dots$$

Therefore:

$$\ln\left(\frac{x_{i,t}}{x_{i,t-1}}\right) \simeq \frac{x_{i,t}}{x_{i,t-1}} - 1$$

If the variable is not too volatile over time, that is if $\frac{x_{i,t}}{x_{i,t-1}} \simeq 1$, then we could also compute the growth rate for the variable i:

$$GR_{i,t} = \ln\left(\frac{x_{i,t}}{x_{i,t-1}}\right) = \ln(x_{i,t}) - \ln(x_{i,t-1}) = \Delta \ln(x_{i,t}) \tag{13}$$