Quantitative methods in finance

Eric Vansteenberghe

November 18, 2021

> Perdre un bonheur rêvé, renoncer à tout un avenir, est une souffrance plus aiguë que celle causée par la ruine d'une félicité ressentie, quelque complète qu'elle ait été : l'espérance n'est-elle pas meilleure que le souvenir ?
>
> La Bourse
> *Honoré de Balzac*

Contents

1 Introduction

In this lecture designed for Master 2 Research students, we propose different applications of quantitative methods in finance including time series analysis, asset pricing models, risk measures, panel analysis, etc.

1.1 Which tools are allowed to be used for this lecture?

Two open source tools are suggested to be used during the class: python and R. While both tools have pros and cons (and their rankings depends on which criteria you value, e.g. based on search engines), they both allow for rapid prototyping of investment strategies and/or academic models which is our main objective. In the industry, once back-tested an investment strategy would be implemented typically in C++.
Students are also allowed to use Julia, Matlab, Stata or SAS for their work as long as they can demonstrate on request they have a valid license if the software requires one. If a pirate software is found on a computer, sanctions will apply.

1.2 What type of computer do I need for this lecture?

We strongly recommend students to invest in a laptop. As of mid-2020, based on the website lesnumeriques it is feasible to have a computer well performing for a budget around 800 €, when choosing a laptop keep in mind that rapid access to the hard drive is recommended (which favour SSD), if one wish to work with big data the RAM will be important (8 Go is a good starting point), then for computation the CPU enters into consideration (1.5 GHz is a good starting point).

1.3 What is expected out of these lectures?

The aim of these sessions is not (yet) to produce innovations, but rather to learn to replicate academic research papers using available data (students should register at the library where they would have access to data sources). Once the quantitative methods are mastered, the students should be able to contribute to the economic research field.

1.4 Disclaimer

The methods and conclusions reached in this documents and codes are:

- the view of the author and do not reflect the view of his past and present employers,
- developed for pedagogic purpose only and would not be recommended for actual investment decision making.

2 Projects and evaluation

2.1 Some notes on ethics

According to the Financial Times of 3rd February 2020, Citigroup has suspended one of its most senior bond traders in London (Paras Shah) after he had **stolen food from the canteen**. Two of his former colleagues told the FT he was a well-liked and successful trader.
In 2016, Japan's Mizuho Bank fired a London banker after he was caught stealing a part from a colleague's bike worth about 5 GBP.
I first found it hard to understand how wealthy bankers would steal a sandwich or a bicycle bell. But then I realized how, in a rush, we might misjudge a situation and be tempted to take short-cuts. It would have taken both employees half an hour to go to a regular shop to buy a sandwich or a bicycle bell. They could both afford spending half an hour on this task, or even 5 minutes to order it online, and spending the money. They did not bother and they lost their jobs and their names are out there, making it hard for them to find another position at a bank.
This is a reminder to be humble and even when it is 20:30 and we are tired not to take shortcuts that can cost our reputation.

2.2 First semester: a first step toward a master thesis

The students will be evaluated via two methods (all submissions are expected in electronic format):

2.2.1 *Coding* grade

During most classes, exercises have to be carried out by the students and can be requested at the end of the lecture to be evaluated and become the *coding* grade. Alternatively, the students can be selected to present a technical subject and programming application during the lecture. If she has not been selected during the semester at the end of a class or for a presentation, the students can have to work during the final session on coding in R or python. Work will be individual. Internet connexion might not be permitted during the whole session (to be determined on the day depending on the subject). In case the student has no computer(or its battery is empty) or its Spyder or RStudio environments are not working she will have to submit her work in paper form, otherwise submission is expected at the end in e-mail format to the supervisor.

2.2.2 Final report for (tbd) midnight

For the final report, one student or a pair can chose to prepare for a summary of the theory (for example re-demonstrate some key equations to clarify how the model work and on which theory it is based), and to reproduce the models and results of research papers to address a question (for this semester, the same question is given to all students and pairs). Each student will have to:

- deliver the python, R, Julia, Stata[1], Matlab or SAS code
- deliver the data source used (in csv format)
- deliver the report in pdf **and also** provide the .tex file they have to use to generate the pdf
  - we suggest to install MikTex and TexMaker to work with LaTeX
  - overleaf[2] can also be used as an alternative to MikTex and TexMaker, you'll still need to

[1] if Stata, Matlab or SAS are used, the students will have to provide on request proofs that all students in the group had a legal version of the software
[2] here is an introduciton to LaTeXfor overleaf users

export and send me a .tex file that can be executed or a link to your project, you can use an example here

- we suggest to use the following as templates: (make sure you have the graph gobronperfhistory.pdf in your folder for this to compute) and `latex_doc_template.tex`

2.2.3 Question to be addressed

To be announced.
We want to test the following skills:

1. write a concise literature review on a specific subject to tackle an imposed research question:
   - from your literature review, determine your data need and main model;
   - this is very important, do not rush to downloading data or writing code, complete a full literature review to be able to determine what data you need and what methods are the most relevant to be applied.
   - try to focus on articles published in category 1 (CNRS) reviews
     - for example, you should be able to trust an article publish in the J*ournal of Applied Econometrics* which is categroy 1, but be careful about non-ranked reviews such as the *Asian Journal of Medicine and Health* where it has been possible for some researcher to publish a fake article pretending that **SARS-CoV-2 was Unexpectedly Deadlier than Push-scooters: Could Hydroxychloroquine be the Unique Solution?** where they wrote

       > Following the methodological rule according to which the smaller the sample, the higher the statistical significance, we decided to stop recruitment as soon as a significant effect at 84% was detected. [...] Joachim Son-Forget, Member of Parliament, who taught us that linear regression starts from 3 points; we soon hope to push the limits and reach the purity of linear regression at 1 point.

2. collect data:
   - it might be difficult to get qualitative and long historical data, so manage your effort and you might decide focusing your study on a sub-sample.
3. prepare the data before modeling.
4. econometric tests, show that you master both their theoretical aspects and applications to data sets.
5. model calibration.
6. writing a report:
   - discuss your results, if you find different patterns than the literature, understand why, if you find similar patterns than the literature, understand why.
   - be concise, the shorter the better;
   - we will judge your work by its precision, cleanliness. Length tends to be negatively related to the final grade. If everything you have to say fits in four pages, with a proper introduction, literature review, results display, discussion and conclusion then great;

- so remember, you can summarize test results in one table in your report, with a legend and a discussion of your results. Don't be frustrated that three days of work end up in a table and a discussion of the results that fits in half a page in your report, this is fine. We know you will spend time on the data collection and on writing the code, I will review both the code and the data sets.

2.2.4 Some remarks about the final report and code

If you want a reference on how to best write your code in python, you can follow PEP 8.
As a minimum, in your python or R codes:

- Start your code with a comment section with the date, authors and a short description of the purpose of the code
- For each equation used, the variables should be defined with names that are concise and meaningful
- Each figure presented should have a caption and some explanation
- Make your report self explanatory as much as possible
- Indicate where you got your data set from (Bloomberg, Macrobond, CRSP, etc.), detail the code name or reference number of the variable you used (e.g. if you downloaded the Global Price of Wheat from FRED, indicate the reference PWHEAMTUSDM)
- Provide the original raw data and make modifications within your code
- Describe the data set used: is it composed of returns, index prices? Define the columns of your DataFrame if the names are not explicitely related to variables from an equation in your paper
- If you have to perform the same actions several times, leverage on functions and loops, define functions with verbs that are meaningful
- Before performing a regression, apply econometric tests on the variables
- In your code, use decipherable variable names and try to avoid hard coding values (typical example: define a variable $nbobs = len(df)$ instead of using the actual value (let's say 1748) of the length of your DataFrame in your code
- Include comments in your code and especially comments that help relating a part of your code to a section of your report
- If your code is made of several files, explain the architecture, the options and how to run your programme

Nota bene:

- it might not be necessary to use as much data as in the original paper, for example, if the paper test a trading strategy over the CRSP, you might not have access to the CRSP, but can test on a smaller sample of stock prices (e.g. choose 100 stocks and test the strategy).
- if you are not sure about some aspects of your results, don't hide, just explain the difficulties you have in detail.

2.3 Second semester: same research question as the first semester

3 Quantitative methods with R or python

3.1 Lecture, books and internet resources for quantitative methods in finance

- Professor Catherine Bruneau lectures.
- Lecture Notes in Financial Econometrics (MSc course). Paul Soderlind.
- Series temporelles, Xavier Guyon
- For the theoretical part, Massimo Guidolin lectures
- Library of Statistical Techniques The "Rosetta Stone" for econometrics between python, R and Stata

3.2 Books and internet resources to learn python and R

If you have never used any tool like R, Matlab, Stata, you might want to start with those step by step exercises we suggest to understand how to use matrices, DataFrames, functions, loops:

- Beginner's guide to use arrays and functions section 5.
- Beginner's guide to use DataFrames and manipulate data set section 6.

The aim of our Quantitative Methods in Finance is to provide students with tools to deploy the economic and mathematical theories to empirical data sets. There are some resources if students want to learn the languages R and python in depth:

- python and/or R
  - The website Stack Overflow is your best ally. You can just ask questions and it is most likely that it will already be answered. Example, if you wonder how to loop through rows of a DataFrame, it is answered here in plain English.
  - Financial risk forecasting. Jon Danielsson
- python
  - SciPy lectures
  - Python for economist. Alex Bell
  - Fichier dans la langue de Molière sur python
  - Introduction to Python for Econometrics, Statistics and Data Analysis. Kevin Sheppard
  - The python official tutorial and the pandas DataFrame tutorials.
  - Some MIT teachers, Peter Wentworth, Jeffrey Elkner, Allen B. Downey and Chris Meyers, wrote a book freely available online: *How to Think Like a Computer Scientist: Learning with Python 3 Documentation*.
  - The economist Thomas J. Sargent has a website with material to learn python: here.
  - Mastering Python for Finance in our bibliography item (Weiming, 2015).
  - Xavier Dupre lecture for ENSAE
  - Python Machine Learning. Sebastian Raschka.

- R
  - An introduction to R by the R Core Team
  - Statistics and Data Analysis for Financial Engineering with R examples. Ruppert, David, Matteson, David S.
  - Cheat sheets for R

3.3 Install Python 3.x and Anaconda

For the purpose of the following lecture, we recommend to install the latest version of Python 3 and a suite that allows to leverage on the power of many packages and offers a visual environment: *Anaconda* .

3.3.1 Step 1

Install lastest Python 3 version e.g. with this link

3.3.2 Step 2

Install *Anaconda* for Python 3 click here to be redirected to their website If installing on a Mac, you might prefer to install it via command line namely to have the conda command in the terminal working properly as explained here

3.3.3 Step 3

- Click on the Anaconda *Navigator*, then launch the Scientific Python Development Environment: *Spyder*.
- Chose a folder in which you want to work and copy-paste the file `types_loops_functions.py`
- In *Sypder* open the source code file
- Execute the source code in Spyder: *Run File (F5)*
- If you want to execute line by line, place the cursor on a line and hit (F9)

To explore more of Spyder, use Help -> Spyder tutorial

3.4 Install R and RStudio

For the purpose of the following lecture section, we recommend to install the latest version of R and a suite that allows to leverage on the power of many packages and offers a visual environment: *RStudio Desktop* (click here to be redirected). NB: although it is technically possible to run RStudio from Anaconda, we do not recommend this as we have experienced issues in the past and recommend to follow the steps below to install R and RStudio as a standalone.

3.4.1 Step 1

Install the latest R version from `https://cran.r-project.org/`

3.4.2 Step 2

Install *RStudio Desktop*

3.4.3 Step 3

Open RStudio

3.4.4 Step 4

- Chose a folder in which you want to work and copy-paste the source code file `helloworld_vansteenberghe.R`
- Create in the above folder a folder data where you place all the data files (`peabis_price.csv`, etc.)
- In *RStudio* open the source code file `helloworld_vansteenberghe.R`
- Change in the code the setwd("...") into the path to your chosen working directory
- Execute the source code in RStudio:
  - Select the all the code *Ctrl-All*
  - Execute the code *Ctrl-Enter*
- You will need to install packages: *Tool- Install packages...* then type the name of the package and press enter. We recommend to do the following: do not install manually packages, rather add them to a list in a R code as we do in the file packages_manager.R, this way if you have to re-install R, you can then run this file and install all required packages for your codes.

4 Python: variables, functions. An introduction

For a full review of the types in python, you can go to the documentation: here for the built-in types and the operations you can apply to them.[3]
We intend to present what you are most likely to manipulate as an economist[4].

4.1 Immutable build-in types

Imagine you want to define a variable with name "growth". This variable could be:

- None, if we don't know what it should be
- bool, a boolean: True if we have growth and False if we don't
- int, if it is an integer[5]
- float, if it is a floating number
- complex, if it is a complex number
- str, it is is a string
- tuple, if it is a tuple[6]

If at any point in time you want to clean the "Variable explorer" (and the variables in memory), type in the IPython console %reset and confirm typing y.

4.2 Mutable build-in types

If you want to compute growth from a time series, you won't just manipulate multiple variables, you can use mutable build-in python types:

- list
- dictionary

4.3 Conditions and loops

We might want to apply an action if a condition is met or loop a variable through different values or while a condition is met.

4.4 Function

We can define our own functions and call that function any time we need to do the same operations as defined in the function and return the results.

[3] No need to learn the operations by heart, if you want to do an operation, go and type into Google: "python concatenate two lists"
[4] types_loops_functions.py
[5] one might wonder why there are int and float, appart from memory management, here is just an illustration
[6] cf. python documentation, Tuples are immutable sequences, typically used to store collections of heterogeneous data

4.5 numpy and pandas

You can also chose to work with libraries: numpy or pandas which are powerful and each has its advantages. Just as an illustration we compare the speed to perform matrix multiplication with numpy or pandas objects.

4.6 Immutable versus mutable types and memory management

As an economist, you might not need to know how python manage immutable and mutable types variables and memory, here are some elements and a comparison with C.
The main takeaway is **aliasing** where two variable refer to the same object, this link can be broken with .copy(). There are some subtleties, however, which is important to understand for some application in pandas and the management of object identity as revealed by some limitation in apply .copy(deep=True) to a DataFrame as detailed here.
But do not worry too much about this, I never encountered this problem and I believe that for your master work you are unlikely to encounter such issues.

5 Python for non-programmers: numpy exercise part 1

We suggest a code[7] to get familiar with some numpy basics features in python.

5.1 Work with numpy

We suggest in this lecture to work with the spyder environment.

5.1.1 Import the package

First we want to import the package numpy.

5.1.2 Work with exponential function

With the package numpy you already have a lot of functions implemented and ready to use.
For example, the function **exponential**:

$$\exp : x \rightarrow \exp(x)$$

read "funtion exponential that to a variable x associates the value "exponential of x".
If we want to know the value of $exp(1.5)$:

5.1.3 y=f(x)

Now we want to use x as a vector containing different values and then for each value of x compute the associated y=f(x), with f being a function (in our case we work with exponential function).
We create a vector called x, ranging from 0 to 9 with steps of 1, either we define manually the array or define a range.
This creates a vector:

$$X = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{pmatrix} \tag{1}$$

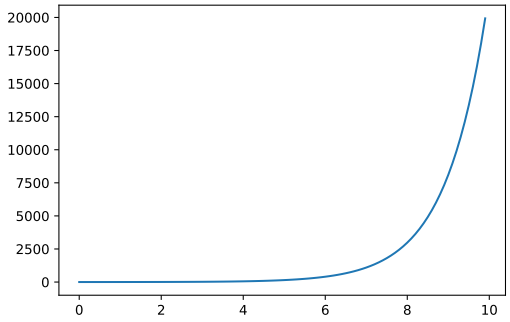
Then we do $y = f(x)$, the function f in our case being the exponential.
This returns a vector as output:

$$Y = \begin{pmatrix} e^0 & e^1 & e^2 & e^3 & e^4 & e^5 & e^6 & e^7 & e^8 & e^9 \end{pmatrix} \tag{2}$$

5.1.4 Plot y=f(x)

Now we want to use a package to be able to plot $y = f(x)$.
We might want a smoother plot, in this case, we will use more computing power and have smaller steps between 0 and 10, as an illustration we want to plot by steps of 0.1:

[7] numpy_exercise_part1.py



5.2 Create you own function

Now you should create your own function, that you could call my_f:

$$\text{my\_f}: x \to x^2 + x - 2$$

Once defined, you want to apply your function to the original x vector to get an output vector y_f and plot it.

5.3 Find x where f(x)=0

Now we want to solve for $f(x) = 0$. We can try if a=b using the operator "==":

my_f(0)==0

we find that this is not the case: FALSE

my_f(1)==0

we find that this is the case: TRUE
Indeed, $x^2 + x - 2 = 0$ when $x = 1$.
Now we can loop through all the element in the vector x $(0, 1, \ldots, 9)$ and print which element solve: f(element)=0.

5.3.1 fsolve

The good news is that there is already a function to do the above: fsolve.
You need to first import the function from the right package. Now to find the solution to $f(x) = 0$, it is in one line:

fsolve(my_f,0)

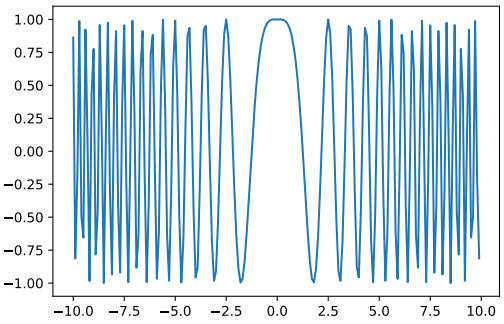
Not that ",0)" here tells fsolve **where** to start searching for the root in the abscisse line.
You can get help on this function using the console:

?fsolve

x0 = 0 is *the starting estimate for the roots of 'func(x) = 0'*

5.3.2 Define another function and find the root closest to x=10

Define a function: $f(x) = \cos(x^2)$, plot it and find the solution of $f(x) = 0$ close to $x = 10$.



5.4 Fixed-point iteration

Given a function f and a starting point x_0, the fixed-point iteration is:

$$x_{n+1} = f(x_n), n = 0, 1, \ldots \tag{3}$$

giving rise to a sequence $x_0, x_1, \ldots$ which is hoped to converge to a point x. If f is continuous, then one can prove that the obtained x is a fixed point of f, i.e. $f(x) = x$

5.4.1 Babylonian method for computing the square root of $a > 0$

We use the function:

$$f(x) = \frac{1}{2}\left(\frac{a}{x} + x\right) \tag{4}$$

which approach the limit $x = \sqrt{a}$ from whatever starting point x_0. We use a sequence of n_{int} iterations. This is a special case of Newton's method for finding roots of a differentiable function f, writing $g(x) = x - \frac{f(x)}{f'(x)}$ we write the fixed-point iteration: $x_{n+1} = g(x_n)$ if this converges to a fixed point x, then x is a root of f.

5.4.2 Banach fixed point theorem

We choose an example satisfying the Banach fixed point theorem: the fixed-point iteration $x_{n+1} = cos(x_n)$ which converges to the unique fixed point of the function $f(x) = cos(x)$ for any starting point x_0. It can be shown that starting from x_0, the error after n steps satisfies:

$$|x_n - x| = \frac{q^n}{1-q}|x_1 - x_0|$$

with $w = 0.85$ when $x_0 = 1$

5.4.3 Fixed-point iteration on "any" function

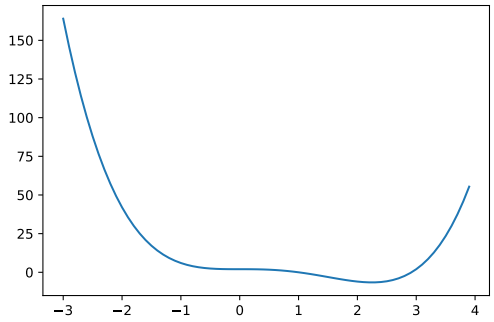
- **Question 1**
  - Apply the fixed-point iteration of the function: $f(x) = 3x^3 - 2x^2 - 0.5$ do you observe convergence? Vary the starting point x_0.

5.5 Gradient descent

We simply follow the Wikipedia example:

> Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. To find a local minimum of a function using gradient descent,

we start at an initial guess x_0, then take steps proportional (γ) to the negative of the gradient ∇f (or approximate gradient) of the function at the current point: $x_{n+1} = x_n - \gamma \nabla f(x_n)$. As an illustration, we use the gradient descent to find the minimum of the function $f(x) = x^4 - 3x^3 + 2$



6 Python for non-programmers: exercise with pandas - part 1

We suggest exercises[8] to get familiar with some pandas basics features in python's pandas library.

6.1 Work with panda DataFrame

6.1.1 Import the package

First we want to import the package panda and create a dataframe we call df:

Index	A	B	C
0	1	1	2
1	3	4	5
2	7	8	9

You might want to sum over rows for each column .sum(axis=0) or sum over columns for each row .sum(axis=1)

6.1.2 Finding location of an element in a DataFrame

The panda's documentation explains how to access specific element of a DataFrame, using either the label or the integer positions. Link here.

6.1.3 Selection By Label

You can use the .loc attribute to access elements in a DataFrame, label-based. For df, that would be:

	A	B	C
0	df.loc[[0],['A']]	df.loc[[0],['B']]	df.loc[[0],['C']]
1	df.loc[[1],['A']]	df.loc[[1],['B']]	df.loc[[1],['C']]
2	df.loc[[2],['A']]	df.loc[[2],['B']]	df.loc[[2],['C']]

You can also create subset of your dataframe, slicing it.

```
slice1 = df.loc[ :,['A', 'C']]
slice2 = df.loc[1,['A', 'C']]
slice3 = df.loc[1:,:'B']
```

Which yields from left to right:

Index	A	C
0	0	2
1	3	5
2	7	9

Index	1
A	3
C	5

Index	A	B
1	3	4
2	7	8

6.1.4 Selection by position (integer)

If you want to be able to select a cell in a DataFrame, regardless of the name of the column or the name of the row, you can use integer based indexing with .iloc attribute. Keep in mind that python indexing convention starts at 0.
Creating the above slices would become:

[8] pandas_exercise_part1.py

slice1i = df.iloc[:,[0, 2]]
slice2i = df.iloc[1,[0, 2]]
slice3i = df.iloc[1:,:2]

Note the subtlety, illustrated by our example slice3i:

- indexing starts at 0
- with python slices, the start is included but not the stop; so $'1:2'$ yields $[0,1]$ and not $[0,1,2]$

6.1.5 Other selection methods

You can also use other selection methods, described in panda's documentation: .at, .iat, .ix

6.1.6 Unstack

We can use the function unstack on the DataFrame to transform it to a list with the combinations of row and column index. We observe for df unstacked:

A,0	df.iloc[0,0]
A,1	df.iloc[1,0]
A,2	df.iloc[2,0]
B,0	df.iloc[0,1]
B,1	df.iloc[1,1]
…	…

6.1.7 Loop through a DataFrame

By using positions or names of columns and/or index, we are able to loop through a DataFrame. The idea behind this is that the elements of the DataFrame df2 are found:

df.iloc[0,0]	df.iloc[0,1]	df.iloc[0,2]
df.iloc[1,0]	df.iloc[1,1]	df.iloc[1,2]
df.iloc[2,0]	df.iloc[2,1]	df.iloc[2,2]

It is also possible to transpose a data frame. The transpose of df2 would be:

df.iloc[0,0]	df.iloc[1,0]	df.iloc[2,0]
df.iloc[0,1]	df.iloc[1,1]	df.iloc[2,1]
df.iloc[0,2]	df.iloc[1,2]	df.iloc[2,2]

6.1.8 DataFrame as matrix for multiplication

We create a vector with its index as the matrix column names and we can then perform a matrix multiplication[9]

$$\begin{bmatrix} 1 & 1 & 2 \\ 3 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix} \times \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 18 \\ 52 \\ 100 \end{bmatrix}$$

[9] It is important that the columns names of the matrix are the same as the index of the vector for the product to be acceptable

6.1.9 Multiplying element by element (Hadamard product)

We might want to multiply element by element, this is also called Hadamard product, that we write here $\odot$.

First we duplicate the column to form a square matrix and multiply the two matrices:

$$\begin{bmatrix} 1 & 1 & 2 \\ 3 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix} \odot \begin{bmatrix} 2 & 4 & 6 \\ 2 & 4 & 6 \\ 2 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 12 \\ 6 & 16 & 30 \\ 14 & 32 & 54 \end{bmatrix}$$

6.1.10 Singular and non singular matrices

A square[10] matrix Δ is said to be non-singular if it is possible to find a matrix Δ^{-1} such that $\Delta * \Delta^{-1} = I$ with I the identity matrix[11]. We can also say that a $n \times n$ matrix is non-singular if its n rows (columns) are linearly independent, or in other words, the matrix has full rank. Non-singular matrices have a non-zero determinant.
Matrices will be quite helpful when working on AR(p) and VAR models.

6.1.11 Compute the eigenvalues and eigenvectors of the DataFrame

When studying the stability of a repeated matrix product with a vector, we might want to compute the eigenvalues and eigenvectors of our matrix.
An eigenvector V_1 for a matrix Δ is a vector such that there exist an eigenvalue λ_1 such that $\Delta V_1 = \lambda_1 V_1$. For a non-singular matrix, it will be the case that all eigenvalues are different than 0.
NB: one can introduce unit test to check that indeed the eigenvalues and eigenvectors we extracted follow the equality.
Let's imagine that we are dealing with a network of exposures in the vein of DebtRank introduced by (Bardoscia et al., 2015), we define the Δ in their paper as dfa.
We apply as in the paper an initial shock of $h(1) = 0.5\%$ on all nodes of the network. The shock propagates over time

$$h(t+1) = \sum_{s=0}^{t+1} \Delta^s h(1)$$

Which if the eigenvalues of Δ are all smaller than one can be computed, after checking that $(I - \Delta)$ is non-singular:

$$h(\infty) = (I - \Delta)^{-1} h(1)$$

with I the identity matrix.
If an eigenvalue of Δ is greater than one, the sum diverge (which in the DebtRank model means a default).

6.2 DataFrame manipulation for population time series

We want to create the following DataFrame with the French population values:

[10] for a matrix, being square mean that it has the same number of rows and columns
[11] a matrix filled with zeros except for its diagonal filled with ones

Septembre	2016	66 790
Aout	2016	66 763
Juillet	2016	66 735
Juin	2016	66 710
Mai	2016	66 688
Avril	2016	66 672
Mars	2016	66 659
Fevrier	2016	66 644
Janvier	2016	66 628

Look carefully at the way your DataFrame is indexed, it starts from 0 and not from 1. Therefore here you have 9 observations indexed from 0 to 8.
As we entered the data in the wrong order, we need to reverse:

pop = pop.iloc[::-1]

6.2.1 Why can we reverse the index using ::-1?

You can look into python's documentation on slice class. Imagine you have a list you call mylist:

- mylist = [0, 1, 2, 3]

Conveniently, we chose the value of each item to be its index position in the list, so 0 is at index 0, 1 at index position 1, etc. Imagine you want to reverse this mylist: we have a list of length 4 so going from index position 0 to index position 3. To enumerate the index, you can do: $start : end : step$ within the [and] signs. So mylist[0:len(mylist):1] will return the full list.
To reverse the list, we want the step to be -1, we omit the start and end position, so that python takes all possible items. Just look back at section 6.1.3 where we used 1 : to select all items starting at index 1 and after.

6.2.2 Indexing with time

This is optional, but you might want to index your DataFrame with time. First you want to create an object with the dates as 'year-month' from 2016-01 to 2016-09 and modify our DataFrame index. Note that we have to range until 2016-10 for the last item of our dates object to be 2016-09.

6.2.3 Renaming the column and plotting

We see that the name of our column is '0' we want to rename it with 'population'. We can have a look at the data by plotting it and choosing a title.

6.2.4 Computing the monthly changes of the population

We want to compute the change of the French population from month m to month $m+1$:

$$\text{change}_{m+1} = \frac{\text{pop}_{m+1} - \text{pop}_m}{\text{pop}_m}$$

In visual terms, that will be:

pop	**pop.shift(1)**	**change_pop**
pop_m		NaN
pop_{m+1}	pop_m	$\frac{\text{pop}_{m+1}}{\text{pop}_m} - 1$
pop_{m+2}	pop_{m+1}	$\frac{\text{pop}_{m+2}}{\text{pop}_{m+1}} - 1$
pop_{m+3}	pop_{m+2}	$\frac{\text{pop}_{m+3}}{\text{pop}_{m+2}} - 1$
…	…	…

7 Python, import and pandas: Import csv data as DataFrame

We suggest[12] to build a DataFrame of the evolution of the French population as taken from INSEE website: here.

7.1 Indicating where your file is located

Now you need to 'tell' python where you file is located.
Import the package Miscellaneous operating system interfaces:

import os

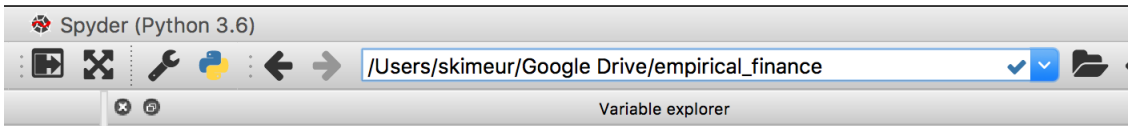
Then you need to tell him where your data folder is located, in my computer, that would be:

os.chdir('/Users/skimeur/Google Drive/empirical_finance/data')

If you are note sure where your file is located, here is a step-by-step guide:

1. make sure you downloaded the data zip file
2. you want to unzip the data folder and place it in the folder where you have your python code for this lecture
3. if you are not sure how to write the path to your data folder, you have a function (inspired from here) to search for your file
4. once you have your location, you can use it to change your working directory
5. you have to be careful and might have to replace '\' signs with '/'
6. if you have special characters in your path, you might need to use "raw strings" and use the prefix r
    - so this could be r'/Users/avoid space and special characters in folder names pl)§se'

Another way to approach this is by using the Spyder working directory browser until you find the folder where your file is located:



7.2 Importing the full data set from INSEE

From the INSEE website here you can choose to import the table as a csv file. It will be called 'valeurs.zip'.
We prefer to work with comma-separated values file, for some of the reasons described here.
First unzip the file to have it a Valeurs.csv
Copy the Valeurs.csv file into your usual data folder.
Follow the indications section 7.1 to indicate where your file is located.

[12] pandas_exercise_part1_import.py

7.2.1 Import the csv file

If you open your file with a text editor[13], it looks like this:

Libell...;;D...mographie -Population au...
IdBank;;001641607
Ann...e;Mois
2016;9;66 790
2016;8;66 763
2016;7;66 735
2016;6;66 710

So because we are working with French data:

- they separate variables with ; instead of ,
- they use some special character as à

Hence we need to tell that:

- the values are separated by ';'
- it is encoded in latin1 style
- the frist two rows are not relevant for our study, so skip them

7.3 Cleaning the imported data set

The data set that has been imported and is stored into the DataFrame named df is not clean.

7.3.1 Reverse the row order

First we want to have an timely ascending series:

df=df.iloc[::-1]

7.3.2 Rename the columns

We want to rename the columns of the DataFrame with Month and Population.

7.3.3 Reset the index

We want to reset the index, knowing that the observations are monthly and start in January 1994:

7.3.4 Delete a column

Now that we have reset the index, we can drop the column with the months that is not bringing much information now, you drop vertically, a column, that is axis = 1.

[13] be carefule, if you open this file with Excel, do not save the modifications

7.4 Plot the DataFrame

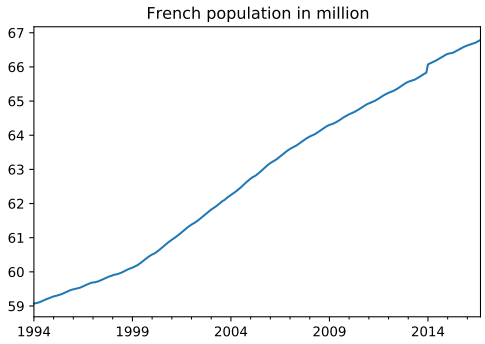
Now try to plot your DataFrame df, it doesn't work and tells you:

Empty 'DataFrame': no numeric data to plot

This means that when importing the data from the .csv file, python did not recognise that the column 'Population' were numeric. We now need to convert those data to numeric. First you need to remove the space in the numbers, then converting strings to float (note the use of regex[14]):

```
df=df.replace({' ': ''}, regex=True)
df=df.astype(float)
```

Now when you try to plot, you should get:



7.5 Resample our DataFrame (from monthly to yearly observations)

There is the possibility to resample our DataFrame from one frequency to another using the resample() function. We need to indicate the desired frequency described in the **Offset Aliases** section of the time series documentation available here. We list some of the more commonly used:

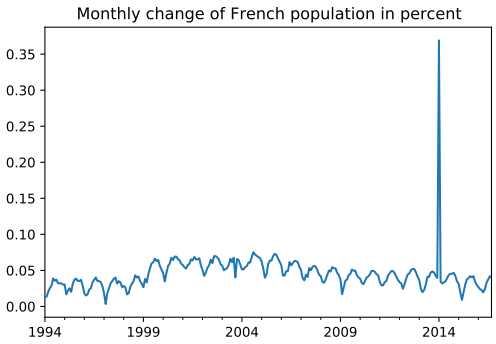
- H hourly frequency
- D calendar day frequency
- W weekly frequency
- Q quarter end frequency
- A year end frequency

The we need to chose a "method" for resampling, for example .sum() or .mean()
In our case, if we want to move to yearly frequency and get the average over the 12 months. If we plot the obtained DataFrame, we see that the plot is slightly smoother.

7.6 Compute monthly changes

Now you can plot and compute monthly changes.

[14] Without regex=True the method will look for exact matches. When you use regex=True it will look for the sub strings too



7.7 Descriptive statistics

You can get some descriptive statistics on the population and the monthly population changes:

	Population in million	Population growth rate in percentage
count	273	272
mean	62.78	0.05
std	2.44	0.02
min	59.07	0.00
max	66.79	0.37

7.8 Exercise with import, DataFrame and matrices: Leontief Input-Output model

Wassily Leontief popularized Input-Output model of the economy, started in his work Leontief (1949), his main published papers on the subject are compiled in a book Leontief (1986).
In this model, the economy is divided into measurable sectors and each sector, to be able to produce it output, requires input from other sectors. Hence, the production of each sector is determined by internal demand (from other sectors) and external demand (from final customers).
In matrix notation, writing P the vector of production of each N sectors $[p_1,\dots,p_N]$, D the vector of external demand for each sector $[d_1,\dots,d_N]$ and M the matrix of internal demand, where $m_{i,j}$ is the demand from sector i of production from sector j:

$$P = MP + D$$

This can be rearranged as:

$$(I - M)\,P = D$$

Then if $(I - M)$ is an invertible matrix, then this has a unique solution, and if this is non-invertible, then this can have none or infinitely many solutions.

7.8.1 Simplified example

Let's imagine we have the simplified economy:

	Agriculture	Manufacturing	Services	External demand, D
Agriculture	35	5	5	40
Manufacturing	5	60	20	60
Services	10	25	40	130
Total gross ouput	85	160	220	

We use the consumption matrix M after dividing by the Total gross output. We can now estimate P:

$$P = (I - M)^{-1} D$$

Now, if there is a one unit increase in the demand for agriculture output, this should have a positive impact of .1 on the manufacturing production.

7.8.2 Input-Output exercise

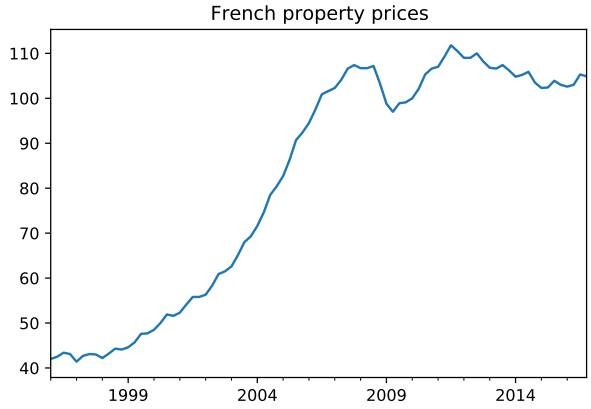
- **Question 2**
  - Import Input-Output Accounts Data from bea.gov website.
    - Use_SUT_Framework_2007_2012_DET.xlsx has been place on your data folder if need be.
  - Shape the matrix M and vector D
    - For the matrix M, keep columns up to T001 and index up to S00900.
    - Drop the element 4200ID, custom duties
    - Make the matrix M square (keep only elements common in both the index and the columns.
    - Consider that D is the total use of product, T019, minus the total intermediate T001.
    - Make M the consumption matrix, so divide each column j of M by element TIO_j of the Total industry output (basic value), T018.
  - Compute P
  - Now, imagine the external demand for Oilseed farming (1111A0) increases by one unit. What is the impact on the production for the Oil and gas extraction (211000)?
    - For this, create a D' filled with zeros and a 1 placed as 1111A0;
    - Now use M to compute the impact vector P': $P' = (I - M)^{-1} D'$
- **Question 3**
  - Perform a similar analysis for the France, using OECD data set
  - Now apply a lock-down like shock (negative) on "Accomodation and food services", what are the impact on the other sectos, which are the mostly impacted sectors?

8 Empirical data: loading files, using API, scraping websites

I would tend to say that before engaging in a long research project, you should explore whether some data will be available to test your model or not.
You will need to change the working directory to be able to load some files,previous section[15] explained step-by-step how to change your working directory.

8.1 Getting time series from Banque de France and INSEE

Banque de France provides historical times series via its webstat portal, here.
We might want to download the historical evolution of property prices in France, this can be found (with some more explanations) under the code RPP.Q.FR.N.ED.00.1.00 (Indices des prix des logements anciens, Ensemble des logements, France metropolitaine)



- **Question 4**
  - Update the property prices data set, do you see a lock-down effect?
    - Suggest a way to measure this effect.

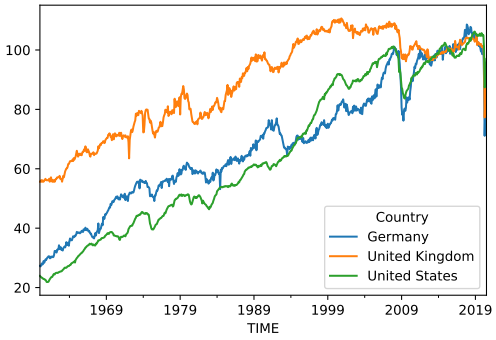
You can access INSEE data from their website. As an illustration, the Demography - Population at begining of a month in France, series ID: 000436387

- **Question 5**
  - Update the French population times series, do you see a pandemic effect?
    - Suggest a way to measure this effect.

8.2 Getting data from OECD and pivot-table

The OECD provides panel data via its websilte here.
You can select "index of Industrial Production". Then you'll need to reorganise the data set, making use of pivot table:

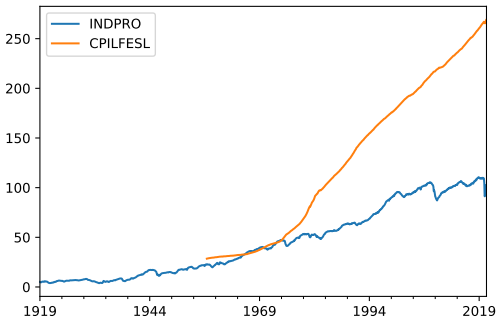
[15] get_data_vansteenberghe.py



8.3 Getting data from the Federal Reserve of Saint Louis - FRED

You can get data from the Federal Reserve of Saint Louis, their Federal Reserve Economics Database, or FRED.
What is convenient is that you can download the time series directly form python if you know the code of the variable you are looking for. For example, we can download the following variables for the US economy, which are the usual variable used in basic macroeconomic models:

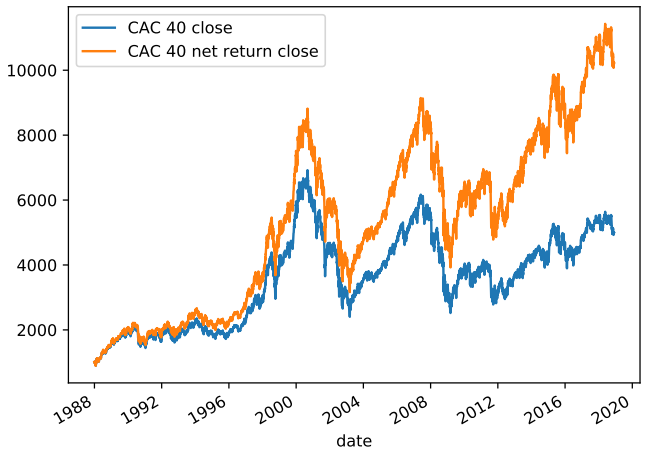
- Real GDP
- Industrial Production
- Core CPI
- Unemployment Rate
- 10 Year Yield
- 1 Year Yield
- Baa Yield
- Aaa Yield



- **Question 6**
  - You can try to reproduce the following with updated data:

8.4 Getting data from Euronext

Sometimes, it is possible to get data directly from providers like Euronext, where it is possible to get index historics here. This enables you for example the return of the CAC index with or without the dividends:



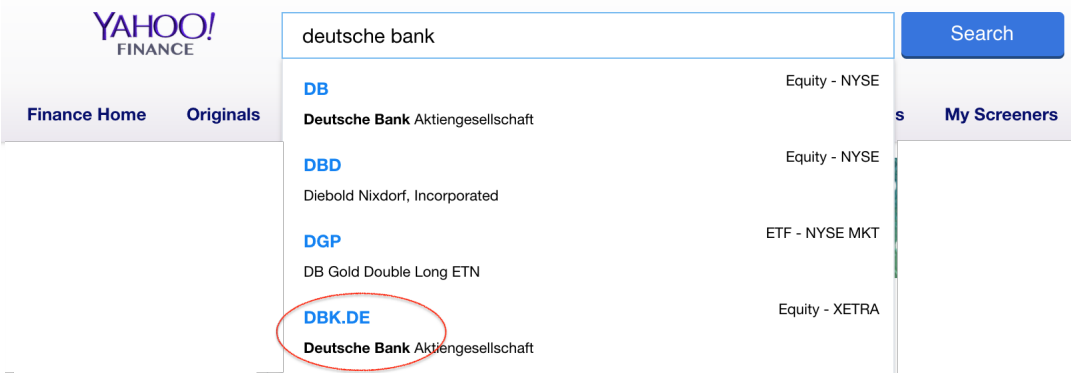
8.5 Installing and getting data from Quandl using an API

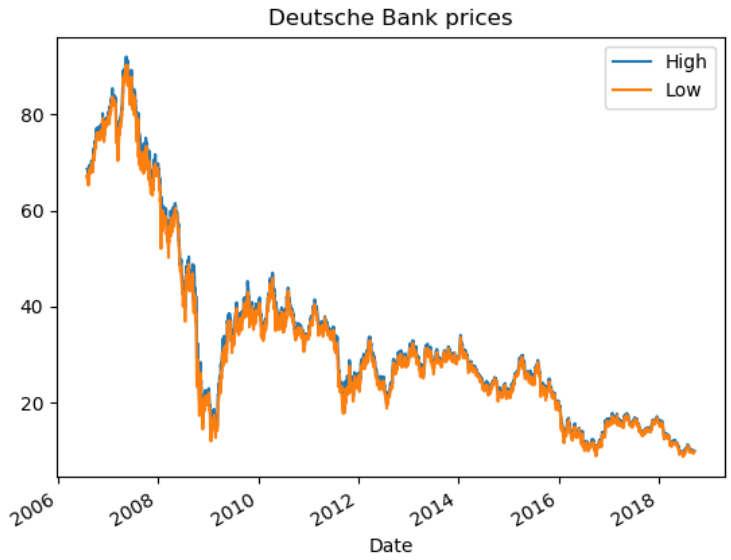
With Quandl, you can "get millions of financial and economic datasets from hundreds of publishers directly into Python". You can start by installing the Quandl library[16], then you can download datasets. You should request your API by creating an account. You can then download massive dataset, as explained here.

[16] if you are using Windows, you might have to use a Command Prompt: C:/Anaconda3/Scripts/pip install quandl

8.6 Getting data from Yahoo finance

In order to get financial time series from Yahoo finance, you first need to define the stock (or list of stocks) you are interested in. For this, you can search at the Yahoo finance URL, here for the name of the underlying company, let's say for illustration purpose Deutsche Bank:





You get historical time series on opening and closing prices, highs and lows as well as volumes.

8.7 Getting data from Bloomberg

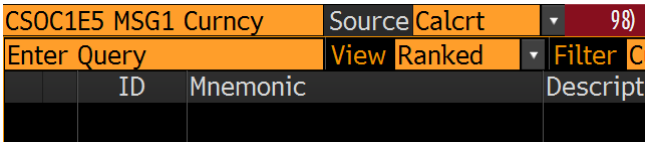
At the University, you have access to a Bloomberg terminal. We suggest that you install the Bloomberg extension for Excel (look into the start menu in window, you should find this installer).

You might want to download some historical time series with Excel. We suggest the following step:

8.7.1 Find the right field

You need to go onto the Bloomberg terminal in order to find the correct field: type FLDS

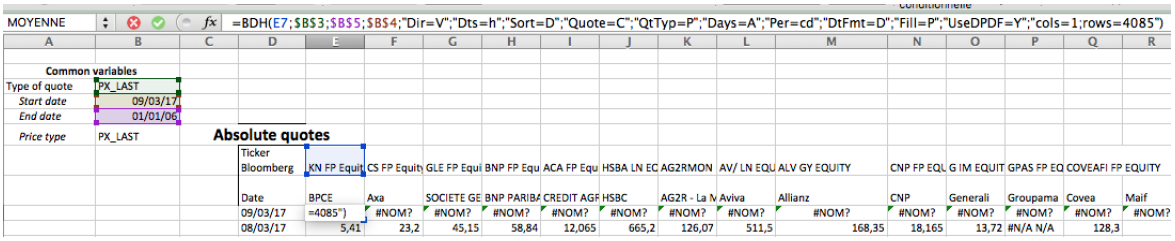
Then you can search for a company name for example, it will give you its Bloomberg ticker:



Once you get the Bloomberg ticker (e.g. "KN FP Equity" for the bank BPCE), you can use this ticker in an excel spreadsheet.

8.7.2 Download the time series in Excel

We suggest to organize your Excel sheet as follow:



Here is an simple example to download stock price data vertically:

```
=BDH(E7;$B$3;$B$5;$B$4;"Dir=V";"Dts=h";"Sort=D";"Quote=C";"QtTyp=P";"Days=A";"Per=cd";
"DtFmt=D";"Fill=P";"UseDPDF=Y";"cols=1;rows=4085")
```

8.8 Getting data from other sources

You can get data from the IMF here, from the ECB at its Statistical Data Warehouse. The World Bank provides time series on World Development Indicators. The Bank for International Settlements (BIS).

- **Question 7**
  - Use data from the BIS and reproduce the graphs in the article about the Anna Karenina Principle of global banking published in a research article. This exercise is for advanced pandas users and you should complete section 6 before doing this exercise.

8.9 Web scraping with python, a use case with IFP

We do not intend to provide here a full lecture on web scraping with python. We provide one use case: getting a list of Fintech that are IFP from ORIAS website.
Lending in France is regulated by law under what is known as monopole bancaire[17] and is regulated at the European level by the CRDIV[18] on taking deposits or other repayable funds and by the CRR[19] for granting credits for its own account. This regulation is a barrier to entry into the market but has

[17] Article L511-5, this monopoly is a mean to protect depositors from institutions insolvency and illiquidity
[18] Article 9-2 of the directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions
[19] Article 4-1 (1) of the regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013

been relaxed by the introduction of new categories of actors in France in 2014[20]. For the French market, the ACPR[21] has the authority to deliver accreditation[22] to intermediaries in participatory financing (IFP). ORIAS provides information on the IFP but to gather the information, one has to research and select the data. We suggest to use python libraries for web scraping to feed a data frame:

	Siren Number	ID	Name	Category	Postcode	City	oriasID	date in	date out
0	539015149	14006008	credit.fr	Crowdfunding Advisor (CIP) Crowdfunding Interm...	92300	Levallois-Perret	103995	2014-10-17	NaN
1	804606796	14006007	Lendopolis	Crowdfunding Advisor (CIP) Crowdfunding Interm...	75010	Paris 10e Arrondissement	104072	2014-10-17	NaN
2	798999983	14006009	Primus Finance	Crowdfunding Intermediary (IFP)	78110	Le Vésinet	103779	2014-10-17	NaN
3	803411982	14006571	Fidelfin	Crowdfunding Advisor (CIP) Crowdfunding Interm...	59510	HEM	104721	2014-11-07	NaN
4	750121311	14006337	Blue Project	Crowdfunding Intermediary (IFP)	75011	PARIS	104399	2014-11-07	NaN

This exercise was done with the help of a StackOverflow member (for more details on the problematic see StackOverflow website) that helped us overcome the CSFR restrictions.
We provide the code: 20190817_orias_scrape.py with little comments, this is for advanced python users, you might skip this part for now and come back on it later on.

- **Question 8**
  - After deciphering the code, apply the same approach to scrape data on Conseiller en investissement financier (CIF) from ORIAS website.

8.10 Data Assurance Quality

When working on Machine Learning projects and (econometrics too), you would soon realize that in fact data assurance quality (data cleaning, preparation, organisation) is in fact 80% of the workload. It is also an important step to perform because it will give you insight into the strength and weaknesses of the data you have.

For this section, we propose to work[23] on the data set of US Adult Census data relating income to social factors such as Age, Education, race, and other features base on the work of (Kohavi, 1996). The data set can be found here.

The first task is to get an overview and describe the data set and spot the missing values:

[20] Ordonnance n° 2014-559 of the 30th May 2014 which is a Dérogation au monopole bancaire introducing the role of conseillers en investissements participatifs and intermédiaires en financement participative

[21] Autorité de contrôle prudentiel et de résolution, the French financial regulator

[22] Ordonnance n° 2014-559 of the 30th May 2014 article 16

[23] ML_US_wages.py

	0
Feature	Missing values count
age	0
workclass	-1836
fnlwgt	0
education	0
education_num	0
marital_status	0
occupation	-1843
relationship	0
race	0
gender	0
capital_gain	0
capital_loss	0
hours_per_week	0
native_country	-583
income_bracket	0

We check whether the NaNs are missing values because the people are not working or if it is incorrectly filled values. In the second case, we will want to remove the rows with those NaNs.

We also look at native countries, and find that there is one named "South" which is not easy to sort a priori. Cross-checking with "race", it is very likely to correspond to South Korea for the most part.

We can also decide to create some new variables, combining capital gains and losses into one, or sorting the marital status between people effectively living alone or in couple.

- **Question 9**
  - Perform:
    1. Cluster with K-means, Hierarchical clustering
    2. Plot and set a threshold for the Dendogram
    3. Apply a logit model and check if you overfit the data with a Lasso logit
    4. Random Forest
    5. Boosting

9 Python: Entertaining application - Perudo game

To apply the skills developed so far, we suggest to play Liar's Dice and play the Perudo[24] version and take the rules from perudo.com, but with some twists to make it simple:

> The object of perudo is to be the last player with a die or more.
>
> Perudo is played in rounds. Each player receives a cup and five dice. Each round begins by all players rolling their dice around in the cup. After shaking the dice, players turn the cups over on a table top, so that the dice are rolled and under the cups. Each player may peek in his own cup.
>
> Players bid, guessing at the number of rolls. When a player believes that another player has over-estimated, they say Dudo, which means "I doubt" in Spanish.
>
> The first player announces a number, and then the next player has the choice of doubting it, by saying Dudo or raising the bid, either by the number of dice or by the value of the dice (or by doing both). For example, if player one bid three twos, then player two could bid three threes, four twos, four fours, or even ten sixes.

The next player can also announce "calza" which means that he believes that the announcement of the player just before him was the correct guess. In fact, any player can call "calza" at any time even on his own announcement. If the player calling "calza" is correct, then we give him another dice, but if he is wrong then he loses one of his dice.

> After player 2 bids, play goes on to the left.
>
> If a player calls Dudo, and is correct, then the player must show the dice, and each player must show their dice to verify whether the number was indeed too high.
>
> If there are enough dice of that number, then the player who called Dudo must place a die in the discard pile. If there are not, then the player who made the last bid must place a die in the middle. In either case, a new round begins.
>
> The player who lost a die in the last round is the first player in the new round. If the player lost his last die, then the player to his left plays first instead.
>
> [...]
>
> The last player left with at least one die wins the game. ("Perudo" is a registered trademark of University Games Corporation, Burlingame, CA.)

9.1 Should you call Calza, raise the bet or call Dudo?

9.1.1 If you don't use the information provided by the players

Let's imagine you have 6 players, this means that you start with 30 dices (ndices). You want to know the most likely number of similar dices. We will do this by a Monte Carlo type approach: we do a lot of random "tirages". So a dice value can be in $[1,6]$. But remember that an ace can take any value between two and six (this is like a joker)!
We find that the average count is 10. But then what about the chance of calling a "Dudo"?
A Dudo is correctly called when the count is lower than the announcement. So we accumulate all the counts that makes the Dudo call correct. For this configuration, when you predecessor announce 13, then if you call Dudo you have more than 80% chance to be correct.

[24] code: vansteenberghe_perudo.py

But in fact, the players before your predecessor gave you information (that could be bluff), so this changes the odds!

9.1.2 If you use the information provided by the players and your dices

Now before you speak, several players did before you, but be careful, they might be bluffing. Let's imagine that you have three dices of 2, one ace and so far two players spoke and told the audience 3 twos and then 12 twos. If you say 13 twos, are the odds still at 80% against you?

- **Question 10**
  - Now consider that you are a player and so you know the values of your 5 dices, this changes the odds, you randomize the other players dices only.
  - Also consider the 5 other players with 5 dices. Some of those other players spoke befor you have to speak. Assign for each player a probability to "bluff".
    - As an illustration, if a player calls "there are 13 twos". If he is not bluffing, then he believes that there are round(25 dices / 3) so 8 twos in the 25 other dices and you can deduct that this player has 13-8 = 5 twos in his hidden hand (or some twos and some aces to be correct). This is highly unlikely, but this player can take into account what two players just before him said, maybe the players before him called "11 twos" then "12 twos" so it can be believed that those two players had 5 twos in their 10 dices and they both believed that the others (except them both) had round(20 dices / 3) = 6 twos... Well, the point is that guesses are iterative.
    - You have to be conservative and believe that sometimes players are too optimistic. So maybe a player that announces "there are 13 twos" actually is bluffing and only has 3 twos in his hidden hand... To take this into account, you can use a random variable taking value between $[0,1]$ that will diminish your estimation of dices based on players bet.

9.2 Illustration: rational and truthful players

For simplicity, let's imagine that the game is slightly different:

- 5 players: A, B, C,D and you (Y);
- each player has 6 dices (this will make odds computation much easier as round);
- all players play rationally and truthfully.

So we still have a game with 30 dices. Let's imagine player A is the first to speak:

$$\begin{bmatrix} A \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} \begin{bmatrix} B \\ \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} Y \\ \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} C \\ \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} D \\ \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix}$$

A has to bet first. He could chose to bet on 4's. He has two fours in his hand, but he has no information on the other 24 dices. So $E_A[\text{count}(4)] = \frac{24}{3} + 2 = 10$.

Player A bets: "There are 10 4's".

Now it is players B's turn to play. He takes into account what A said and as A played rationally and truthfully, he knows that player A has two 4's. In his hands, B has three 4's, so $E_B[\text{count}(4)] = \frac{18}{3} + 2 + 3 = 11$

A	B	Y	C	D
1 or 4	1	?	?	?
1 or 4	2	?	?	?
2, 3, 5 or 6	4	?	?	?
2, 3, 5 or 6	4	?	?	?
2, 3, 5 or 6	5	?	?	?
2, 3, 5 or 6	6	?	?	?

Player B bets: "There are 11 4's".

Now it is your turn to play. A and B being rational and truthful, you believe that they have 2 and 3 fours. You have 2 fours in your hand so you guess: $E_Y[\text{count}(4)] = 2+3+2+\frac{12}{3} = 11$.

A	B	Y	C	D
1 or 4	1 or 4	1	?	?
1 or 4	1 or 4	2	?	?
2, 3, 5 or 6	1 or 4	3	?	?
2, 3, 5 or 6	2, 3, 5 or 6	4	?	?
2, 3, 5 or 6	2, 3, 5 or 6	5	?	?
2, 3, 5 or 6	2, 3, 5 or 6	6	?	?

You want to call Calza and believe that there are 11 4's around the table.

9.3 Back to human players

Let's go back to the actual game

- 6 players: A, B, C,D, E and you (Y);
- each player has 5 dices.

So we have a game with 30 dices. Let's imagine player A is the first to speak:

A	B	C	Y	D	E
1	?	?	?	?	?
2	?	?	?	?	?
3	?	?	?	?	?
4	?	?	?	?	?
5	?	?	?	?	?

A has to bet first. He could chose to bet on 4's. He has two fours in his hand, but he has no information on the other 25 dices. So $E_A[\text{count}(4)] = \frac{25}{3} + 2 = 10.3$. Player A has no incentive to bet too high.

Player A bets: "There are 10 4's".

Now, the situation for player B is trickier:

$$\begin{bmatrix} A \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} B \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} C \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} Y \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} D \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} E \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix}$$

Player B knows that player A can bluff, so he is not sure whether player A has 1 or 2 4's. Let's imagine player B believe that player A has 2 4's: $E_B^A[\text{count}(4)] = 2$. Then: $E_B[\text{count}(4)] = 2+2+\frac{20}{3} = 10.6$. If B believes that A has only one 4: $E_B[\text{count}(4)] = 9.6$ he should call a Calza on A's bet.
It can make sense for B to bet 11 4's:

Player B bets: "There are 11 4's".

$$\begin{bmatrix} A \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} B \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} C \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} Y \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} D \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} E \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix}$$

Player C has two 4's and player A and B spoke before him. Let's imagine that C believe that A and B are extra conservative and so he believes that player B only bets 11 if the odds were 11.6 (so if B had 3 4's):

$$E_C[\text{count}(4)] = E_C^A[\text{count}(4)] + E_C^B[\text{count}(4)] + E_C^C[\text{count}(4)] + E_C^{Y,D,E}[\text{count}(4)]$$

$$E_C[\text{count}(4)] = 2+3+2+\frac{15}{3} = 12$$

Player C bets: "There are 12 4's".

It is your turn to play Y:

$$\begin{bmatrix} A \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} B \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} C \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} Y \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} D \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} E \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix}$$

$$E_Y[\text{count}(4)] = E_Y^A[\text{count}(4)] + E_Y^B[\text{count}(4)] + E_Y^C[\text{count}(4)] + E_Y^Y[\text{count}(4)] + E_Y^{D,E}[\text{count}(4)]$$

$$E_Y[\text{count}(4)] = \left(10-\frac{25}{3}\right) + \left(11-\left(10-\frac{25}{3}\right)-\frac{20}{3}\right) + \left(12-\left(11-\left(10-\frac{25}{3}\right)-\frac{20}{3}\right)-\frac{15}{3}\right) + 2 + \frac{10}{3}$$

$$E_Y\left[\text{count}(4)\right] = 1.6 + 2.6 + 4.3 + 2 + 3.3 = 14$$

So now B's boldness inflated the expectation and you might be tempted to bet 13 and even 14 4's. Whereas in fact, so far 8 4's are present out of 20 dices and so you would need to have six 4's out of the last 10 dices of player D and E for you bet to be correct, rather unlikely:

$$\begin{bmatrix} A \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} B \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} C \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} Y \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} D \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} E \\ \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix}$$

You could be conservative about players' bets and introduce some discount (let's call it bluff discount $\delta^i \in [0,1]$) for each player, e.g. as you know player B to be bold, $\delta^B = 0.5$ (let's keep $\delta^A = \delta^C = 1$ for simplicity here):

$$E_Y\left[\text{count}(4)\right] = \delta^A 1.6 + \delta^B 2.6 + \delta^C 4.3 + 2 + 3.3 = 12.6$$

Now, you might want to either call Calza on 12 4's or raise to 13 4's maximum. This is still too high, as you can see that we have 8 4's and for the last two players (D and E), we can expect to have $\frac{10}{3} = 3.3$ so 11.3 4's. Therefore the calibration of bluff discounts is important (but not trivial).

10 Python for non-programmers: numpy exercise part 2

We suggest a code[25] with further exercises to get familiar with python's numpy library, we will approximate functions, their derivatives and integrals.

10.1 Approximating the exponential function

One last visual application: we might want to convince ourselves that:

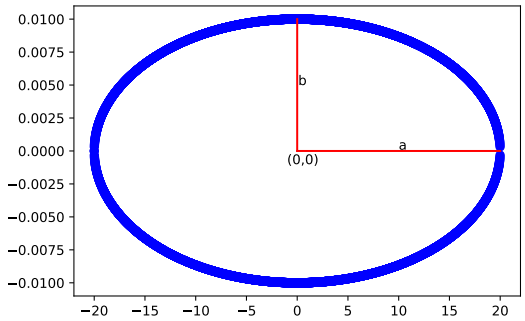
$$\exp(x) = \lim_{n\to\infty}\left(\left(1+\frac{x}{n}\right)^n\right)$$

For this, we can simply plot the exponential function against $\left(1+\frac{x}{n}\right)^n$ for different values of n, for example $[10, 100, 1000]$.
Then finally, we can convince ourselves visually that:

$$\exp(i\pi) = -1$$

10.2 Compute the area of an ellipse

We draw an ellipse:



The equation of such a standard ellipse centred at the origin, with height 2b and width 2a is:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{5}$$

It can be shown that the area inside the ellipse is: πab
As an exercise we will compute an approximate of this area by integral. We reorganize equation 5:

$$y(x) = b\sqrt{1-\frac{x^2}{a^2}} \tag{6}$$

We can compute the area of the semi-ellipse (above part) by:

$$\int_{-a}^{a} b\sqrt{1-\frac{x^2}{a^2}}\,dx \tag{7}$$

[25] numpy_exercise_part2.py

10.2.1 Riemann integral

We want to approximate integral 7 on the interval $[-a, a]$. We can use the lower and upper Darboux sums which is a simple approach in the spirit of the Riemann integral:

1. partition the interval $[-a, a]$ into $-a = x_0 < x_1 < x_2 < \ldots < x_n = a$
   - here for simplicity we evenly split the interval into n intervals
2. define the lower Darboux sum:

$$\sum_{i=0}^{n-1} \inf_{x \in [x_i, x_{i+1}]} y(x)(x_{i+1} - x_i) \tag{8}$$

3. define the upper Darboux sum:

$$\sum_{i=0}^{n-1} \sup_{x \in [x_i, x_{i+1}]} y(x)(x_{i+1} - x_i) \tag{9}$$

Those two sums, as $y(x)$ is continuous, converge to the integral we want to compute.

- **Question 11**
  - Work on Kepler's laws.

10.3 Mittag-Leffler function

A Generalized Mittag-Leffler function can be defined as:

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)} \tag{10}$$

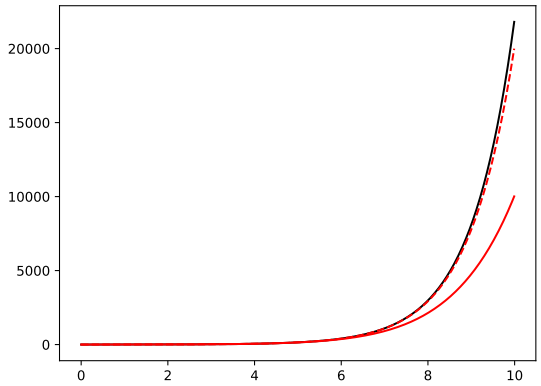
with $\alpha, \beta \in \mathbb{C}, \mathscr{R}(\alpha) > 0, \mathscr{R}(\beta) > 0, z \in \mathbb{C}$ and the function Gamma: $\Gamma(r) = \int_0^{+\infty} u^{r-1} e^{-u} du$.

10.3.1 Mittag-Leffler function to approximate the exponential function

Equation 10 can be set to approximate the exponential function:

$$E_{1,1}(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}$$

We have in practice to replace the sum to a limit $nlim$, instead of ∞, that we set to 10 (red line) and 15 (dashed red line) for illustration purpose:



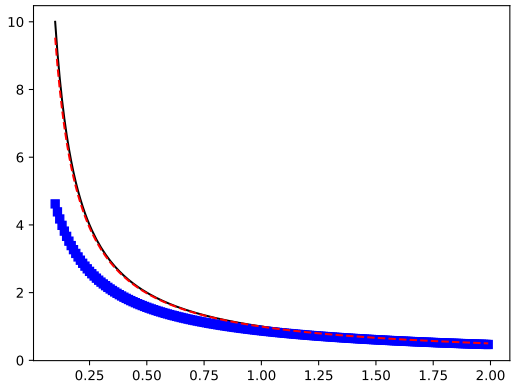
For example, Mittag-Leffler type functions combined with Laplace transform are useful to find explicit expressions of insurance companies ruin probability as in (Constantinescu et al., 2018)

10.4 Approximate the derivative of a function

Remember that the derivative $f'(x)$ of a function $f()$ at x is in fact a rate of change defined as:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

We can convince ourselves that $\ln'(x) = \frac{1}{x}$
We can set our own function to approximate the derivative and check for decreasing values of h if indeed our results are fitting with the function $\frac{1}{x}$, which as expected is the case.



10.5 Special case: approximate the integral of the logarithm

We might want to test with python that:

$$\int_1^2 \ln(x)dx = \ln(2) - \ln(1) = \ln(2)$$

And we can approximate:

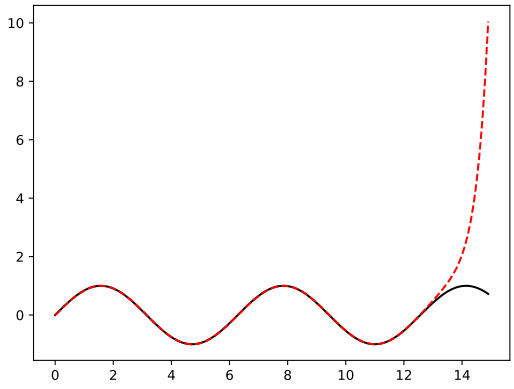
$$\int_1^2 \frac{1}{x}dx = \sum_{i=0}^{N} \gamma(1+h_i)^{-1}$$

with $\forall i > 0, h_i - h_{i-1} = \gamma$ and $h_N = 1$.

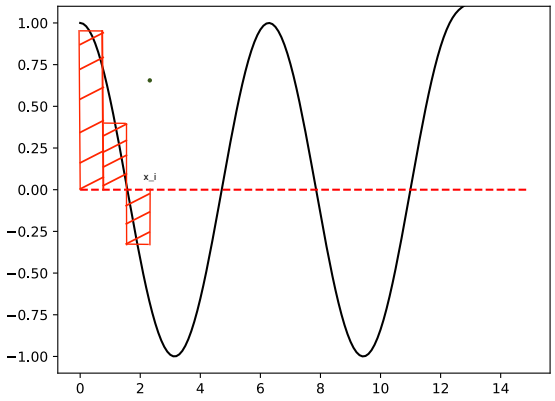
10.6 Approximate the sinus function with Mittag-Leffler integral

Finally, we can approximate the sinus function with an integral of a Mittag-Leffler function:

$$sin(x) = \int_0^x E_{2,1}(-s^2)ds$$



To decompose, in black we draw the $E_{2,1}(x)$ function. To compute $sin(x_i)$ we create intermediary points (3 here) from 0 to x_i and we add (or subract when negative) the area of the rectangles. This then gives us $sin(x_i)$ represented by a green dot:



This has to be done for every x_i between 0 and 14 with a chosen step number (the more steps the more precise but the longer it will take to compute).

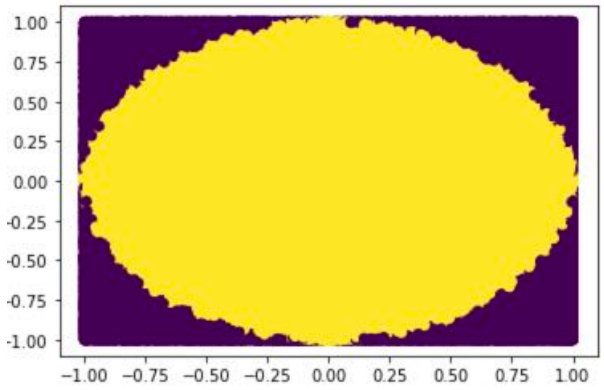
- **Question 12**
  - Show visually (plot) that $E_{1,2}(x) = \frac{e^x - 1}{x}$.

10.7 Monte Carlo method and uniform distribution: estimate the value of π

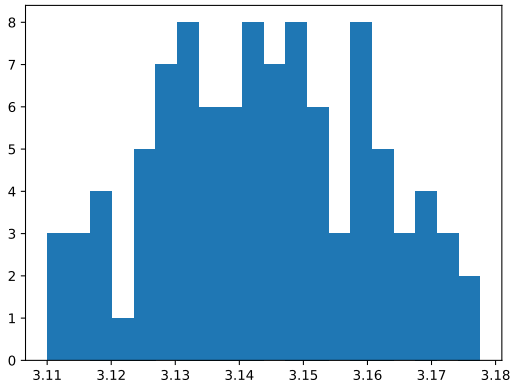
The main idea behind a Monte Carlo approach is to repeat N times random sampling in order to estimate some coefficient. In this simple example, we wish to approximate the value of π.
We define a circle of radius $r = 1$. Then its surface is known to be πr^2. We draw x and y from the uniform distributions $a = -1$ and $b = 1$. We write the probability density function of the uniform distribution:

$$f(x;a,b) = \begin{cases} \frac{1}{b-a}, & \text{for} \quad a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

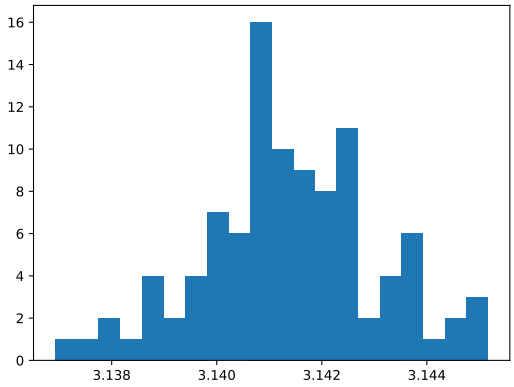
The square surface will be $2 \times 2 = 4$ and the area enclosed by the circle is π:



A recurring question is on the choice of the number of draws for the Monte Carlo method. We are not aware of a definite answer, but here is an illustration of the histogram of 100 estimates of π we found with 10^4 draws:



compared with 100 π estimates we found with 10^6 draws:



- **Question 13**
  - Create a list of number of draws $N = [10, 10^2, 10^3, \ldots, 10^n]$, chose n depending on your laptop computing power and create a list of the standard deviation of the error terms of the π estimates compared to the numpy π value for 100 estimates, is it a monotonic relationship? Can you advise for a value of n, the sample size?

10.8 Monte Carlo integration: estimating the integral of the exponential function

We want to estimate the integral $I = \int_0^1 e^x dx$. We know that the expected value is: $I = e - 1$. The Monte Carlo integration method[26] consists in choosing random points at which the integrand will be evaluated. The volume of the subset $[0, 1]$ is $V = \int_0^1 dx = 1$. The naive Monte Carlo approach is to draw from the uniform distribution as in equation 11 and set $a = 0$ and $b = 1$ to build a sample of size Ns, $x_1, \ldots, x_{Ns}$. With the law of large numbers it can be demonstrated that:

$$\lim_{Ns \to \infty} V \frac{1}{Ns} \sum_{i=1}^{Ns} e^{x_i} = I \tag{12}$$

The variance of the estimation is $V^2 \frac{Var(exp)}{Ns}$ and the errors can be estimated as the standard deviation, so it decreases with rate $\frac{1}{\sqrt{Ns}}$.

- **Question 14**
  - Plot the Monte Carlo integration errors as a function of the sample size to demonstrate the rate of decrease is as expected.

So to formalize what we did in section 10.7, we used the function:

$$H(x, y) = \begin{cases} 1, & \text{if} \quad x^2 + y^2 = 1 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

The volume of the subset $[-1, 1] \times [-1, 1]$ is $V = \int_{-1}^1 \int_{-1}^1 dx dy = 4$ and we computed the approximation

$$V \frac{1}{Ns} \sum_{i=1}^{Ns} H(x_i, y_i)$$

[26] some more elements here

11 Python : Population, samples, parametric distribution, Central Limit Theorem and outlier detection

In the following we will study the probability laws governing random variables: their distributions. We are interested in their probability density function $P(X = x)$ and their cumulative distribution function $P(X \leq x)$.

11.1 Concepts of Data-Generating Process and iid

For any data set, as an empiricist, we are making assumptions. The strongest and least credible assumptions we make are on the data-generating process (DGP)[27] where we assume that there is an underlying mechanism that is generating the observations under study. In our example, this means that we assume that the CAC 40 daily returns are generated by a process, we will introduce possible distributions in the following sections. We further initially assume that the return at date t is drawn from the DGP independently than the outcome at date $t-1$, but for time series we will need later to tackle the fact that this hypothesis doesn't hold.
Further, when applying an ordinary least square method to estimate a linear regression, we will assume (and test) that the residuals have mean 0, were generated by the same underlying distribution, and are independent of each other. For the last two conditions, we say that the residuals are independent and identically distributed. This is often violated: residuals are serially correlated or their variance differ for some observations (heteroskedasticity).

11.1.1 In a data rich world, should we drop our a priori?

(Armatte et al., 2017) review the work of Edmond Malinvaud and his position on "*A Priorism versus Empiricism*":

> Is economic science inductive? Does it begin from observation? Is it not simply a theoretical structure whose coincidence with reality is doubtful?" He answers them in the following manner: "yes, economic science has become inductive; the knowledge of economic phenomena is developed in the same way as that of physical phenomena, that is, through a continual back and forth [va-et-vient] between observation and theory; given the particular conditions in which it operates, economics employs an inductive approach which is, in part, original, and thanks to which it progressively ensures a better correspondence with reality.

he stresses the

> three steps which were preconditions to the progress of an inductive economic science: the accumulation of data, the elaboration of descriptive syntheses such as series of indices and national accounting, the diffusion of a method use to establish laws on the basis of observed data, namely, econometrics. [...] What naivety, is there not, in supposing, as I have seen it done repeatedly, that the abundance of data, certainly a necessary condition, is also sufficient for understanding phenomena and for good forecasting.

And he gives this recommendation:

[27] Davidson et al. (1978) consider that "one could characterise 'econometric modelling' as an attempt to match the hypothetical data generation process postulated by economic theory with the main properties of the observed data."

> In order to make the best use of the observations at his disposal, the economist must first draw all the consequence of what he already knows, that is, construct a model expressing his initial knowledge, in terms of which he can circumscribe properly the supplementary information to be gleaned from the data. Effective research demands this type of approach. To advance on the path of knowledge, the economist can and must first proceed with a deductive analysis of the potential causes of phenomena. The model is precisely the result of this deductive analysis. [...] To gain knowledge of reality, the economist must continually rationalize and refine his theoretical constructions as he deciphers the information contained in the increasingly abundant data at his disposal.

11.2 Distribution function, mean, variance

We follow section 8.4 and use the history of CAC 40 index including dividends (net returns)[28]. Now if as an investor I believe that the history or daily returns are representative of the likely upcoming daily returns, I might want to know to likelihood of a future daily return of 0.05 (i.e. 5%). I can approach this question two ways:

1. I build an empirical histogram
2. I fit a parametric[29] distribution (normal, student, etc.) on the empirical data

11.2.1 Empirical histogram

If we build an empirical histogram, we have to chose the granularity of the bins. If doing it manually, we can start by limiting our representation to 2 bins: $]-\infty,0]$ and $]0,+\infty[$. In our representation, we have:

	neg	pos
count	3762	4072

We can normalize this so that the sum of the bins is equal to 1 (100%):

	neg	pos
count	0.48	0.52

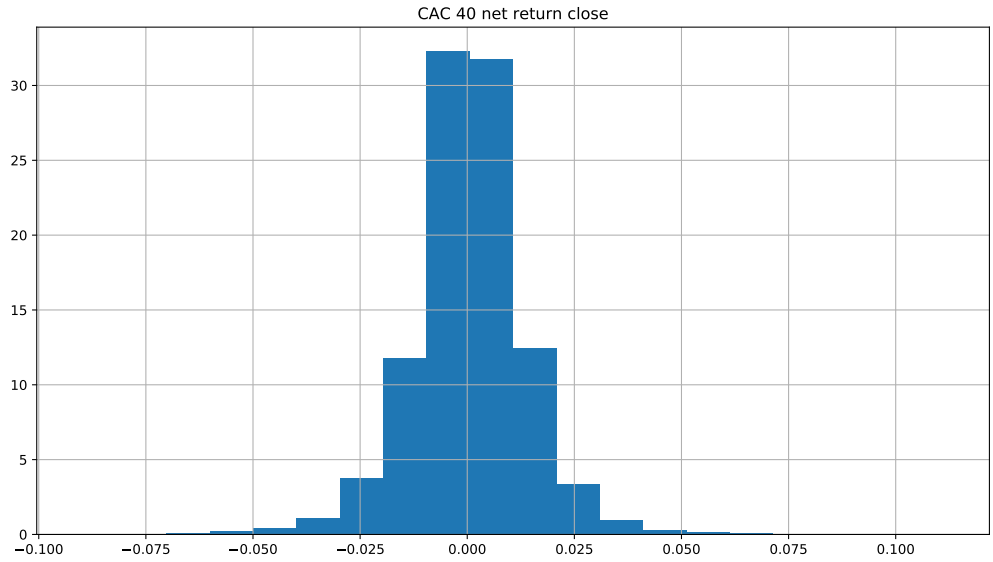
In other words, empirically, I have 52% of probability to have a positive daily return[30].

We can use directly the python command and chose both the number of bins (20 in this example) and whether we normalize the histogram or not:

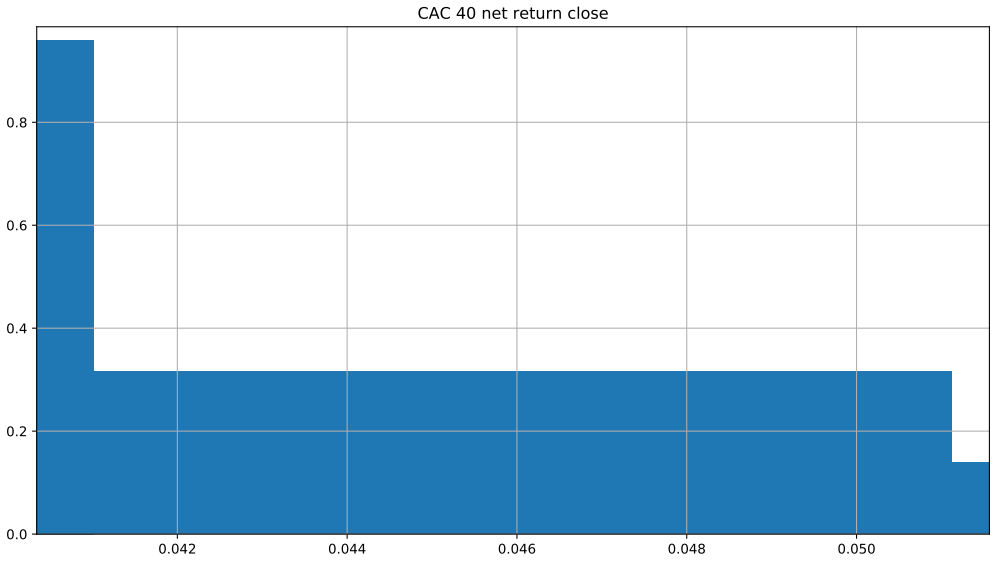
[28] QMF_exercise_samples.py

[29] usually, the density function is not known, here we select an arbitrary function and determine the appropriate parameter for this function to describe the distribution of the sample

[30] before concluding, we would need to perform statistical test to see if this is indeed statistically different than 50%



So to answer our initial question on the likelihood to have a daily return of 4.5%, we can zoom in:



We visually observe a probability of 0.3% which we confirm by counting how many daily return fell inside the bucket $[0.04, 0.05]$ of size 0.01 over the total count of daily return: 0.38%.

11.3 Parametric density estimation

We will present various parametric distributions in the coming sections[31], we need to believe that it is possible to fit a "function" $P(X = x)$ that gives us the probability that the daily return is x. Note that the goodness of fit of that function will be relevant if we want to trust the answer on the probability, for this there exist various tests, one of which is presented section 12.1.3.

11.3.1 Mean

We define the mean as:

$$E(X) = \int_{-\infty}^{\infty} xP(X = x)dx$$

in our case it is 0.038%.

11.3.2 Standard Deviation

We define the standard deviation as:

$$s(X) = \sqrt{\int_{-\infty}^{\infty} [x - E(X)]^2 P(X = x)dx}$$

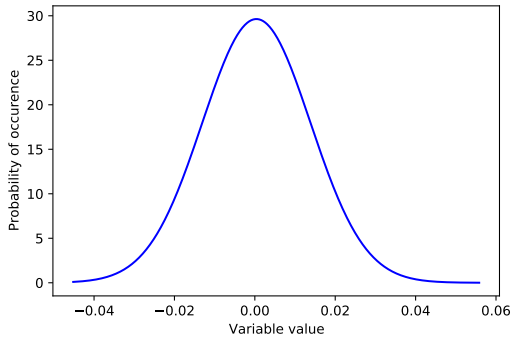
in our case we find 0.013

11.4 Normal distribution

We introduce the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Its probability density function is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

in simple words this function gives us the probability that a random variable X takes the value x if the random variable follow the normal distribution $\mathcal{N}(\mu, \sigma^2)$. For illustration, we take the parameter μ and σ as the empirical mean and standard deviation observed on the CAC 40 daily returns:



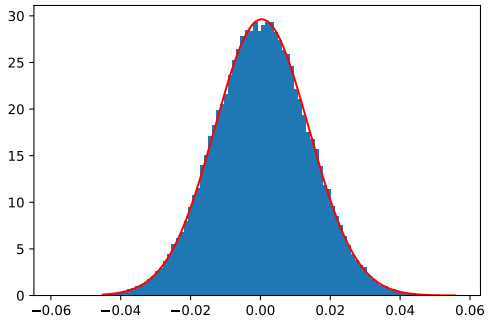
The mean or first moment of the distribution can be computed following section 11.3.1:

$$\mu^1 = \int_{-\infty}^{+\infty} xf(x; \mu, \sigma)dx = \mu \tag{14}$$

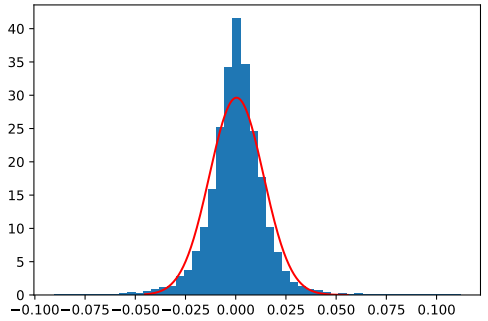
[31] all of those are unimodal (that is, have a single local maximum), meaning that if you work with a multimodal distributed sample you will need to use other technics e.g. KDE we cover in section **??**

11.4.1 Random draws

We can draw randomly from a normal distribution (let's say 10^4 draws) and then plot the histogram against the probability density:



The empirical data doesn't fit the theoretical normal distribution, visually you can observe fatter tails:



11.4.2 Introduction to Maximum Likelihood Estimation

To estimate the parameters μ and σ we can use the Maximum Likelihood Estimation. The probability density function for a sample of n independent identically distributed normal random variables is the likelihood we want to maximize:

$$L(\mu,\sigma) = f(x_1,\ldots,x_n;\mu,\sigma) = \prod_{i=1}^{n} f(x_i;\mu,\sigma)$$

As the logarithm function is continuous strictly increasing and it is easier to deal with the log-likelihood, we maximize:

$$\log\big(L(\mu,\sigma)\big) = -\frac{n}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

The first order condition yields:

$$\hat{\mu} = \sum_{i=1}^{n}\frac{x_i}{n}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

It can be demonstrated that the estimator of the variance is biased:

$$E\big(\hat{\sigma}^2\big) = \frac{n-1}{n}\sigma^2$$

11.4.3 Estimate the probability of occurence of a given return

With our DGP and our empirical observations, we can answer the previous question for the probability to see a daily return of 4.5%, or more specifically, the probability that our CAC 40 daily return will fall between 4 and 5%:

- from empirical distribution, the probability is 0.38%;
- from a parametric normal distribution, the theoretical probability is 0.12% (and we estimated it at 0.15% with a Monte Carlo method).

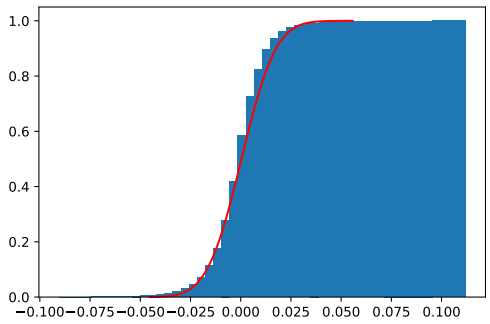
With a normal distribution we tend to underestimate the probability of having absolute large daily returns.

11.5 Cumulative distribution function

The cumulative distribution function (cdf) answers the question: "What is the probability that my returns are below or equal to x?"

$$P(X \leq x) = \int_{-\infty}^{x} P(X)dX$$

We can apply a normal law cdf to the CAC 40 daily return:



With the empirical distribution function, for the cumulative up to 4.5% daily return, we obtain 99.6%, but for the normal distribution we find 99.95% where again with the normal distribution we underestimate the probability of having daily returns exceeding 4.5%: in practice we had 0.4% of the daily returns above 4.5% but the normal distribution function estimates that probability at 0.05%.

- **Question 15**
  - Find the return x_i so that at this return, the probability to have daily return below is 0.4%.
    - We'll see section 18.1, that you are then searching for a Value at Risk (VaR).

11.6 What is the mean of the absolute of a normally distributed random variable?

Now let's imagine that we have a random variable X that follows $\mathcal{N}(0, \sigma^2)$, we might wonder the mean of the $|X|$ random variable, $E(|X|)$. Writing $x = \sigma Z$ where conveniently $Z \sim \mathcal{N}(0, 1)$, we have

$E(|X|) = \sigma E(|Z|)$. And:

$$E(Z) = \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi} e^{-\frac{x^2}{2}}} dx$$

$$E(|Z|) = \frac{2}{\sqrt{2\pi}} \int_{0}^{+\infty} x e^{-\frac{x^2}{2}} dx$$

$$E(|Z|) = \sqrt{\frac{2}{\pi}} \left[-e^{-\frac{t^2}{2}} \right]_{O}^{\infty} = \sqrt{\frac{2}{\pi}}$$

Finally, $E(|X|) = \sigma \sqrt{\frac{2}{\pi}}$
We can verify this by drawing a sample and computing the formula to compare.

11.7 What is the largest observation you expect to have in a sample?

Let's imagine that we have a random variable X that follows $\mathcal{N}(\mu, \sigma^2)$. If you have a sample of size N, what is the largest value you expect to observe in this sample?

- **Question 16**
  - Show that this is close to the formula[32]

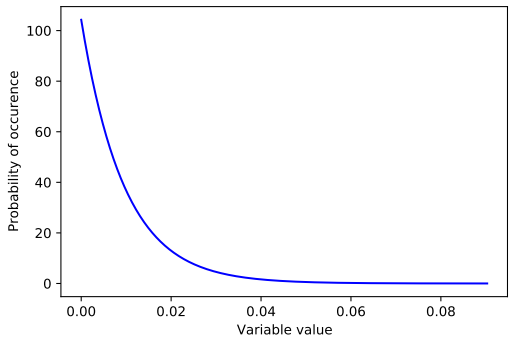
$$\mu + \sigma \sqrt{2 \log N} - .5\sigma \frac{\log\log N}{\sqrt{2 \log N}}$$

11.8 Exponential distribution

We write the probability density function of the exponential distribution:

$$f(x; \alpha) = \begin{cases} \alpha e^{-\alpha x}, & \text{for} \quad x \geq 0 \\ 0 & \text{for} \quad x < 0 \end{cases}$$

For illustration here the exponential probability density function:



The likelihood function on a sample of size n is:

$$L(x_1, \ldots, x_n | \alpha) = \alpha^n \exp\left(-\alpha \sum_{i=1}^{n} x_i \right)$$

[32] inspired by Jon Keating's lectures

Deriving this expression with respect to α to find an optimum α_o:

$$n\alpha_o^{n-1}\exp\left(-\alpha_o\sum_{i=1}^{n}x_i\right)-\sum_{i=1}^{n}x_i\alpha_o^n\exp\left(-\alpha_o\sum_{i=1}^{n}x_i\right)=0$$

an optimum is thus find for the reciprocal of the sample mean:

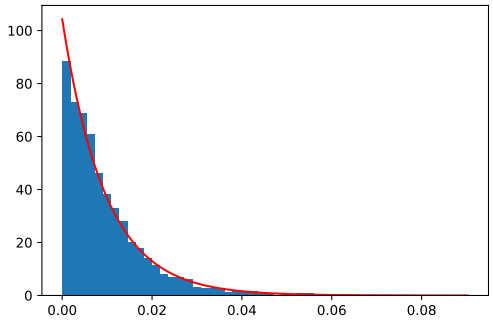
$$\alpha_o=\frac{n}{\sum_{i=1}^{n}x_i}$$

For the exponential law the first moment (the mean) is:

$$\mu^1=\alpha\int_0^{+\infty}xe^{-\alpha x}dx=\alpha\left[e^{-\alpha x}\left(\frac{-\alpha x-1}{\alpha^2}\right)\right]_0^{\infty}=\frac{1}{\alpha}$$

and we can demonstrate that the variance of the exponential law is $\frac{1}{\alpha^2}$.

We can fit this to our negative CAC 40 return (we need to take the absolute values to be able to fit our exponential distribution):



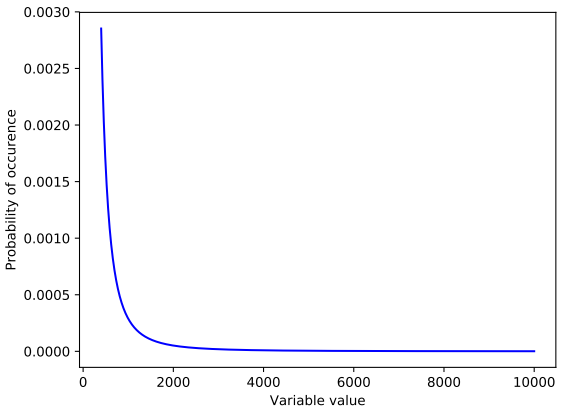
Typically, in insurance claims modelling (non-life), we model the magnitudes of claims as an exponential law of parameter α.

11.9 Pareto distribution

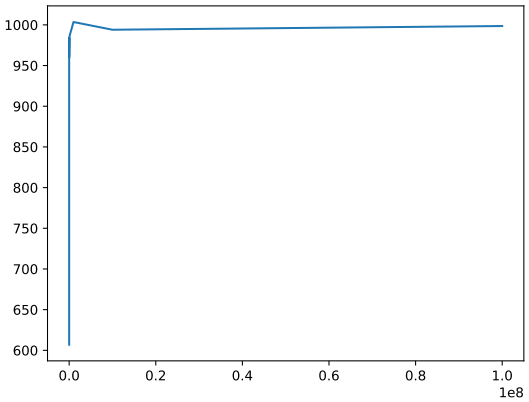
The Pareto distribution probability distribution function writes:

$$f(x;\alpha,x_m)=\alpha\frac{x_m^{\alpha}}{x^{\alpha+1}}\quad,x>x_m$$

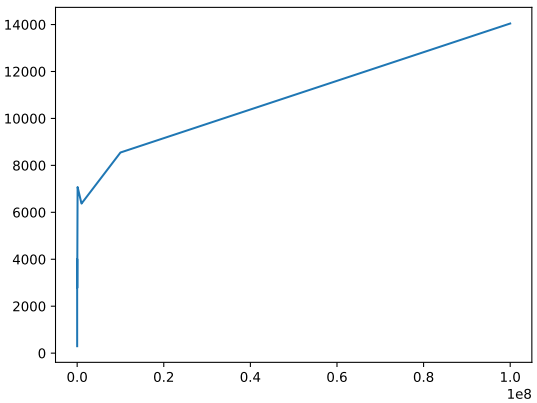
we chose $1<\alpha=1.95<2$ (for fire losses, (Rytgaard, 1990) suggests $\alpha=1.5$), as $\alpha>1$ the mean of this distribution is $\frac{\alpha x_m}{\alpha-1}$ so we chose $x_m=\frac{\alpha-1}{\alpha}*1000$, however as $\alpha<2$ this distribution has no finite variance so in the words of Gourieroux the "pseudo-true value of the parameter" doesn't exist.
This is illustrated:



As the population size increases (x axis), the mean of the distribution converges:



But as the population size increases (x axis), the variance of the distribution doesn't converge:



To fit a Pareto distribution on a sample, the maximum likelihood estimation of x_m is $\min_i(X_i)$ and of α is:

$$\frac{n}{\sum_{i=1}^{n} \ln\left(\frac{X_i}{\hat{x}_m}\right)}$$

where we assume that $X_1, \ldots, X_n$ are independent and identically distributed.

11.10 Generalized extreme value distribution

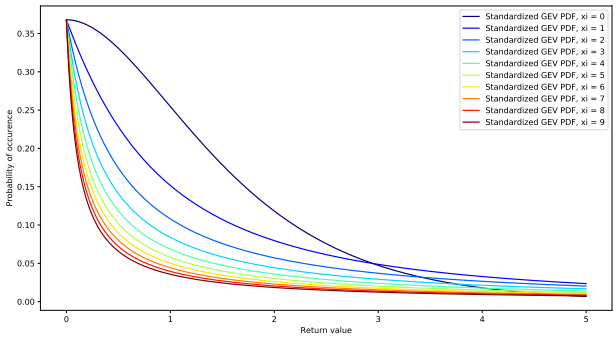
We write the probability density function of the exponential distribution:

$$f(x;\mu,\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left[1+\xi\frac{x-\mu}{\sigma}\right]^{-1-\frac{1}{\xi}} \exp\left[-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right], & \text{for} \quad x \neq 0 \\ \frac{1}{\sigma}\exp\left[-(1+\xi)\frac{x-\mu}{\sigma}\right]\exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] & \text{for} \quad x = 0 \end{cases}$$

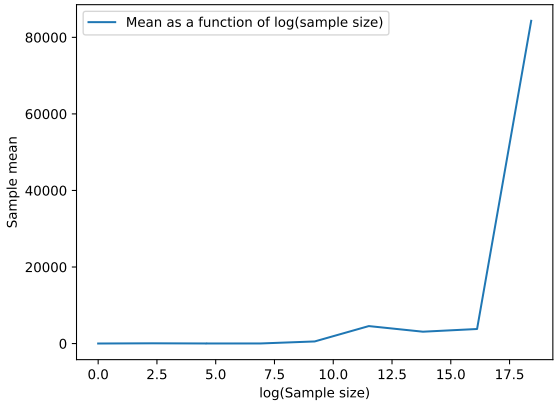
For the standardized distribution:

$$f(x;\xi) = \begin{cases} [1+\xi x]^{-1-\frac{1}{\xi}} \exp\left[-(1+\xi x)^{-\frac{1}{\xi}}\right], & \text{for} \quad x \neq 0 \\ \frac{1}{\sigma}\exp\left[-(1+\xi)x\right]\exp\left[-\exp(-x)\right] & \text{for} \quad x = 0 \end{cases}$$

For illustration here the standardized GEV probability density function for different values of ξ:



For $\xi \geq 1$ this distribution has neither finite mean nor variance, illustration wih $\xi = 1.5$:



11.11 Compound Poisson-Exponential distributions

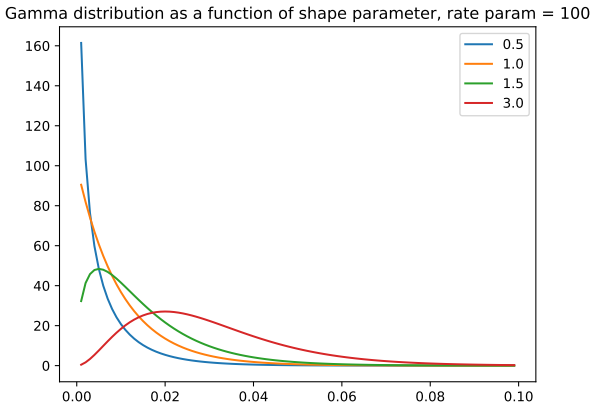
Typically, in insurance claims modelling (non-life), we model the occurrence of claims as a Poisson[33] process of parameter λ and the magnitudes of claims as an exponential law of parameter α. Writing that the claim epoch follow a Poisson process is equivalent to writing that the inter-occurrence time

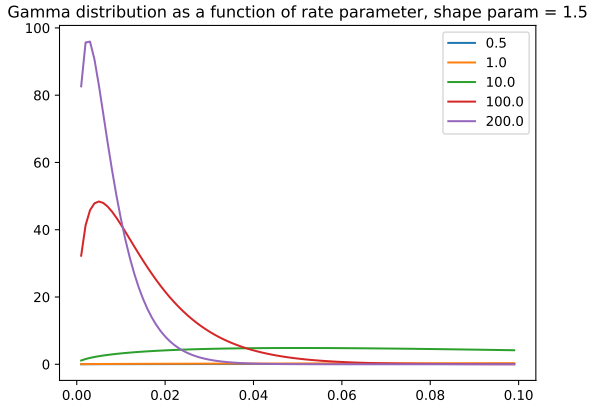
[33] the probability density function of a Poisson distribution is $P(N_t = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad \lambda > 0$

follow an exponential distribution with rate parameter λ and a probability density function of $\lambda e^{-\lambda x}$ for $x \geq 0$.
We model an insurance company facing claims arriving over time with a Poisson law N_t and which claim size C_i follow and exponential law. We have the compound process: $X_t = \sum_{i=1}^{N_t} C_i$ and its estimate $\hat{X}_t = \sum_{i=1}^{N)_t} \hat{C}_i$.
We have: $E[X(t)] = E[N(t)]E[C] = \frac{\lambda}{\alpha}$ and $Var[X(t)] = E[N(t)]E[C^2] = \lambda(\sigma^2 + \mu^2) = \lambda(\frac{1}{\alpha^4} + \frac{1}{\alpha^2})$

11.12 Gamma distribution

We write the Gamma distribution with distribution function $f(x) = \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x}$, $x > 0$, $r > 0$ the shape parameter, $\alpha > 0$ the rate parameter and $\Gamma(r) = \int_0^{+\infty} u^{r-1} e^{-u} du$, we have the property[34] $\Gamma(r+1) = r\Gamma(r)$ and $\frac{\Gamma(r+1)}{\beta^{r+1}} = \int_0^{+\infty} u^r e^{-\beta u} du$. Such distribution can be useful in insurance modelling (claims size distribution). if $r > 1$ then it is alike a log-normal distribution:





For a random variable following a gamma distribution, we have:

$$E(X) = \int_0^{+\infty} \frac{\alpha^r}{\Gamma(r)} u^r e^{-\alpha u} du = \frac{\alpha^r}{\Gamma(r)} \int_0^{+\infty} u^r e^{-\alpha u} du$$

$$E(X) = \frac{\alpha^r}{\Gamma(r)} \frac{\Gamma(r+1)}{\alpha^{r+1}} = \frac{r}{\alpha}$$

For the variance, we have $Var(X) = E(X^2) - (E(X))^2$ and $E(X^2) = \frac{r(r+1)}{\alpha^2}$ so $Var(X) = \frac{r}{\alpha^2}$.

[34] some elements of demonstration here

11.13 Sample statistics, Law of Large Numbers and Central Limit Theorem

11.13.1 Mean and Standard Deviation in theory

If we have a sample of size n from a population, we consider the random variable X which have the outcome values x (that would be $x \in [0,\infty[$ for a stock price), each outcome having the probability of occurring of $p(x)$. We can consider that this random variables are taken from a greater population with mean μ and standard deviation σ. In practice, we do not have access to the full population and have to work with finite samples.
Not that the outcome values X are considered a random sample of independent and identically distributed random variables with the same probability distribution function. We will see in section 18 possible distributions function for financial returns.
In practical terms, if over 5 days we observe the following price values x_t for a stock X, we have the probabilities:

t	1	2	3	4	5
x_t	1.1	1.2	1.3	1.2	1.2

$$p(1.1) = \frac{1}{5} \quad p(1.2) = \frac{3}{5} \quad p(1.3) = \frac{1}{5}$$

This is a very basic example and in practice you would need much more observations to work with plausible probability values.
The expected value of your random variable X is called the sample mean, we can show that it is unbiased compared with the population true mean:

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

We can also write:

$$E\left(\bar{X}_n\right) = \sum_x x p(x)$$

In our example, this would be: $E(X) = 1.1 \times \frac{1}{5} + 1.2 \times \frac{3}{5} + 1.3 \times \frac{1}{5} = \frac{6}{5} = 1.2$
We can state the weak law of large numbers:

11.13.2 Weak law of large numbers

If we have n iid random variables taken from a population, with $E(X_i) = \mu$ and $\sigma(X_i) = \sigma$, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$.

The variance of the mean of our sample is biased:

$$V(\bar{X}_n) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n V(X_i) = \frac{\sigma^2}{n}$$

If we want an unbiased standard deviation[35] measure from our sample, one can demonstrate that it can be written as:

$$S_n^2 = \frac{1}{n-1} E\left(\left[X_i - \bar{X}_n\right]^2\right)$$

[35] this can be measured using the standard deviation function with a degree of freedom of 1 (ddof = 1)

The variance of the sample is:

$$V(X) = E\Big(\big[X - E(X)\big]^2\Big) = \sum_x p(x)\big[x - E(X)\big]^2$$

In our example, this would be: $V(X) = (1.1-1.2)^2 \times \frac{1}{5} + (1.2-1.2)^2 \times \frac{3}{5} + (1.3-1.2)^2 \times \frac{1}{5} = 0.004$

11.13.3 Central Limit Theorem

If we have n iid random variables taken from a population, with $E(X_i) = \mu$ and $\sigma(X_i) = \sigma$, then if we define $\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}$, then $\bar{Z}_n$ converges in distribution, $n \to \infty$ to the standard normal distribution $\mathcal{N}(0, 1)$.

We suggest to illustrate this with samples taken from a population that has a uniform distribution function, with its probability distribution function:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for} \quad a \geq x \geq b \\ 0 & \text{for} \quad x < a \quad \text{or} \quad x > b \end{cases}$$

we take $a = 0$ and $b = 1$.

- **Question 17**
  - Use a standardized generalized extreme value distribution with $\xi \geq 1$, as this distribution has no finite mean, is the central limit theorem still applicable?

11.13.4 Gosset and its Student

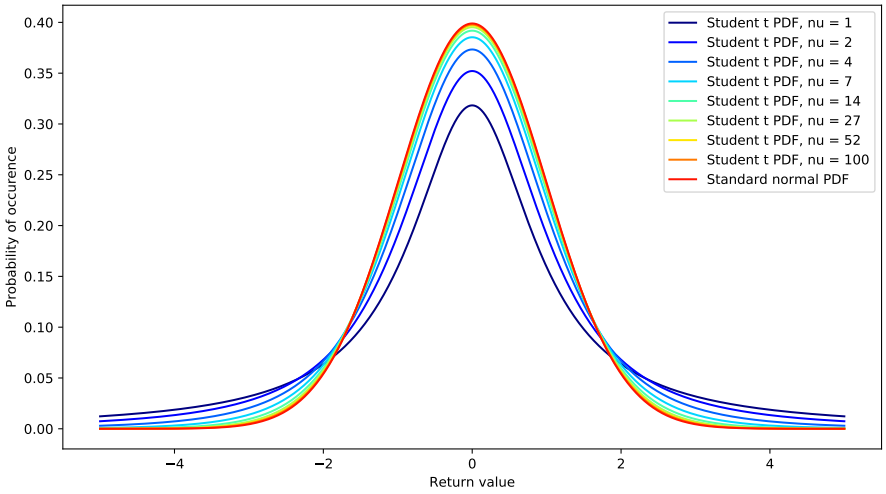
Gosset has established under the pseudonym Student that:

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1)$$

and t_{n-1} being the t-Student distribution with $n-1$ degrees of freedom:

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{S_n^2}{n}}} \sim t_{n-1}$$

And remember that for a t-Student distribution with ν degrees of freedom t_ν, then when $\nu \to \infty$, then it becomes a standard normal distribution $\mathcal{N}(0, 1)$:



This means that the greater our sample size n, the closer to the actual population mean and standard deviation we will measure. This can be rewritten and is called the Central Limit Theroem:

11.13.5 Skewness, Kurtosis and covariance in theory

The standard deviation of our sample is defined as:

$$\sigma_X = \sqrt{V(X)}$$

its skewness is defined as:

$$S(X) = E\left[\left(\frac{X - E(X)}{\sigma_X}\right)^3\right]$$

its kurtosis is defined as:

$$K(X) = E\left[\left(\frac{X - E(X)}{\sigma_X}\right)^4\right]$$

The covariance between two random variables X and Y indicates how likely they are to occur together:

$$Cov(X,Y) = E\left[(X - E(X))\,(Y - E(Y))\right] = E(XY) - E(X)E(Y)$$

11.13.6 Mean, Standard Deviation and Correlation in practice

If we look at an observation vector $X = [x_0, x_1, \ldots, x_{T-1}]$ (like the close price of a company's stock), we can call $E(X)$ the expected price, which in practice will be the average over the sample:

$$E(X) = \frac{1}{T}\sum_{i=0}^{T-1} x_i$$

The standard deviation is a measure of risk expressed in %:

$$\sigma(X) = \sqrt{\frac{1}{T}\sum_{i=0}^{T-1}(x_i - E(X))^2}$$

The covariance of X with another vector of observations $Y = [y_0, y_1, \ldots, y_{T-1}]$ can be computed as:

$$cov(X,Y) = \frac{1}{T}\sum_{i=0}^{T-1}[x_i - E(X)]\left[y_i - E(Y)\right]$$

The Pearson coefficient of correlation between X and Y is then:

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

11.13.7 Correlation Measures techniques

We can detail the correlation measures:

- Pearson correlation coefficient $\in [-1,1]$:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2)\sum(y_i - \bar{y})^2)}}$$

- Kendall rank correlation coefficient:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

- Spearman rank correlation coefficient:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

with:

- n: number of value in each data set
- n_c: number of concordant, a pair (x_i, y_i) is said to be concordant with the pair (x_j, y_j) if either both $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$
- n_d: number of discordant, a pair (x_i, y_i) is said to be discordant with the pair (x_j, y_j) if either $x_i < x_j$ and $y_i > y_j$ or $x_i > x_j$ and $y_i < y_j$
- d_i: the difference between the ranks[36] of corresponding values x_i and y_i

One might wonder: should correlation be applied to prices or returns? You can find detailed advice to apply correlation on asset returns here, here and here.

[36] think of ranking from first, second to last, with possibly similar ranking when the values are equal, then you need to adjust for ties (take the mean rank not the upper nor lower rank of the tie)

12 Python: Weight and height distributions

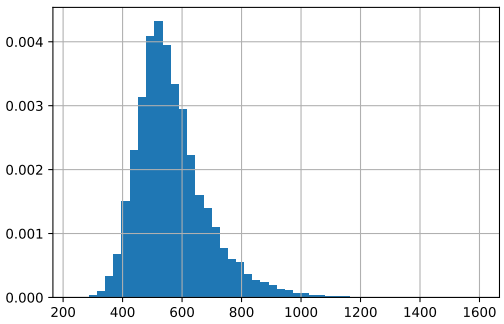
(Hermanussen et al., 2001) found that

> Body weight is not normally distributed, but skewed to the right. Also power transformation was inadequate to sufficiently describe the shape of this distribution. The right tail of weight distributions declines exponentially, beyond a cut-off of +0.5 standard deviations.

We further explore the concept of DGP introduced secion 11.1. Now, the weight and height of an individual is likely to be related to his parent's, but appart from those relations, we are closer to an i.i.d. process than with a time series.
In the following[37], we use Indian 2005-6 Demographic and Health Surveys and would like to thank Professor Dean Spears for his help on collecting the data:

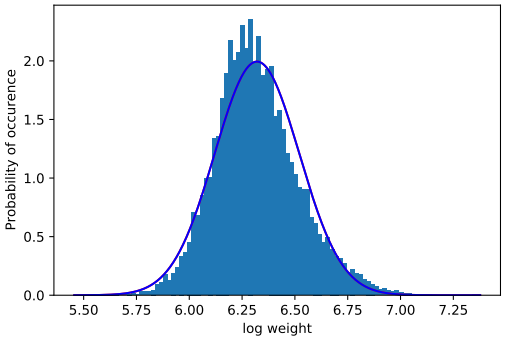
	age	weight in .1 kg	height in mm
count	15494	15494	15494
mean	32.2	567.5	1573.6
std	9.4	121.1	67.8
min	15.0	234.0	1267.0
25%	24.0	486.0	1530.0
50%	32.0	546.0	1573.0
75%	40.0	625.0	1617.0
max	49.0	1599.0	1967.0

Our sample body weights are not normally distributed, but skewed to the right:

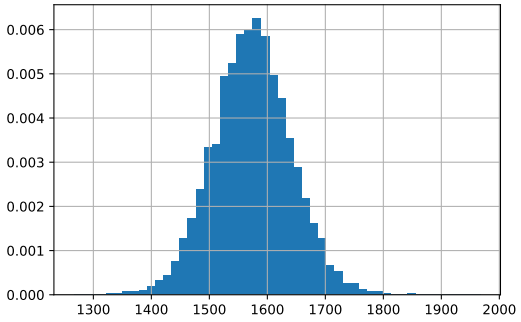


We apply a tow-sided t-test and we do not reject the null hypothesis H0: the expected value of this sample made of (presumably) independent observations is equal to $\mu = 567.5$.
We apply a two-sided Kolmogorov-Smirnov (cf. section 12.1.3, alternatively we can use the Anderson-Darling test) test on our sample with the null hypothesis: H_0 the sample was taken from a normal distributed population and we reject the null hypothesis.
If taking the log of the weight seems closer to a normal distribution, it still fails the KS test:

[37] code: vansteenberghe_height_normal_distrib.py



Our sample body height is almost normally distributed:



12.1 Are the weights normally distributed?

We apply a two-sided Kolmogorov-Smirnov (cf. section 12.1.3) one-sample test on our sample with the null hypothesis: H_0 the sample was taken from a normal distributed population and we reject the null hypothesis. But if we apply a two-sample test with a randomly generated normal sample, we do not systematically reject the null hypothesis at the 1% threshold that both sample were taken from the same distribution[38].

12.1.1 Jarque-Bera test

We can test normality of the distribution with a Jarque-Bera test, with the test statistics that follows asymptotically a χ^2_2:

$$JB = \frac{T}{6} skewness^2 + \frac{T}{24}(kurtosis - 3)^2$$

The statistics is above the critical χ^2_2 threshold and we reject H_0 the null hypothesis of normal distribution.

12.1.2 Chi-Square Goodness-of-Fit Test

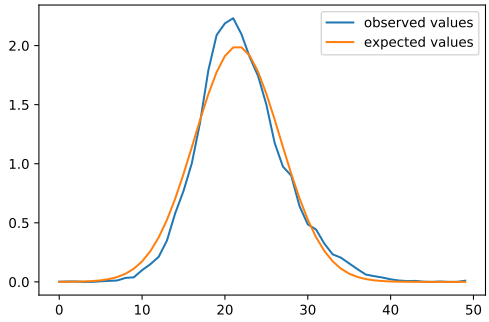
The chi-square test tests if a sample of data came from a population with a specific distribution. The null hypothesis of the test is H_0: the data follow a specified distribution.

[38] we later move on to the Lilliefors test

For the chi-square goodness-of-fit computation, the data are divided into k bins and the test statistic is defined as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{15}$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i.
We reject H_0, and visually:



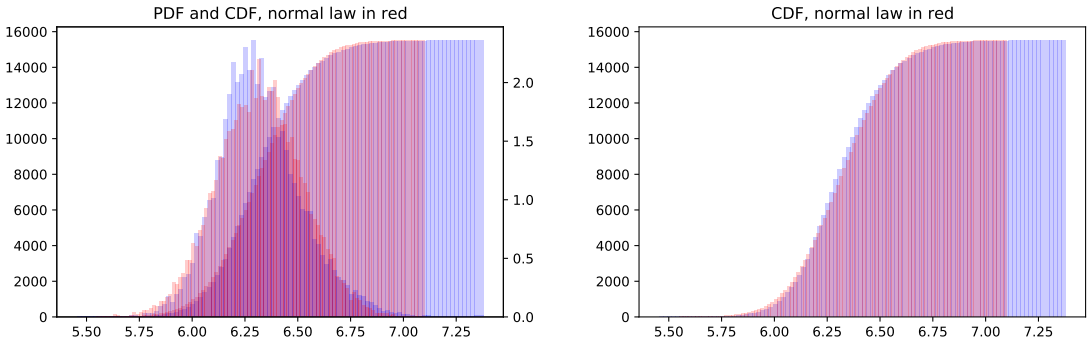
12.1.3 Kolmogorov-Smirnov tests

The Kolmogorov-Smirnov test allows us to compare the distribution of two observed returns series. We can implement a two-sample Kolmogorov-Smirnov test, where the null hypothesis is that both groups were sampled from populations with identical distributions.
The two-sided test uses the maximum absolute (vertical) difference between the cumulative distribution function (CDF, denoted $\hat{F}_1$ and $\hat{F}_2$ here) of the distributions of the two data vectors. The test statistic is:

$$KS = \max_x |\hat{F}_1(x) - \hat{F}_2(x)|$$

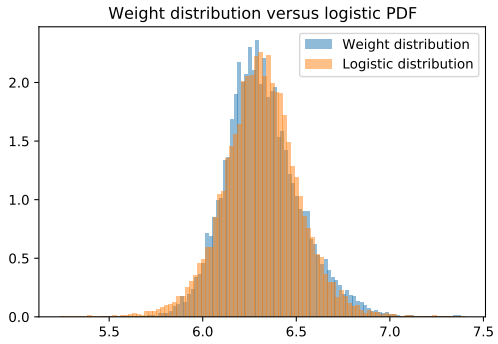
If the K-S statistic is small (but be cautious in how you define small as we talk about differences in distribution functions) or the p-value is high, then we cannot reject the hypothesis H_0 that the distributions of the two samples are the same.

We reject the hypothesis that the daily mean return of the CAC 40 index are normally distributed:



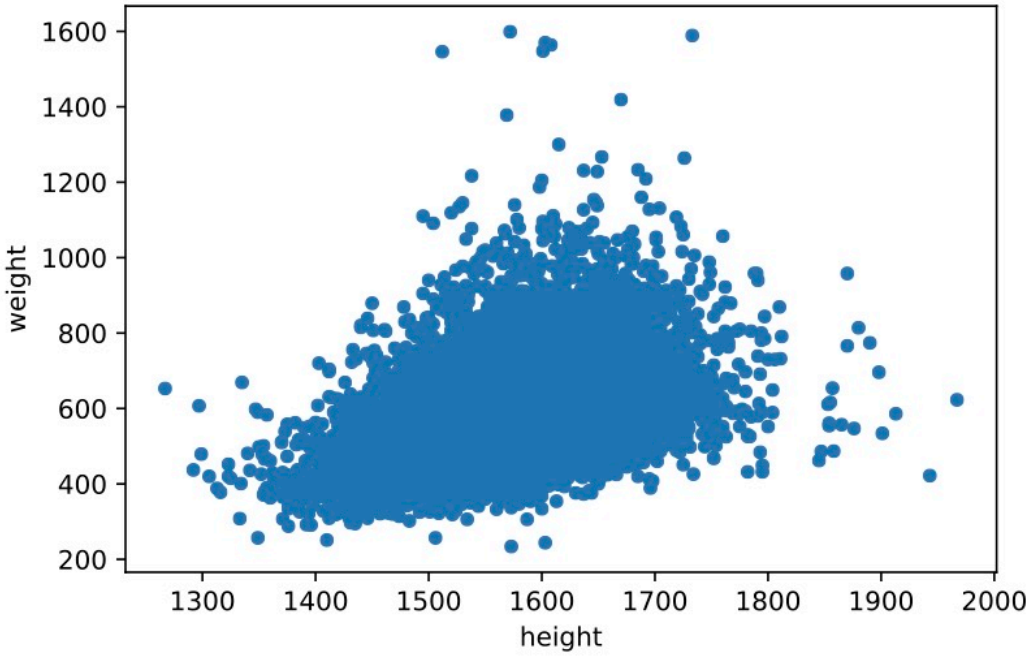
12.2 Searching for a more suitable distribution

We run a loop searching for alternative distributions and based on the p-values of Kolmogorov-Smirnov test, this can be very specific to the time frame we look at and might not be generalized, we select:

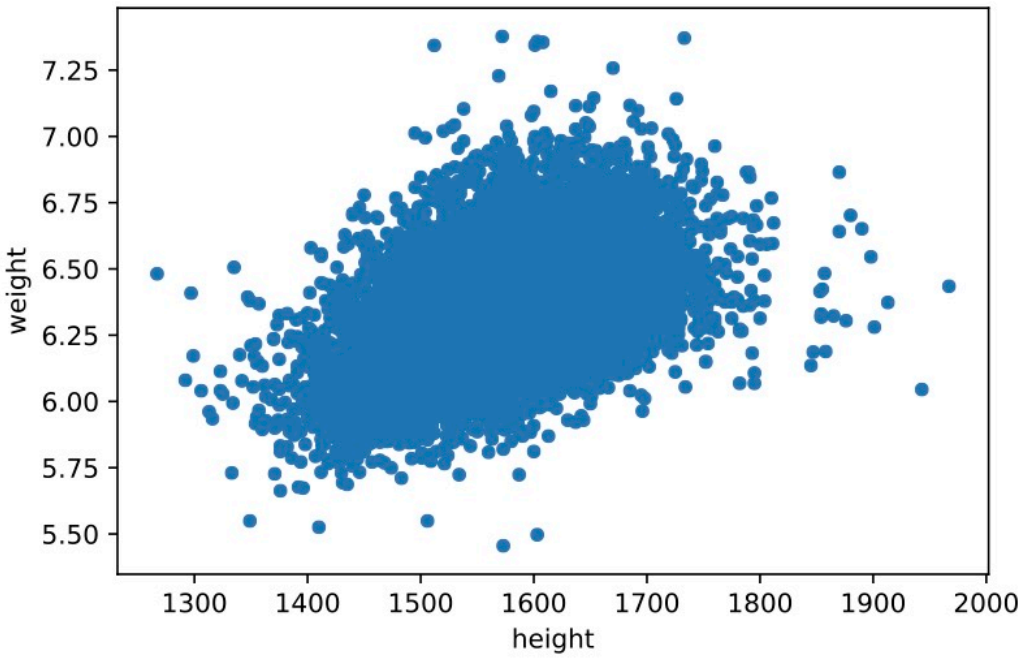


12.2.1 Linear regression

We might want to link the weight of a person with her height:



As we will describe in more detail section **??**, we do not seem to validate the homoskedaticity hypothesis, we take the log of weight so our data set better fit this hypothesis:



We assume that $\log(w)$ and h follow given Data Generating Processes and might want to fit a linear regression:

$$\log(w_i) = \alpha + \beta h_i + \epsilon_i \tag{16}$$

From the literature on the subject, we know we are missing explanatory variables (parents' height and weight at a minimum, but we do not have that information).

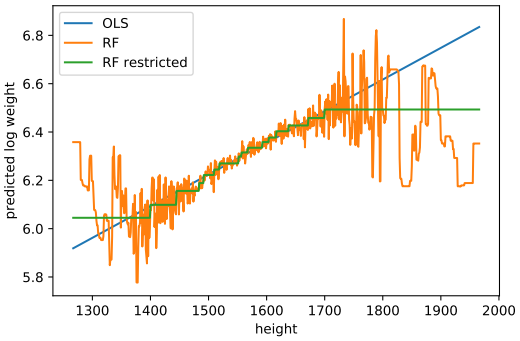
Dep. Variable:	logv437	**R-squared:**	0.197
Model:	OLS	**Adj. R-squared:**	0.197
Method:	Least Squares	**F-statistic:**	3801.
Date:	Wed, 25 Mar 2020	**Prob (F-statistic):**	0.00
Time:	16:16:21	**Log-Likelihood:**	4646.1
No. Observations:	15494	**AIC:**	-9288.
Df Residuals:	15492	**BIC:**	-9273.
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	4.2587	0.033	127.231	0.000	4.193	4.324
v438	0.0013	2.13e-05	61.654	0.000	0.001	0.001

Omnibus:	1317.052	**Durbin-Watson:**	1.664
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1972.605
Skew:	0.669	**Prob(JB):**	0.00
Kurtosis:	4.125	**Cond. No.**	3.66e+04

Our linear model predicts that an increase of 1 unit of height will result in an increase of .1% of the weight.

We can also use machine learning techniques, but these will be oriented toward answering the question: given a height h_i what is the predicted weight w_i, but the model being nonparametric, we cannot easily derive a sensitivity.
As an illustration, we fit a Random Forest regression model on the sample, we might want to restrict our fit not to overfit:



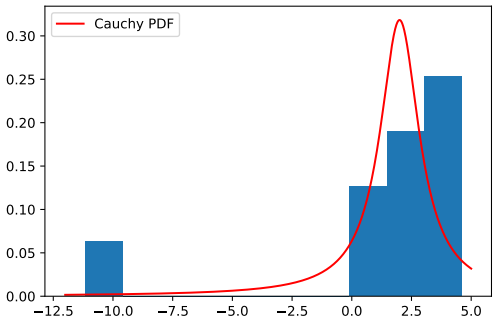
knowing its limitation, the linear model allows an interpretation of the impact of height on weight.

13 Python: Median versus mean and outlier detection

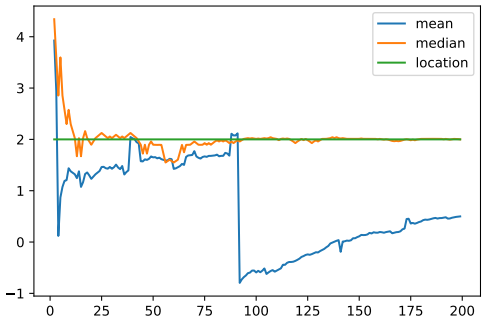
The mean of a sample is vulnerable to outlying observation, we talk about breakdown point as in (Leys et al., 2013) which is

> the maximum proportion of observations that can be contaminated (i.e., set to infinity) without forcing the estimator to result in a false value (infinite or null in the case of an estimator of scale). For example, when a single observation has an infinite value, the mean of all observations becomes infinite; hence the mean's breakdown point is 0. By contrast, the median value remains unchanged. The median becomes absurd only when more than 50% of the observations are infinite. With a breakdown point of 0.5, the median is the location estimator that has the highest breakdown point.

We take[39] for example a Cauchy distribution of location $\mu = 2$ as we know it is symmetric and interestingly it has no defined mean nor variance. We draw 10 random points to form a sample where we observe a clear outlier at -10:



If we observe the evolution of the mean and median, from a sample ranging from 2 to 200 draws, we observe that the median is robust around the location while the mean, depending on the random draw will be "polluted" by outliers and in fact increasing the sample size will not enable us to has a mean that converges:



So at this stage we might wonder why using the mean (which is at the core of the OLS we will see in the following sections) and not the median. We find an answer in the efficiency of the two measures. According to the central limit theorem, the mean of a sample with variance σ^2 has a variance of $\frac{\sigma^2}{n}$.

[39] code: 20200410_median_mean.py

Let's demonstrate the the median of a sample with variance σ^2 has a variance of $\frac{\pi\sigma^2}{2n}$. For this, we follow Dr David A. Stephens demonstration:
Suppose $X_1, \ldots, X_n$ are i.i.d. continuous random variables from distribution with cumulative distribution function F_X and we define:

$$Y_n(x) = \frac{1}{n}\sum_{i=1}^{n} Z_i(x)$$

with

$$Z_i(x) = \begin{cases} 1 & \text{if} \quad X \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

Then $Z_i(x)$ follows a Bernoulli distribution, hence:

$$E\left[Z_i(x)\right] = F_X(x)$$

$$Var\left[Z_i(x)\right] = F_X(x)\left(1 - F_X(x)\right)$$

Using the central limit theorem:

$$\sqrt{n}\left[Y_n(x) - F_X(x)\right] \xrightarrow{d} \mathcal{N}\left(0, F_X(x)\left(1 - F_X(x)\right)\right)$$

Applying the Delta method:

$$\sqrt{n}\left[F_X^{-1}Y_n(x) - x\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{F_X(x)\left(1 - F_X(x)\right)}{\left[f_X(x)\right]^2}\right)$$

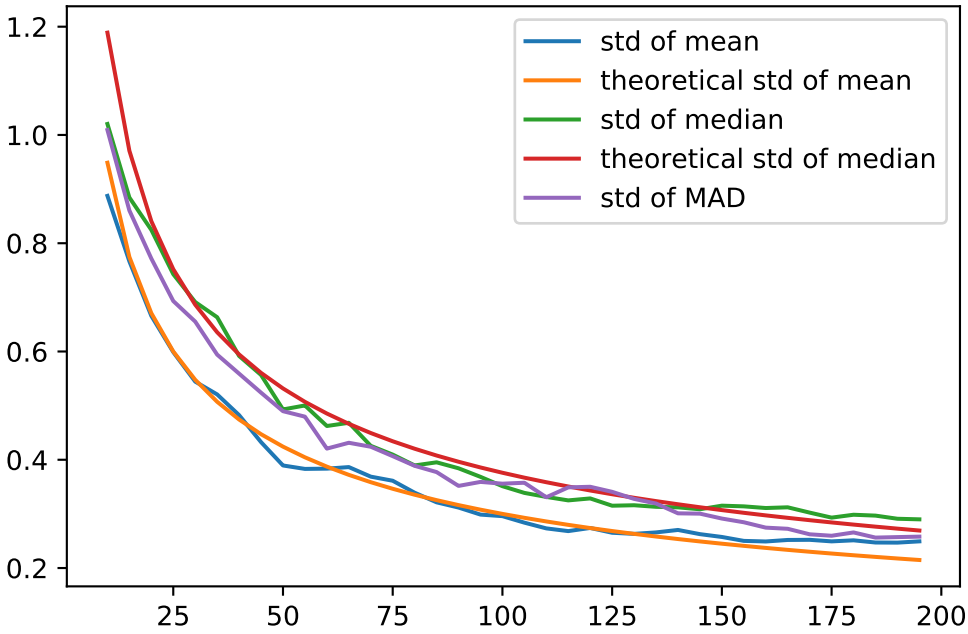
Writing $p = F_X(x)$we have the random variable $F_X^{-1}Y_n(x)$ that is lying between the $(p-1)$ and p^{th} sample quantile. So if we apply this to the median $p = \frac{1}{2}$ of a symmetric distribution with a defined standard deviation and a true median θ, then the median $\tilde{X}_n$ of the sample follows:

$$\sqrt{n}\left[\tilde{X}_n - \theta\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4\left[f_X(\theta)\right]^2}\right)$$

and if we assume that the population has a normal distribution of mean μ, then $\theta = \mu$ and $f(\mu)^2 = \frac{1}{2\pi\sigma^2}$:

$$\sqrt{n}\left[\tilde{X}_n - \theta\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{\pi\sigma^2}{2n}\right)$$

We suggest an empirical application, growing 100 samples from a normal distribution of mean 2 and standard deviation 3, we also plot the Median Absolute Deviation we introduce section 13.1.1 as a benchmark of a measure which is known to be robust and efficient or as (Huber, 1981) put it the "single most useful ancillary estimate of scale":



We indeed find in this well-behaved example that the mean of the sample with a theoretical variance of $\frac{\sigma^2}{n}$ is more efficient than the median of the sample with a theoretical variance of $\frac{\pi\sigma^2}{2n}$. The lack of efficiency the median is what leads to consider M estimators. But before, let's consider methods to detect outliers.

13.1 Outlier detection

Let's imagine[40] that we have at our disposal a small sample of size N from a population that is almost normally distributed with mean $\mu = 0$, but with standard deviation σ that changes randomly based on a latent variable (sigchoice) following a combination of discrete uniform distributions that determine whether the standard deviation is 1 or 100 for any given observation and we make it so that the low standard deviation is more likely than the high one. If the sample size is large enough, high and low observations from the population will compensate and the mean should be unbiased.
Now if we take a small sample from this population: $-173.29, -1.03, -14.63, 0.3, 0.4, 0.65, 1.88, -0.2, 0.46, -1.79$ then the mean is strongly biased at -18.7 and less so the median at 0.05.
One option to detect an outlier is to remove the extreme 1%, with our limited sample, we trim our sample and reject from it the extreme 10% which is simply in our case with 10 observations the observation with the highest absolute value.
Another approach is to discard outliers which are defined as observations which are further than plus or minus two standard deviation of the sample mean. With our example, we fail to reject the negative outlier which is in line with (Leys et al., 2013)

> the mean and standard deviation are strongly impacted by outliers. [...] this method (the mean plus or minus three standard deviations) is very unlikely to detect outliers in small samples.

[40] code: 20200410_outlier_detection.py

13.1.1 Median absolute deviation

In (Leys et al., 2013), they state that:

> A survey revealed that researchers still seem to encounter difficulties to cope with outliers. Detecting outliers by determining an interval spanning over the mean plus/minus three standard deviations remains a common practice. However, since both the mean and the standard deviation are particularly sensitive to outliers, this method is problematic. We highlight the disadvantages of this method and present **the median absolute deviation**, an alternative and more robust measure of dispersion that is easy to implement.

As we experienced just above, in outliers detections, we have three main issues: firstly, distributions might not be normal. Secondly, the mean and standard deviation are strongly impacted by outliers. Thirdly, this method is very unlikely to detect outliers in small samples.

The median absolute deviation (MAD) is defined as the median of the absolute deviations from the sample median and is a robust alternative to the standard deviation measure of a sample:

$$\text{MAD} = b \quad \text{median}\left(|X_i - \text{median}(X)|\right) \tag{17}$$

where b is linked to the distribution assumption of the population and is taken to be the inverse of the 0.75 quantile of that underlying distribution, which for a normal is $b = 1.4826$.
We finally detect outliers by determining an interval spanning over the median plus/minus three MAD. The trimmed sample presents "well behaved" mean and median:

- MAD trimed sample mean 0.08
- MAD trimed sample median 0.35

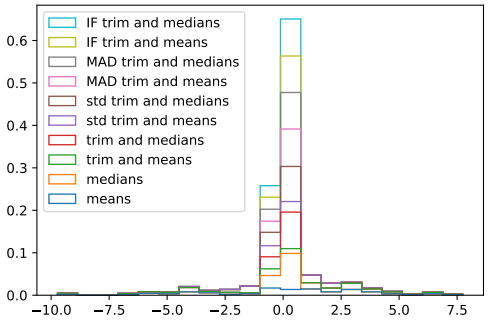
13.1.2 Isolation Forest

(Liu et al., 2012) detail the use of isolation Forest (IF), a method rather suited toward "big data". We nevertheless implement it here to compare with our previous results.

13.1.3 Outlier isolation comparison

Finally, we generate 100 samples of size 100 and compare the different methods:

	mean	std
means	0.204	3.400
medians	-0.010	0.136
trim and means	0.227	3.454
trim and medians	-0.010	0.136
std trim and means	-0.039	1.713
std trim and medians	-0.009	0.139
MAD trim and means	-0.015	0.104
MAD trim and medians	-0.012	0.137
IF trim and means	-0.012	0.144
IF trim and medians	-0.015	0.137



To go further about robust statistics, we recommend (Maronna et al., 2019) who recommend the bisquare M-estimator for location with MAD for scale. And for machine learning application, isolation forest has been extended in (Hariri et al., 2019).

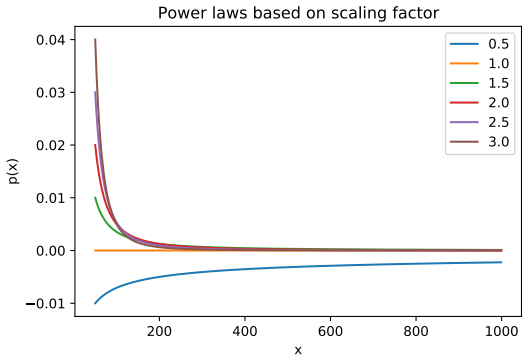
- **Question 18**
  - increase progressively the size of the generated samples and comment on the convergence of the measures and/or their efficiency.

14 Python: introduction to power laws

A a quantity x obeys a power law if it is drawn from a probability distribution

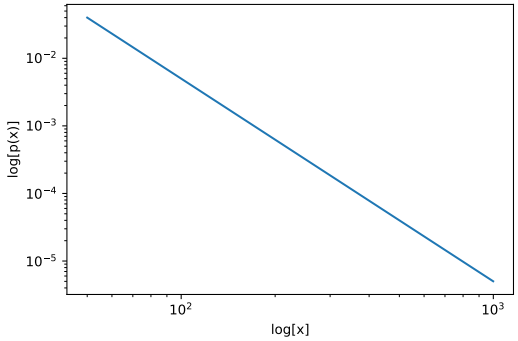
$$p(x) = Cx^{-\alpha} \tag{18}$$

where both α and c are constant. Distribution of the form 18 are said to follow a power law with an exponent α also called scaling parameter, typically between 2 and 3. C is not interesting, as once α is calibrated, it is chosen so that $p(x)$ integrates to 1, $C = (\alpha - 1)x_{min}^{\alpha-1}$ provided $\alpha > 1$.
We assume in the following that the scaling factor is greater than 1:



Provided that the scaling factor is greater than 1, when replotted with logarithmic horizontal and vertical axes, the histogram follows quite closely a straight line. Let $p(x)dx$ be the fraction of observations between x and $x + dx$, then we have a relationship, reformulated from equation 18:

$$\ln p(x) = -\alpha \ln x + c$$



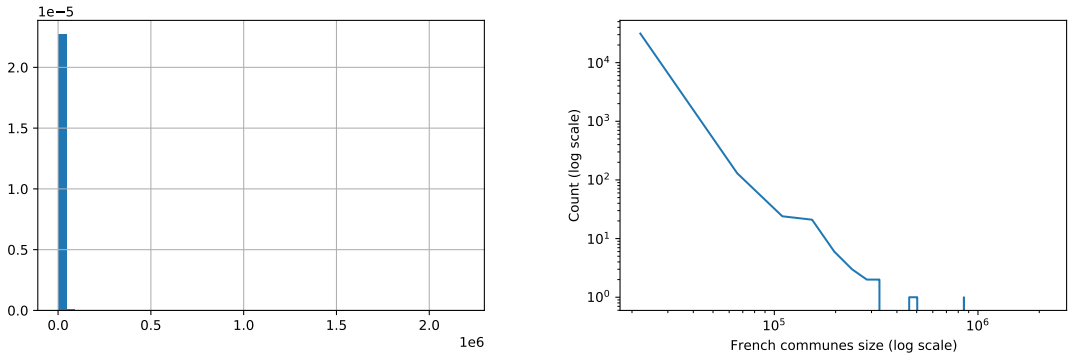
Most of the time, only the tail of the distribution follows a power law, in other words the power law applies only for values greater than some minimum x_{min}.

14.1 Power laws in nature

(Newman, 2005), (Pinto et al., 2012) and (Gabaix, 2016) suggested applications and data sets for power laws observable in nature. We follow their work and apply it to a data set we manually download next section.

14.2 Application to French city sizes

As in (Newman, 2005), we already studied the distribution of human heights section 12. Now as claimed, we check that the distribution of French city sizes[41] follow a power law[42]. (Gabaix, 1999) argues that this is more applicable to agglomerations which is coherent with our data set. We find similar histogram as FIG. 2 of that paper, with a highly right-skewed histogram:



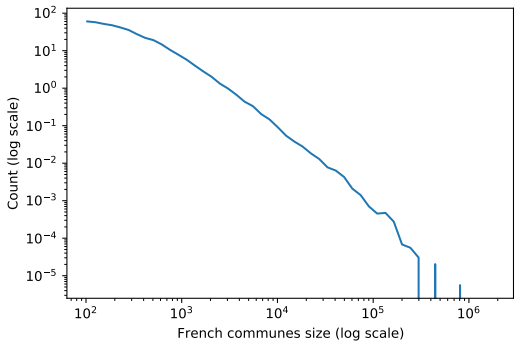
To construct the log-log histogram, we first have to construct an empirical probability distribution function:

- we chose a number of bins for this histogram, nbins
- we divide our observation segment from x_{min} to x_{max} into nbins segments of equal sizes
- we count for each segment the number of observations that falls within
- we report at each point in the middle of a segment how many observations where counted.

We can see that the right-hand end of the distribution is noisy because of sampling errors if we keep the bin size evenly distributed. We logarithmically spaced bins:

$$t = \left(\frac{x_{max}}{x_{min}}\right)^{\frac{1}{nbins}}$$

and for $i \in [1, nbins]$, the bin range is $[x_{min}t^{i-1}; x_{min}t^{i}]$:



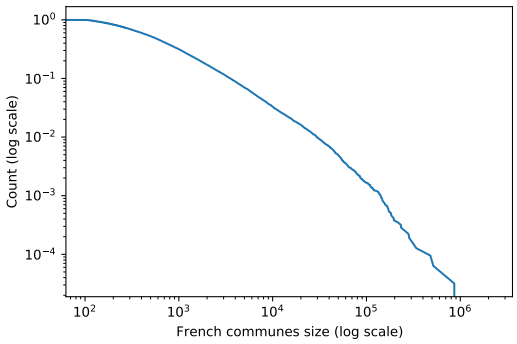
[41] data taken from INSEE

[42] code: 20200811_powerlaw_introduction.py

We also see that we might have to increase the minimum threshold under which we censor the data set.
If we want to avoid modifying the bin width, we can use the complementary cumulative distribution function: $P(X \geq x) = 1 - P(X \leq x)$ and we have

$$P(X \geq x) = C \int_x^{\infty} x^{-\alpha} dx = \frac{C}{\alpha - 1} x^{1-\alpha} \tag{19}$$

thus $P(X \geq x)$ follows a power law but of exponent $\alpha - 1$:



- **Question 19**
  - As in (Gabaix, 1999), draw a log-rank against log-size. Do you get a straight line, what is the slope?
- **Question 20**
  - manually implement a fake data set from a power law distribution as suggest in (Newman, 2005) footnote 3 to reproduce all four sub-figures of FIG. 3

If we try a very basic OLS fit with the pdf and cdf, method dating back to the 19$^{\text{th}}$ century work of Pareto:

	Dependent variable:	
	equation 18	equation 19
Intercept	37.443***	3.609***
	(3.765)	(0.007)
x	-2.875***	-0.729***
	(0.307)	(0.001)
Observations	11	34,077
R^2	0.907	0.934
Adjusted R^2	0.896	0.934
Residual Std. Error	1.002(df = 9)	0.257(df = 34075)
F Statistic	87.445*** (df = 1.0; 9.0)	479684.588*** (df = 1.0; 34075.0)
Note:		*p<0.1; **p<0.05; ***p<0.01

We lose less observations with equation 19 so we might be more precise, but we know that fitting an OLS is a poor method. Applying the step by step procedure introduced in (Clauset et al., 2009), we adapt the x_{min} threshold of our data set and improve the exponent estimation.

14.2.1 Step by step approximation

We first have to assume that $\alpha > 1$, otherwise we would need apply another method and our power law distribution would lose its interest.
Given empirical observations sorter $x_1, \ldots, x_N$ so that $x_1 \leq x_2 \leq \ldots \leq x_N$, we compute the Kolmogorov-Smirnov (KS, see section 12.1.3) statistic KS_i considering that $x_{min} = x_i$:

$$D_i = \max_{x \geq x_i} |S(x) - P(x)| \tag{20}$$

where $S(x)$ is the empirical CDF and $P(x)$ is the CDF of the power law that best fit the data choosing $x_{min} = x_i$. We chose x_min that minimizes D_i.
At each step, to fit $P(x)$, we use the Maximum Likelihood Estimation to estimate α, method that we introduced section 11.4.2. The probability density function for a sample of n independent identically distributed normal random variables is the likelihood we want to maximize:

$$L(\alpha = f(x_1, \ldots, x_n; \alpha) = \prod_{i=1}^{n} f(x_i; \alpha) = \prod_{i=1}^{n} (\alpha - 1) x_{min}^{\alpha - 1} x_i^{-\alpha}$$

As the logarithm function is continuous strictly increasing and it is easier to deal with the log-likelihood, we maximize:

$$\log(L(\alpha)) = n \ln(\alpha - 1) - n \ln x_{min} - \alpha \sum_{i=1}^{n} \ln \frac{x_i}{x_{min}}$$

The first order condition yields:

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}$$

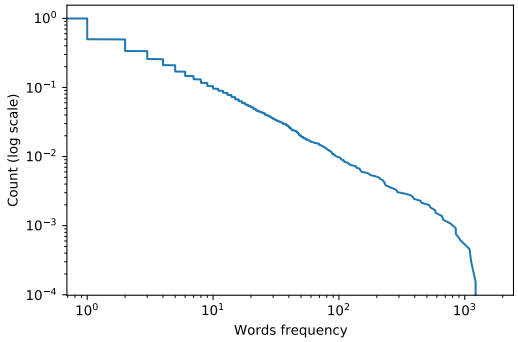
Finally, (Clauset et al., 2009) suggest using a goodness-of-fit test where the p-value is the fraction of synthetic distances (taken from the estimated power law distribution) that are greater than the observed empirical distance. They suggest to rule out the power law distribution if $p \leq 1$. This goodness-of-fit is not implemented in the powerlaw package and (Alstott J, 2014) rather argue to compare the power law fit with competing distribution such as the exponential.

14.2.2 Step by step approximation applied to French city sizes

We now use the step by step procedure suggested in (Clauset et al., 2009), we find an exponent of 2.13, which is close to the US city sizes exponent of 2.3 reported in (Newman, 2005) and 2.37 reported in (Clauset et al., 2009).
Using equation 19 and the fact that $\alpha \simeq 2$, then as in (Gabaix, 1999) we can say that cities with population greater than S is proportional to $\frac{1}{S}$.

14.3 Application to word frequency in Marcel Proust's work

We import Marcel Proust's *A la recherche du temps perdu - Du côté de chez Swann* and find that the word frequencies follow a power law of exponent 2.1:



14.4 Why power-law emerges for city sizes, an explanation by Xavier Gabaix

When $\alpha = 2$, then equation 19 simplifies into Zipf's law:

$$P(X \geq x) \propto \frac{1}{x} \tag{21}$$

(Gabaix, 1999) suggest that random growth process can be an explanation for the emergence of Zipf's law, if the following hypotheses are met:

- Gibrat's law: the growth rate is independent on x
- there are common mean and variance of growth rate process
- **Question 21**
  - Following (Gabaix, 1999):
    - Take N cities
    - Chose initial city sizes of $S_0^i = \frac{1}{N}$
    - Impose a growth rate of $2\% \pm .2\%$
    - Normalize the city sizes $\sum_{i=1}^{N} S_t^i = 1$
  - Show that city sizes converges to a Zipf's law

14.5 Fit power law to GDP or stock market returns

(Pagan, 1996) reviewed the econometric literature on financial markets up to 1996, suggesting:

- unit root in asset prices;
- financial series are not independently distributed over time;
- stock returns are not normally distributed;
- stock returns density have fat tails, exponent $\alpha > 2$ both left and right and for daily or monthly returns.
  - but as the returns are not i.i.d., the estimation of α suffers from lack of precision as demonstrated in (Kearns and Pagan, 1997);
  - (Jondeau and Rockinger, 2003) tested and found that left and right tails are not statistically different.

- A theoretical foundation for fat tailed return densities was explored by (Lux and Sornette, 2002).

(Cont, 2001) reviewed the literature on extreme stock returns with a power-law tail exponent: $2 < \alpha(T) \leq 5$ while warning on the significant digits that can be estimated. (Plerou et al., 2001) found for stock returns of 5 minutes intervals:

$$\alpha = \begin{cases} 3.10 \pm 0.03, & \text{positive tail} \\ 2.84 \pm 0.12, & \text{negative tail} \end{cases} \tag{22}$$

- **Question 22**
  - Try to fit power law to GDP or stock market returns, take the frequency (daily, quarterly, yearly) into account before comparing their exponent.
  - Which power law have higher exponent, what does it mean?

14.6 Alternative distribution with fat tails

(Mantegna and Stanley, 1995) found that the probability distribution of the Standard & Poor's 500 can be characterized by a Lévy stable process for the central part of the distribution and an exponential in the tails.
Nonetheless, (Williams et al., 2017) stated that

> A consensus has developed in the literature that the distribution of GDP growth rates can be approximated by the Laplace distribution in the central part and power-law distributions in the tails.

they study a panel of GDP growth rates and find that the distribution of GDP growth rates can be fitted using the heavy-tailed Cauchy distribution for almost all countries. We can identify several shortcomings in the approach: there are a maximum of 61 data points for each countries, which would make it difficult to fit power-law in the tails. They do within-country growth distribution.

- What about between-country growth distribution per year?
- What about you use all data to fit a generic law? To fit a power law in the tail?

(Fagiolo et al., 2008) approximate output growth-rate distribution by symmetric exponential power (EP) densities whose functional form reads:

$$f(x; b, a, m) = \frac{1}{2ab^{\frac{1}{b}}\Gamma\left(1+\frac{1}{b}\right)} e^{-\frac{1}{b}\left|\frac{x-m}{a}\right|^b} \tag{23}$$

they explain that

> The shape parameter is the crucial one for our analysis: the larger is b, the thinner are the tails. In fact, the EP density encompasses both the Laplace and the Gaussian distributions. If $b = 2$, the distribution reduces to a Gaussian. If $b < 2$, the distribution displays tails fatter than those of a Gaussian (henceforth 'super-Normal' tails). If $b = 1$, one recovers a Laplace. Finally, values of b smaller than one indicate tails fatter than those of a Laplace ('super-Laplace' tails in what follows). The above property is the value-added of the EP density, as it allows one to precisely measure how far the empirical distribution is from the normal benchmark and how close it is instead to the Laplace one. Another important property of the EP density is that it is characterized by exponentially shaped tails, which are less thick than those of power-law distributions.

Their data set is available here and they used the package subbotools

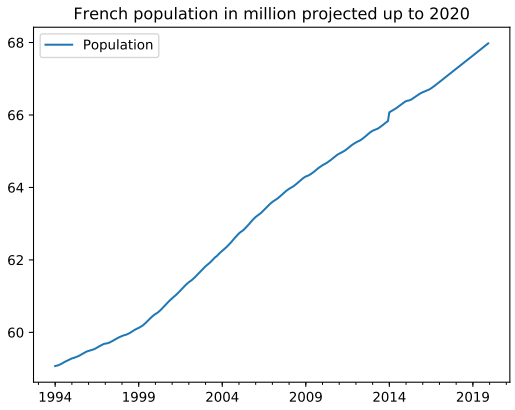
15 Python for non-programmers: exercise with pandas - part 2

We suggest further exercises[43] to get familiar with some pandas basics features in python's pandas library.

15.1 An introduction to forecasting and confidence intervals

15.1.1 Plot some previsions

From the average monthly French population change, append to your DataFrame some projections until 2020, plot this projection.



- **Question 23**
  - stick to the monthly seasonality, compute a monthly average population change and use it to project the population up to 2020.

15.1.2 Show a "confidence" interval for your forecast

If we assume that the French population growth is constant (avgchg) and its evolution is subject to errors which cannot be predicted $\epsilon_m(h)$, h being the horizon of forecast. If our model is well specified, then the error terms should have no trend, no autocorrelation and a mean of 0.

$$\text{pop}_{m+1} = \text{pop}_m \left(1 + \text{avgchg}\right) + \epsilon_m(1)$$

For the 273 observation we have, we can compute the error terms $\epsilon_m(h)$.
The confidence interval can be computed with the root mean square errors[44]:

$$\text{pop}_{m+1} = \text{pop}_m \left(1 + \text{avgchg}\right) \pm Z_{\alpha/2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \epsilon_m(1)^2}$$

where $Z_{\alpha/2}$ is the inverse of the cumulative distribution function of the normal law (percent point function) and α is a confidence level we chose (it is divided by 2 as it is two-sided).

[43] pandas_exercise_part2.py
[44] if the original series is stationary, then the confidence interval limits converges to a finite value

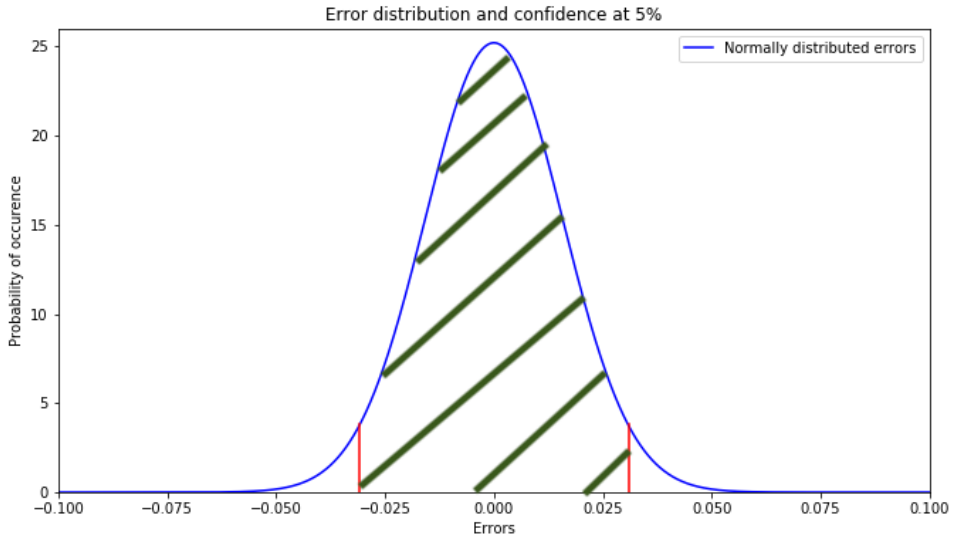
We can assume that the model errors are normally distributed with zero mean and compute their standard deviation $\sigma_m(1)$, that is our errors are independent and identically distributed $\epsilon_m \sim \mathcal{N}(0, \sigma_m(1)^2)$. Then for a $1-\alpha$ % interval of confidence, it would be, with an approximate formula[45]

$$\text{pop}_{m+1} = \text{pop}_m \left(1 + \text{avgchg}\right) \pm Z_{\alpha/2}\sigma_m(1)$$

and we approximate the residual variance σ_ϵ^2 by:

$$s(1)^2 = \frac{\sum \epsilon_m(1)^2}{n-k}$$

with n the number of observations and k the number of independent variables.
As an illustration, we plot the normal distribution of the error terms and the shaded area equals 95%, we plot similarly the boundaries on the time series plot forecast:



For a 5% two-sided confidence interval we take a t distribution with n degrees of freedom (which we approximate by a normal distribution as $n >> 120$): $Z_{\alpha/2} = 1.96$. But with this, we consider that our last projection was correct and start from this projection point and only add the uncertainty.
Intuitively, we can expect that our projection to a horizon of $h >> 1$ will be less precise:

$$\text{pop}_{m+h} = \text{pop}_m \left(1 + \text{avgchg}\right)^h + \epsilon_m(h)$$

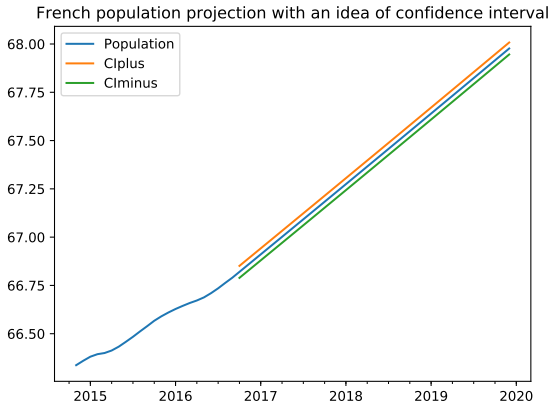
This is very similar[46] as assuming a process of the form:

$$\text{pop}_{m+1} = \psi \text{pop}_m + \epsilon_m(1)$$

[45] For more information, there is a detailed lecture on this here
[46] but actually we will first need to learn about stationarity to understand why our current model is naive

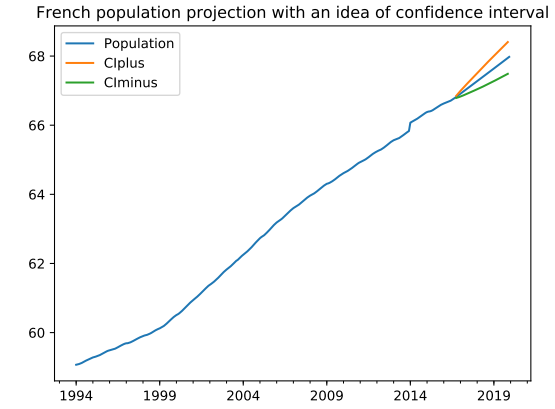
It can be shown that[47] if the time series pop_m is stationary (we discuss further these concepts section 16.1):

$$\sigma_m(h)^2 = \sigma_m(1)^2 \frac{1-\psi^{2h}}{1-\psi^2}$$



In fact, we are not certain that our model is correct, we can compute in-sample the values of our error for different forecast horizon h. Then it is possible to compute our confidence interval as:

$$\text{pop}_{m+h} = \text{pop}_m \left(1 + \text{avgchg}\right)^h \pm Z_{\alpha/2}\sigma_m(h)$$



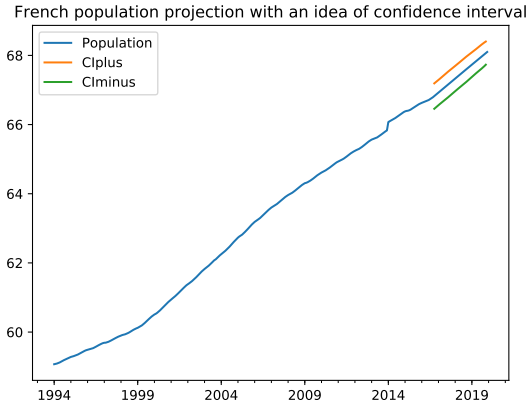
15.1.3 Trend and AR(1) component

We can suggest a third model:

$$\text{pop}_{m+1} = \alpha + \beta m + \phi \text{pop}_m + \epsilon_m(1)$$

We estimate the model with the techniques we will see further in this chapter, and we can forecast, this time, with such a model the confidence interval seems stable, this is an incentive to develop the stationarity concepts latter in this chapter:

[47] in fact, as long as avgchg << 1, then the difference between the left and the right hand side are negligible



15.1.4 Augmenting our projections with the error terms

One easy way to improve our projections is to take into account the error term from the previous period:

$$\text{pop}_{m+1} = \text{pop}_m \left(1 + \text{avgchg}\right) + \epsilon_{m-1}(1) + \epsilon_m(1)$$

We can compare the standard deviation of the error terms (a good proxy for the RMSE of the models) and find that this is minimized with our model corrected with past error terms.

15.1.5 Variable projection horizon

Another methods is to calibrate our model based on the projection horizon h:

$$\text{pop}_{m+h} = \alpha(h) + \beta(h)\, m + \phi(h)\text{pop}_m + \epsilon_m(h)$$

We compute the coefficient for each desired projection horizon h.

16 Python for non-programmers: exercise with pandas - part 3

Stationarity, Dickey-Fuller test (original and augmented), auto-regressions[48].

16.1 Concept of stationarity

Before regressing (or for AR), we need to go step by step on the integration order and the cointegration[49] of the variables. A time series y_t is (weakly) stationary if the mean, variance and autocovariance are finite and constant over time, $\forall t > s \in \mathbb{N}$:

$$\begin{cases} E(y_t) = \mu < \infty \\ Var(y_t) = \gamma_0 < \infty \\ Cov(y_t, y_{t-s}) = \gamma_s < \infty \end{cases}$$

It can be shown[50] that the following empirical moments are consistent estimators of the above moments when $T \to \infty$:

$$\begin{cases} \hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t \\ \hat{\gamma_0} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu})^2 \\ \hat{\gamma_s} = \frac{1}{T-s}\sum_{t=1+s}^{T}(y_t - \hat{\mu})(y_{t-s} - \hat{\mu}), \text{with} \quad T >> s \end{cases}$$

The most basic weakly stationary process are white noises, introduced in section 17.3 and moving average introduced in section 17.7.
y_t is strictly[51] stationary if the joint distribution of $(y_t, \ldots, y_{t-k})$ is independent of t for all k.
If Δ is the difference operator ($\Delta y_t = y_t - y_{t-1}$, $\Delta^2 y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$, etc.), the we say that then y_t is integrated of order d or I(d) if y_t, Δy_t, ..., $\Delta^{d-1} y_t$ are non-stationary and $\Delta^d y_t$ is stationary[52].

16.2 Wold theorem

If y_t is a stationary process it can be written as:

$$y_t = c + \sum_{i=0}^{\infty} a_i \epsilon_{t-i} \tag{24}$$

with c a constant, $\epsilon \sim \text{WN}(0, \sigma^2)$ and $\sum_{i=0}^{\infty} |a_i| < \infty$.

16.2.1 Impulse Response Function

Impulse Response Function aims at answering how much our process y_t is affected at time t by an innovation (a shock) that happened j periods before t. From the Wold theorem, for a stationary process the innovation at time $t - j$ is ϵ_{t-j} and the Impulse Response Function if $\frac{\partial y_t}{\partial \epsilon_{t-j}}$ which is thus θ^i.

[48] pandas_exercise_part3.py
[49] For the cointegration test, the null hypothesis is no cointegration
[50] with the law of large number that can be applied as y_t is stationary
[51] with financial time series, strict stationarity is rare but can be taken as an assumption then discussed with robustness checks
[52] the concept is detailed in (Engle and Granger, 1987). Definitions vary, (Alexander, 2008) defines integrated of order d it as $\Delta^d y_t$ having a stationary ARMA representation, keep in mind that integration of order greater than two are not very relevant for economic applications

16.3 AR(1) stationary process

We consider a stationary AR(1):

$$y_t = \theta y_{t-1} + \epsilon_t \tag{25}$$

with $0 < |\theta| < 1$ and $\epsilon \sim \mathcal{N}(0,1)$. We can rewrite:

$$y_t = \sum_{i=1}^{t-1} \theta^i \epsilon_{t-i} + \theta^t y_0$$

we have:

$$\begin{cases} E\left(y_t\right) = \theta^t y_0 \to 0 \\ Var\left(y_t\right) = \sum_{i=1}^{t-1} \theta^{2i} \sigma^2 \to \frac{1}{1-\theta^2}\sigma^2 \\ Cov\left(y_t, y_{t-s}\right) = \theta^s \frac{\sigma^2}{1-\theta^2} \end{cases}$$

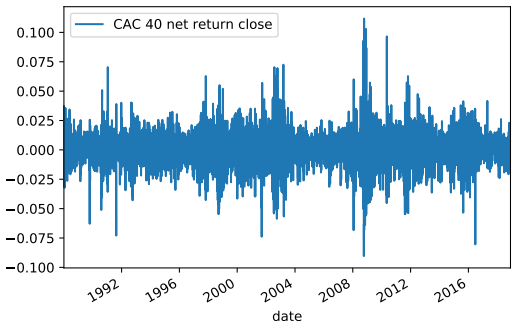
We say that an AR(1) is asymptotically stationary, and usually just say stationary.
Another way to look at it, is that if this process is to be weakly stationary, then, we need:

$$\begin{cases} \mu(1-\theta) = 0, \text{ as } \quad E\left(\epsilon_t\right) = 0 \\ \gamma_0(1-\theta^2) = \sigma^2 \\ Cov\left(y_t, y_{t-s}\right) = \gamma_s \end{cases}$$

which has solution if $0 < |\theta| < 1$, as a variance is necessarily positive. We can rewrite:

$$\begin{cases} \mu = 0 \\ \gamma_0 = \frac{\sigma^2}{1-\theta^2} \\ Cov\left(y_t, y_{t-s}\right) = \gamma_s \end{cases}$$

If $\theta = 1$ either we have no solution, either $\sigma = 0$, which would then mean that the error term is a constant $\forall t, \epsilon_t = 0$ and then $\forall t, y_t = y_0$ which is stationary but presents no interest for study.
From this first approach to AR(1) model, we can get the intuition that if we work with asset prices, to respect $E\left(y_t\right) \to 0$ we will not work with the price levels (we would not want to suggest a model where all prices tend to 0 or equivalently where all companies default), rather we will work on price returns and will de-trend to respect this condition $E\left(y_t\right) \to 0$. Then for the condition of constant volatility, if we take CAC 40 index time series we introduced in section 8.4, we can visually see that there are volatility clusters which mean that an AR model will not be satisfactory and we will want to introduce a model on the volatility evolution over time (ARCH model):



16.4 Finite order lag polynomial

We can rewrite equation 25 with a lag polynomial:

$$(1-\theta L)y_t = \epsilon_t$$

The root of the lag polynomial $(1-\theta L)$ is $\frac{1}{\theta}$ and if $0 < |\theta| < 1$, then the lag polynomial[53] is invertible:

$$(1-\theta L)^{-1} = \sum_{i=0}^{\infty} \theta^i L^i$$

we can rewrite equation 25 as:

$$y_t = \sum_{i=1}^{\infty} \theta^i \epsilon_{t-i}$$

as $\sum_{i=0}^{\infty} |\theta^i| < \infty$, y_t is a stationary process.
This can be generalize with a lag polynomial of order p:

$$(1+\phi_1 L + \ldots + \phi_p L^p)y_t = \Phi(L)y_t = \epsilon_t$$

with roots λ_k that respect $\forall k, |\lambda_k| > 1$, $\Phi(L)^{-1} = \Pi_{i=1}^{p}(1-\lambda_i L)^{-1} = \sum_{i=0}^{\infty} a_i L^i$ and it can be demonstrated that $\sum_{i=0}^{\infty} |a_i| < \infty$ and the AR(p) with the lag polynomail $\Phi(L)$ is a stationary process.

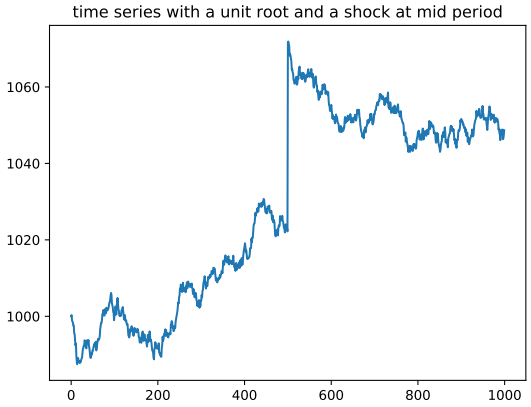
16.5 Process with a unit root: a non-stationary process

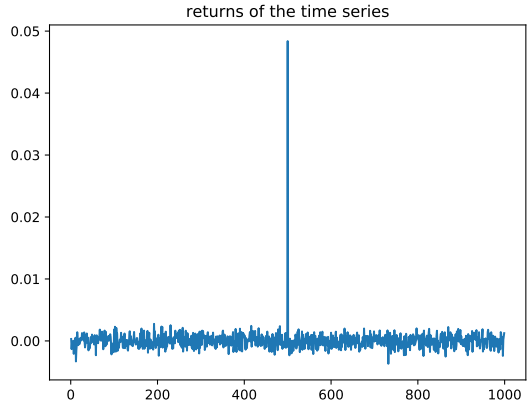
If we follow section 16.3 but now set $\theta = 1$, $y_t = y_{t-1} + \epsilon_t$ which is a Random Walk and chose $y_0 = 100$.
We can rewrite the process with L the lag operator, $(1-L)y_t = \epsilon_t$, the root of the equation $1-x$ is 1, hence we say that this process has a unit root.
We can rewrite $y_t = \sum_{i=1}^{t-1} \epsilon_{t-i} + y_0$ and find that $Var\left(y_t\right) = t\sigma^2 \to \infty$ which is against our definition of weak stationarity.

- **Question 24**
  - create a stationary AR(1) ($\theta < 1$) and a unit root one, with 1000 observation for each. Compute the variance of the series and compare them to the expected variance as described above.
  - for the stationary AR(1), introduce at time 500 an additional shock (innovation) and evaluate the impulse response function at time 505.

In simple terms, a time series has a unit root, or follows an integrated process, if shocks have permanent effects on the level ,but not on the rate of change change in the variable. Indeed, if at a time t_{choc} the time series experience a shock $\Delta = 50$, then the future level of y_t will be impacted ($\forall t \geq t_{\text{choc}}$) but not the rate of change $\frac{y_{t+1}-y_t}{y_t}$.
For illustration, we apply a shock Δ at mid-period to our time series:

[53] $\phi(L)$ is also called the autoregressive polynomial of y_t





16.6 Illustration of a spurious regression

We can define to times series as in (Granger and Newbold, 1974):

$$y_t = y_{t-1} + \epsilon_t$$

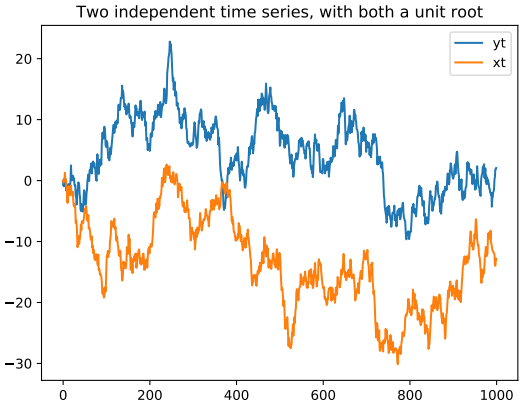
$$x_t = x_{t-1} + \nu_t$$

with ϵ_t and ν_t two independent white noise from a normal law and we set $y_0 = 0$ and $x_0 = 0$. When regressing:

$$y_t = \alpha + \beta x_t + \gamma_t$$

one would expect the β not to be significantly different than 0 and the R^2 of the regression to be low. This is not the case as both series present a unit root.

We illustrate this with an example of two draws and their R^2, coefficients, t-values and Durbin-Watson statistics:



R^2	19%
Intercept coefficient	9.1
Intercept coefficient t-value	25.3
β	0.4
β t-value	15.2
Durbin-Watson statistics	0.038

The R^2 seems "attractive"[54], both the intercept and β seems to have high t-value which means that they seems significantly different than 0, but there is a low Durbin-Watson statistics which suggest that the residuals of the regressions are strongly autocorrelated. These are the symptoms of two I(1) series not cointegrated being regressed on one another, a spurious regression where two independent time series seems related. So in fact the true value of β is zero but as demonstrated by Philips (1986) the estimate of β doesn't converge in probability to zero (estimates of the regression coefficients are inefficient). Also the usual significance tests on the coefficients are invalid. This condition (high R^2 and low Durbin-Watson statistics) is not a necessary condition for a regression to be spurious as demonstrated in (Granger and Newbold, 1974).

16.7 AR(2) process with a unit root

We define an AR(2) process for illustration[55], its stationarity depends on β_1, β_2:

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \epsilon_t \tag{26}$$

with ϵ_t a white noise as define in section 17.3, which is by definition stationary.

- **Question 25**
  - check if the mean and variance of this process is constant or vary over time, applying a small impulse shock ϵ_0 and chosing $\beta_1 + \beta_2 = 1$, for example $\beta 1 = 1.6$ and $\beta 2 = -0.6$.

With L the lag operator, we can rewrite equation 26:

$$(1 - \beta_1 L - \beta_2 L^2) y_t = \epsilon_t$$

[54] and the time series have a correlation of 43%

[55] we include here an interecept but one might argue that it is not necessary as detailed here, autoregressive models (AR) are introduced in section 17.6

If we call λ_1 and λ_2 the roots of the polynomial $\beta_2 x^2 + \beta_1 x - 1$[56], we can further rewrite equation 26:

$$\left(1 - \frac{L}{\lambda_1}\right)\left(1 - \frac{L}{\lambda_2}\right) y_t = \epsilon_t$$

If both $|\lambda_i| > 1$, then we can invert the AR into an MA[57] that is stationary and the series converges in mean square:

$$y_t = \left(1 - \frac{L}{\lambda_1}\right)^{-1}\left(1 - \frac{L}{\lambda_2}\right)^{-1} \epsilon_t$$

This is equivalent to say that the matrix $\begin{bmatrix} \beta_1 & \beta_2 \\ 1 & 0 \end{bmatrix}$ has all eigenvalues lying inside the unit circle.

And also, this is equivalent to say that $\left(I - \begin{bmatrix} \beta_1 & \beta_2 \\ 1 & 0 \end{bmatrix} L\right)$ has all roots lying outside the unit circle.

The equation 26 can be rewritten in a multivariate form, a VAR:

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} + \begin{bmatrix} \beta_1 & \beta_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \epsilon_t$$

or equivalently, with L the lag operator:

$$\left(I - \begin{bmatrix} \beta_1 & \beta_2 \\ 1 & 0 \end{bmatrix} L\right) \begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} + \epsilon_t$$

For our example, the conditions for stationarity are met.

- **Question 26**
  - take the first difference $y_t - y_{t-1}$, is this new process stationary?
  - take some other values for $\beta_1 + \beta_2 \neq 1$ so that the root of the polynomial are greater than one, take for simplicity $\alpha = 0$, and the AR(2) can be inverted into an MA
  - take some values for $\beta_1 + \beta_2 \neq 1$ so that the root of the polynomial are lesser than one, is the process explosive?
  - going back to what we discussed about applying correlation to prices or return in section 11.13.7, generate two independent times series $y_t = y_{t-1} + \epsilon_t$ and $z_t = z_{t-1} + \gamma_t$, with $\epsilon \sim \mathcal{N}(0,1)$ and $\gamma \sim \mathcal{N}(0,1)$, $t \in [0,100]$ and compute the correlation of their level and of their returns.

16.8 Dickey-Fuller test

We can follow the unit root test introduced by David Dickey and Wayne Fuller in 1979.
In its simplest form, the Dickey-Fuller test the null hypothesis H_0: $\rho - 1 = 0$ with a Student test over the equation, with no autoregressive process:

$$\Delta y_t = (\rho - 1) y_{t-1} + \epsilon_t \tag{27}$$

Which can be re-written:

$$\Delta y_t = \gamma y_{t-1} + \epsilon_t \tag{28}$$

[56] Remember that for such a polynomial of order two there are two roots: $\frac{-\beta_1 \pm \sqrt{\beta_1^2 + 4\beta_2}}{2\beta_2}$, for a polynomial $ax^2 + bx + c = 0$, the roots are $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

[57] moving average models are introduced in section 17.7

The statistics $\frac{\hat{\gamma}}{\text{Standard Error}(\hat{\gamma})}$ can be tested again the threshold of the standard t-distribution, this is discussed and confirm in (MacKinnon, 2010) for the simple version of the test.
To find the critical value, you cant import the t-distribution function from the stats package and define a probability threshold and degree of freedom[58].
For the French GDP, our statistic > critical value $\implies$ we reject the null hypothesis of the test: hence we suggest that $\rho - 1 \neq 0$ and suggest that there is no unit root. But this test is not powerful enough to have full confidence as there is likely to be autoregressive component in the GDP: use an augmented test.
If you include an intercept in regression 28, then the critial value for rejecting the null of $\rho = 1$ at the 5% level changes from -1.95 with no intercept to -2.86 in large samples, you can refer to (Dickey and Fuller, 1979).

16.9 Augmented Dickey-Fuller test

The Augmented Dickey-Fuller test can be used to test whether a series is stationary, and take autoregressive effects into account. H_0: $\gamma = 0$.
The formula being used for the augmented Dickey-Fuller test is of the form:

$$\Delta y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \gamma y_{t-1} + \sum_{j=1}^{p-1} \delta_j \Delta y_{t-j} + \epsilon_t \tag{29}$$

(MacKinnon, 2010) discuss the number of lags, selected via information criteria methods, and discuss the test statistics for the "no-constant" ('nc'), "no-trend" ('c'), and "with-trend" ('ct') version of this test with reference to the table to be used.

With the python function adfuller in statsmodels, you can chose the parameter of the regression:

- 'c' : constant only (default), $\beta_1 = \beta_2 = 0$
- 'ct' : constant and trend $\beta_2 = 0$
- 'ctt' : constant, and linear and quadratic trend
- 'nc' : no constant, no trend $\beta_0 = \beta_1 = \beta_2 = 0$

The output of the test are:

- The statistics
- The p-value
- The number of lags used for the test
- The number of observations used
- The critical values

[58] for 268 degree of freedom and probability 95% the critical value is 1.65

with the null hypothesis:
H_0: the time series has a unit root.

To apply this test, you need to have a "testing strategy" as suggested in (Elder and Kennedy, 2001) which "exploits known information about the growth character of the variable under investigation". So:

- if your original times series is evolving erratically around a non-zero mean like the unemployment rate, i.e. it doesn't grow in the long run ($\beta_0 \neq 0$), then you do not assume any *drift* in the first difference and on conduct an F test on the joint null that $\beta_0 = 0$ and $\gamma = 0$;
- if the original times series has a constant rate of growth ($\beta_1 \neq 0$), we say that the series have a *drift*, therefore conduct a F test on the joint null that $\beta_1 = 0$ and $\gamma = 0$;
- (Elder and Kennedy, 2001) discard the option $\beta_1 \neq 0$ as it implies an ever increasing (or decreasing) rate of change.

Additional data, you can refer to the book (Lütkepohl and Krätzig, 2004) and the detailed answer posted here.
(Elder and Kennedy, 2001) stress the importance of the correct model as

> if the trend term is erroneously omitted, the tests are biased toward finding a unit root. This is easy to explain. If there is a trend but no trend term included in the regression, the only way the regression can capture the trend is by estimating a unit root and using the intercept (drift) to reflect the trend. On the other hand, including a trend term in the regression when it is inappropriate reduces the power of the unit root tests, for the same reason that including irrelevant explanatory variables increases variance.

The monthly variables of unemployment, population and GDP are integrated of order 1.

16.9.1 Thorough augmented Dickey-Fuller test

If no obvious conclusion can be taken from observations of the data set, then we can resort to a thorough procedure for the augmented Dickey-Fuller test, with three competing models:

$$\Delta y_t = \beta_0 + \beta_1 t + \gamma y_{t-1} + \sum_{j=1}^{p-1} \delta_j \Delta y_{t-j} + \epsilon_t \tag{30}$$

$$\Delta y_t = \beta_0 + \gamma y_{t-1} + \sum_{j=1}^{q-1} \delta_j \Delta y_{t-j} + \epsilon_t \tag{31}$$

$$\Delta y_t = \gamma y_{t-1} + \sum_{j=1}^{r-1} \delta_j \Delta y_{t-j} + \epsilon_t \tag{32}$$

For each model, the lag order (p, q or r) can be selected by t-values of information criteria and one should check that with the chosen lag order the residuals have no serial correlation.
Then one starts following (Dolado et al., 1990) with the model 30 and tests for the null of β_1 if the null is rejected, the augmented Dickey-Fuller test can be performed with this model. If the null is not rejected, then one move on to model 31. Note that for the null tests we must use the asymptotic distributions tabulated in (Dickey and Fuller, 1981).
With the model 31, one tests for the null of β_0, if the null is rejected, the aumented Dickey-Fuller test can be performed with this model. If the null is not rejected, then one move on to model 32 and perform the augmented Dickey-Fuller test with this model.

16.10 Building an AR(p) process for the French population

We found that the time series of the French population was I(1), so we now want to fit an AR(p) process on the French population growth rate.
A first naive approach is to fit an AR(1), we de-trend the growth rate, $\tilde{y}_t = y_t - \bar{y}$, with $\bar{y}$ the average growth rate observed over the period and then perform an ordinary least squares estimation of the linear regression:

$$\tilde{y}_t = \theta \tilde{y}_{t-1} + \epsilon_t$$

As we are working with a finite sample of size T and exogeneity can be safely assumed for shocks on past population growth rates but the exogeneity assumption doesn't hold for future population growth rate, then the OLS estimator $\hat{\theta}$ of θ is biased, $E\left(\hat{\theta}\right) - \theta \neq 0$. The demonstration will not be provided here, but you could try to convince yourself with the Example 2 of this lecture notes. To put it simply:

$$\hat{\theta} = \frac{\sum_{t=0}^{T} y_t y_{t-1}}{\sum_{t=0}^{T} y_{t-1}^2}$$

is unbiased if $E\left[\epsilon_t | y_0, \ldots, y_t, \ldots, y_T\right] = 0$ which for time series is not the case.
But it can be shown that in case $|\theta| < 1$, then the series is strictly stationary and ergodic and the the OLS estimator is consistent, meaning that as the size of the sample increases (as $t \to \infty$) the estimate tends to the quantity being estimated.

- **Question 27**
  - plot projection with this AR(1) and confidence interval using the OLS coefficient and the residuals distribution.

Log-likelihood estimate of an AR(1) if we assume that errors are normally distributed:
If we assume that $y_t - \theta y_{t-1} \sim \mathcal{N}\left(0, \sigma^2\right)$ and
For the first observation y_0, we have a specific estimation of its distribution: $y_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\theta^2}\right)$ (remember that $Var(y_t) = \theta^2 Var(y_{t-1}) + \sigma^2$).
as seen in section 11.4.2 on maximum-likelihood estimation, the log-likelihood is additive

$$log\left(L(\theta)\right) = -\frac{T-1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T-1}\left(y_t - \theta y_{t-1}\right)^2 - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{\sigma^2}{1-\theta^2}\right) - \frac{1}{2}\frac{y_0^2}{\frac{\sigma^2}{1-\theta^2}}$$

We implement the log-likelihood estimate for this AR(1) on the French population and compare this estimate with the parameter from the OLS.

- **Question 28**
  - Now that you have estimated an AR(1) on what you assume to be a weakly stationary time series, go back to the confidence interval introduced in section 15.1.2 and forecast and show a confidence interval with this simple model. Although the AR(1) is in returns, do all the plots in level for visual comfort and ease of interpretation.

We might want here to try manually to fit a AR(3) and look at the significance of the regression coefficients:

$$\tilde{y}_t = \theta_1 \tilde{y}_{t-1} + \theta_2 \tilde{y}_{t-2} + \theta_3 \tilde{y}_{t-3} + \epsilon_t$$

We will see starting section 17.8 how to select the order p of the AR, also look at moving average (MA) component and how to fit the model to the data. Methods used are maximum likelihood estimations and method of moments.

In simple terms, the Partial Autocorrelation Function (PACF) of order k can be estimated as P_k in the following AR(k) estimations:

$$\tilde{y}_t = P_1 \tilde{y}_{t-1} + \epsilon_{1,t}$$

$$\tilde{y}_t = \theta_{1,2} \tilde{y}_{t-1} + P_2 \tilde{y}_{t-2} + \epsilon_{2,t}$$

$$\tilde{y}_t = \theta_{1,3} \tilde{y}_{t-1} + \theta_2 \tilde{y}_{t-2} + P_3 \tilde{y}_{t-3} + \epsilon_{3,t}$$

$$\dots$$

The significance level at 95% for a series with T observations is[59] $\pm\Phi^{-1}\left(\frac{(1+0.95)}{2}\right)\frac{1}{\sqrt{T}}$ with Φ^{-1} the inverse of the cumulative distribution function of the normal distribution.
To conclude on this AR models, to chose the order p, you can introduce and minimize the Akaike's Information Criterion (AIC) which you compute at each order p:

$$AIC_p = -\frac{2}{T} log\left(L(\theta, p)\right) + \frac{2}{T} p$$

or the Schwarz-Bayesian information criterion (BIC):

$$BIC_p = log\left(\hat{\sigma}^2(p)\right) + \frac{log(T)}{T} p$$

with $\hat{\sigma}(p)$ the estimate of the innovation standard deviation from the maximum likelihood of the AR at order p. BIC tends to penalize more lag order $log(T)$ for large data set.
We usually strive for model parsimony (limiting the order p in our example), the first reason is the Occam's razor principle where the least assumptions you make the more likely your explanation. The second is the risk of overfitting where if you have T observations, you could reach $p \to T$ and we can mention in that sense (Ledolter and Abraham, 1981) that argued that each unnecessary parameter increases the variance of the prediction error by a factor of $\frac{\sigma^2}{T}$.
One might wonder which criteria to use between AIC and BIC, but we are not aware of one that outperform the other in every applications. BIC might lead to model that are too parsimonious, this is why as a practitioner we compare both. We discuss further lag order selection in section 24.2.
We use the procedure to fit an ARMA (although restricting MA order q to null for this exercise).
If our assumptions are correct, the residuals from the model should be a white noise. We apply a Ljung-Box test[60] removing the order p of the AR to the degree of freedom of the chi-square distribution.

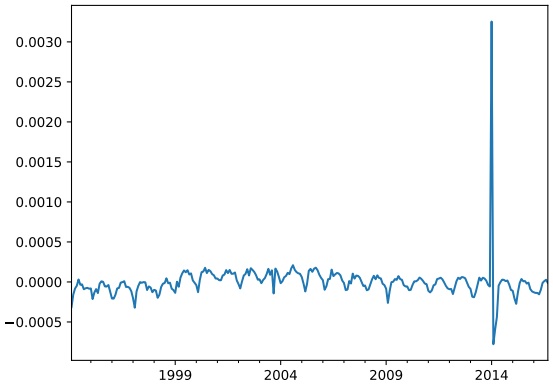
16.11 Are residuals independently distributed? Ljung-Box test

To test for serial correlation with a maximum lag h in the residual time series returns, we compute the Ljung-Box q-statistic against a χ^2_h distribution as in (Ljung and Box, 1978):

$$Qstat = T(T+2) \sum_{k=1}^{h} \frac{\tau_k^2}{T-k} \tag{33}$$

with τ_k the sample autocorrelation at lag k.
For residuals of an AR model, we remove the order p of the AR to the degree of freedom of the chi-square distribution.
The H_0 of the Ljung-Box test is: no serial correlation at lag max.
With first a visual inspection of the AR(3) model residuals, they don't seem to be iid:

[59] which can be approximated directly with $\pm\frac{2}{\sqrt{T}}$
[60] see section 16.11 for more detail on this test



With this test, we reject the hypothesis that residuals are independently distributed, but we conclude that residuals are rather serially correlated.
Our model is not satisfactory.

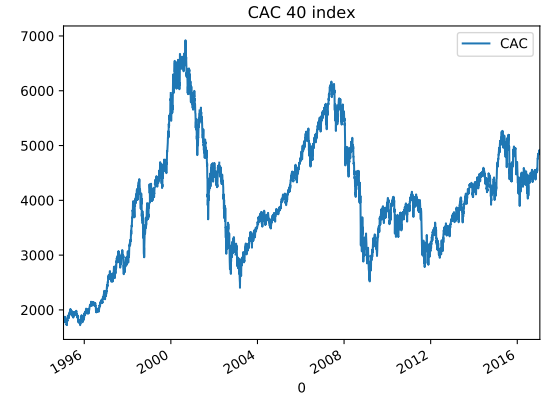
- **Question 29**
  - Remove seasonality in the time series and apply again an AR(3) model.
  - After seasonality is removed, use the AIC and BIC to select order p and q of an ARMA(p,q). Are residuals serially correlated?

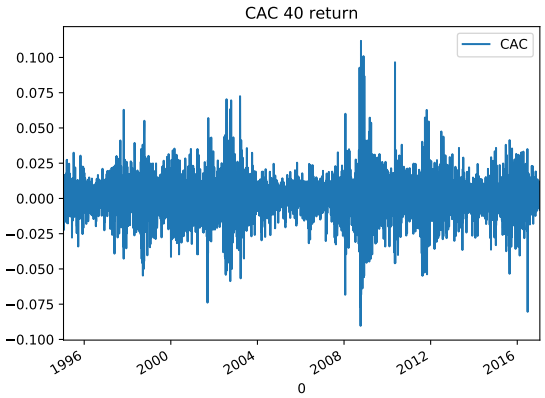
In fact, seasonality for birth rate in France is a well documented fact as INED showed, as such the 23rd of September has 5% additional births as a usual days, this day is exactly 256 days after the 31st of December.

17 Python: financial returns distributions and forecast

17.1 CAC 40 index data: mean and two-sided t-test, histogram

We fetch[61] CAC 40 index historical data from 1995-1-16 to 2017-1-16. We compute the daily return on business days only.





17.2 Modelling and forecasting the CAC 40 daily return

One of the simplest question we could be interested in when it comes to financial time series is: "what is its value most likely to be tomorrow?"
If we assume (wrongly) that the daily returns are iid and follow a Gaussian DGP, then we only need to look at the histogram of the observed daily returns and can follow the logic presented section 11.2.1, we model the CAC 40 return as a white noise section 17.3.
But if we call $y_t = \frac{P_t}{P_{t-1}} - 1$ the daily returns, we can expect that the realization of y_t will depend on the realization of $y_{t-k}, k \in \mathbb{N}$.
One would take the historical data available for this time series, in our example, we have 5 638 observations from 1995-1-16 to 2017-1-16. We could choose the parameters based on the first 60% of our data set (train our model) and then verify with the remaining 40% observation that our model performs well (test our model). We might also want to cross-validate it, more on this topic can be found here.

[61] code: ARIMA_vansteenberghe.py

An I(1) time series can typically decomposed as:

$$y_t = \mu + D_t + ST_t + C_t$$

with:

- μ a constant
- D_t a drift or deterministic trend (e.g. μt)
- ST_t a stochastic trend, the Random Walk part
- C_t a cyclical part

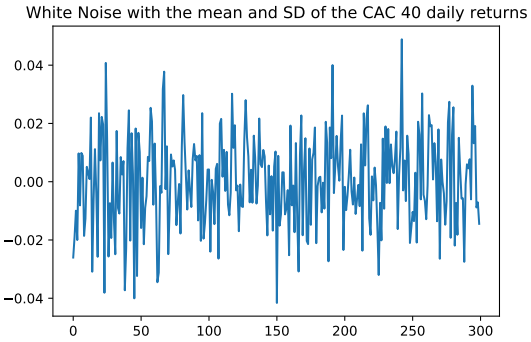
Once the deterministic trend has been removed, we can try to model our time series with an ARIMA, those concept are studied below.

17.3 White Noise: model and tests

A white noise is a is a sequence of independent and identically distributed random variables y_t with finite mean and variance. Most of the time we take the assumption that y_t is normally distribution
We use the simple model:

$$y_t = \mu + \epsilon_t$$

with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ and as we are working on modelling and forecasting the CAC 40 returns: μ is the mean of the daily returns



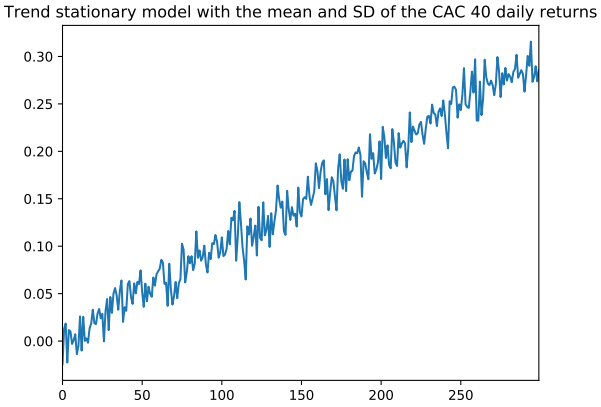
As defined in section 16.1, this series is by definition stationary and as expected, when we perform an augmented Dickey-Fuller test: the series comes out as stationary.

17.3.1 Trend stationarity

We modify our model with a trend:

$$y_t = \mu + \delta t + \epsilon_t$$

with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.
We take $\delta = 0.1\%$.



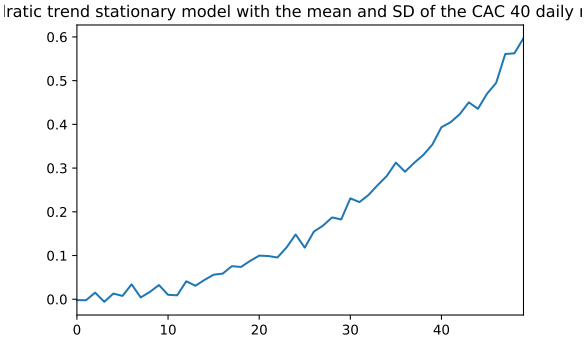
We perform an augmented Dickey-Fuller test taking the trend into account: the series is stationary. If we are interested in the significance of the trend component of this test, we might have to do it by hand. Let's consider:

$$\Delta y_t = \beta_0 + \beta_1 t + \gamma y_{t-1} + \delta \Delta y_{t-1} + \epsilon_t$$

We find that in this case, the trend component is significant.

We might want to remove the trend. We can start with a first difference.

We also study quadratic trend model.



17.4 Random Walk

If we think that the CAC 40 returns are following a random walk, then:

$$y_t = y_{t-1} + \epsilon_t$$

s.t.

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Here it is just that we use the previous business day's return to predict today's return: $E(y_t) = y_{t-1}$
This brings us to some of the measurements of forecasting errors:

17.5 Forecast error measures

- Mean Forecast Error: average the differences between the forecast and the actual data

$$\text{Mean Forecast Error} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_t - y_t)$$

- Mean Absolute Forecast Error

$$\text{Mean Absolute Forecast Error} = \frac{1}{N}\sum_{i=1}^{N} \mid \hat{y}_t - y_t \mid$$

- Root Mean Square Forecast Error

$$\text{Root Mean Square Forecast Error} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_t - y_t)^2}$$

17.6 Auto Regressive model

We follow section 16.3 that introduced AR(1) models.
Before starting, intuitively, an AR(1) model would mean that y_t and y_{t-1} related, therefore we can test for serial correlation of our daily returns, e.g. with a Ljung-Box test introduced section 16.11.
There are two reasons to ditch the CAC 40 returns for an AR(1) model (no serial correlation, our series looks more like a withe noise) and the AR(1) parameters are not significant. For a further study of ARCH effect in CAC 40 returns, see section **??**.
We introduce a stationary AR(1) on time series of better returns candidates and for stationary, we check that: $0 < |\theta| < 1$:

$$y_t = c + \theta y_{t-1} + \epsilon_t$$

with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, thus by definition (or assumption in our model if we apply it to a real time series) we assume that y_t and ϵ_t are uncorrelated.
We can search for the optimal value of ϕ, once we re-ordered the AR(1), detrending the y_t time series to get rid of the intercept in the AR(1):

$$y_t - \mu = \phi(y_{t-1} - \mu) + \epsilon_t$$

If $\phi < 1$ we can iterate backward :

$$y_t - \mu = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$$

If we note $x_t = y_t - \mu$, $E_t(x_t) = \phi x_{t-1}$. That is why AR(1) is usually a benchmark model against which one assess the relevance of a more complicated model. If a complex model cannot beat an AR(1) forecast, then we might just stick to an AR(1) model to base our predictions on.

$$\gamma(h) = cov(x_{t+h}, x_t) = \frac{\sigma^2 \phi^h}{1 - \phi^2}$$

Hence the Autocorrelation Function (ACF) of an AR(1) is: $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h$, $h \geq 0$
If $|\phi| > 1$ the AR(1) process is explosive.

- **Question 30**
  - as defined in section 16.1, show numerically that this process is weakly stationary for a certain range of values for ϕ.

17.6.1 AR(p)

The definition of an AR(p) model is simply:

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t$$

17.7 Moving Average model

MA(1) model:

$$y_t = \epsilon_t + \theta \epsilon_{t-1}$$

with

$$E(y_t) = \mu$$
$$\epsilon \sim \mathcal{N}(\mu, \sigma^2)$$

If we note: $\gamma(h) = cov(y_{t+h}, y_t)$

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma^2, & \text{if} \quad h = 0 \\ \theta\sigma^2, & \text{if} \quad h = 1 \\ 0, & \text{if} \quad h > 1 \end{cases} \tag{34}$$

Hence the ACF of an MA(1) is:

$$\begin{cases} \frac{\theta}{1+\theta^2}, & \text{if} \quad h = 1 \\ 0, & \text{if} \quad h > 1 \end{cases} \tag{35}$$

We can demonstrate that there is a problem of non-uniqueness of a MA(1) model (take θ or $\frac{1}{\theta}$ and an adjusted value of σ for example).

17.7.1 MA(q)

The definition of an MA model of order q is simply:

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$$

17.8 ACF and PACF

The autocorrelation function (ACF) of a stationary time series is:

$$\rho_k = corr(y_t, y_{t+k})$$

We call it a partial autocorrelation function (PACF) of a stationary process when we filter out the effect of the random variables $y_{t+1}, \ldots, y_{t+k-1}$. The PACF is used to identify the order (p) of an autoregressive model, i.e. AR(p).

$$\Phi_k = corr(y_{t+k+1} - \hat{y}_{t+k+1}, y_{t+1} - \hat{y}_{t+1}) \tag{36}$$

and

$$\Phi_1 = \rho_1 \tag{37}$$

with $\hat{y}_{t+k+1}$ (respectively $\hat{y}_{t+1}$) being the orthogonal projection of y_{t+k+1} (respectively y_{t+1}) on $y_{t+1}, \ldots, y_{t+k}$. We can use the ACF and the PACF as visual tools to identify respectively the order of the Moving Average and the Auto Regressive one:

- if for a time series the auto-correlation are non significant $\forall h > q$, we choose q as the order of the MA component of our model.
- if for a time series the partial auto-correlation are non significant $\forall h > p$, we choose p as the order of the AR component of our model.

17.9 ARMA model

A process y_t is an ARMA(p,q) if it is stationary and:

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$$

with $\phi_p \neq 0$ and $\theta_q \neq 0$
We can follow the Box-Jenkins methodology:

1. identify the order of the model p and q, this can be done with the help of ACF and PACF following section 17.8. We should also investigate for seasonality, see section **??**.
2. estimate the model with the maximum likelihood method[62].
3. model residuals diagnosis, e.g. with Breusch-Godfrey test described in section **??**.

For our time series, we reject H_0 for the Breusch-Godfrey test and residuals seems to be autocorrelated, our model is not acceptable and we would need to search for a better model.

Topics to further explore: parameters redundancy, causality, invertibility

17.10 ARIMA

Most time series can be decomposed in a nonstationary trend component and a zero-mean stationary composant:

$$y_t = \mu_t + x_t$$

such that

$$\mu_t = \beta_0 + \beta_1 t$$

x_t is stationary thus $\Delta y_t = \beta_1 + \Delta x_t$ is stationary.

A process x_t is said to be ARIMA(p,d,q) if $\Delta^d x_t$ is ARMA(p,q).

17.11 R: Box-Jenkins methodology for the ARIMA

We follow[63] the Box-Jenkins methodology to fit an ARIMA model as introdued in (Box and Jenkins, 1970).

[62] by default, the ARIMA funcion maximizes the conditional sum of squares likelihood
[63] 20190625_Box_Jenkins_methodo.R

17.11.1 Identification

1. To identify the order of the ARIMA(p,d,q), we plot the ACF and PACF.
2. We difference the time series until the ACF looks like one of a stationary process.
3. We use the inverse autocorrelation function to check for over-differencing.
    - Following the details here, we compute the dual model with the Yule-Walker method. The ACF of this model is called the inverse autocorrelation function, if this latter decays slowly it is likely we over-differenced the original time series.

17.11.2 Parameters estimation

Once we identified the orders p,d and q of the ARIMA, we estimate the parameters, e.g. using AIC criterion and we can simulate our model. Again, here we are not looking for a perfect reproduction of the historical time series, rather we want to check that our model can replicate the behavior of the original time series.

17.11.3 Checking the model

Once estimated, we check the error terms using the Ljung-Box[64] test to test the null hypothesis: H_0: the error terms are independent.

- **Question 31**
    - Us the long history of U.S. GDP available here, apply the tests and then an ARIMA model. Discuss the performance of your model.

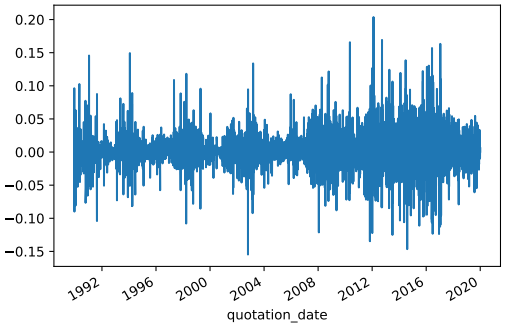
17.12 Main take away and the need for ARCH

We want to forecast the the returns and for this explore a DGP that would be 'behind' the observations. If the observations were iid, then $E\left(y_t|y_{t-1}\right) = E\left(y_t\right) = \bar{y}$. Here this is not the case and a simple AR(1) model beat a forecast only based on the mean. We next need to explore the residuals of this AR(1) model and check if they are iid with a 'well behaved' DGP.
If we observe the residuals ξ_t of an ARMA(1,1) model of the daily returns, we observe as in Mandelbrot (1963) writes it:

> large changes tend to be followed by large changes –of either sign– and small changes by small changes

and following Black (1976) volatility tends to rise with "bad" news and fall with "good" news:

[64] see section 16.11 for more detail on this test



To account for those properties in our residuals, we follow Nelson (1991) for the model notations.

17.13 Testing for the presence of ARCH

17.13.1 ARCH(1) model

If y_t represents the returns of a financial time series, then to model those returns as an ARCH as in (Engle, 1982):

$$y_t = \mu_t + \sigma_t \epsilon_t = \mu_t + a_t$$

with $\epsilon_t \sim \mathcal{N}(0,1)$ and $\mu_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \sigma_{t-j}\epsilon_{t-j}$. In an ARCH model we focus on the volatility and write:

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \ldots + \alpha_m a_{t-m}^2$$

In the case of an ARCH(1), $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2$
We have $E[a_t] = 0$ and $var(a_t) = E[(\sigma_t\epsilon_t)^2] = E[\sigma_t^2] = \alpha_0 + \alpha_1 E[a_{t-1}^2]$ thus: $var(a_t) = \frac{\alpha_0}{1-\alpha_1}$ which is constant over time.
We usually test this on model residuals u_t:

$$\hat{u}_t^2 = \alpha + \beta \hat{u}_{t-1}^2 + \nu_t \tag{38}$$

We can test for the presence of ARCH with 1) a Ljung-Box test presented section 16.11 or 2) a **Lagrange Multiplier test** (LM test) as in (Engle, 1982).
If the residuals are iid, then so are the squared residuals, if not then we need to address this in our model to get iid innovations (ν_t if an ARCH(1) is sufficient).

17.13.2 Li and McLeod test for heavy-tailed finite samples

For heavy-tailed finite samples, the Ljung-Box test may fare poorly and we prefer the (Li and McLeod, 1981) test statistic against a $\chi^2(k^2 m)$ distribution :

$$LMcstat = T^2 \sum_{i=1}^{m} \frac{1}{T-i} b_i'(\hat{\rho}_0^{-1} \otimes \hat{\rho}_0^{-1}) b_i$$

with the sample cross-correlation matrix $\hat{\rho}_i$ and $b_i = vec(\hat{\rho}_i')$.

17.14 GARCH models

We follow Nelson (1991) notations:

- ξ_t a model's prediction error;
- σ_t^2 the variance of ξ_t given information at time t

1. $\xi_t = \sigma_t z_t$
2. $z_t \sim$ iid from GED[65] density with $E(z_t) = 0$ and $Var(z_t) = 1$

a GARCH model writes

$$\sigma_t^2 = \omega + \sum_{q=1}^{q} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{p} \alpha_j z_{t-j}^2 \sigma_{t-j}^2 \tag{39}$$

where $\omega, \beta_i, \alpha_j$ are nonnegative.

17.15 AR(1)-ARCH(1)

Nelson (1991) details how to use maximum likelihood methods to fit an ARMA(p,q) (and potentially exponential) ARCH model and discuss the regularity conditions to be checked, but as in this paper, we will not check this as this stage ans assume that the ML estimator is consistent and asymptotically normal.
We fit an AR(1)-ARCH(1) $\sigma_t^2 = \omega + \alpha_1 z_{t-1}^2 \sigma_{t-1}^2$ and an AR(1)-GARCH(1,1) $\sigma_t^2 = \omega + \beta_i \sigma_{t-1}^2 + \alpha_1 z_{t-1}^2 \sigma_{t-1}^2$ model on the CAC 40 daily returns.

17.16 AR(1)-EGARCH

To address:

- the negative correlation between current returns and future returns volatility;
- parameters restrictions of GARCH model that are often violated;
- shock persistence interpretation difficulty with GARCH models,

Nelson (1991) suggest the EGARCH model:

$$\begin{cases} R_t = a + bR_{t-1} + c\sigma_t^2 + \xi_t \\ \xi_t = \sigma_t z_t \\ \ln\left(\sigma_t^2\right) = \omega + \sum_{i=1}^{p} \alpha_i \left(|z_{t-i}| - \sqrt{\frac{2}{\pi}} \right) + \sum_{j=1}^{o} \gamma_j z_{t-j} + \sum_{k=1}^{q} \beta_k \ln \sigma_{t-k}^2 \end{cases} \tag{40}$$

as in Nelson (1991), we keep $c = 0$ and we would need to select the order p, j, and q with information criterion.

- **Question 32**
  - Are the standardized residuals z_t of the AR(1)-GARCH or EGARCH model iid?

In equation 40, we have a linear form of "GARCH-in-mean" with the constant c. We could also include any type of function (possibly non-linear) $g()$ and update equation 40 with $R_t = a + bR_{t-1} + g(\sigma_t) + \xi_t$. We will explore section **??** how with the package rugarch we can use the options in "mean.model" with "archm Whether to include ARCH volatility in the mean regression" and "archpow Indicates whether to use st.deviation (1) or variance (2) in the ARCH in mean regression".

[65] Generalized Error Distribution

17.17 GARCH model checking

17.17.1 Significance of the coefficients

Beyond checking coefficients p-values, there is the Lagrange Multiplier Test (LM Test), for a parametric model with true parameters θ_0, $H_0 : R\theta_0 = r$, where we can test that some coefficients in θ_0 are null (imposing some 0s in R and r).
Following Francq and Zakoïan (2009), use the Wald, LM or Likelihood Ratio statistics.

17.17.2 Test on the residuals and squared residuals

Ljung-Box test on the residuals (auto-correlation), squared residuals (ARCH) and ARCH LM test (presence of ARCH).

17.17.3 Test of second order stationarity

To test the second order stationarity, you want to test $H_0 : \sum_i \beta_i + \sum_j \alpha_j < 1$, with the parameters taken from equation 39, against $H_1 : \sum_i \beta_i + \sum_j \alpha_j \geq 1$.

18 Python: assets risk measures, VaR and ES

Before investing in financial assets an investor might want to first study the risk it exposes itself to. If one believes that past performances are a good indication of future performances, one can use historical return data. A simple approach is to consider that returns of an assets followed a given distribution in the past and will follow a similar distribution in the future. In the following we will test the adequacy of fitting a normal, t Student or Levy stable law to a list of asset returns.
Once we consider that a distribution is suitably fitted to each asset, we can then either extract random draws from these distributions to assess the portfolio total risk, and/or compute the Value at Risk (VaR) of Expected Shortfall (ES) defined hereafter.
We suggest[66] to use daily price data from 3 667 assets over the period 01/01/1998 to 06/08/2018. This is an interesting period as one captures the "dot com" and the "sub-prime" crises.
As defined in (BIS, 2009), VaR is used by central bank regulator to determine the capital charge[67] required for certain activity

> the trading book capital charge for a bank using the internal models approach for market risk will be subject to a general market risk capital charge (and a specific risk capital charge to the extent that the bank has approval to model specific risk) measured using a 10-day value-at-risk at the 99 percent confidence level and a stressed value-at-risk.

and the financial institutions

> will also be required to use hypothetical backtesting at least for validation.

Having in mind the historical or normal approach to compute a VaR, one might wonder if there is the need for a more complex approach. (Danielsson et al., 2018) provide some answers to this question. "[E]stimate the worst case by taking the most negative outcome in the historical sample" or "estimate the lower tail of the distribution by semi-parametric methods", they find "that either method is best, depending on how heavy the tails are and their specific shape. Generally, for the heaviest, the semiparametric approach is best, and as it thins, the historical minima eventually becomes better."
If we define the **rate of loss** of an asset i over time by $R_{i,t} = -\frac{P_{i,t}-P_{i,t-1}}{P_{i,t-1}}$.

18.1 Value at Risk

For a portfolio composed of assets with iid returns, the Value-at-Risk (VaR) summarises the worst possible loss of a portfolio at a given confidence level $q \in (0,1)$ over a given time period $[t, t+1]$. VaR_q is such that the rate of loss of the portfolio won't exceed that level over that period with a probability inferior or equal to $1-q$:

$$VaR_q = \inf\{x : \mathbb{P}(R > x) \leq 1-q\} \text{ or}$$
$$VaR_q = F^{-1}(q) = \inf\{x : F(x) \geq q\}$$

where $F(x) = \mathbb{P}(R \leq x)$, $x \in \mathbb{R}$ is the distribution function of the rate of loss R and F^{-1} the generalized inverse of F.

[66] VaR_ES_vansteenberghe.py
[67] for simplicity, a capital charge can be understood as the capital a financial institution has to "keep safely aside" to be able to perform some risky activities

For example, in a year with 252 trading days we might want to know the worst return we are likely to face one day out of those 252, that is VaR at 99.6% over a year.
There are two categories of methods to compute VaR: those using historical rate of loss directly, called historical simulation or nonparametric methods, and those relying on statistical models, called parametric methods as they rely on parametric distribution.
Historical simulations rely on the assumption that the data set is representative of the future behaviour of the asset. If we consider the sorted realizations of a random variable $(R_{n,1} \leq R_{n,2} \leq ... \leq R_{n,n})$ then the VaR can be estimated as:

$$\widehat{VaR}_q = R_{n,s}, \text{ with } s = [nq] + 1$$

where $[x]$ represents the integer part of x.
The main drawback of the historical computation is that it is not possible to make the VaR depend on the current state of the market, namely to update volatility.
Some parametric models explicitly include the volatility to predict the VaR, the exposure to risk is conditional on the period of observation.

18.1.1 VaR: parametric methods

1. we suppose that the return are following a law
2. we compute the mean $\mu = E(X)$ and the standard deviation σ
3. if it is a Student t law, we need to choose a degree of freedom ν
4. we use a given quantile of the law [68]

If we assume that R follows a Gaussian distribution: $R \sim \mathcal{N}(\mu, \sigma^2)$, then its VaR_q can be estimated as

$$\mathbb{P}(R \leq VaR_q) = \mathbb{P}\Big(\frac{X-\mu}{\sigma} \leq \frac{VaR_q - \mu}{\sigma}\Big) = q.$$

If z_q is the q-quantile of a gaussian law $\mathcal{N}(0,1)$, then VaR_q is defined as:

$$VaR_q = \mu + \sigma z_q.$$

We can estimate VaR_q replacing μ and σ^2 by the estimations: $\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$ and $\bar{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R}_n)^2$, then VaR_q can be estimated as:

$$\widehat{VaR}_q = \bar{R}_n + \bar{\sigma}_n z_q.$$

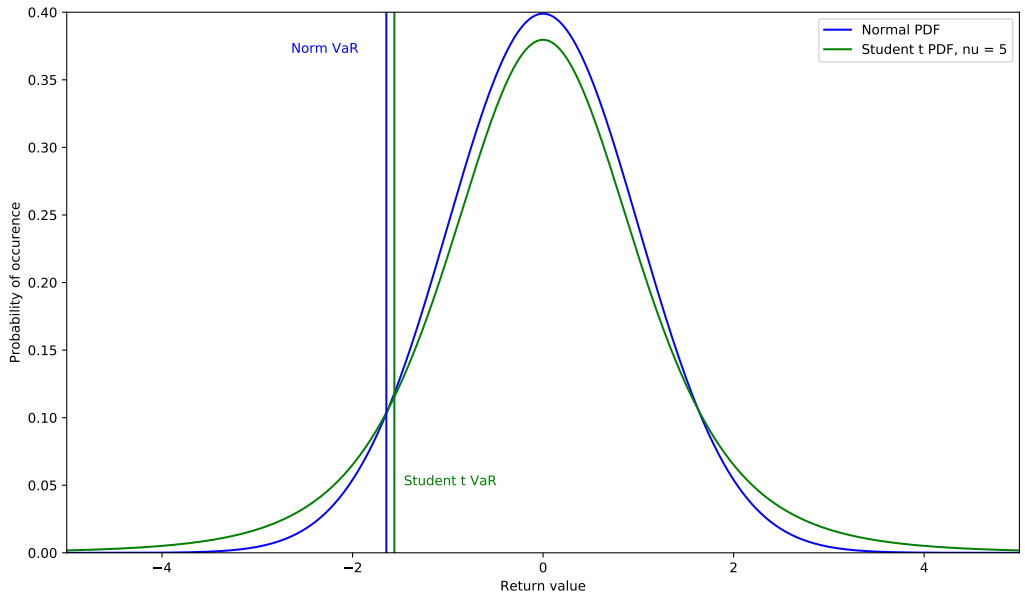
If we denote by Φ the percentage point function which is the inverse of the cumulative distribution function[69] is:

$$\text{VaR}_\alpha = \mu - \Phi(1-\alpha)\sigma$$

For the generalized Student's-t distribution, we use μ' the location, σ' the scale and ν the degree of freedom[70]. For $\nu > 2$, we have $\mu_{\text{sample}} = \mu'$ and $\sigma_{\text{sample}} = \sigma'\sqrt{\frac{\nu}{\nu-2}}$, hence with t the percentage point function of the Student's-t:

$$\text{VaR}_q = \mu + t(1-q)\sigma\sqrt{\frac{\nu-2}{\nu}}$$

[68] for the normal law, it is 1.65 for 95$^{\text{th}}$, 2.33 for 99% and 2.65 for 99.6%
[69] in simple terms, $-\Phi(1-\alpha)$ find the return level z on the left of which the area under the pdf is α
[70] For the details of the calculus, cf. (Alexander, 2009)



Monte Carlo approach: If we are not certain about the inversion of the distribution function, we can generate a large number of returns from the distribution and then compute the VaR and ES for this sample.
Some asset managers declare their VaR policy in the funds brochure, e.g. from SEEYOND EQUITY VOLATILITY STRATEGIES:

> the Global Exposure Risk is managed using the absolute Value-at-Risk approach (VaR approach). The VaR approach measures the maximum potential loss at a given confidence level (probability level) over a specific time period, under normal market conditions. The absolute VaR for the Sub-fund cannot exceed 20% of its net asset value in a confidence interval of 99% for a 1-month holding period (20 available working days).

18.1.2 historical VaR

1. we rank the observed returns
2. we consider the 0.4% worst for the VaR 99.6% or the 5% worst for the VaR 95%

As in (Dominique Guegan and Li, 2017), let $X_1, \ldots, X_n$ be a sequence of losses. The historical VaR at the p-quantile is: $VaR_p = X_m$ where $m = np$ if np is an integer and $m = [np] + 1$ otherwise[71].

18.2 Expected Shortfall

The **Expected Shortfall** is the mean of the returns once the VaR has been crossed, over a given period. It is also named C-VaR or Tail-Loss:

[71] The operator [] is the largest integer inferior

$$ES(\alpha) = E[r|r < VaR(\alpha)]$$

in practice, with ψ the standard normal density function[72]

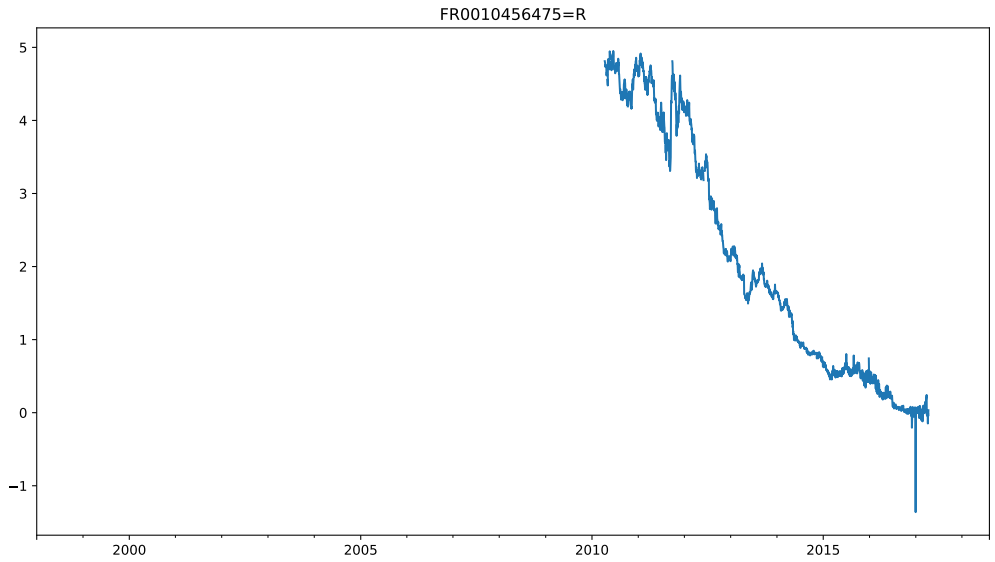
$$\mathrm{ES}_\alpha = \mu - \alpha^{-1}\psi\left(\Phi(\alpha)\right)\sigma$$

18.3 Other mean to compute the VaR and ES from a distribution

If we are not certain about the inversion of the distribution function, we can generate a large number of returns from the distribution and then compute the VaR and ES for this sample.

18.4 Data preparation

Before exploiting our data set, we need to prepare the data.
First, we might want to limit our observations to trading days in order to make sure that we won't capture null returns on week-ends and overweight 0 in our histogram. We filter our data frame rows on the trading days.
Second, we want to distinguish between real 0's and missing values. It can happen that instead of a NA the value 0 is given in the data set. More problematic are negative price values. We investigate the minimum price value per asset. For those with negative and 0 values we do a visual inspection. For example we can have one observation at 0, in which case we migth want to keep the asset in the data set but remove the observation with a 0 value (we might replace that 0 with an average between the previous and the next observation:



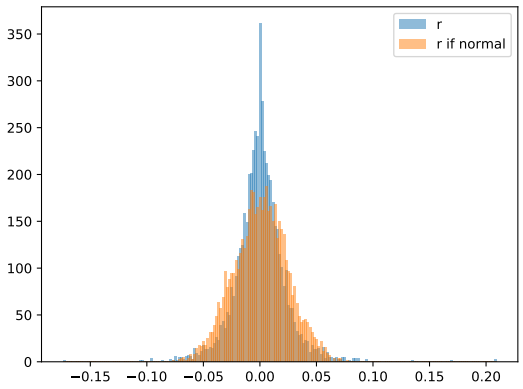
We can also investigate price time series which are either stuck at 0 by computing successive 0's in the time series or for returns stuck at 0, then by visual inspection we can decide if we keep the asset or not.

[72] for the Student's-t ES, one can look into the details here

If the data set is so large that we cannot commit to visual inspection we can be more brutal in dropping potential anomalous series.
We can drop series of which observation are below a given threshold compare to the rest of the data set.
We can also deal with anomalous return values by winsorizing the returns data set. Winsorizing is simply putting a positive and negative limit on the values the returns can take, for example if we winsorize returns between -1 and 1: $R_{\text{winsorised},t} = \max\left[\min\left(1, R_{\text{original},t}\right), -1\right]$. This approach is different than just getting rid of series where there exist a return for which $|R_{\text{original},t}| > 1$, if our data set is large enough, we might prefer to get rid of series with anomalous returns as winsorizing can introduce bias.

18.5 Beyond normal law: t-Student, Levy stable

Now we want to check whether those returns are normally distributed, meaning that the daily return. We do a first visual check by comparing the histogram of actual BNP's daily returns and draws from the normal distribution:



Visually it seems that both distributions differ, we need to use a statistical test to determine whether both distributions can be consider similar or different: we use here the Kolmogorov-Smirnov[73] test where the null hypothesis H_0 is that both groups were sampled from populations with identical distributions.
The two-sided test uses the maximum absolute difference between the cdfs of the distributions of the two data vectors. The test statistic is:

$$KS = \max_x |\hat{F}_1(x) - \hat{F}_2(x)|$$

The p-value of the test is low enough that we can reject H_0 and the normal distribution is not adequate to model BNP's daily returns over the period.

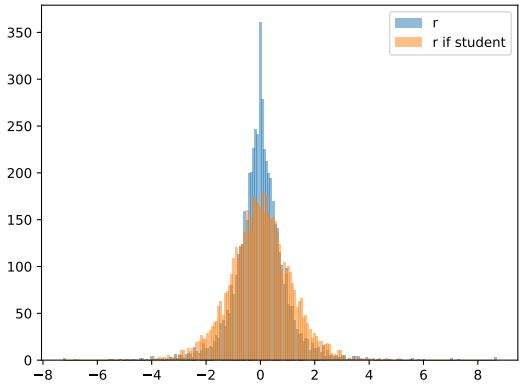
18.5.1 Student t distributions and stable laws

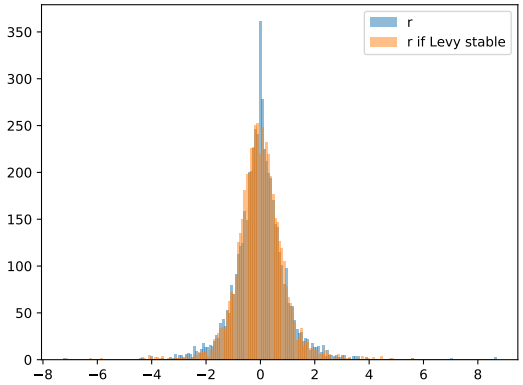
The normality of financial asset returns distribution has been question since the work of Mandelbrot and Fama in the 60's. Other distributions have been suggested and we explore two:

- Student t distributions that present fat tails and variance;

[73] we suggest another normality test in section **??**

- Stable laws that present fat tails and skewness.





We fit the normal, t Student and Levy stable distributions on the empirical returns for each of the 3 667 assets and compute the p-value of the two sided Kolmogorov-Smirnov test[74]. We keep respectively 82, 71, and 419 candidates for the normal, t Student and Levy stable laws.

18.6 Parametric VaR and ES when considering normal, Student's and Levy distributions

For each of the 2 866 assets we compute the parametric and historical VaR and ES. We find that for a large share of the assets the parametric VaR and ES based on normal assumption are less conservative than the historical observation. This is likely due to a poor fit of the normal distribution, namely the existence of fat tails.
We find that the median VaR and ES are too optimistic in the normal and Student's t case while too extreme for the stable law:

[74] a preliminary remark: we would need to generate several random samples from the distributions and compute each time the p-values in order to approach the real p-value of the test

historical VaR	-0.028
parametric VaR normal	-0.021
parametric VaR student	-0.024
MC VaR levy	-0.049
historical ES	-0.043
parametric ES normal	-0.023
parametric ES student	-0.031
MC ES levy	-0.347

We can observe that the Levy stable law can be too conservative while the normal and Student's t law are not conservative enough.

- **Question 33**
  - As established by Artzner et al. (2007), show some numerical applications that the VaR does not verify the subadditivity property (incoherent risk measure).

19 Python: Manipulating financial data and some investment strategies

19.1 Share price history

We wrap up what we've learn so far in simple examples[75]:

- get data from the internet or from a file
- *clean* the dataset
- indexing and joining, merging a DataFrame
- plot and save the data
- create a function (help on this or here) to improve the harvest of data
- select and search through data via for and/or while loops (example)
- observe descriptive statistics
- simulate investment decisions, assess their performance

19.1.1 Data set cleaning

We inspect our data set and clean for missing data or "suspicious" data points (this can be peaks that cannot be explained). In our first approach, sort each stock price time series per maximum deviation factor to its mean, the deviation factor is computed as:

$$\max \frac{P_{i,t} - \bar{P}}{\sigma}$$

with $\bar{P}$ the mean of the price over the observed period and σ the standard deviation of the price over the observed period. If we find some anomalies or outliers[76], we need to first remove those anomalies before proceeding to our analyses. Also, we can check for prices or returns stuck at 0, another indicator of missing data.
Nota bene: if after some period the price is stuck at a given level, this could be a good indicator of a default of a company. In this exercise we are likely to remove such asset from the sample which give a positive bias to any trading strategy as we could have picked a company that would have defaulted and lost our position.

19.1.2 Daily returns

We can now compute the daily returns for the stock i:

$$\text{return}_{i,t} = R_{i,t} = \frac{P_{i,t}}{P_{i,t-1}} - 1$$

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

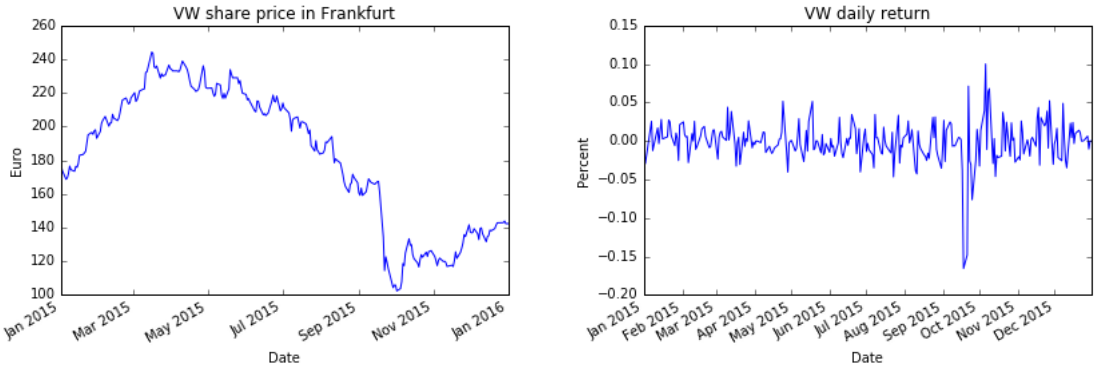
$$\ln(x) = \ln(1) + \frac{x-1}{1} - \frac{(x-1)^2}{2} + \dots$$

[75] financial_data_vansteenberghe
[76] you can find section 13.1 some further approaches to detect outliers

If the price is not too volatile over time, that is if $\frac{P_{i,t}}{P_{i,t-1}} \simeq 1$, then we could also compute[77] the daily returns for the stock i:

$$R_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right) \tag{41}$$

We can then plot the evolution of the share price and its daily return, we clearly observe the emission scandal effect on Volkswagen share price and returns:



We provide 10 years of price data for stock which would be eligible for a French "PEA": a selection of 495 shares in `pea_price.csv` and `pea_price_2.csv`, with time series from 2006 to 2016.

19.1.3 Share's return in practice

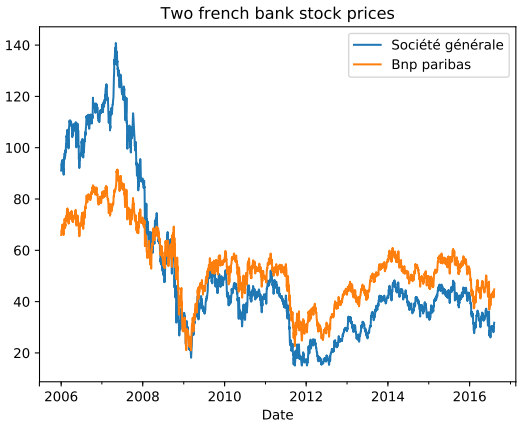
In practice, as an investor, you are not only interested in the growth of the share price, but also in the dividend D_i the company will pay[78]

$$\text{return}_{i,t} = R_{i,t} = \frac{P_{i,t} - P_{i,t-1} + D_{i,t}}{P_{i,t-1}}$$

There are also some split decision where share can be split or repurchase. This is why we work with **Adjusted Closed** values which account for both the dividends and the splits.

19.1.4 A simple regression, but a useless one

We might want to work on two french banks' stock price evolution, BNP and Societe Generale (GLE):



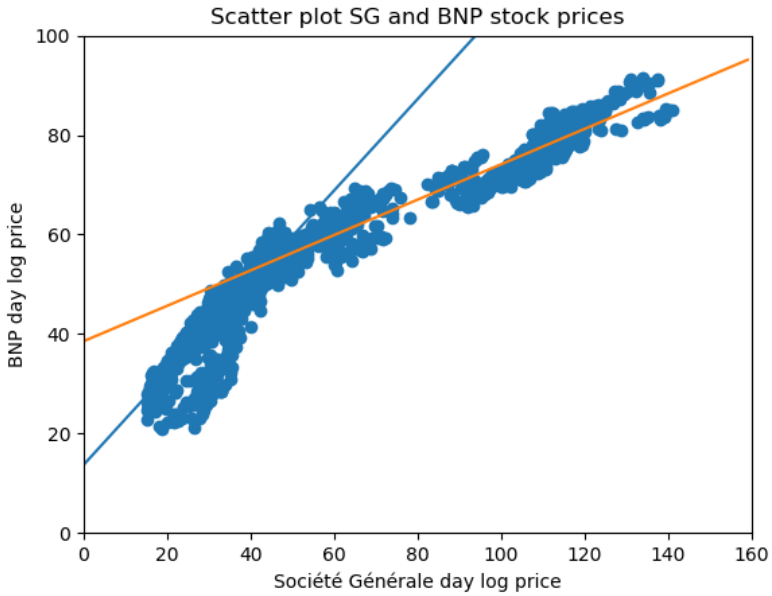
[77] $\ln\left(\frac{x_{i,t}}{x_{i,t-1}}\right) \simeq \frac{x_{i,t}}{x_{i,t-1}} - 1$

[78] we saw in section 8.4 the importance of dividends for the total return for an investor in the stock market

Next, just as illustrations, we suggest a simple regression (Ordinary Least Square) of BNP daily log prices on Societe General daily log prices, with a constant:

$$y_{\text{BNP},t} = \alpha + \beta x_{\text{Societe Generale},t} + \epsilon_t \tag{42}$$

we find a β of 0.45, but in fact this should not be interpreted as a sensitivity between the two time series.
We might want to separate and execute two regressions, one where the Societe Generale stock price ies below 50 and the rest, we obtain a plot that looks better:



It looks as if we had something, but it is a catastrophe in terms of econometrics. You can see in section **??** why (think about stationarity, cointegration, etc.). In other terms, we might be looking at a spurious regression.

- In fact, we perform the following
  1. ADF test on the series in level and in first differences.
  2. Both series are I(1)
  3. We do a cointegration test over the full sample and the time series are not cointegrated. Doing a cointegration test for the dates after 2008 the time series seems to be cointegrated.

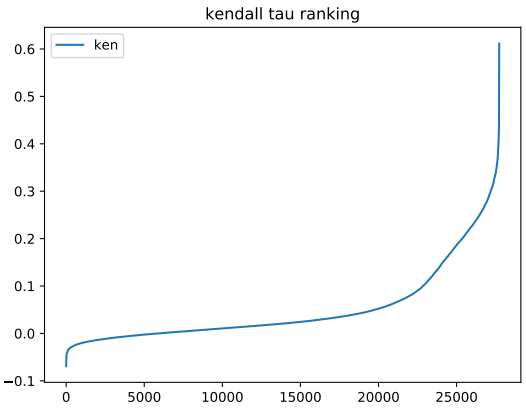
- **Question 34**
  - Perform an ECM for the dates after 2008.
  - measure correlations of prices and returns of both times series. What can you say with respect to section 11.13.7

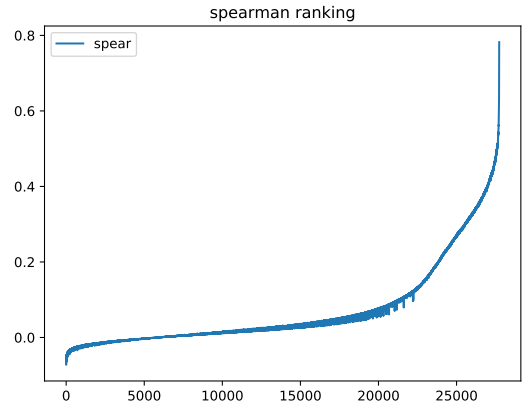
19.1.5 Comparing correlation measures

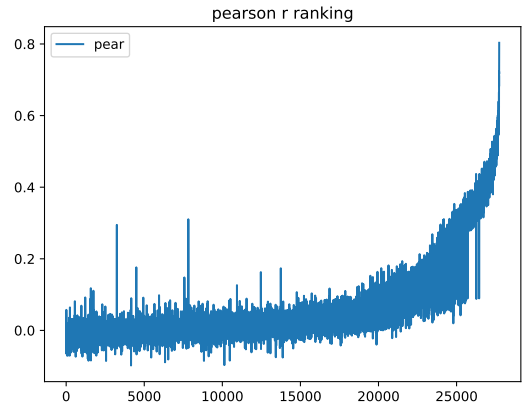
On our samples, we are comparing our correlation measures on returns[79] Pearson, Kendall, and Spearman, following the notions introduced in 11.13.

[79] we also suggest reading these elements

We sort our DataFrame according to the Kendall coefficient and check whether this is also monotonous for the Pearson and Spearman:







There is some "noise" in the Pearson plot, meaning that our sorting based on Kendall is not the same as it would be for Pearson.
As this is intensive in terms of computing power, we suggest to run this via a command line[80]. With MacOS, you can run a Terminal and try

[80] from more tips on how to run a python script via the command prompt, see this web page

```
python '/pathtofile/filename.py'
```

With WindowsOS, you can run cmd.exe, and give a command line:

```
C:\pathtoAnaconda3\python.exe C:\pathtofile\filename.py
```

19.1.6 Manipulating those concepts on our sample

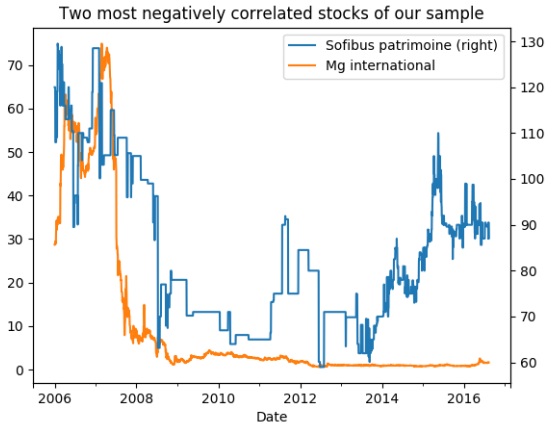
As a first simple approach of risk deversification, we can look through our 496 stocks which pairs are historically the most positively or negatively correlated. For this, we first compute the correlation matrix which looks somethings like this:

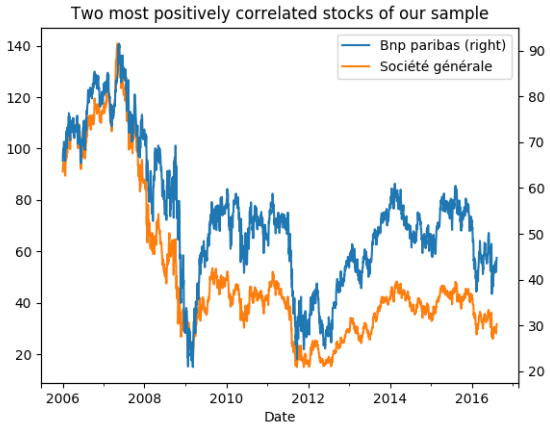
Index	FR0000079691 ISIN	FR0000062184 ISIN	FR0000066680 ISIN	FR0004035913 ISIN
FR0000079691 ISIN	1	0.489	0.838	0.0982
FR0000062184 ISIN	0.489	1	0.347	-0.224
FR0000066680 ISIN	0.838	0.347	1	0.0944
FR0004035913 ISIN	0.0982	-0.224	0.0944	1

We ignore the diagonal component, unstack this matrix to have a list of all pairs in this correlation matrix:

('FR0000079691 ISIN', 'FR0000062184 ISIN')	0.489
('FR0000079691 ISIN', 'FR0000066680 ISIN')	0.838
('FR0000079691 ISIN', 'FR0004035913 ISIN')	0.0982
('FR0000079691 ISIN', 'FR0010086371 ISIN')	-0.4
('FR0000079691 ISIN', 'FR0000120859 ISIN')	0.628

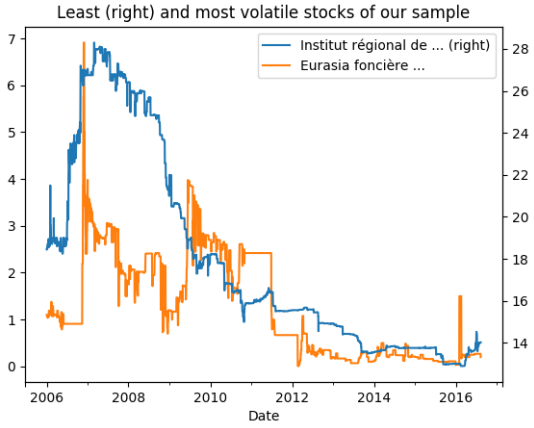
We are then in a position to sort this list from minimum to maximum and we can, as an illustration, select the first and last elements of this list to plot the two most negatively and positively correlated stocks of our 495 stocks sample:





In our sample, the two most correlated stock prices are the ones of the companies: Essilor International and Dassault Systemes.

With a similar approach we can detect the least and most volatile stock:



Is there a correlation between stock capitalisation and volatility in our sample?

1. We compute the average capitalisation of each firm over the period.
2. Then compute the standard deviation of each stock over the period.
3. Finally we regress the average capitalisation over the standard deviation and we reject the model.

19.1.7 Computing the return of a portfolio - an introduction

If we are working with Adjusted Close values of stocks, we can focus on the returns of the holdings (assuming that the dividends, splits, and other effects are taken into account in the prices).

We make the following assumptions for an unsophisticated investor:

- the initial cash to invest is 100,000 €
- each transaction cost (buy or sell) is 7 €

- we assume that based on a given criteria, our investor purchase the 10 best performers
- our investor dispatch uniformly its investment
- all the position are sold at the end of the investment period and the return is computed

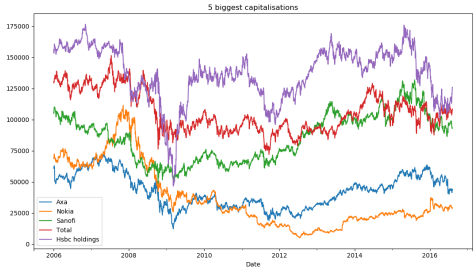
That is, if we have a sample of $N = 10$ best performers, a transaction cost TC the number of shares our investor will want to purchase for a performer i is:

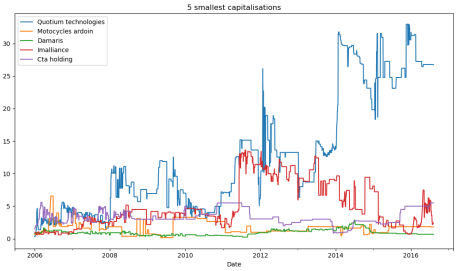
$$\frac{\text{Cash}_{t_0} - \text{TC} \times N}{N} \times \frac{1}{P_{i,t_0}}$$

We have to keep in mind that our cash is limited, hence we have to round this to the inferior unit and keep whatever is left in the form of cash.

19.1.8 Defining a scoring to allow basic selection

Historical correlation and volatility can be very basic selection criteria for a portfolio, the capitalisation as well. With the suggested sample, we can plot the biggest and smallest capitalisation (in K EUR) and observe different behaviours:





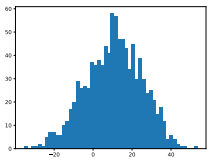
Based on what we have done so far, we can compute for each stock its daily returns. Then we compute the average return over the period, its volatility, and the average capitalisation of the company.
An investor could want to invest in stocks that historically showed positive return, low volatility, and high capitalisation (the stock is generally more liquid). We can define a naive scoring for a stock i out of our N sample:

$$\text{score}_i = \alpha \frac{E(R_i)}{\frac{1}{N}\sum_{j=0}^{N-1} E(R_j)} + \beta \frac{E(Cap_i)}{\frac{1}{N}\sum_{j=0}^{N-1} E(Cap_j)} + \gamma \frac{\sigma(R_i)}{\frac{1}{N}\sum_{j=0}^{N-1} \sigma(R_j)} \tag{43}$$

- **Question 35**
  - suggest some values for α, β and γ based on investor's risk aversion.
    1. Start with simple values of α, β and γ, sort your scores and compare with the returns one could expect with this selection.
    2. Separate the sample into a training and a test sample.
    3. Calibrate the weights on the training sample and compute the portfolio return (of the 10 best stocks selected) on the test sample. Is it convincing? Remember that the financial crisis happened during the training sample.
- **Question 36**
  - Introduce the skewness and kurtosis into the selection criteria.

To follow up on this questions, you might want to go to section **??**.

19.1.9 Testing a strategy

- **Question 37**
  - Define a strategy over time to trigger buying and selling orders.
    1. Using CAC 40 data (investment in CAC4 0 ETF is meant) compare the performance of your portfolio against the return of investing in a CAC 40 index.
    2. Simulate x times (100 or more) a random selection and create a list of returns. Compare this to the performance of a portfolio selected with your method. The distribution looks something around 10% like: 
    3. Try to vary the dates you buy and sell based on some criteria to improve the performance of your portfolio.

19.2 Share price history: focus on French biotech

As a practical exercise[81], we suggest to focus on the French and Benelux biotech sector over 5 year since September 2012 with a capitalization over EUR 100 million. Use data directly from Euronext and perform the following exercises:

- **Question 38**
  - Now you want to keep the information of each day's lowest (Valeur Basse) and highest values (Valeur Haute) taken by the stock
  - You start with a initial cash value of EUR 100,000, you buy or sell by amount of 2,000 maximum per stock per day, each transaction costs EUR 7 (buying or selling)
  - Each day, you are allow to choose to place an limit order, buy or sell, on all stocks based on your strategy, for the following day
  - If your buy order is above the lowest value of the following day, or your sell order below the highest value of the following day, your order is executed
  - Compute your return over the 5 years for your strategy and deliver an history day per day of how much of each stock you hold over the 5 years (this is for a future work on "**herding**")
  - Construct a portfolio selecting the 5 lowest volatility in the last 20 trading days, compare its performance with the performance above

- **Question 39**
  - Identify days and stock with "extreme variation", chose a criteria until you are left with a minimum of 10 event
  - On those dates and for the companies involved, search for relevant news that could explain the variations

[81] biotech_price_histories

19.3 Efficient Markets Theory: an introduction

If we go back to Eugene Fama in 1964 and confront his statement mentioned in (Shiller, 2016) that

> the average correlation coefficient between successive day' log price changes over the thirty Dow Jones Industrial Average stocks between 1957 and 1962 was only 0.03

19.3.1 The debate: contrarian versus momentum strategies and the efficient market hypothesis

When investing in the stock market, should one buy low and sell high or buy high and sell even higher? Momentum and contrarian strategies are still debated in the press like in this FT article. In this section, we won't pretend to lead to an answer, but we will explore some aspects with the data we have at hand: a selection of 495 shares in `pea_price.csv` and `pea_price_2.csv`, with time series from 2006 to 2016.
De Bondt and Thaler in their 1985 paper (Bondt and Thaler, 1985) show that over 3-to 5-year holding periods stocks that performed poorly over the previous 3 to 5 years achieve higher returns than stocks that performed well over the same period.
Jegadeesh and Titman 1993 paper (Jegadeesh and Titman, 1993)

> investigates the efficiency of the stock market

and provides an analysis of relative strength trading strategies, meaning buying stocks that showed high return over 3- to 12-month horizons. They find that

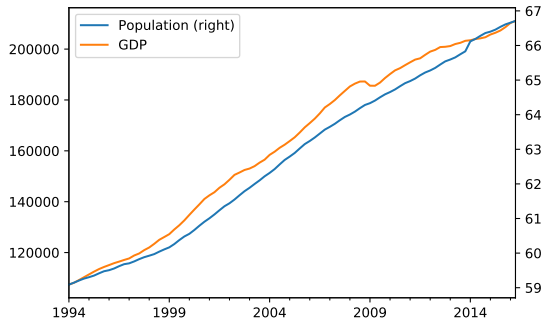
> the portfolio formed on the basis of returns realized in the past 6 months generates an average cumulative return of 9.5% over the next 12 months but loses more than half of this return in the following 24 months.

- **Question 40**
  - Compute the average correlation coefficient between successive daily returns.
  - Perform a quick check of the De Bondt and Thaler theory in (Bondt and Thaler, 1985) on another data set.
  - Describe and replicate the 16 strategies in Jegadeesh and Titman (Jegadeesh and Titman, 1993) (only Panel A and up to section II of the paper, leave aside the t-statistics for now).
  - What is your most successful zero-cost strategy? Compare it with the paper (Jegadeesh and Titman, 1993) findings.
  - Consider buying the losers versus the winner decile portfolios.

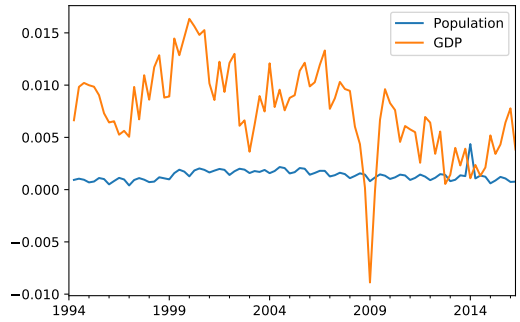
20 Python for non-programmers: exercise with pandas part 4

20.1 Before regressing or computing correlation, plot your data

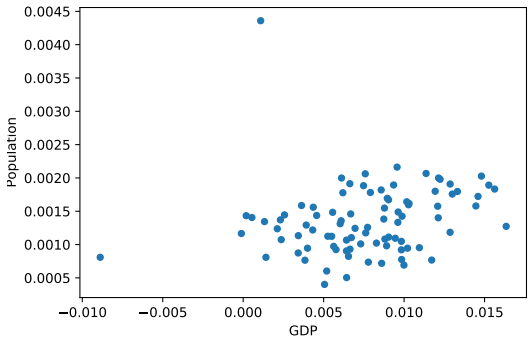
Before[82] applying any type of regression model to a data set, you should plot the data set.
Let's now download the French GDP from the INSEE website, here.
We do the same manipulations we did on the population csv file, except that now we have quarterly data.
Now we want to concatenate the GDP and the population time series, on the dates, as explained here.
We can then plot, using a secondary y axis on the right for the population.
You should get:



We can compute the quarterly changes and we expect to have a low correlation between both series:



We inspect a scatter plot of the GDP growth rate with respect to the Population growth rate:



[82] we suggest the code pandas_exercise_part4.py

We can clearly identify two points as outliers: one with a significant negative GDP growth rate and one with a significant positive Population growth rate. We will discuss in section13.1 how to deal with outliers.

20.1.1 Correlation

Following the work detailed in section 11.13.6, we compute the correlation between the GDP and Population growth rates: we find a low correlation of 19%.

20.2 Unit Root test - Dickey-Fuller test

We test "manually" the GDP for a unit root following the test described in section 16.8.
We test that both the French population and GDP time series are I(1).

20.3 Cointegration, ECM and linear regression

20.3.1 Cointegration test

After we differenced the variables, we cannot check for their **long term** relation in level, when differenced the variables are only compared in short term models.
With two I(1) series[83] y_t and z_t, level comparison via linear regression would be difficult because subject to spurious regression. But two I(1) series are said to be cointegrated, if there exist a linear combination[84] of them (e.g. $\beta_0 + \beta_y y_t + \beta_z z_t$) that is I(0). This means that y_t and z_t share a common stochastic trend and the two time series, after having applied some coefficients, cannot drift too far apart from one another in the long-run[85]. If there exist[86] β_y and β_z such that $E(\beta_0 + \beta_y y_t + \beta_z z_t) = 0$, then there is a long term relationship between y_t and z_t and one can test the following model:

$$y_t = \alpha + \beta z_t + \epsilon_t \tag{44}$$

and from our statements above, one can deduct that the cointegrating residuals ϵ_t have to be $I(0)$ if y_t and z_t are cointegrated[87].
We can then test for cointegration between variable with a cointegration test: the null hypothesis is no cointegration. The second output give in the p-value of the test.
For our variables, we do not reject the null hypothesis and assume no cointegrations.

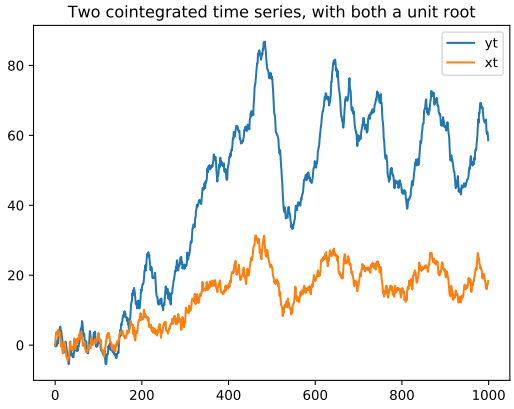
20.3.2 Building two cointegrated time series

We suggest to build two cointegrated time series with the system of equations:

$$\begin{cases} y_t = \beta_y y_{t-1} + \beta_z z_t + \epsilon_t \\ z_t = z_{t-1} + \nu_t \end{cases} \tag{45}$$

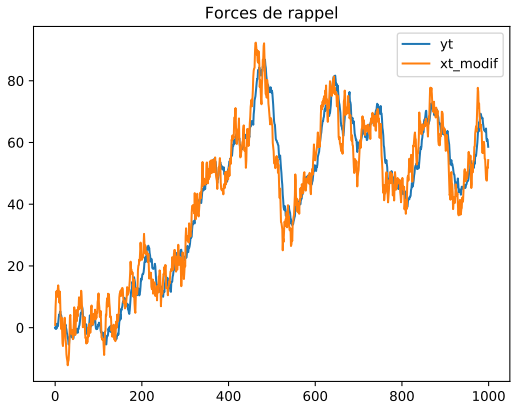
with both ϵ_t and ν_t white noises.
We can show that both series are I(1) and cointegrated. Visually the link might not be obvious:

[83] we define by I(n) integrated series of order n, more details here
[84] called long-run equilibrium relationship
[85] unless the "forces" that kept them together in the past do not hold any longer
[86] β_y and β_z are not unique (one can add a constant and the relationship would still hold), also note that β_0 can be null and the following still applies
[87] but as demonstrated in Phillips and Ouliaris (1990), one cannot use the traditional Augmented Dickey-Fuller for this test because of the spurious nature of the regression, furthermore, the the t-statistics for β will not be asymptotically normal if ϵ_t are serially correlated, we will introduce the Johansen procedure in section **??**



Once we applied the cointegrating regression and found the coefficients, we can show the "forces de rappel" where there seem to be some forces that brings y_t to "follow" x_t:

$$y_t = \gamma_0 + \gamma_1 z_t + \chi_t \tag{46}$$



We can then study the following Error Correction Model, with a grain of salt[88] with regard to the consistency of the estimation by OLS as detailed in (Zivot and Wang, 2003):

$$\begin{cases} \Delta y_t = \alpha_1^y + \alpha_2^y \chi_{t-1} + \alpha_3^y \Delta y_{t-1} + \alpha_4^y \Delta z_{t-1} + \epsilon_t^y \\ \Delta z_t = \alpha_1^z + \alpha_2^z \chi_{t-1} + \alpha_3^z \Delta y_{t-1} + \alpha_4^z \Delta z_{t-1} + \epsilon_t^z \end{cases} \tag{47}$$

We provide an empirical example on stock prices and dividends in section **??**

20.3.3 Cointegrated variables - example

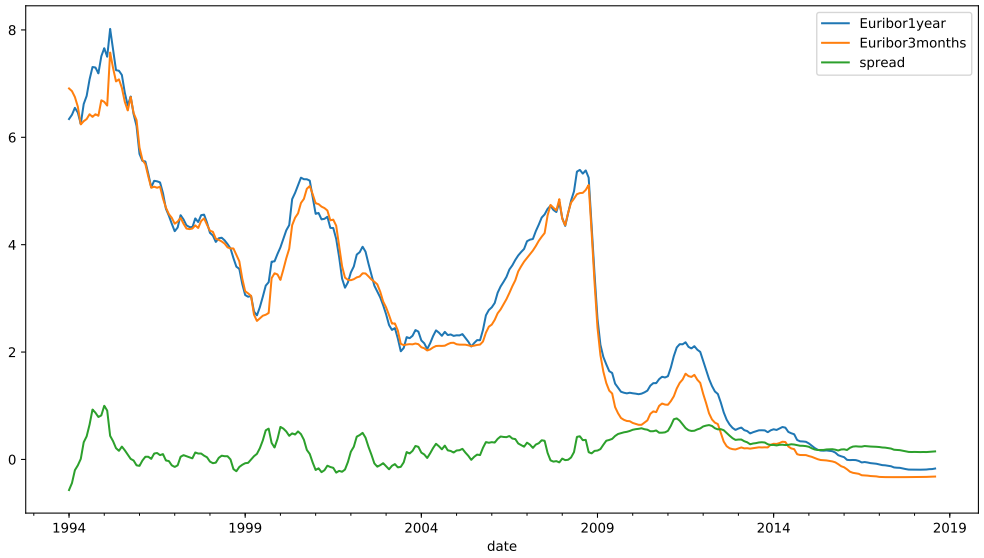
We take the Euribor 1 year (FM.M.U2.EUR.RT.MM.EURIBOR1YD_.HSTA) and the Euribor 3 months (FM.M.U2.EUR.RT.MM.EURIBOR3MD_.HSTA) from the ECB statistical warehouse, check for cointegration between the two time series and build an ECM if relevant.

[88] e.g. "ECM system may be estimated by seemingly unrelated regressions (SUR) to increase efficiency if the number of lags in the two equations are different"

Noting that both time series are I(1) and cointegrated, there is no unit root in the error terms ϵ_t of the regression:

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

Remember that if our model is correct, then $E(\epsilon_t) = 0$



20.3.4 Error Correction Model - examples

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:

$$\hat{\epsilon}_t = Y_t - \hat{\alpha} - \hat{\beta} X_t$$

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

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

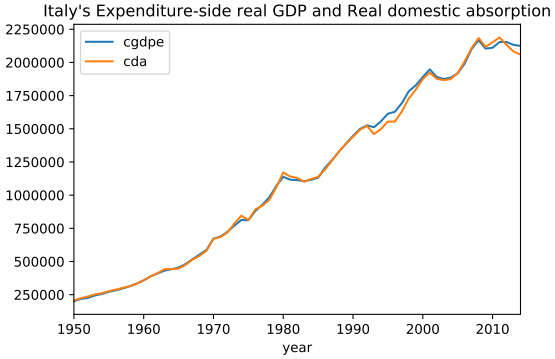
We can say that our model explains 78% of the variance of the observed changes. And that **for a** 1% **increase in the Euribor 3 months rate there is a** 0.92% **increase in the Euribor 1 year rate** ($\gamma_1 = 0.92$ and is significantly different than 0 with a t-value of 31 well above 1.96).

- **Question 41**
  - In the ECM above, add the lagged variables of the expenditure and absorption, are the results still holding? $\Delta Y_t = \gamma_0 + \gamma_1 \Delta X_t + \gamma_2 \hat{\epsilon}_{t-1} + \gamma_3 \Delta Y_{t-1} + \gamma_4 \Delta X_{t-1} + \nu_t$

20.3.5 Error Correction Model - consumption and income

Noting that the permanent income model implies cointegration between consumption and income, we use some proxy data from the Penn World Table and focus on data for Italy:

- 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)



20.3.6 Before regressing two I(1) time series, test for cointegration

We test the French population and GDP for cointegration. As they don't seem cointegrated we perform a linear regression.

20.3.7 Linear regression

Now we want to perform a simple linear[89] regression of population change on French GDP change:

$$\Delta Y_{\text{Pop},t} = \alpha + \beta \Delta X_{\text{GDP},t} + \epsilon_t \tag{48}$$

We perform an Ordinary Least Square method.
If both series have 0 mean, the one can also test a model with no constant:

$$\Delta Y_{\text{Pop},t} = \beta_{\text{no constant}} \Delta X_{\text{GDP},t} + \epsilon_t \tag{49}$$

But in our example, both return series have a mean statistically different than 0 hence we need to have a constant in the regression.

20.3.8 R-square and t-student "interpretation"

For a simple regression model: $y_i = \beta x_i + \epsilon_i$, T the length of our data set, our Ordinary Least Square programme is:

$$\min_{\beta} \sum_i (y_i - \beta x_i)^2$$

With first order conditions, we can find the estimate of β and we could prove that it is linear, unbiased, and efficient:

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} \tag{50}$$

[89] For non linear regression, the reader can start with this reference

We can demonstrate that, if $\text{var}(\epsilon) = \sigma^2$:

$$\text{var}(\beta) = \frac{\sigma^2}{\sum_i x_i^2}$$

The best estimate for the standard deviation of the residuals is:

$$\hat{\sigma}^2 = \frac{\sum_i (\hat{y}_i - y_i)^2}{T-2}$$

And the standard deviation for the estimated parameters are[90]

$$\hat{\sigma_{\beta_0}}^2 = \hat{\sigma}^2 \frac{\sum_i xi^2}{T\sum_i x_i^2 - \left(\sum_i x_i\right)^2}$$

$$\hat{\sigma_{\beta_1}}^2 = \hat{\sigma}^2 \frac{T}{T\sum_i x_i^2 - \left(\sum_i x_i\right)^2}$$

By decomposing the variance and with enough observations, it can be demonstrated that the total variance is decomposed by the variance of the model and the error terms:

$$\text{var}(y_i) = \text{var}(\hat{y}_i) + \text{var}(\hat{\epsilon_i})$$

we then define R^2 as the ratio between the variance of the model and the total variance:

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2}$$

$R^2 \in [0,1]$, the closest to 1, the more explanatory our model (if our econometric methods were correct[91]), meaning the ability of our model to explain the movements of the dependent variable.
The t statistic for a coefficient is defined as:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

This statistics follow a Student law with $N-k$ degree of freedom (N being the number of observations and k the number of estimated parameters).
The null hypothesis H_0 is : $\beta = 0$, we can use a threshold of 5%, as it is two-sided, we search for t^* such that:

$$\text{Prob}(X < t^*) = 0.025$$

If $|t| > t^*$, then we can reject H_0, again, if our econometric methods were correct, x has a significant influence over y.

For a model with an intercept, the results become:

$$\hat{\beta}_1 = \frac{\sum_i x_i (y_i - \bar{y})}{\sum_i x_i (x_i - \bar{x})}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

And we can show that:

$$R^2 = corr(yi, xi)^2$$

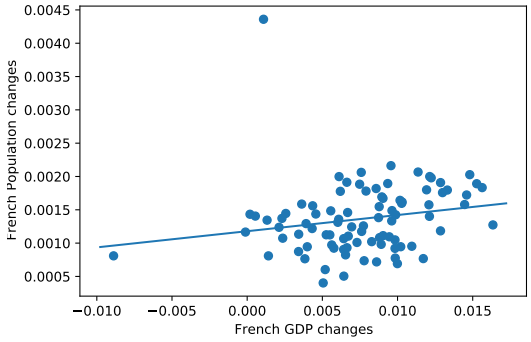
- **Question 42**
  - Compute the variance of β, the R^2 and the t statistic. Compare the results from the linear regression with and without constant term (intercept).

[90] more details can be found here

[91] for example, if the series have not a zero mean and we nonetheless force the intercept to be null, then the R^2 will seems larger than what the explanatory power of the model

20.3.9 Visuals

We can store the α and β value of that regression for future use and display our regression line over a scatter plot of the original data:



In our visual of the regression, we can identify visually at least two outliers. We would need to check for the robustness of our model if we remove those two outliers as a first robustness check.

20.3.10 Outlier detection and regression coefficient robustness

We identified visually two potential outliers[92]. We want to determine if they influence the regression coefficients out of proportion. We identify one outlier with the minimum GDP growth rate and one with the maximum population growth rate.
We create a dummy variable Dummy that has the value 0 for all observation except for the outlier one where the value is 1, we then do a new regression:

$$\Delta Y_{\text{Pop},t} = \alpha + \beta \Delta X_{\text{GDP},t} + \gamma \text{Dummy}_t \tag{51}$$

We test for both outliers if the γ is significantly different to 0. Instead of using the typical t-Student threshold $t_{1-\frac{\alpha}{2},n-p-1}$ we rather use the Bonferroni correction t-Student threshold: $t_{1-\frac{\alpha}{2n},n-p-1}$
We find that one of the outliers coefficient is significantly different to 0 and we remove it from the observations before running a new linear regression and notice that the GDP growth rate sensitivity to the population growth rate is significantly different and the explanatory power of our model is improved.

20.4 Q-Q plots

In our approach, we assume that we are observing realisations x_t from a random variable X with theoretical probability distribution function F, a quantile of order p is defined as[93]

$$Q(p) = F^{-1}(p) = \min x \in \mathbb{R} : F(x) \geq p = \min x \in \mathbb{R} : Pr(X < x) \geq p$$

If we have n observations, we can write:

Value	x_1	x_2	...	x_n
Probability	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

[92] To go further on the topic of outliers and regressions, refer to this lecture and Krasker et al. (1983)
[93] when the random variable X is continuous, then $Pr(X < x) = Pr(X \leq x)$

If there are i, j such that $x_i = x_j$, then we regroup them into $y_1, \dots, y_k$ for $k \leq n$, we now have, with n_j the number of x's regrouped in y_j:

Value	y_1	y_2	$\dots$	y_k
Probability	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\dots$	$\frac{n_k}{n}$

We have the empirical repartition function:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}$$

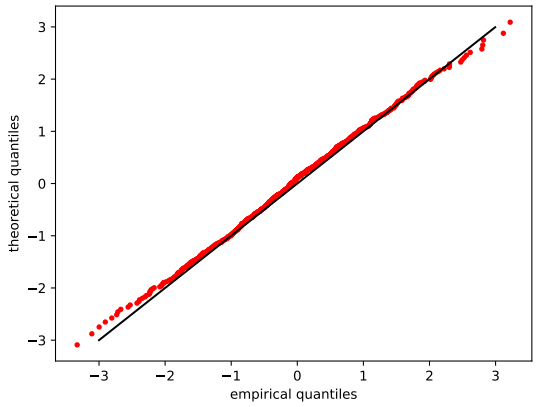
According to the law of large number[94], this empirical function converges to the theoretical repartition function in probability.
We can now define the empirical quantile of order p:

$$\hat{Q}_n(p) = \min\{x \in \mathbb{R} : \hat{F}_n(x) \geq p\}$$

$\hat{Q}_n(p)$ converges in probability to $Q(p)$.

20.4.1 Q-Q plot from a standard normal population

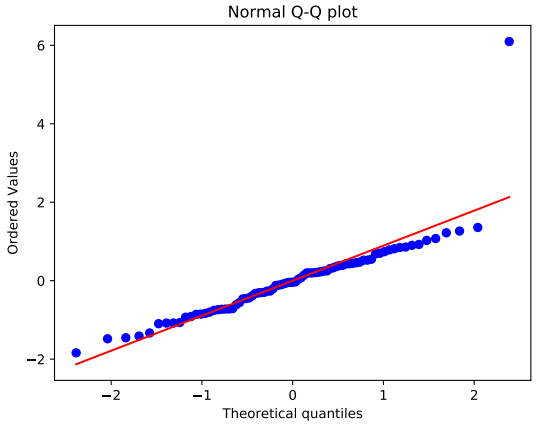
If we take n observations from a standard normal population, we can order those observations (we should also regroup the observations of same values). Then for each ordered observation i, we can compute the quantile of the theoretical standard normal distribution of probability $\frac{i}{n}$.



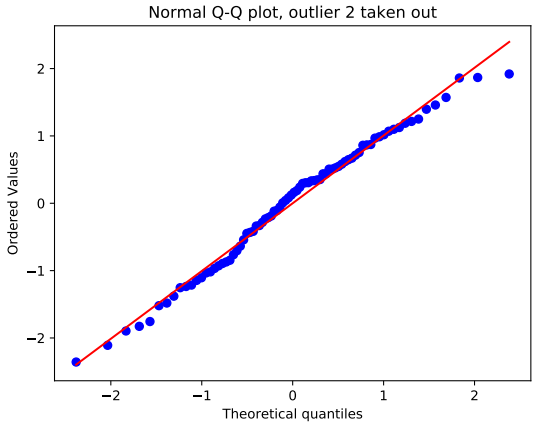
20.4.2 Q-Q plots of our linear regressions' normalized residuals against a normal law

From the orginal regression, we can check whether the residuals are normally distributed:

[94] cf. section 11.13

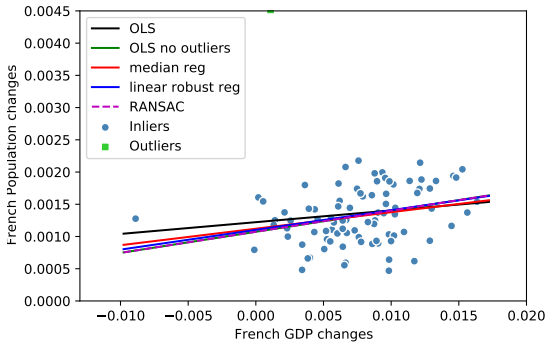


We can see that if we remove the outlier number 2, then our residuals seems to be normally distributed and the model improved:



20.5 Robust regression

We apply robust regression methods[95]. Table 1 show the OLS regression and table 2 the quantile regressions (median) columns (1) and (2) and we also apply robust linear models with support for the M-estimators columns (3) and (4). We apply a dummy variable that is equal to 1 for the outlier where French population change is higher:



[95] code: 20200411_robust_regressions.py

Section 20.3.7 we saw the OLS progam for a model $y_i = \alpha + \beta x_i + \epsilon_i$ to be $\min_\beta \sum_i f(\epsilon_i)$ with $f(x) = x^2$, now for robust linear models, we use the Huber objective function is:

$$f(x) = \begin{cases} .5x^2 & \text{for} |x| \leq k \\ k|x| - .5k^2 & \text{for} |x| > k \end{cases} \tag{52}$$

with k linked to a tuning parameter, in python's statsmodels, this is taken to be 1.345 so here k = 1.345/0.6745MAD

	Dependent variable:	
	(1)	(2)
GDP	0.024^*	0.038^{***}
	(0.013)	(0.01)
Intercept	0.001^{***}	0.001^{***}
	(0.0)	(0.0)
dummy		0.003^{***}
		(0.0)
Observations	89.0	89.0
R2	0.037	0.465
Adjusted R2	0.026	0.452
Residual Std. Error	0.001(df = 87.0)	0.0(df = 86.0)
F Statistic	3.376^* (df = 1.0; 87.0)	37.338^{***} (df = 2.0; 86.0)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 1: OLS comparison

With the following regression, we could be tempted to state that when we see 1% increase in the French GDP we expect to see a 0.03% increase in the French population. The elasticity from the three estimates from an OLS on trimmed data set, median regression and robust linear regression being comparable. We compare our result with the RANdom SAmple Consensus (RANSAC) to fit a robust regression model which almost 'agrees', with a sensitivity of 0.04 but note that the RANSAC results are very sensitive to the residual threshold choice.

	Dependent variable:			
	(1)	(2)	(3)	(4)
GDP	0.034^{**}	0.033^{**}	0.035^{***}	0.038^{***}
	(0.016)	(0.018)	(0.011)	(0.012)
Intercept	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}
	(0.0)	(0.0)	(0.0)	(0.0)
dummy		0.003^{***}		0.003^{***}
		(0.0)		(0.0)
Observations	89.0	89.0		
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2: Quantile and linear robusts regressions

21 Pairs trading and statistical arbitrage

Pairs trading, the ancestor of statistical arbitrage, has over 400 citations in the literature, for which a review is done by (Krauss, 2017).

21.1 Pairs trading strategy based on cointegration

The following[96] is inspired by the academic paper (Rad et al., 2016) and the work here. Cointegration test and some economic examples have been introduced in section 20.3.1.(Gatev et al., 2006) found that between 1962 and 2002

> find[ing] two stocks whose prices have moved together historically. When the spread between them widens, short the winner and buy the loser. If history repeats itself, prices will converge and the arbitrageur will profit. [...] we show that our profits are not caused by simple mean reversion as documented in the previous literature.

Such a strategy yield up to 11% average annualized excess returns for self-financing portfolios of top pairs. Noting that "If the U.S. equity market were efficient at all times, risk-adjusted returns from pairs trading should not be positive."

Two $I(1)$ time series x_t and y_t are said to be cointegrated if there exist a linear combination of them (e.g. $\alpha + y_t + \beta_x x_t$) that is I(0). If $E(\alpha + y_t + \beta_x x_t) = 0$, then there is a long term relationship between y_t and x_t, with an equilibrium relationship: $y_t = -\alpha - \beta_x x_t$.
We can then test for cointegration between variable with a cointegration test:

the null hypothesis H_0 is no cointegration.

The second output of the python test gives the p-value of the test.

In simple terms, two cointegrated time series have a long term relationship: if they are departing from one another in level, one can expect the series multiplied by some coefficients to "meet again" (if past behaviours would be a good predictor for future behaviours). If you believe in this, then you might spot some investment opportunities (inspired from here).

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

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

To get an intuition, you can think of a system:

$$\begin{cases} \Delta Y_t = \beta_Y \epsilon_{t-1} + \nu_{Y,t} \\ \Delta X_t = \beta_X \epsilon_{t-1} + \nu_{X,t} \end{cases} \tag{53}$$

Then if we imagine that we have significant coefficients $\beta_Y < 0$ and $\beta_X > 0$, then in case the spread depart from its long-term mean (e.g. $\epsilon_{t-1} > E(\epsilon_{t-1})$) then the stock price Y_t is expected to decrease and the stock price X_t is expected to increase to correct and bring the spread to its long-term mean.
We can also think of an error-correction model by OLS:

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

[96] pyton code: cointegration_trading_vansteenberghe

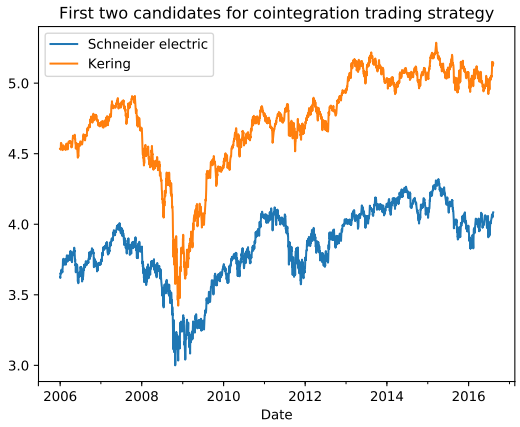
with

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

We search for cointegrated time series candidates:

1. we focus on the biggest capitalization to avoid liquidity issues when trading
2. we work with series in log to avoid level issues
3. we keep only $I(1)$ time series
4. we rank the couples of time series by cointegration test p-values

Here is an example with two candidates:



Then at time t:

1. we estimate a coefficient β of sensitivity between y and x based on all the history of the time series, the value at time t not included.
2. we estimate the spread between both time series:

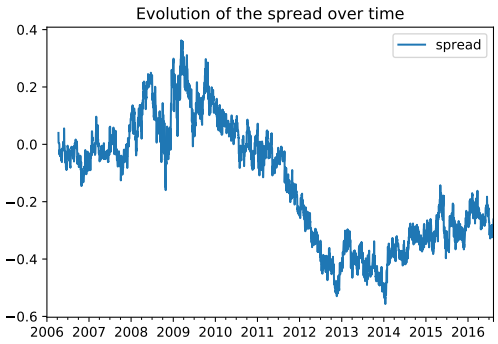
$$\text{Spread}_t = y_t - \hat{\alpha} - \hat{\beta} x_t$$

3. if the value of the Spread at time t is away from its historical mean μ by more than 2 times its historical standard deviation σ, we are:
   - short y and long x if $\text{Spread}_t > \mu + 2\sigma$
   - long y and short x if $\text{Spread}_t < \mu - 2\sigma$
4. we exit a trade once we are only only or less than 2 standard deviations away from the mean.

Nota Bene: as we assume both series to be cointegrated, we expect:

$$E(\text{Spread}_t) = \mu$$

hence if the Spread in absolute value is significantly different than μ we consider that there is a trading opportunity.



- **Question 43**
  - Compute the return of this strategy. Apply this strategy on some other pair candidates, is the strategy robust?
  - In fact, to form your decision on whether the series are cointegrated or not, you can only work on past historical data, change the code to reflect this.
  - To reduce the number of short/long orders you might want to initiate order only when crossing the 2 standard deviations distance from the spread mean.

21.2 Introduction to copulas

We follow[97] the illustration for the paper (Stander et al., 2013):

- First we rank the stock log return pairs per Kendall τ
- Use the Clayton copulas: $\theta = 2\tau(1-\tau)^{-\theta}$
- Assume the pdf: $C(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}$
- Identify if you are in the 95% confidence interval to determine trading opportunities

[97] copula_vansteenberghe

22 python and R: Some portfolio performance measures

22.1 Downloading data and importing it

Import data from the Kennth French website: here.

1. We are using the the 30 Industry Portfolios in csv format, link here.
2. It is already in the data forlder under the name `30_Industry_Portfolios_Daily.CSV`
3. Import the data as a data frame, for that to work, you might want to use the first column as index, those are the dates[98]. These are already returns so you can directly work and regress them (after doing ADF test).
4. You might want to convert the dates into date format.

22.1.1 Plotting the data

1. Plot one of the 30 industries daily returns.

22.2 Risk and performance measures

Nota bene: in this section, for simplicity, we consider that $R_{f,t} = 0, \forall\, t$.

22.2.1 Risk: variance and standard deviation of returns

Compute the sample variance of each of the 30 industry returns as a risk measure. The historical variance of an asset i over the time horizon T is measured as:

$$\hat{\sigma}_i^2 = \frac{1}{T-1} \sum_{t=1}^{T} (R_{i,t} - \overline{R}_i)^2 \tag{54}$$

with $\overline{R}_i$ the average return over the entire period T. Here you can use the standard deviation function directly to compute $\hat{\sigma}_i$.

22.2.2 Sharpe ratio

For an asset i, the Sharpe ratio is defined as:

$$SR(R_i) = \frac{\overline{R_i - R_f}}{\sigma(R_i - R_f)} \tag{55}$$

You might want to decompose this in two steps:

1. Compute the average return for each column of the data set
2. Divide this average return per asset over its standard deviation computed before

You should find the best Sharpe Ratio for the "Food" at 0.0467.

[98] index_col = 0 if you use python

22.2.3 Semi-volatility

As an investor, we are mostly concerned about negative returns. Therefore we might want to compute the semi-volatility, which only takes into account the returns below the mean return over the horizon:

$$sv(R_i) = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T} \max(0, \overline{R_i} - R_{i,t})^2} \tag{56}$$

For this, you might want to follow these steps:

1. Create a new data frame dfneg where for each column you remove its mean return
2. Take the minus of this data frame
3. Replace negative values by 0
4. Take the squares of the values of this matrix
5. Take the mean for each column
6. Take the square roots of those values

The worst semi-volatility you should find is for Servs at 1.66

22.3 Sortino ratio

The Sortino ratio is a performance measure computed as:

$$SoR(R_i) = \frac{\overline{R_i - R_f}}{sv(R_i - R_f)} \tag{57}$$

Compute it for each column. The best Sortino ratio is for Foods at 0.065.

22.4 Drawdown

22.4.1 Maximum Drawdown

The Maximum Drawdown at time t for a given time period s (e.g. 252 for a given trading year) for a portfolio with liquidation value P_t is the percentage largest peak to bottom within that time period:

$$\mathrm{MD}_{t,s} = \max_{t-s \le i < j \le t} \left(\frac{P_i - P_j}{P_j} \right) \tag{58}$$

If the definition is quite simple, one need to define a "scanning" function over the time period, an idea to proceed:

1. loop i from s to 1 and at each point compute the ratio from the portfolio value at time $t-i$ and the minimum portfolio value found in the range $[t-i, t]$
2. select the minimum ratio

22.4.2 Length

The length of the Maximum Drawdown is the time between the peak to bottom.

22.4.3 Recovery time

Here we define the recovery time as the time it take, once the bottom of the Maximum Drawdown has been reach, to recover at least the portfolio value at the time of the peak of the Maximum Drawdown.

22.5 List of assets

From our first data frame, each column represent an asset, you can create a list of asset by converting these columns names into a list, call it assetlist[99].

22.6 Fama-French factors

Nota bene: in this section, we now have $R_{f,t} \neq 0$.

- Import the daily times series of Fama-French factors, this is in the data folder under the file `F-F_Research_Data_5_Factors_2x3_daily.csv`
- Import the data as a data frame you name df2, for that to work, you might want to use the first column as index, those are the dates[100].
- You might want to convert the dates into date format.

22.6.1 Merge both data set

You now join both data frame df and df2 into a unique data frame keeping only days for which you have the full information (no NAs in the row)[101].

22.6.2 Market returns and CAPM

From your merged data frame, you can consider that the market returns are the column 'Mkt-RF'. In the following we call these returns $z_{m,t}$.

22.6.3 CAPM for each asset

For each asset, you can compute the following linear regression:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i z_{m,t} + u_{i,t} \tag{59}$$

You might want to proceed as follow:

1. Create a data frame dfCAPM with the assets as index and two columns, alpha and beta
2. Loop through all assets and compute $R_{i,t} - R_{f,t} = \alpha_i + \beta_i z_{m,t} + u_t$
3. Store the α_i and β_i of each regression for each corresponding asset.

[99] with python, you can use list(df.columns)
[100] index_col = 0 if you use python
[101] with python you can use pd.concat and then dropna(), df here is the data frame with the industry returns

22.6.4 Treynor ratio

We can now compute the Treynor ratio as:

$$TrR_i = \frac{\overline{R_i - R_f}}{\beta_i} \tag{60}$$

The best Treynor ratio is for Beer at 0.079.

22.6.5 Jensen's alpha

Jensen's alpha characterise the over- or under-performance of an asset return compared with the return expected with the CAPM:

$$\hat{\alpha}_i = \overline{R_i - R_f} - \hat{\beta}_i \overline{z_m} \tag{61}$$

The highest Jensen's alpha is for Smoke at 0.029.

22.6.6 5-factor model

In our data frame we have 5 factors which enable us to estimate a 5-factor model:

1. 'Mkt-RF': Market
2. 'SMB': Size (Small Minus Big)
3. 'HML': Book-to-market (High Minus Low)
4. 'RMW': Profitability (Robust Minus Weak)
5. 'CMA': Investment (Conservative Minus Aggressive)

The first step is to perform a linear regression:

$$R_{i,t} - R_{f,t} = \alpha_i + \sum_{k=1}^{5} \beta_{i,k} f_{k,t} + u_{i,t} \tag{62}$$

22.6.7 Arbitrage Pricing Theory (APT)

With no arbitrage opportunity, we impose $\alpha_i = 0$, hence estimate the regression:

$$R_{i,t} - R_{f,t} = \sum_{k=1}^{5} \beta_{i,k} f_{k,t} + u_{i,t} \tag{63}$$

You can follow the following steps:

1. Loop through each asset and estimate the $\beta_{i,k}$
2. Store the estimated $\beta_{i,k}$ in a data frame dfAPT

22.6.8 Risk premia under the APT

Once the $\beta_{i,k}$ have been estimated, we can estimate the risk premia γ_k as:

$$\overline{R_{i,t} - R_{f,t}} = \beta_{i,k} \gamma_k \tag{64}$$

Perform this regression and provide the risk premia.

23 Python: Monte Carlo method and econometric tests

We suggest[102] Monte Carlo methods to help understanding standard econometrics hypothesis testing. In section 20.3.7 we performed ordinary least square to estimate linear regressions of the form:

$$y_t = \alpha + \beta x_t + \epsilon_t$$

our estimate $\hat{\beta}$ of the true parameter β will depend on the size of the samples and on the validity of assumptions we made about underlying data generation processes for ϵ and x.
We assumed that ϵ was iid and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$ and $x \sim \mathcal{N}(\mu_x, \sigma_x)$. Even if those assumptions are true, we have a finite and relatively small number of observations so the precision of our estimate will be limited.
The t statistic for a coefficient is defined as:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

This statistics follow a Student law with $T-k$ degree of freedom (T being the number of observations and k the number of estimated parameters).
The critical threshold can be computed at a given confidence level with the inverse of the cumulative Student t distribution, the percent point function (ppf) which in the 5% case is t^* such that: Prob($X \leq t^*$) = 0.025
The null hypothesis H_0 is : $\beta = 0$, if $|t| > t^*$, then we can reject H_0, again, if our econometric methods were correct, x has a significant influence over y.

23.1 Monte Carlo method for coefficient significance of a simple OLS

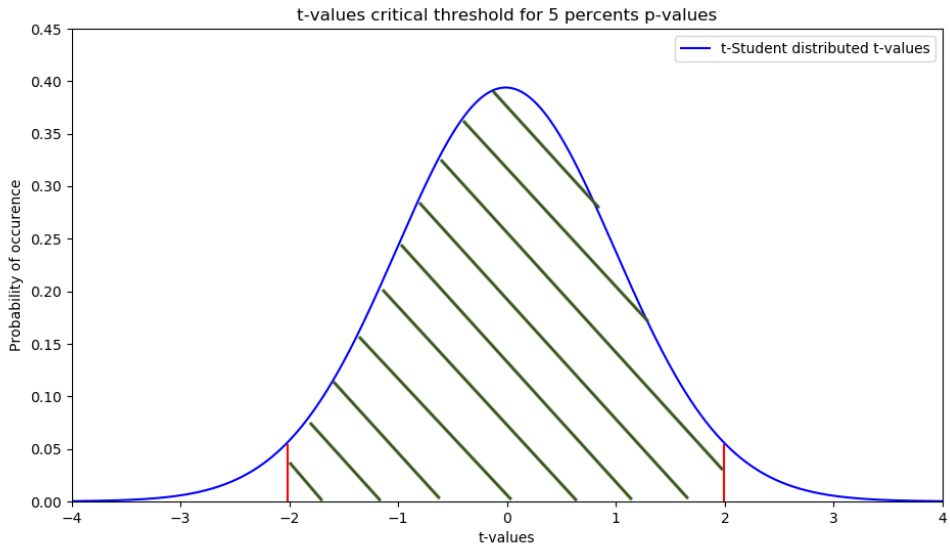
We know that the t statistic for β follows a Student law with $T-k$ degree of freedom. We can generate a "large" number of simulations (e.g. MCrun = 1000), drawing the error and observable from the assumed distributions: $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$ and $x \sim \mathcal{N}(\mu_x, \sigma_x)$ and extract for each simulation the t-value of the β. So step by step:

1. we estimate $y_t = \alpha + \beta x_t + \epsilon_t$ and extract: $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}_\epsilon$, $\hat{\mu}_x$, and $\hat{\sigma}_x$;
2. we generate MCrun simulations generating $\tilde{\epsilon}_t$ and $\tilde{x}_t$ sample of length T, the original series' length;
3. we compute each simulation i's $\tilde{y}$ series: $\tilde{y}_t^i = \hat{\alpha} + 0 \times \tilde{x}_t^i + \tilde{\epsilon}_t^i$, so under the null hypothesis, $\tilde{x}_t^i$ has no "influence" over $\tilde{y}_t^i$
4. from each simulation, we estimate the regression: $\tilde{y}_t^i = \delta + \gamma \tilde{x}_t^i + \nu_t$, from each regression we store in a list the t-statistic of the estimated coefficient $\hat{\gamma}$;
5. we are "sure" that $\hat{\gamma}$ should be 0 rather than $\hat{\beta}$, hence 100% of our t-statistics from the previous step should be in the H_0 acceptance region and are distributed as a t-Student with $T-2$ degree of freedom;
    - at this point, we can compute the critical threshold in a similar manner as a parametric VaR for a t-Student distribution as we develop in section 18.1;

[102] pandas_exercise_Monte_Carlo.py

6. we extract the β t-statistic from the original estimation $y_t = \alpha + \beta x_t + \epsilon_t$ and count how many t-statistics we generated are above this value, because our test is two-sided, this is half the probability to wrongly reject H_0. In other words, if we are almost outside of the empirical distribution of t-values we generated we can safely reject H_0 and (as long as our assumptions are correct) β is significantly different from 0, but if we are almost in the middle of the empirical distribution, we cannot reject H_0 and it is likely that $\beta = 0$.

Below is the distribution of the β t-values under the null hypothesis. We draw left and right the threshold so that the green area in between is 95%. So for our test, we reject the null hypothesis at the 5% confidence level is our t-value is outside of this zone:



23.2 Monte Carlo method for augmented Dickey-Fuller test p-value

We are now interested in the augmented Dickey-Fuller test we perform as in section 16.9. We apply this with the no constant (option 'nc') on the demeaned GDP growth rate and the comparison is simple as the test select no lags, hence equation 29 simplifies into:

$$\Delta y_t = \gamma y_{t-1} + \epsilon_t$$

We want to test the null hypothesis H_0 under which $\gamma = 0$.
For this, as in previous section, we will generate a "large" number of series that respect this null hypothesis, and compare our inital regression t-value of γ with the distribution of t-value of γ obtained, with the underlying assumption that $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$:

1. we estimate $\Delta y_t = \gamma y_{t-1} + \epsilon_t$ and extract: $\hat{\sigma}_\epsilon$;

2. we generate MCrun simulations generating $\tilde{\epsilon}_t$ sample of length T, the original series' length;

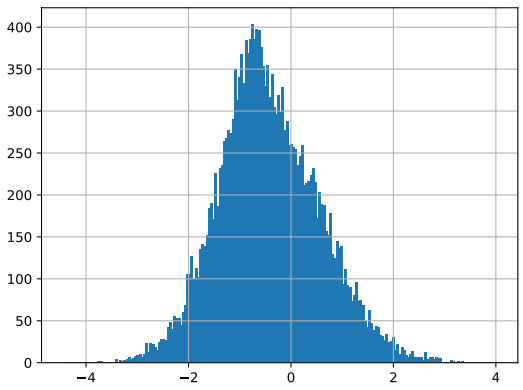
3. we compute each simulation i's $\tilde{y}$ series: $\tilde{y}_t^i = 1 \times \tilde{y}_{t-1}^i + \tilde{\epsilon}_t^i$, so under the null hypothesis, $\tilde{y}$ has a unit root

4. from each simulation, we estimate the regression: $\Delta\tilde{y}_t = \gamma\tilde{y}_{t-1} + \nu_t$, from each regression we store in a list the t-statistic of the estimated coefficient $\hat{\gamma}$;

5. we are "sure" that $\hat{\gamma}$ should be 0 (our series has a unit root), hence 100% of our t-statistics from the previous step should be in the H_0 acceptance region;
    - be careful: as demonstrated in (MacKinnon, 2010), this t-values are not distributed as a traditional t-Student distribution, we can use the tables in (MacKinnon, 2010) and compute the critical value following as in the paper $\beta_\infty + \beta_1/T + \beta_2/T^2 + \beta_3/T^3$ or in our case we use a Monte Carlo method which doesn't need to take any assumption on the distribution of the t-values of $\hat{\gamma}$;

6. we extract the γ t-statistic from the original estimation $\Delta y_t = \gamma y_{t-1} + \epsilon_t$ and count how many t-statistics we generated are below this value and this is the probability to wrongly reject H_0. In other words, if we are almost outside of the empirical distribution of t-values we generated we can safely reject H_0 and (as long as our assumptions are correct) γ is significantly different from 0, but if we are almost in the middle of the empirical distribution, we cannot reject H_0 and it is likely that $\gamma = 0$ and our series has a unit root (as long as our assumptions on the constant, drift, trend, and quadratic terms were correct!).

Note that the empirical distribution of the $\hat{\gamma}$ t-values is not "well-behaved", in the sense that for example even with 20 000 runs the distribution is not smooth and this can introduce some noise in our computed p-value:



Hence to get a powerful test we would need to perform "many" Monte Carlo simulations, here we have an estimated p-value of 0.06% meaning that we expect to have 60 t-values below our t-statistics out of 100 000 simulations (so in that respect, even 100 000 seems not enough to be confident over the results). In (MacKinnon, 2010) they ran 200 000 simulations.

24 Python for non-programmers: exercise with pandas - part 5

So far we applied univariate models or regressions, we will now introduce[103] multivariate models such as Vector autoregressions (VARs) that were introduced into empirical economics by Sims (1980).

24.1 VAR model

Model equation 48 links the the French population growth with the GDP growth. This models that immigration and birth rate simultaneously increase when the economy is performing well. If we agree to discard contemporaneous relationships between variables, we are limited to model that the GDP growth is linked to *lagged* immigration and *lagged* birth rate, so:

$$\begin{cases} \Delta Y_{\text{Pop},t} & = b_1 + c_{1,1}\Delta Y_{\text{Pop},t-1} + c_{1,2}\Delta X_{\text{GDP},t-1} + e_t^0 \\ \Delta X_{\text{GDP},t} & = b_2 + c_{2,1}\Delta Y_{\text{Pop},t-1} + c_{2,2}\Delta X_{\text{GDP},t-1} + e_t^1 \end{cases} \tag{65}$$

which can be written in matrix form as:

$$\begin{bmatrix} \Delta p_t \\ \Delta g_t \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix} \begin{bmatrix} \Delta p_{t-1} \\ \Delta g_{t-1} \end{bmatrix} + \begin{bmatrix} e_{p,t} \\ e_{g,t} \end{bmatrix} \tag{66}$$

With:

- p_t the population at quarter t
- g_t the GDP at quarter t
- $e_{i,t}$ idd error terms $\sim \mathcal{N}(0, \sigma_i)$

and

$$e_t \sim \mathcal{N}(0, \Omega)$$

with Ω a diagonal matrix, if this condition is not respected, then before applying shock, you need to consider Cholseky decomposition and SVAR. Non-diagonal correlation of the residuals can be a reflection of the contemporaneous relationships between the VAR model variables.
If we write:

$$\Phi = \begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix}$$

we can rewrite equation 66

$$y_t = b + \Phi y_{t-1} + e_t \tag{67}$$

then if $|\text{eigenvalue}(\Phi)| < 1$ then the VAR is stable[104] and the reduced form VAR presented in equation 66 can be consistently estimated by OLS equation by equation. VAR models can also be estimated with maximum likelihood methods (as of 2019, this option wasn't implemented in the python tsa package fit() method). For a VAR model, stability implies covariance stationarity.

24.1.1 Get the data

As we did with the GDP series, we fetch the unemployment series from INSEE website and import the data.

[103] we suggest the code pandas_exercise_part5.py

[104] this is the same as checking that not root lies strictly outside the unit root circle, if an eigenvalue or a root lies on the unit circle, we need to look for cointegrating vectors

24.1.2 Preparing the VAR model

We prepare the VAR model, first we import the corresponding package and then resample the population series in quarterly. Then we concatenate the time series into a DataFrame, take the differences related to the integration order of each variable and drop the rows where there are some missing values (NaN). VAR models are not good to deal with seasonality nor trend, we need to remove those effects before applying the model. Finally we apply the VAR model and print the results summary.

24.2 Select the lag order

For a VAR model of the type: $y_t = c + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t$, the order of the VAR is the value p such that $A_p \neq 0$ and $A_k = 0, \forall k > p$.
We suggest two main references (Lütkepohl, 2005) and (Kilian and Lütkepohl, 2017) to select the lag order, keeping in mind as Luetkepohl mentions:

> we know that the approximate Mean Square Error matrix of the 1-step ahead predictor will increase with the order p. Thus, choosing p unnecessarily large will reduce the forecast precision of the corresponding estimated VAR(p) model

- **Question 44**
  - As in (Lütkepohl, 2005), we want to experiment fitting a VAR model of higher order than the data generating process:
    1. consider the following VAR model:

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix} = \begin{bmatrix} 17.12 \\ -12.86 \end{bmatrix} + \begin{bmatrix} 0 & -0.61 \\ -0.17 & 0.29 \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}$$

    2. consider $\epsilon_1 \sim \mathcal{N}(0, 0.1)$ and $\epsilon_2 \sim \mathcal{N}(0, 0.9)$
    3. generate 200 time series for $T = 30$ and $T = 100$
    4. on the generated series, fit a VAR(1), a VAR(2) and a VAR(3), and for each model compute the average squared forecast errors at horizon $h = 1$ and $h = 3$.

We create the VAR model and select the lag, search for up to lag order 15.
If the chosen lag order is unnecessarily large, the forecast precision of the corresponding VAR(p) model will be reduced.
The VAR order selected is 1, the statistics for four criterio are given:

1. Akaike's Information Criterion (AIC)
2. Bayesian information criterion (BIC) or Schwarz information criterion
3. Final Prediction Error (FPE)
4. Hannan-Quinn Criterion (HQIC)

Following (Lütkepohl, 2005), there is no systematic rule to chose which coefficient to base your decision on: HQIC and BIC are consistent but

> In small samples, AIC and FPE may have better properties (choose the correct order more often) than HQ and SC [BIC]. Also, the former two criteria are designed for minimizing the forecast error variance. Thus, in small as well as large samples, models based on AIC and FPE may produce superior forecasts although they may not estimate the orders correctly. [...] AIC and FPE asymptotically overestimate the true order with positive probability and underestimate the true order with probability zero
>
> In a simulation study based on many other processes, Luetkepohl (1985) obtained similar results. In that study, for low order VAR processes, the most parsimonious SC criterion was found to do quite well in terms of choosing the correct VAR order and providing good forecasting models. Unfortunately, in practice we often don't even know whether the underlying data generation law is of finite order VAR type. Sometimes we may just approximate an infinite order VAR process by a finite order model. In that case, for moderate sample sizes, some less parsimonious criterion like AIC may give superior results in terms of forecast precision. Therefore, it may be a good strategy to compare the order estimates obtained with different criteria and possibly perform analyses with different VAR orders.

VAR Order Selection

lag order	aic	bic	fpe	hqic	
0	14.58	14.65	2.138e+06	14.60	* Minimum
1	14.27*	14.48*	1.572e+06*	14.35*	
2	14.33	14.69	1.683e+06	14.47	

AIC and BIC are (Information Criteria) statistics that measure the distance between observations and the model. We wish to minimize that distance to fit the data generating process. IC have two additive parts:

- a goodness-of-fit measure (e.g. minus the maximized likelihood);
- a penalty that increases with the model's complexity.

A study of lag order selection for VAR Impulse Response Analysis is proposed by (Ivanov and Lutz, 2005).
(Hirano and Wright, 2017) suggest to add bootstrap aggregation (bagging) :

> It is common practice to select the model based on pseudo out-of-sample fit from a sequence of recursive or rolling predictions. Parameters are then estimated over the whole sample period. The idea of using an out-of-sample criterion was discussed by Clark (2004) and West (2006), and has a long history, going back to Wilson (1934), Meese and Rogoff (1983), and Ashley, Granger, and Schmalensee (1980). Instead, one might select the model based on in-sample fit, but adjust for potential overfitting by using an information criterion, such as the Akaike information criterion (AIC) (Akaike (1974)), as advocated by Inoue and Kilian (2006).
>
> We consider various methods of model selection and forecasting, including: using in-sample fit with the AIC information criterion; selecting the model based on recursive pseudo out-of-sample forecast accuracy and then using the whole data set for parameter estimation; and splitting the sample into two parts, using one part for model selection and the other for parameter estimation. We call this last method the split-sample approach.

24.3 Apply a VAR model and checking for impact

We apply a VAR(1) to our data. We perform test on our model (all eigenvalues of Φ have modulus less than 1, autocorrelations and cross-correlations of model residuals, normality tests of residuals, ARCH-LM test on residuals).
The coefficient $c_{2,1}$ seems to be significantly different from 0, so we expect that a shock on the GDP will have an impact on the population dynamics. That seems to be the only non diagonal element different from zero. We can apply a Granger causality test (more detail on this test can be found section 25) and also visually check that with an impulse response function (irf).
But for this model, we had to assume no contemporaneous relationships, which might not be realistic and we might not want to trust our results so far! Especially, for the irf to be interpreted, you need to verify that Ω is a diagonal matrix, otherwise a shock on a given variable contemporaneously impact the other variables and you cannot isolate this effect.

24.4 (structural) VAR model

To apply a simple VAR model, we had to assumes no contemporaneous relationship between our variables. Now if we want to add contemporaneous relationships, we need to identify the following set of simultaneous equations:

$$\begin{cases} \Delta Y_{\text{Pop},t} &= a_1 + \beta^0 \Delta X_{\text{GDP},t} + d_{1,1}\Delta Y_{\text{Pop},t-1} + d_{1,2}\Delta Y_{\text{GDP},t-1} + \epsilon_t^1 \\ \Delta X_{\text{GDP},t} &= a_2 + \beta^1 \Delta Y_{\text{Pop},t} + d_{2,1}\Delta Y_{\text{Pop},t-1} + d_{2,2}\Delta Y_{\text{GDP},t-1} + \epsilon_t^2 \end{cases} \tag{68}$$

If both relationships hold, then the model is not identified. To identify the following equations with n variables, on top of assuming that ϵ^1 and ϵ^2 are uncorrelated we need $\frac{n(n-1)}{2} = 1$ restrictions[105], we impose $\beta^0 = 0$: we believe that people cannot "nowcast" GDP. We believe that the current population does contribute to the current GDP, $\beta^1 \neq 0$. We will show section 24.4.2 that this is equivalent with a Cholesky decomposition with a careful ordering of variables.
We estimate the following structural[106] VAR:

$$\begin{bmatrix} 1 & -\beta^0 \\ -\beta^1 & 1 \end{bmatrix} \begin{bmatrix} \Delta p_t \\ \Delta g_t \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} d_{1,1} & d_{1,2} \\ d_{2,1} & d_{2,2} \end{bmatrix} \begin{bmatrix} \Delta p_{t-1} \\ \Delta g_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{g,t} \end{bmatrix} \tag{69}$$

We assume that ϵ^0 and ϵ^1 are iid, orthogonal and exogenous. If this orthogonality condition is not respected, then changes in ϵ^0 affects ϵ^1 and then changes in $X_{\text{GDP},t}$ will be due to changes in $Y_{\text{Pop},t}$ and in ϵ^1, it is then not possible to identify (or isolate) the change in $X_{\text{GDP},t}$ caused by $Y_{\text{Pop},t}$.

We write: $A = \begin{bmatrix} 1 & -\beta^0 \\ -\beta^1 & 1 \end{bmatrix}$
As long as A^{-1} exists[107], the system of equations can be written in **reduced form** exactly as in equation 66. But now regressors are correlated with $e_t = \begin{bmatrix} e_{p,t} \\ e_{g,t} \end{bmatrix}$ is the forecast error and both forecast errors are affected by both shocks ϵ_p and ϵ_g.
If our assumptions are correct that B^{-1} exists and the innovations follow:

$$\epsilon_t = \begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{g,t} \end{bmatrix} \sim \mathcal{N}(0, D)$$

[105] those assumptions cannot be tested
[106] structural because it is assumed to be derived by some underlying economic theory
[107] so at least $\prod_i \beta^i \neq 1$

with D a diagonal matrix.
The main issue we have with the reduced VAR form is that we do not identify the β^i, we only have correlation of variables with forecast and cannot properly apply orthogonal shock ϵ_i, that is why we implemented the **structural** VAR, in our example, this is equivalent with a Cholesky decomposition (provided we have the right ordering of variables). Our main reference text book is (Kilian and Lütkepohl, 2017).

24.4.1 Getting convinced about the necessity for parameter restrictions

Let's imagine we want to work with the simple model:

$$\begin{cases} \Delta Y_{\text{Pop},t} & = a_1 + \beta^0 \Delta X_{\text{GDP},t} + \epsilon_t^1 \\ \Delta X_{\text{GDP},t} & = a_2 + \beta^1 \Delta Y_{\text{Pop},t} + \epsilon_t^2 \end{cases} \tag{70}$$

If $\Delta Y_{\text{Pop},t} = a_1 + \beta^0 \Delta X_{\text{GDP},t} + \epsilon_t^1$, then $\Delta X_{\text{GDP},t} = -\frac{a_1}{\beta^0} + \frac{1}{\beta^0} \Delta Y_{\text{Pop},t} - \frac{1}{\beta^0} \epsilon_t^1$, so you do not have the liberty to identify β^0 and β^1 separately.

24.4.2 Cholesky or LDL decomposition, aka short-run restrictions

We have made the assumptions:

$$\epsilon_t = \begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{g,t} \end{bmatrix} \sim \mathcal{N}(0, D)$$

with D a diagonal matrix and

$$e_t \sim \mathcal{N}(0, \Omega)$$

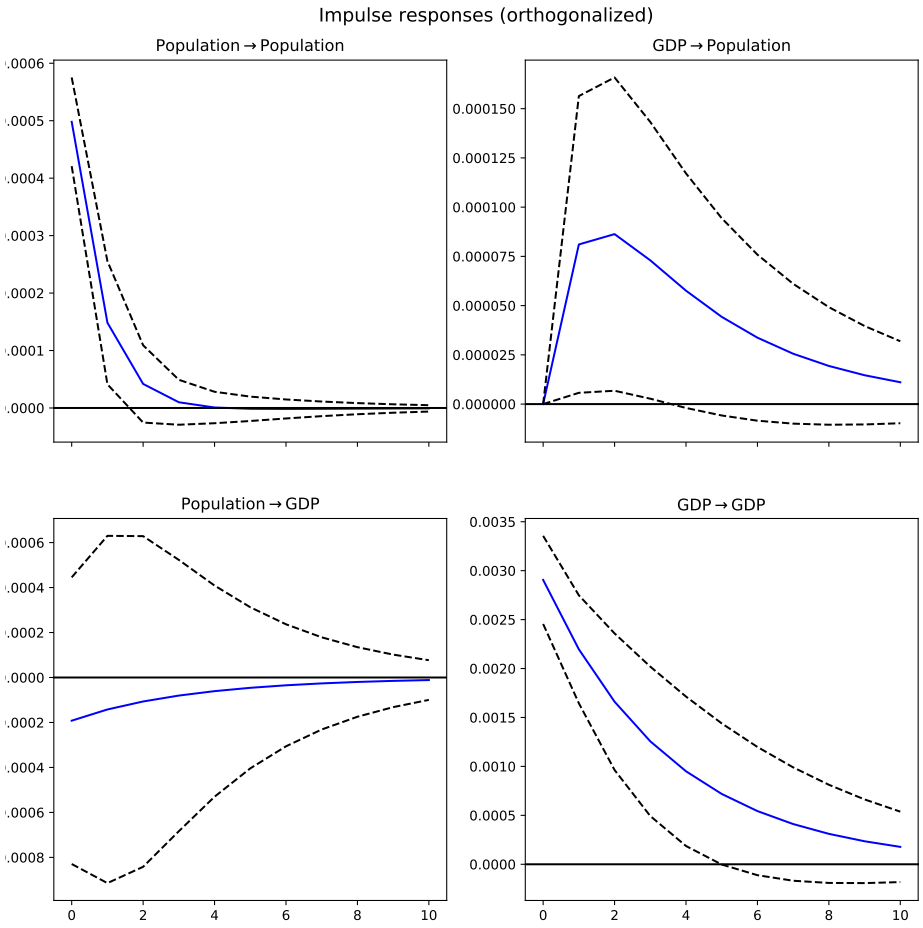
In order to identify the model, we can make the assumption that the population get knowledge of the GDP growth only with a lag, thus the immigration or birth rate is only "reacting" to the GDP growth rate with a lag and we can assume a lower triangular matrix, further we assume that:

$$\Omega = A^{-1} D (A^{-1})'$$

which enables us to identify A, the inverse of a lower (resp. upper) non-singular triangular matrix is a lower (resp. upper) triangular matrix. Note that the variable ordering is very important before you apply the Cholesky decomposition.
It is then possible to do a moving average or Wold representation of the VAR to determine the impact of shocks on the variables.

24.4.3 Impulse response function

We can also plot the impulse responses: these are the responses to a unit impulse to a variable that our VAR model predicts. In our case, we can believe that it is possible to have an isolated shock on the population that is not contemporaneous with a shock on the GDP, hence it makes economic sense to study such an isolated shock.
As we expected from our Granger causality test, a shock on the GDP has an impact on the population significantly different from 0 (but marginally):



- **Question 45**
  - An economist can also imagine that the unemployment rate both is cause and will cause effect on the population growth of France. Add the unemployment, do a three-variable SVAR (which restrictions do you suggest?):

$$\begin{bmatrix} \Delta p_t \\ \Delta g_t \\ \Delta u_t \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} \\ c_{2,1} & c_{2,2} & c_{2,3} \\ c_{3,1} & c_{3,2} & c_{3,3} \end{bmatrix} \begin{bmatrix} \Delta p_{t-1} \\ \Delta g_{t-1} \\ \Delta u_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{g,t} \\ \epsilon_{u,t} \end{bmatrix} \tag{71}$$

- **Question 46**
  - Perform a Granger causality test to test if the GDP is Granger causing the unemployment.
  - We have built a VAR with 3 variables. Compare the forecast precision of such a complex model with the simpler AR(1) model we try to manually design section 16.10. After you covered the sections on ARMA models, try to conclude on the model you would chose as an economist to predict French population "next year".

24.5 "The Demise of Granger Causality Tests in Macroeconomics"

As (Kilian and Lütkepohl, 2017) put it:

> Initially, economists thought that bivariate Granger causality tests established that money leads income, but that result weakened once more variables were included in the VAR model. It also proved highly sensitive to different forms of detrending and to changes in the model specification. Moreover, it was shown that apparent Granger noncausality from money to income may simply reflect the omission of a third variable, whereas a finding of bivariate Granger causality may likewise reflect the omission of a third variable, calling into question even further the usefulness of Granger causality tests [...]. Granger causality tests were replaced by the new idea that we are not interested in the contemporaneous correlation between money and income growth or the lead-lag pattern, but in the question of how income **responds to unanticipated changes** in money growth (also known as innovations or shocks). The hope was that these shocks would be exogenous, even if the underlying money growth time series was not.

24.6 Short-run and long-run restrictions

So far, we only imposed zero short-run restrictions which is simply done by imposing some parameters in A to be null, this is equivalent with a Cholesky decomposition. This can be justified, if we take:

$$y_t = My_{t-1} + \epsilon_t$$

where y_t is the time series vector of GDP growth rate and inflation. Economists will assume that the structural innovations represent supply s_t^S and demand s_t^D shocks and can restrict that a demand shock as no contemporaneous effect on the GDP growth rate, which correspond to the Cholesky decomposition. We check[108] ex-post our assumptions:

- a positive supply shock should increase the GDP and decrease prices (violated).
- a positive demand shock should increase the GDP and increase prices

We might also want to impose[109] zero long-run restrictions, introduced in (Blanchard and Quah, 1989) where they take GDP growth rate and unemployment rate. They assume that the structural innovations represent supply s_t^S and demand s_t^D shocks, but restrict that the demand shock has no long-run effect on the GDP.
Economic theory posit that money is neutral: the level of money has no long-run impact on the level of GDP. A VAR model of the macroeconomy is introduced section **??**, such a model should include: GDP, consumption, investment, employment, nominal wages, money, prices and nominal interest rates.

24.7 A macroeconomic model

As a basic macro-economist, assume the GDP growth g_t, the nominal interest rate i_t and the inflation rate π_t follow:

$$\begin{cases} g_t &= \alpha_g + \beta_{i,g} i_{t-1} + \beta_{\pi,g} \pi_t + u_{g,t} \\ i_t &= \alpha_i + \beta_{i,i} i_{t-1} + \beta_{\pi,i} \pi_t + u_{i,t} \text{ Taylor rule} \\ \pi_t &= \alpha_\pi + \beta_{\pi,\pi} \pi_{t-1} + \beta_{g,\pi} g_{t-1} + u_{\pi,t} \text{ Phillips curve} \end{cases} \tag{72}$$

where the structural shocks u_t are uncorrelated.

[108] SVAR_vansteenberghe.py
[109] SVAR_long_run_restrictions_vansteenberghe

- **Question 47**
  - Write the model in reduced form $Y_t = \Phi Y_{t-1} + \epsilon_t$ show that $\epsilon_t = Bu_t$ and describe the restrictions on B you impose based on your macroeconomic model 72
  - get data from a country of your choice and estimate the model. What is the impact of a positive (orthogonal) monetary policy shock $u_{i,t}$

24.8 How oil price shocks affect U.S. real GDP and inflation?

- **Question 48**
  - Following (Kilian and Lütkepohl, 2017), import the data on the author's website and import "Responses of the U.S. economy to an unexpected increase in the real price of oil (Cholesky decomposition)" data:
    - WTI price of crude oil;
    - the U.S. GDP deflator inflation rate;
    - U.S. real GDP growth.
  - estimate a structural VAR, how oil price shocks affect U.S. real GDP and inflation?

24.9 A VAR model from a textbook

It is not easy to find $I(0)$ data and to model a VAR that makes sense.
We use the data in Bourbonnais' Econometrie, Dunod. You can download the file online here.
Demand Y_1 and prices Y_2 of some commodity are given.

24.9.1 Import the data

We import the data and clean it.

24.9.2 Verify that the series are stationary

We verify that both given series are stationary. We find indeed that both series are $I(0)$. We define a function that print the result of the test and that allow to chose for a threshold for the p-value.

24.9.3 Estimate the VAR coefficients

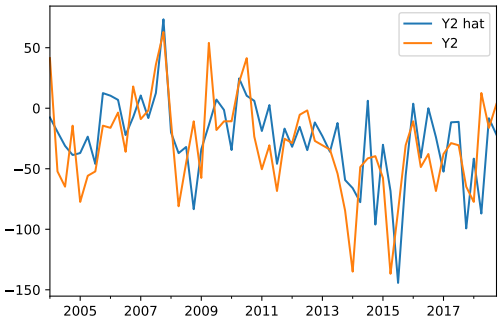
Now that we selected the lag order 1, we can estimate the VAR coefficients. This time we find all the coefficients significative except one:

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix} = \begin{bmatrix} 17.12 \\ -12.86 \end{bmatrix} + \begin{bmatrix} 0 & -0.61 \\ -0.17 & 0.29 \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}$$

The impact of Y_2 on Y_1 seems to be significantly different than 0. We again check this with an impulse response function.

24.9.4 A rolling regression

If we want to perform a rolling regression of Y_2 on Y_1 with a window of 10 quaters, we can simply use the function or define the rolling regression ourselves. We get:



And apply our function. Note that we need to add some zeros at the beginning of our beta list with the size of our window.

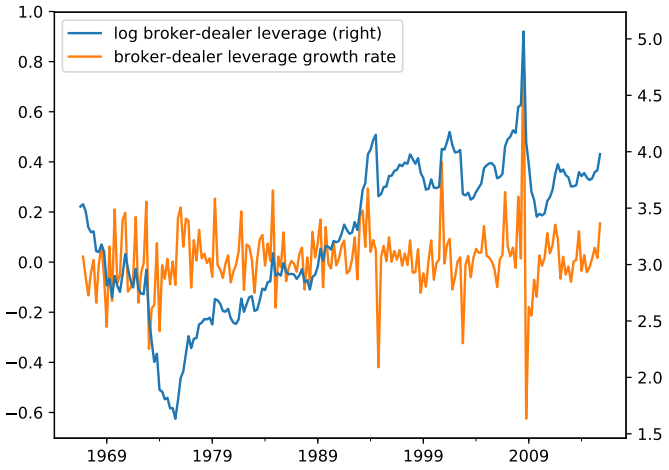
24.10 Rice and wheat prices related to world supplies of rice and wheat

We use data from FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS:

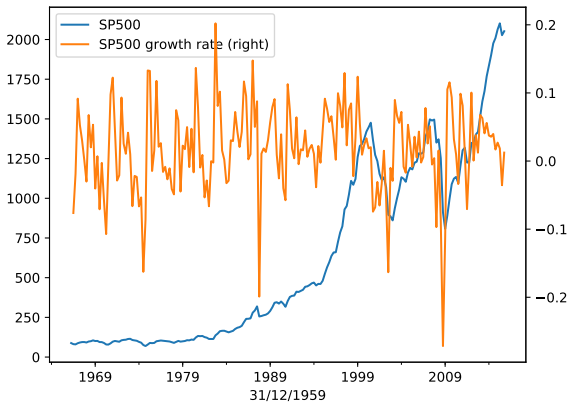
- **Question 49**
  - Based on the exercises done so far, perform:
    1. apply the same approach above to rice and wheat prices and world supplies.
    2. add lagged variables.
    3. once the section 25 has been covered, apply Granger causality tests to this data set.

25 Granger causality test: academic paper replication

In this section[110], with data taken from the Federal reserve website here, that provides historical data of financial accounts of the United States, namely for the broker-dealers, we suggest to replicate a Granger causality test done in the paper (Serletis and Istiak, 2018). Keep in mind that a cause cannot come after the effect, information on one variable helps to predict another one but this do not imply any sort of economic causality (e.g. a banks changes its business model, its stock price moves).
(Serletis and Istiak, 2018) suggested to study the causality between broker-dealer leverage and the stock market in the United States, using quarterly data since 1967. Leverage defined as (total assets) / (total assets - total liabilities) is a measure of how much debt an investor assumes in making an investment. The main idea is that a higher leverage means a higher amount of cash in circulation that should drive the asset prices up.



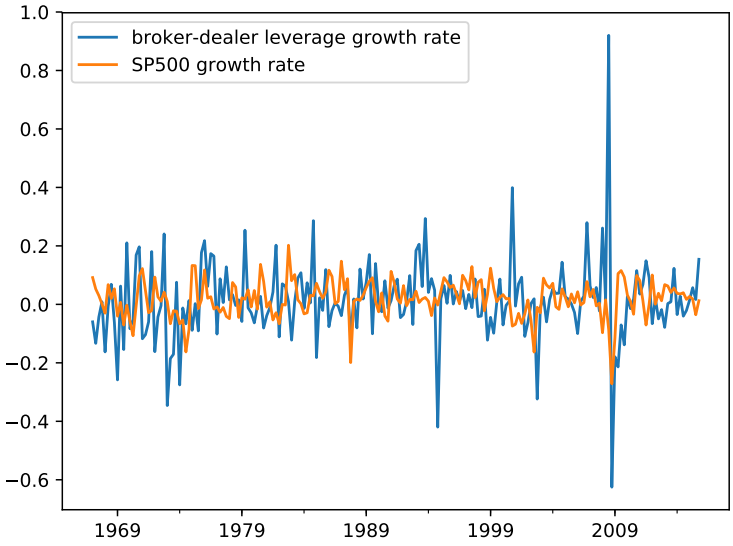
Note that as a first difference in this exercise, we do not divide the S&P 500 by the CRB BLS spot index[111]



[110] broker_dealer_leverage_vansteenberghe.py

[111] we tried to divide with publicly available data, like the US GDP or the consumer price, but it did not help reach the same conclusion for the lag order selection

From the naked eye, it would be very difficult to observe which returns seems to cause the other:



25.1 ADF and cointegration test

As in the paper, we perform ADF and cointegration test. We find that the variables are cointegrated. Hence we can build an ECM to study the two time series.

25.2 Granger causality definition and test

We can say that a variable x_t for which we have a stationary time series of length T does not Granger cause another stationary time series y_t of length T if the information we have on that variable does not improve our ability to predict the other variable. In fact the Granger causality test is just a test whether certain autoregressive coefficients are zero (the null hypothesis is H_0: $\forall i, \beta_i^2 = 0$ and $\gamma = 0$). We do a restricted (r) and unrestricted (u) regression presented in this order, adding the lagged estimated residuals $\hat{\epsilon}_{t-1}$from the cointegration regression as the two time series are cointegrated:

$$\begin{cases} y_t = \sum_{i=1}^{m} \alpha_i^1 y_{t-i} + c^1 + \nu_t^1 \\ y_t = \sum_{i=1}^{m} \alpha_i^2 y_{t-i} + \sum_{i=1}^{n} \beta_i^2 x_{t-i} + \gamma \hat{\epsilon}_{t-1} + c^2 + \nu_t^2 \end{cases} \tag{73}$$

The order of lag, m and n, are determined by minimizing the Schwarz information criterion from the various regressions.

We compute an F statistics with the sums of squared residuals as in the Wald test[112]

$$F : \frac{\frac{SSR_r - SSR_u}{n+1}}{\frac{SSR_u}{T-m-n-1-1}}$$

We need to compute the critical value[113] and then determine:

[112] for further explanation, the reader can refer to (Verbeek, 2017) sections "2.5.4 A joint test of significance of regression coefficients" and "3.2.2 selecting regressors"

[113] F statistic with a numerator degree of freedom of n and a denominator degree of freedom of $T - m - n - 1 - 1$

- our statistic < critical value: fail to reject the null hypothesis
- our statistic ≥ critical value: reject the null hypothesis

In our case, we look in a F-statistic table $F_{4,192}$ for both variables and our statistic is above the critical value, we reject the null hypothesis. Hence we assume the broker-dealer leverage Granger cause the stock price and vice-versa. Note that our lag selection with the BIC leave us with a different model for the tests on the stock market. As noted in the paper, this is a **linear** test and in times of crises, the relationship might present some non-linearities.

Also note that in their 2006 paper, Diks and Panchenko indicates that their test might under-reject their causality for sample with less than 500 observations which is our case here.

26 Recap on econometric models

Griliches (1983):
A model is a simplified representation of an actual phenomenon, such as an actual system or process. The actual phenomenon is represented by the model in order to explain it, to predict it, and to control it, goals corresponding to the three purposes of econometrics, namely structural analysis, forecasting, and policy evaluation.
Let's consider a system of g independent and consistent (i.e. mutually compatible) equations in the g endogenous variables, $y_1, y_2, \ldots, y_g$, the k exogenous (or lagged endogenous) variables, $x_1, x_2, \ldots, x_k$, and the m parameters, $\delta_1, \delta_2, \ldots, \delta_m$. The model can be written in vector notation:

$$f\left(y, x, \delta\right) = 0 \tag{74}$$

where f is a column vector of g functions.
Assuming the functions are differentiable and that the Jacobian matrix of first-order partial derivatives is non-singular at a particular point:

$$\left|\frac{\partial f}{\partial y}\right| \neq 0 \text{ at } (y, x) \tag{75}$$

the implicit function theorem implies that at this point it is possible to solve the system of equations 74 for the endogenous variables as differentiable functions of the exogenous variables and parameters: $y = \Phi(x, \delta)$ where Φ a column vector of g functions.
Econometric models are generally algebraic models that are stochastic in including random variables (as opposed to deterministic models which do not include random variables). The random variables that are included, typically as additive stochastic disturbance terms, account in part for the omission of relevant variables, incorrect specification of the model, errors in measuring variables, etc. The general econometric model with additive stochastic disturbance terms can be written as the non -linear structural form system of g equations: $f\left(y, x, \delta\right) = \epsilon$ where ϵ is a vector of stochastic disturbance terms. If the conditions of the implicit function theorem are met these equations can be solved for the endogenous variables as differentiable functions of the exogenous variables and parameters, with the stochastic disturbance terms included as additive error terms. The resulting non -linear reduced form is the system of g equations: $y = \Phi(x, \delta) + u$, where u is the vector of the stochastic disturbance terms in the reduced form. The econometric model uniquely specifies not the endogenous variables but rather the probability distribution of each of the endogenous variables, given the values taken by all exogenous variables and given the values of all parameters of the model. Each equation of the model, other than definitions, equilibrium conditions, and identities, is generally assumed to contain an additive stochastic disturbance term, which is an unobservable random variable with certain assumed properties, e.g. mean, variance, and covariance. The values taken by that variable are not known with certainty; rather, they can be considered random drawings from a probability distribution with certain assumed moments. The inclusion of such stochastic disturbance terms in the econometric model is basic to the use of tools of statistical inference to estimate parameters of the model.
The basic econometric model is a linear stochastic structural form:

$$y\Gamma + xB = \epsilon \tag{76}$$

There is a trivial indeterminacy in the structural equations in that multiplying all terms in any one of these equations by a non-zero constant does not change the equation. This indeterminacy is eliminated by choosing a normalization rule, which is a rule for selecting a particular numerical value for one of the non-zero coefficients in each question. A convenient normalization rule is that which

sets all elements along the principal diagonal of the Γ matrix of coefficients of endogenous variables at -1, $\gamma_{h,h} = -1$.
Certain stochastic assumptions are typically made concerning the g stochastic disturbance vectors:

- zero mean, $E(\epsilon_i) = 0$
- same covariance matrix of ϵ_i at each observation, $cov(\epsilon_i) = \Sigma, \forall i$
- uncorrelated over the sample, $E\left(\epsilon_i' \epsilon_j\right) = 0, \forall i \neq j$

These conditions are satisfied is ϵ_i are iid[114]. Under these general assumptions, while the stochastic disturbance terms are uncorrelated over the sample, they can, by be correlated between equations. This latter phenomenon of correlation between stochastic disturbance terms in different equations (due to the fact that there is usually more than one endogenous variable in each equation) is an essential feature of the simultaneous-equation system econometric model and the principal reason why it must be estimated using simultaneous-equation[115] (or VAR models) rather than single-equation techniques.
It is usually assumed that Γ is non-singular, se we can write the model 76 as the reduced form:

$$y = x\Pi + u \tag{77}$$

where $\Pi = B\Gamma^{-1}$ and $u = \epsilon\Gamma^{-1}$. We can write the assumptions:

- zero mean, $E(u_i) = 0$
- same covariance matrix of u_i at each observation, $cov(u_i) = \Omega, \forall i$
- uncorrelated over the sample, $E\left(u_i' u_j\right) = 0, \forall i \neq j$

These assumptions summarize the stochastic specification of the reduced-form equations. Under these assumptions the conditions of both the Gauss-Markov Theorem and the Least Squares Consistency Theorem are satisfied for the reduced-form equations, so the least squares estimators: $\hat{\Pi} = \left(X'X\right)^{-1} X'Y$. The problem of identification is that of using estimates of reduced-form parameters Π and Ω to obtain estimates of structural-form parameters Γ, B, and Σ. Once estimated, we have $\hat{B} = \hat{\Pi}\hat{\Gamma}$ and $\hat{\Sigma} = \hat{\Gamma}'\hat{\Omega}\hat{\Gamma}$. But we can imagine a "bogus" system multiplied by a matrix R: $y\Gamma R + xBR = \epsilon R$, then for this system, the estimates of Π and Ω would be the same as the one form model 76. We say that the system is identified in that all structural parameters can be determined from the reduced-form parameters. In this case, we need a priori information, which are restrictions on the structural parameters imposed prior to the estimation of the reduced form.

[114] Sometimes the further assumption of normality is made: $\epsilon_i \sim \mathcal{N}(0, \Sigma)$
[115] Hausman (1983)

27 Python: Data and model of French hospitals deaths with Covid-19

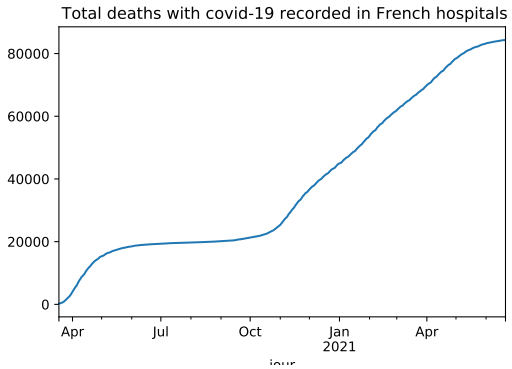
27.1 Data and limitations

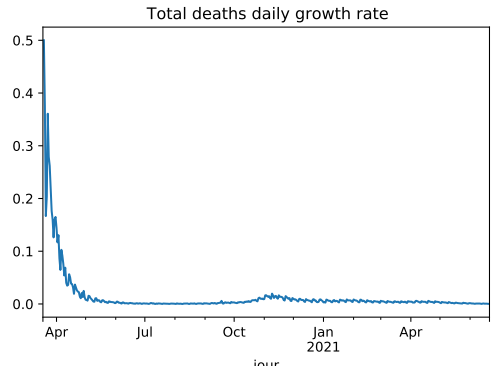
We use[116] data from Sante publique France published each day at 19h00. We focus on the number of recorded deaths in French hospitals as we would not rely on confirmed cases as the test campaign is not systematic.
Nota bene:

- under-reporting:
  - as some Covid-19 people are transferred via trains to other regions, this is creating a downward bias in our data, so for the Parisian and East of France data death toll would be higher
  - there are limited capacities in hospitals, meaning that in highly infected areas, hospitals might not be able to treat all infected people which are then either transferred to other regions' hospital or left to struggle outside of hospitals' monitoring (at home, in the street...)
- we talk about "death with covid-19" which seems more precise than "death due to covid-19", we have no information if this would make a difference in our analysis

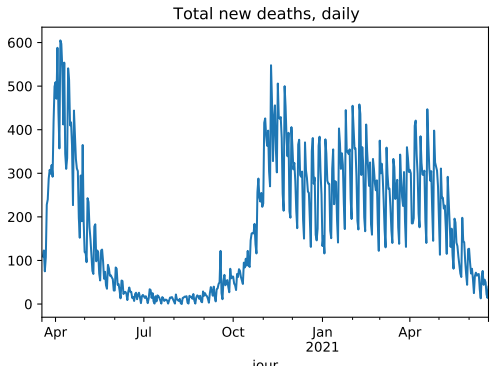
27.2 Overview of death with covid-19 in France

We first have a look at the total recorded death in French hospitals with covid-19 since the 18th of March 2020:





[116] code: 20200807_COVID_19_FR.py



27.3 The SIR model

Since the work of (Kermack et al., 1927), many models have been introduced to model the spread of epidemics. We will limit our analysis here to this SIR model. We write:

$$d_c(t) = \alpha I(t) \tag{78}$$

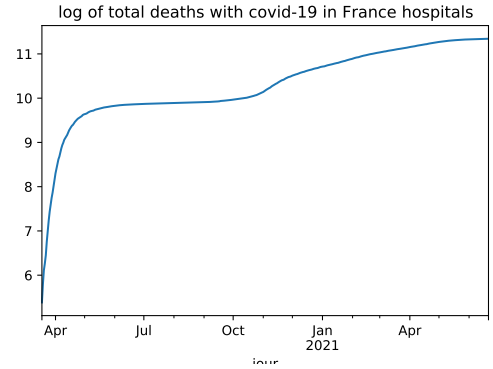
with $I(t)$ the infectious compartment at time t, $d_c(t)$ the cumulative deaths at time t and α the mortality rate of the epidemics, assumed constant. We do not develop the model here, but then it can be approximated in the early phase of the epidemics:

$$\frac{\partial d_c(t)}{\partial t} = \alpha\mu(R_0 - 1)\exp\left[\mu(R_0 - 1)t\right] \tag{79}$$

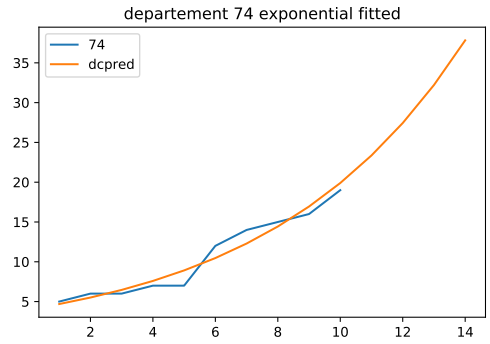
which is to say that at an early stage of the epidemics, the death dynamics in log should fit a linear regression:

$$\log(d_c(t)) = \beta_0 + \beta_1 t + \epsilon_t \tag{80}$$

If the equation 80 is correct over the full sample, then simply taking the log of total death would be a linear function over time:



If it seems linear in the early phase, then maybe the lock-down measures starting mid-March transformed the dynamics.
We will see that in very early days of the epidemics, an exponential model was worryingly fitting:



27.3.1 Fit SIR early epidemics on total French data

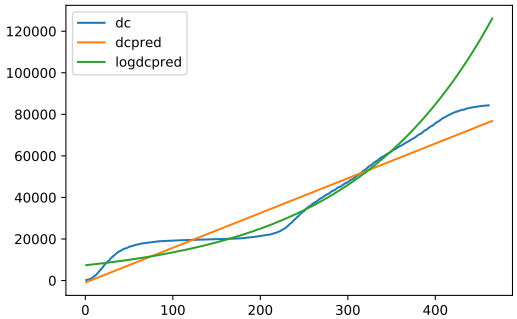
We fit our data to the model 80:

Dep. Variable:	logdc	**R-squared:**	0.522
Model:	OLS	**Adj. R-squared:**	0.518
Method:	Least Squares	**F-statistic:**	136.7
Date:	Fri, 07 Aug 2020	**Prob (F-statistic):**	8.68e-22
Time:	09:36:42	**Log-Likelihood:**	-125.45
No. Observations:	127	**AIC:**	254.9
Df Residuals:	125	**BIC:**	260.6
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	8.1724	0.117	69.895	0.000	7.941	8.404
day	0.0185	0.002	11.690	0.000	0.015	0.022

Omnibus:	57.150	**Durbin-Watson:**	0.012
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	154.861
Skew:	-1.785	**Prob(JB):**	2.36e-34
Kurtosis:	7.064	**Cond. No.**	148.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
We would not want to put too much meaning on the estimates of β_0 nor β_1, what we do in the following is to fit an OLS on the first 90% of the points and then see if the following 10% follows this fitted exponential or deviates from it:

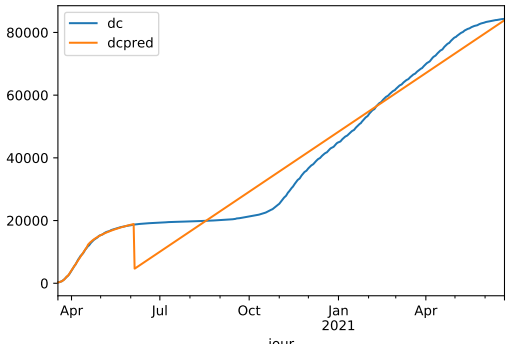


In both cases, we see a very poor fit our the 80 or simple OLS model.

27.4 Model by parts

As the dynamic of the COVID-19 pandemic evolved over time in France, we might want to decompose the period:

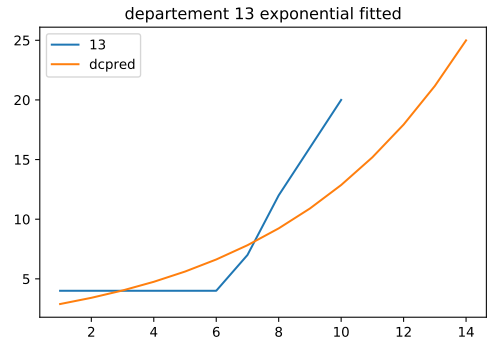
1. an exponential phase as model 80 from 2020-03-18 to 2020-03-28
2. a linear phase from 2020-03-28 to 2020-04-18
3. a almost-logarithm model
4. then a last linear model.

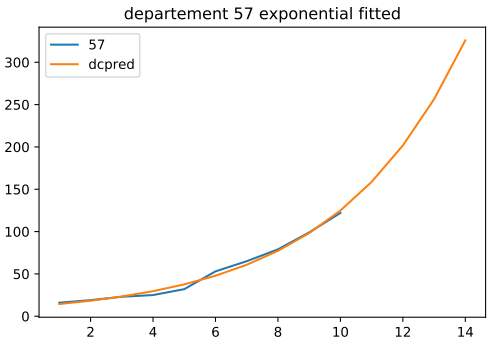


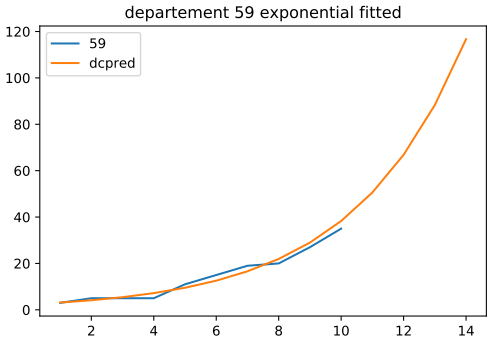
27.5 Fit SIR early epidemics on some French departments

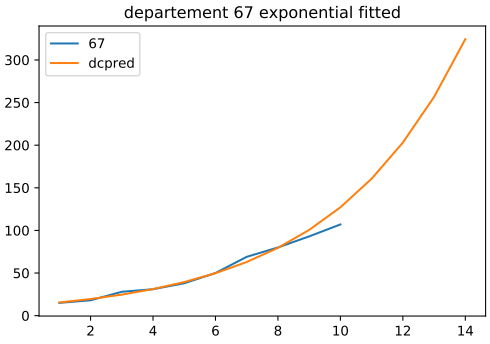
We next focus on some of the most impacted regions (departement) in France:

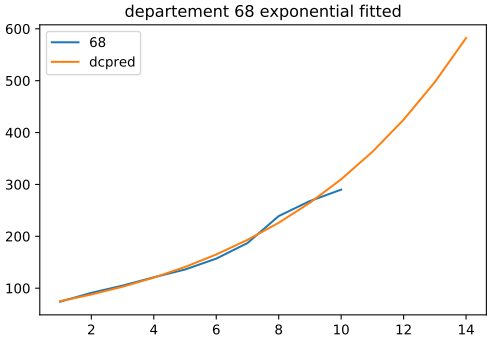
- 13 Bouches-du-Rhone
- 57 Moselle
- 59 Nord
- 67 Bas-Rhin
- 68 Haut-Rhin
- 74 Haute-Savoie
- 75 Seine (Paris)
- 76 Seine-Maritime
- 93 Seine-Saint-Denis
- 94 Val-de-Marne

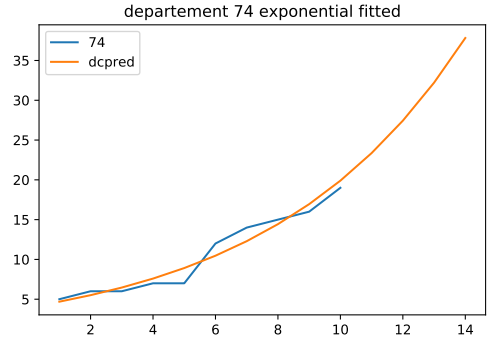

departement 13 exponential fitted
13
dcpred

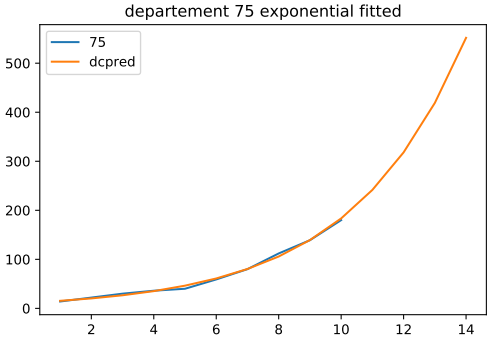

departement 57 exponential fitted
57
dcpred

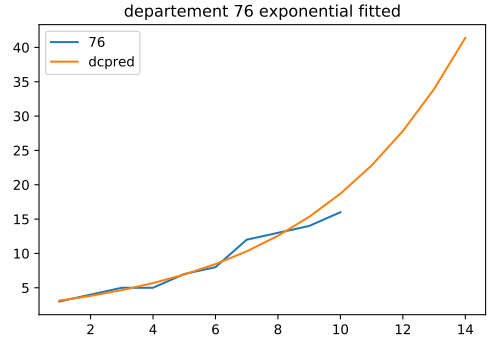

departement 59 exponential fitted
59
dcpred

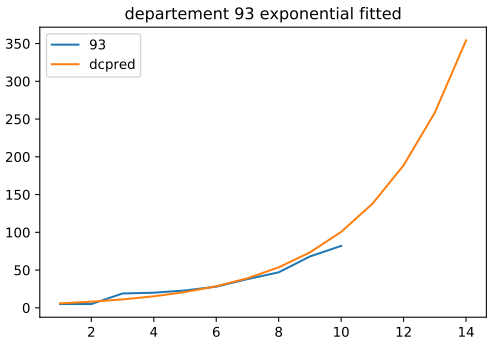

departement 67 exponential fitted
67
dcpred

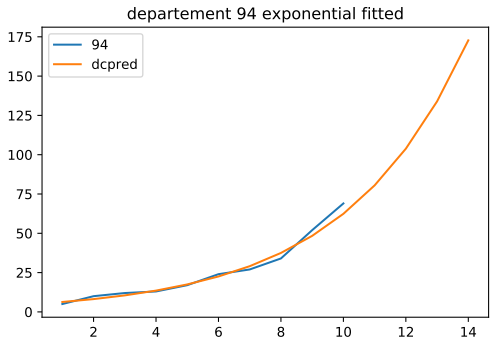

departement 68 exponential fitted
68
dcpred


departement 74 exponential fitted
74
dcpred


departement 75 exponential fitted
75
dcpred


departement 76 exponential fitted
76
dcpred





27.6 Conclusion on modelling the COVID-19 death evolution in French hospitals

While it might be early to conclude, one can say while the death toll with covid-19 rises in France, the lock-down measures could be part of the explanation while the dynamics doesn't fit the expected exponential trend in the early phase and seems to follow a linear from day 20 after the start of the lock-down period, then even a decreasing trend.
(Atkeson et al., 2020) challenge the efficacy of lock-down measure, finding that over several locations,

> the growth rates of daily deaths from COVID-19 fell from a wide range of initially high levels to levels close to zero within 20-30 days after each region experienced 25 cumulative deaths

and this could be attributed to voluntary social distancing, the network structure of human interactions, and the nature of the disease itself.

28 python: Nonlinear time series model: an introduction

The Covid-19 brought both behavioral changes (citizens changed their habits when it came to working, shopping, etc.) and structural changes (in places where a lock-down was imposed, some business were imposed to shut down). Such changes can be sudden and nonlinear. We introduce methods to test the nonlinearities and some models which allow a switch from different regimes.

28.1 BDS test for nonlinearity: non iid time series

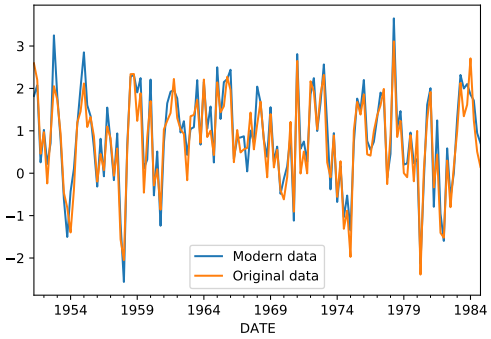
Broock et al. (1996) developed a test for nonlinearity with the null hypothesis H_0: independent and identical distribution of a random variable. If H_0 can be rejected on residuals from a fitted linear time series model, then the linear model is likely to be mis-specified.
For the quarterly US GNP growth rates[117], we reject H_0, hence they are not iid.

28.2 Markov-switching model of growth rates

Hamilton (1989) explored discrete shifts in regime where the economy might be either in a fast or slow growth phase. Taking y_t as the quarterly growth rate of the real US GNP in percent:

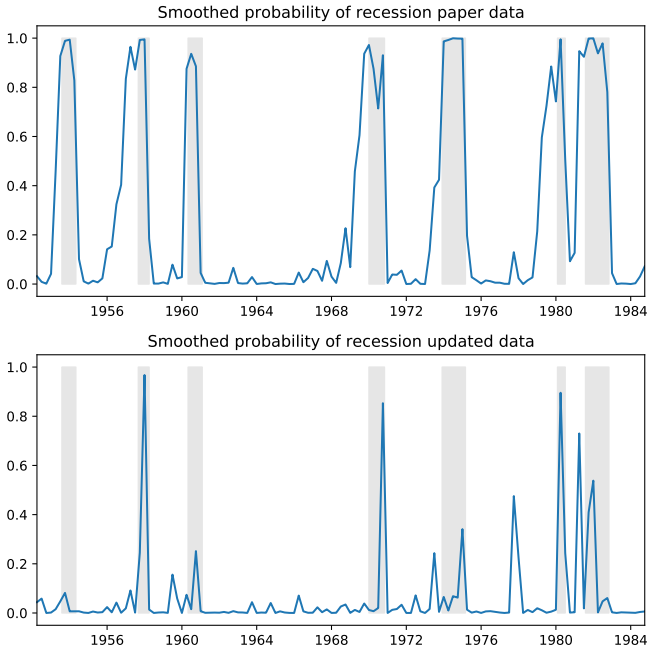
$$\begin{cases} y_t = \alpha_1 S_t + \alpha_0 + z_t \\ z_t = \sum_{i=1}^{r} \phi_i z_{t-i} + \epsilon_t \\ \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right) \\ P\left(S_t = s_t | S_{t-1} = s_{t-1}\right) = \begin{bmatrix} q & 1-p \\ 1-q & p \end{bmatrix} \end{cases} \tag{81}$$

The original data set was taken from the *Business Conditions Digest* we compare it with the modern Federal Reserve Bank of St. Louis data, we observe some discrepancies while the observations are consistent both in terms of sign and relative magnitude:

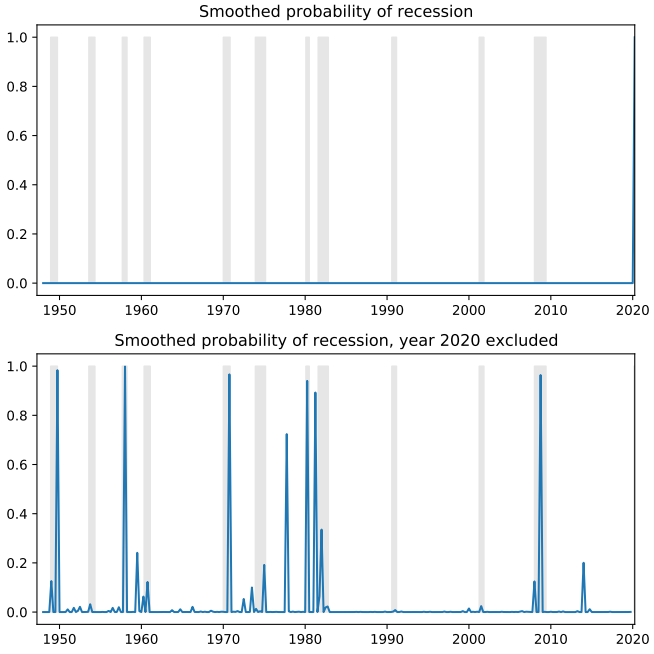


With the original data set, we are able to replicate the model calibration and the inferred probability that the economy was in the falling GNP state at date t. But we find that this model is very sensitive to the data, with the updated data set, we do not find a model that is as precise to inferred probability was in the falling GNP state:

[117] vansteenberghe_MSAR.py



If we now take the full period up to end of 2019, fitting an AR(3) as suggest by the AIC, we are able to infer correctly some states (1950, 1970, 2008) but with some false positives. If now we include the year 2020 in our data set to calibrate our model, 2020 was such an outlier that we cannot infer any other dates for the falling state:



This paper was discussed with some suggestions, namely by Hansen (1992).

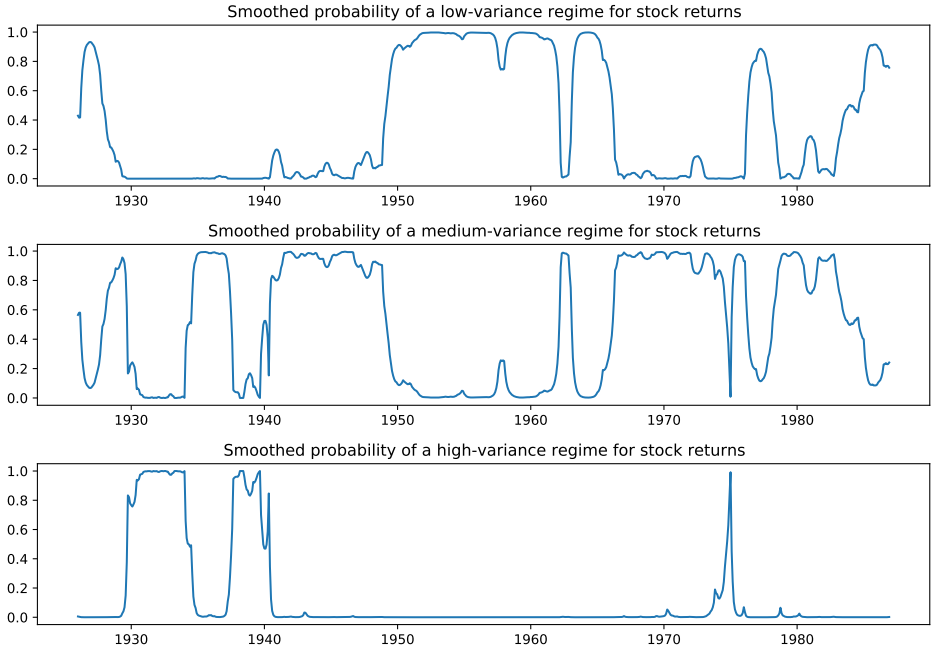
28.3 Three state Markov-switching process

Kim et al. (1998) estimates conditional distributions of parameters as well as the unobserved state they consider as random variables. Following Hamilton and Susmel (1994), they model the demeaned

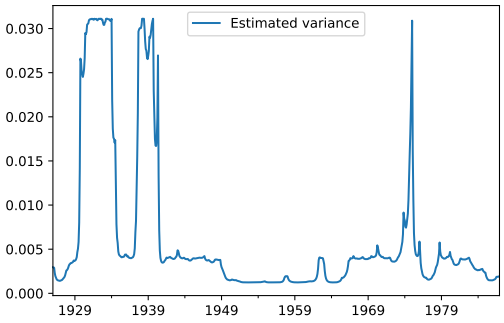
monthly stock market returns y_t as a three state Markov-switching process following:

$$\begin{cases} y_t \sim \mathcal{N}\left(0, \sigma_t^2\right) \\ \sigma_t^2 = \sum_{i=1}^{3} \sigma_i^2 S_{it} \\ S_{kt} = 1 \text{ if } S_t = k, \text{ and } S_{kt} = 0 \text{ otherwise; } k = 1,2,3 \\ Pr\left[S_t = j | S_t = i\right] = p_{ij} \; i,j = 1,2,3 \end{cases} \tag{82}$$

Applying[118] a BDS test on y_t we reject that the returns are iid.
We find the same parameters as in the paper with the following smoothed probability of the variance to be in each of the three states:

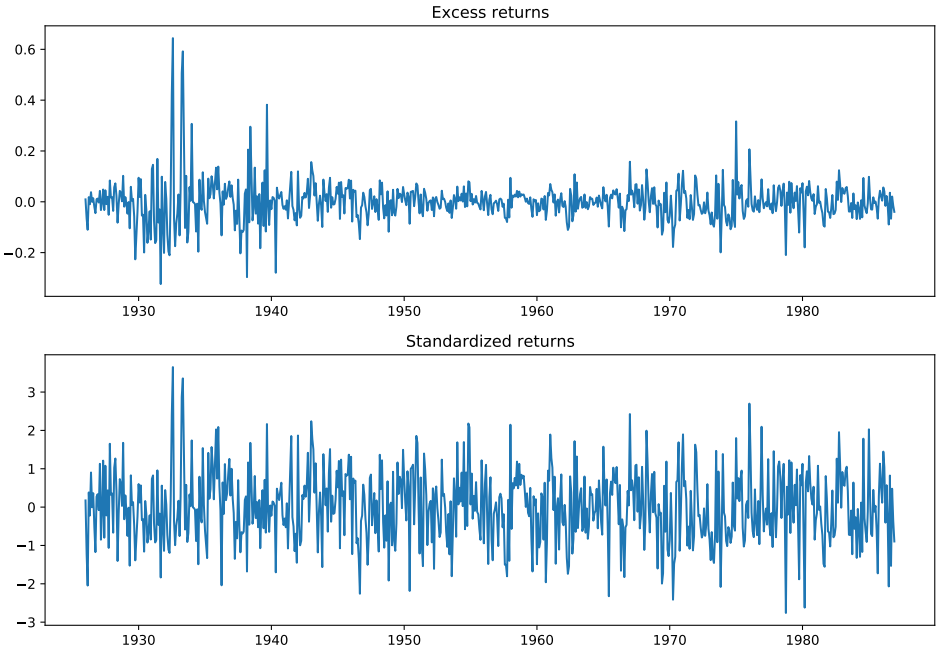


with the estimated variance:

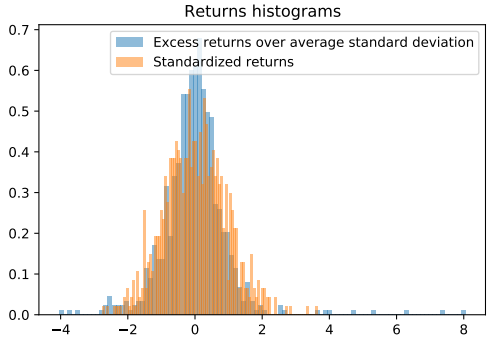


the standardized returns $\frac{y_t}{\sigma_t}$ now looks more iid and a BDS test doesn't reject this assumption:

[118] vansteenberghe_variance_switching.py



and the histograms:

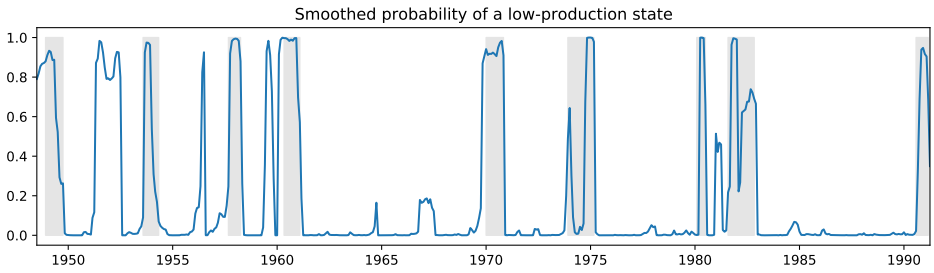


28.4 Time-Varying Transition Probabilities

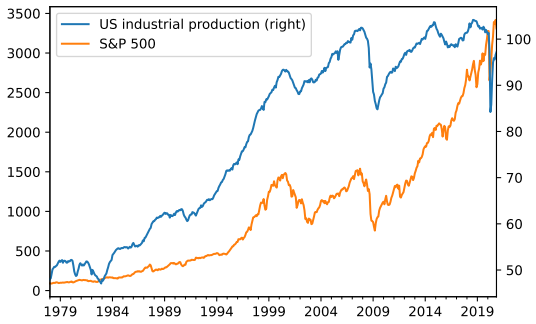
Filardo (1994) also assumes that the state of the economy cannot be known with certainty and allow for a time-varying transition probability between states. He modifies equation 81 with:

$$P(S_t = s_t | S_{t-1} = s_{t-1}) = \begin{bmatrix} q(z_t) & 1 - p(z_t) \\ 1 - q(z_t) & p(z_t) \end{bmatrix} \tag{83}$$

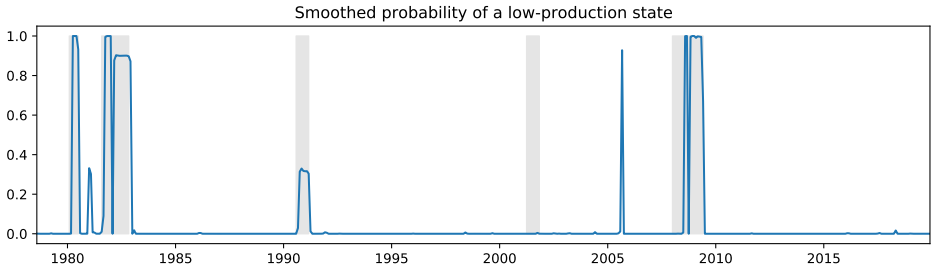
where z_t is an economic-indicator variable and applies it to the seasonally adjusted US industrial production (NBER recessions in grey):



We apply this model[119] to more recent data, seasonally adjusted US industrial production and use the S&P 500 as an economic-indicator variable (as one configuration in Filardo (1994):



If we leave year 2020 out of our sample, the outcome can be interpreted although it is difficult to avoid false positives:



But if we include year 2020 in our sample, then we cannot really interpret the model as it predicts low production state for most of the observation period. In 2020, the production dropped severely while the financial markets faced low turbulence, hence our model become more (too?) sensitive.

28.5 Mean and variance switching states

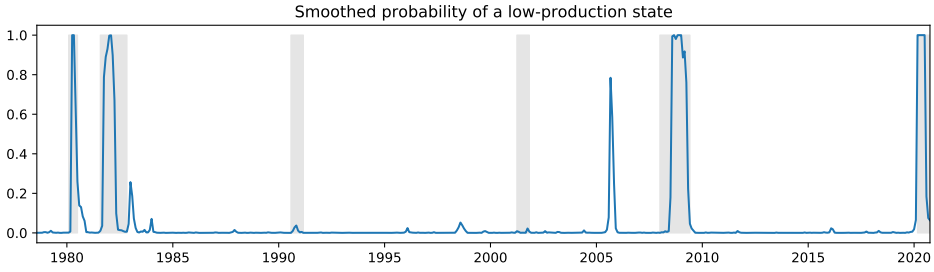
McConnell and Perez-Quiros (2000) combined mean and variance state switching to model US output

[119] 20201104_TVTP.py

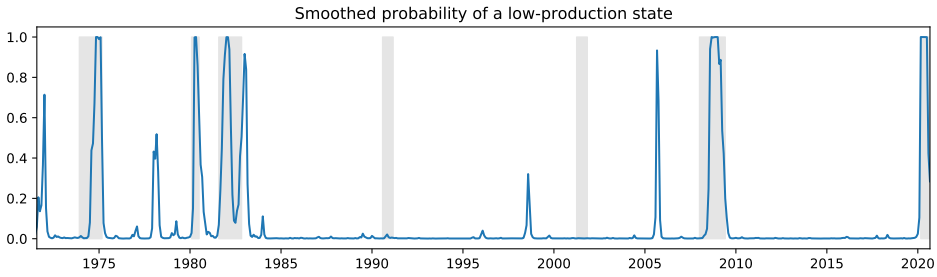
fluctuations and identify periods or recessions and expansions:

$$\begin{cases} y_t = \mu^{S_t} + \Phi\left(y_{t-1} - \mu_{S_{t-1},V_{t-1}}\right) + \epsilon_t \\ \epsilon_t \sim \mathcal{N}\left(0, \sigma^2_{V_t}\right) \\ P\left(S_t = s_t | S_{t-1} = s_{t-1}\right) = \begin{bmatrix} q & 1-p \\ 1-q & p \end{bmatrix} \\ P\left(V_t = v_t | V_{t-1} = v_{t-1}\right) = \begin{bmatrix} u & 1-v \\ 1-u & v \end{bmatrix} \end{cases} \tag{84}$$

In our application[120], we merge S_t and V_t into one latent variable and we find that low-production state are associated with higher variance. We find that the Hamiltonian model is not fit to identify low-production state[121]. Our implemented version of the McConnell and Perez-Quiros (2000) model allows to identify low-production states probability in line with observed recessions, with still some false positives and false negatives:



We might want to use another economic-indicator variable, such as the National Financial Conditions Index (NFCI), the conclusions are similar:



Following these works, Adrian et al. (2019) argue

> that the entire distribution of GDP growth evolves over time, with the left tail of the distribution positively correlated with slack in financial conditions.

[120] 20201105_McConnell.py

[121] in the Hamiltonian model, equation 84 is modified with V_t removed

29 Python: Event studies

There is an abundant academic literature on event studies. The main application is to measure the effects of an economic event on the value of a firm. Methods are described in (MacKinlay, 1997).

29.1 Event study on fines impact on banks

As in the 2016 paper on regulatory sanctions impact in the UK (John Armour and Polo, 2017), we suggest to work on the impact of US fines on banks.

29.1.1 Methodology

We recap the main methodology for an event study.
The abnormal return for a given bank i at time t is:

$$AR_{i,t} = R_{i,t} - \alpha_{i,t} - \beta_i R_{m,t} \tag{85}$$

With $R_{i,t}$ and $R_{m,t}$ the daily returns of the bank i and of the market m. In Europe, the market can be proxied using Eurostoxx index.
The α and β are obtained with a simple OLS regression on a one-year window, so from $t-261$ to $t-2$.
This measure can help decide if on a specific date for a specific bank i there was a significant event for the market.
In order to assess whether all the N banks were impacted, we can compute the average abnormal return:

$$AR_t = \sum_{i=i}^{N} \frac{AR_{i,t}}{N} \tag{86}$$

Finally the cumulative average abnormal return for the period $[t_1, t_2]$ is:

$$CAR(t_1, t_2) = \sum_{t=t_1}^{t_2} AR_t \tag{87}$$

As in the paper (John Armour and Polo, 2017), we encourage to winsorize the abnormal returns to 90% percentile of data.
Finally, it is possible to compute a reputational loss as :

$$\text{Reputational Loss} = \Delta V_t \text{ - Fine}$$

With ΔV_t the change in market capitalisation in the event window around the announcement.

29.1.2 Data

The journal *Le Monde* has released a list of bank fines here.
The data range from 2007 to 2016.
We suggest to focus on banks impacted by descending order of fines.
So let say: Bank of America, JPMorgan, BNP, Deutsche Bank.
You will have to define the exact day of the announcement of the fine.
Start with daily returns, then weekly or monthly and identify if the announcement had an impact and generated an abnormal return on the bank, and what about other banks on the same day?

29.2 Event study on stress test results impact on banks

For an event study on the effects of the publication of stress test results on bank stock prices, we suggest to follow[122] the same method presented section 29.1.
We focus on 53 quoted banks over a 10 year horizon. The window for the regression is one year, so 365 calendar days.
The red dotted line is the day of announcement of the Brexit, 24th of June 2016 and the green dotted line indicates the day of the announcement of the EBA stress test results.
We can illustrate the data with the stock prices and returns for Deutsche Bank AG.
We "zoom" on Deutsche Bank stock prices around the dates of the Brexit and EBA stress test announcements:
Note that the code can take some time to run as we apply 200,000 regressions and computations on returns.
Then we can apply a threshold, for example 4% and observe the selected dates related news that could explain such an abnormal return across the board:

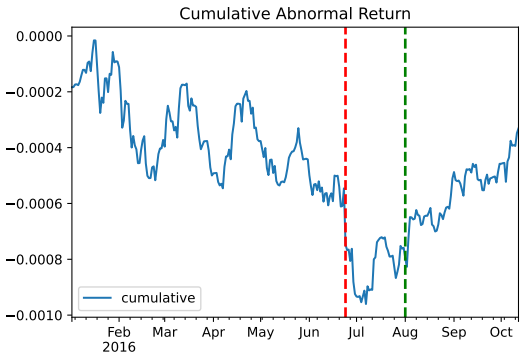
- 18 Sep 2008 Les banques centrales des grands pays industrialises se sont concertees jeudi et ont a nouveau injecte massivement des liquidites sur les marches financiers
- 18 Jan 2012 World Bank fears Europe's crisis could set off deeper global slump than Lehman collapse

Now if we apply a negative threshold of -4%, the day of the announcement of the Brexit clearly appears.
The trading day after the EBA stress test results publication, on the first of August 2016, the observed average abnormal return is around -1%. The day after, on the second of August 2016, the observed average abnormal return is around -3.4%. Either the market needed some time to assess the results or those effects would rather be "dominant" compared with the results publication:

- 2nd August 2016: Deutsche Bank and Credit Suisse booted from top stock market index
- 2nd August 2016: Monte dei Paschi rescue deal over its 10 billion euro worth of non-performing loans was "agreed" but "investors are starting to doubt just how viable the rescue plan" according to the Business Insider and Commerzbank results were published and disappointing

We could also observe bank by bank to identify which day saw significant abnormal returns and see if the trading day following the publication of the stess test results, so the first of August 2016 saw significant abnormal return or see the applicable threshold and how many trading days saw similar or above abnormal return over the last 9 years.
Finally, we plot the cumulative average abnormal returns over a window of one year (365 days) and "zoom" on early 2016, placing the date of the EBA stress test results announcements end of July 2016 or the day of the Brexit announcement on 24th of June 2016.

[122] code: market_reaction_ST_vansteenberghe.py


Cumulative Abnormal Return
0.0000
−0.0002
−0.0004
−0.0006
−0.0008
−0.0010
cumulative
Feb
2016
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct

30 First steps with R - functions, loops, imports and exports

We suggest in the code `helloworld_vansteenberghe.R` to get familiar with some R basics features essential before starting a research economics project:

- use the **help** of R to use existing functions
- set your working directory to access your data sets and save figures and outputs
- install and load **libraries**
- assign values to variables, work with strings
- perform mathematical operations
- **plot** your data
- learn to define your own **functions**
- **if- else** conditions
- use of sapply(), eval(), and fsolve()
- learn to implement **for** and **while** loops
- work with arrays, matrices, **data frames**, lists
  - R data frames are tightly coupled collections of variables which share many of the properties of matrices and of lists, used as the fundamental data structure by most of R's modelling software; a matrix-like structure whose columns may be of differing types (numeric, logical, factor and character and so on)
- handle errors (and bypass them while going through data)
- use of random data generation according to distribution (e.g. normal, uniform, binomial)
- export and import data from external csv files
- perform basic linear regressions

30.1 Some frequent questions on R

30.1.1 How to indicate the path to files to be imported

With RStudio, there is a simple way to indicate the path to files that have to be imported:

1. File > Import Dataset > From Excel...
2. use Browses and select the a file in the folder where all the data sets are located
3. clikc on "Import", the path to your file is displayed and can be used from now on to change the working directory with setwd()

30.1.2 How to convert dates data to Date in R

To convert dates data to R date format, one can follow the steps here. In simple terms, one need to "indicate" R the format the dates have, e.g. if the date is of the format `31/12/1959`, then in the format is in fact `day/month/yearwith4digits` which in R is written `%d/%m/%Y`. And the column "Date" of the dataframe df can be converted with the command:

```
df$Date <- as.Date(df$Date, format = "%d/%m/%Y")
```

References

Tobias Adrian, Nina Boyarchenko, and Domenico Giannone. Vulnerable growth. *American Economic Review*, (4), 2019.

Carol Alexander. *Market Risk Analysis, Volume II, Practical Financial Econometrics*. Wiley, 2008.

Carol Alexander. Market risk analysis, volume iv, value at risk models. *Wiley*, page 492, 2009.

Plenz D Alstott J, Bullmore E. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 2014.

Michel Armatte, Annie L. Cot, Jacques Mairesse, and Matthieu Renault. Edmond malinvaud and the problem of statistical induction. *Annals of Economics and Statistics*, 2017.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, David Heath, and Hyejin Ku. Coherent multiperiod risk adjusted values and bellman's principle. *Annals of Operations Research*, 2007.

Andrew G. Atkeson, Karen Karen Kopecky, and Tao Zha. Four stylized facts about covid-19. *Federeal Reserve Bank of Atlanta Working Paper*, 2020.

Marco Bardoscia, Stefano Battiston, Fabio Caccioli, and Guido Caldarelli. Debtrank: A microscopic foundation for shock propagation. *PLOS ONE*, 2015.

BIS. Revisions to the basel ii market risk framework. *BIS*, 2009.

Fischer Black. The pricing of commodity contracts. *Journal of Financial Economics*, 1976.

Olivier Jean Blanchard and Danny Quah. The dynamic effects of aggregate demand and supply disturbances. *The American Economic Review*, 1989.

Werner F.M. Bondt and Richard Thaler. Does the stock market overreact? *The Journal of Finance*, 1985.

George Box and Gwilym Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

W. A. Broock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron. A test for independence based on the correlation dimension. *Econometric Reviews*, 1996.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 2009.

Corina Constantinescu, Gennady Samorodnitsky, and Wei Zhu. Ruin probabilities in classical risk models with gamma claims. *Scandinavian Actuarial Journal*, 2018.

Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 2001.

Jon Danielsson, Lerby Ergun, and Casper de Vries. Challenges in implementing worstcase analysis. *Bank of Canada Staff Working Paper*, 2018.

James E. H. Davidson, David F. Hendry, Srba Frank, and Stephen Yeo. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the united kingdom. *The Economic Journal*, 1978.

David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 1979.

David A. Dickey and Wayne A. Fuller. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 1981.

Juan J. Dolado, Tim Jenkinson, and Simon Sosvilla-Rivero. Cointegration and unit roots. *Journal of Economic Surveys*, 1990.

Bertrand Hassani Dominique Guegan and Kehan Li. Measuring risks in the tail: The extreme var and its confidence interval. *Risk and Decision Analysis*, page 15, 2017.

John Elder and Peter E. Kennedy. Testing for unit roots: What should students be taught? *The Journal of Economic Education*, 2001.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 1982.

Robert F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 1987.

Giorgio Fagiolo, Mauro Napoletano, and Andrea Roventini. Are output growth-rate distributions fat-tailed? some evidence from oecd countries. *Journal of Applied Econometrics*, 2008.

Andrew J. Filardo. Business-cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics*, 1994.

Christian Francq and Jean-Michel Zakoïan. Testing the nullity of garch coefficients: Correction of the standard tests and relative efficiency comparisons. *Journal of the American Statistical Association*, 2009.

Xavier Gabaix. Zipf's law for cities: An explanation. *The Quarterly Journal of Economics*, 1999.

Xavier Gabaix. Power laws in economics: An introduction. *Journal of Economic Perspectives*, 2016.

Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 2006.

C. W. J. Granger and P. Newbold. Spurious regressions in econometrics. *Journal of Econometrics*, 1974.

M.D. Intriligator Z. Griliches. Handbook of econometrics, volume 1. *Handbook of econometrics*, 1983.

James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 1989.

James D Hamilton and Raul Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 1994.

B. E. Hansen. The likelihood ratio test under nonstandard conditions: Testing the markov switching model of gnp. *Journal of Applied Econometrics*, 1992.

S. Hariri, M. Carrasco Kind, and R. J. Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

Jerry A. Hausman. *Chapter 7 Specification and estimation of simultaneous equation models*. Elsevier, 1983.

M. Hermanussen, H. Danker-Hopfe, and GW Weber. Body weight and the shape of the natural distribution of weight, in very large samples of german, austrian and norwegian conscripts. *International Journal of Obesity*, 2001.

Keisuke Hirano and Jonathan H. Wright. Forecasting with model uncertainty: Representations and risk reduction. *Econometrica*, 2017.

Peter J. Huber. *Robust Statistics*. Wiley, 1981.

Ventzislav Ivanov and Kilian Lutz. A practitioner's guide to lag order selection for var impulse response analysis. *Studies in Nonlinear Dynamics and Econometrics*, 2005.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 1993.

Colin Mayer John Armour and Andrea Polo. Regulatory sanctions and reputational damage in financial markets. *Journal of Financial and Quantitative Analysis*, page 15, 2017.

Eric Jondeau and Michael Rockinger. Testing for differences in the tails of stock-market returns. *Journal of Empirical Finance*, 2003.

Phillip Kearns and Adrian Pagan. Estimating the density tail index for financial time series. *The Review of Economics and Statistics*, 1997.

William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 1927.

Lutz Kilian and Helmut Lütkepohl. *Structural Vector Autoregressive Analysis*. Cambridge University Press, Cambridge, 2017. ISBN 9781107196575. doi: DOI:10.1017/9781108164818. URL `https://www.cambridge.org/core/books/structural-vector-autoregressive-analysis/DAF4217439EA585D10902D58A8849E06`.

Chang-Jin Kim, Charles R. Nelson, and Richard Startz. Testing for mean reversion in heteroskedastic data based on gibbs-sampling-augmented randomization1c.-j.k. acknowledges assistance from non-directed research fund, korea research foundation, 1996.1. *Journal of Empirical Finance*, 1998.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

William S. Krasker, Edwin Kuh, and Roy E. Welsch. *Chapter 11 Estimation for dirty data and flawed models*. Elsevier, 1983.

Christopher Krauss. Statistical arbitrage pairs trading strategies: review and outlook. *Journal of Economic Surveys*, 2017.

Johannes Ledolter and Bovas Abraham. Parsimony and its importance in time series forecasting. *Technometrics*, 1981.

Wassily Leontief. Structural matrices of national economies. *Econometrica*, 1949.

Wassily Leontief. *Input-Output Economics*. Oxford University Press, 1986.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 2013.

W. K. Li and A. I. McLeod. Distribution of the residual autocorrelations in multivariate arma time series models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1981.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 2012.

G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 1978.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

Helmut Lütkepohl and Markus Krätzig. *Applied Time Series Econometrics*. Cambridge University Press, 2004.

Thomas Lux and Didier Sornette. On rational bubbles and fat tails. *Journal of Money, Credit and Banking*, 2002.

A. Craig MacKinlay. Event studies in economics and finance. *Journal of Economic Literature*, 1997.

James G. MacKinnon. Critical values for cointegration tests. *Queen's Economics Department Working Paper*, 2010.

Benoit Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 1963.

Rosario N. Mantegna and H. Eugene Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 1995.

Ricardo A. Maronna, R. Douglas Martin, Victor J. Yohai, and Matías Salibián-Barrera. *Robust Statistics: Theory and Methods (with R)*. Wiley, 2nd edition, 2019.

Margaret M. McConnell and Gabriel Perez-Quiros. Output fluctuations in the united states: What has changed since the early 1980's? *American Economic Review*, 2000.

Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 1991.

Mark Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 2005.

Adrian Pagan. The econometrics of financial markets. *Journal of Empirical Finance*, 1996.

Carla M. A. Pinto, A. Mendes Lopes, and J. A. Tenreiro Machado. A review of power laws in real life phenomena. *Communications in Nonlinear Science and Numerical Simulation*, 2012.

V. Plerou, P. Gopikrishnan, X. Gabaix, L. A. N. Amaral, and H. E. Stanley. Price fluctuations, market activity and trading volume. *Quantitative Finance*, 2001.

Hossein Rad, Rand Kwong Yew Low, and Robert Faff. The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance*, 2016.

Mette Rytgaard. Estimation in the pareto distribution. *ASTIN Bulletin*, 1990.

Apostolos Serletis and Khandokar Istiak. Broker-dealer leverage and the stock market. *Open Economies Review*, 2018.

Robert J. Shiller. *Irrational Exuberance*. Princeton University Press, 2016.

Christopher Sims. Macroeconomics and reality. *Econometrica*, 1980.

Yolanda Stander, Daniel Marais, and Ilse Botha. Trading strategies with copulas. *Journal of Economic and Financial Sciences*, 2013.

Marno Verbeek. *A Guide to Modern Econometrics*. Wiley, 2017.

James Ma Weiming. *Mastering Python for Finance*. Packt, 2015.

Michael A. Williams, Grace Baek, Yiyang Li, Leslie Y. Park, and Wei Zhao. Global evidence on the distribution of gdp growth rates. *Physica A: Statistical Mechanics and its Applications*, 2017.

Eric Zivot and Jiahui Wang. *Cointegration*. Springer, 2003.