1 taking a look at your data

Before setting off into econometrics, it's always worth to study the data. In this session, we calculate sample statistics, create histograms, fit distributions and plot extensively – all for the sake of better understanding the dataset at hand.

1. Open RStudio, create a new R script and save it in a dedicated folder. Set your working directory accordingly (*Session → Set Working Directory → To Source File Location*).

2. Load the following R packages: `ggplot2`, `forecast` and `qqplotr` (for plotting); `tseries` (for testing the normality of returns); `xts` (for dealing with time series objects); `fBasics` (for detailed summary statistics); `tseries` (for testing the normality of returns). We will need them at some point during this session.

3. Load `s1_data.txt` into RStudio. This dataset contains the following variables:

name	description	source
SP500	S&P500 index	R. Shiller's website
SPDIV	dividend per "share" of the S&P500 index, annualized	
USRF	3-Month T-Bill rate on the secondary market, annualized	St. Louis Fed
MSCIE	MSCI Europe index, in USD	Datastream
GOLD	Price of troy ounce of gold on LME, in USD	
SMIUSD	Swiss Market index, in USD	
BORD	Liv-ex Bordeaux 500 index, in USD	

4. Transform your data so as to get a time series object (using the `xts` package). Explore the properties of your time series.

5. Plot the index values of *SP500* and *SMIUSD* together. We see that this is not informative since the series do not start at the same date and have different scales.

6. Compute the logarithmic returns for *SP500*, *SMIUSD* and *MSCIE*. These are the three indices we will deal with during the remainder of the session.

7. Let us look at the period since 1990-01. Subsample the dataset accordingly.

   Henceforth we will only deal with returns. Also, we pretend that dividends do not exist for the time being.

8. Risk-free rates are usually (not on Kenneth French's website though) quoted in percent per annum: the number "0.23" in December 2015 thus (roughly) means that you could have earned 0.23% over the next 12 month, which is tantamount to say that in December alone you could have cashed in only 0.23%/12=0.02%. Transform the US risk-free rate into monthly values.

   Hint: you should see value of 0.3533 in January 2006

9. Compute the Sharpe ratios of the three risky assets. Define the Sharpe ratio of series i as:

$$SR(r_i) = \frac{\mathbb{E}[r_i - r_f]}{\sigma[r_i]}$$

where r_f is the risk-free rate. Interpreting it as a risk-adjusted return, which of the three assets is the best investment?

10. Calculate the correlation matrix of the returns of *SP500, SMIUSD, MSCIE*, and *rf*.

11. The *SP500* return of 0.6% per month that an American investor could have earned since 1990 is something like 8% per year, after compounding. Can we be sure – say, 99% sure – that this is not due to luck and that the true expected value of *SP500* monthly return is not zero? Run a simple t-test assuming that returns are not autocorrelated at any lag.

12. Can we be sure – say, 99% sure – that the market has on average outperformed the T-Bills since 1990? In other words, is the probability of rejecting the null hypothesis

$$H_0 : \mu_{SP500} - \mu_{\text{rf}} = 0 \tag{1}$$

lower than 1% given the evidence since 1990? Perform a t-test.

13. Look at the descriptive statistics of the three risky assets. How much could you have lost in a single month from an investment in each? What is the probability of losing this much or more if returns admitted a normal distribution with mean and variance equal to the sample estimates since 1990?

14. Study the skewness and kurtosis of the three series and identify the most normally distributed one. Plot a histogram of this series and test the hypothesis that the underlying DGP[1] admits a Gaussian density.

15. Go on and create a Q-Q plot of the returns of *SMIUSD*.

16. Create a scatter plot of *smiusd* versus *sp500*, with *smiusd* on the x-axis. Add a linear regression line on top.

[1] data-generating process