MFIN 4328
Applied Financial Econometrics
Assignment 1

Prof. Riccardo Sabbatucci

Due: February 14, 2022

*****Please submit your solution in a standalone PDF saved as "YourStudentID_YourFirstName_YourLastName" report that contains the code and output from your R script. You could either do this by clicking "-File-Knit Document-PDF" in R Studio or using R markdown following instructions at this link *****

1. (4 points): Consider the simple monthly stock returns of the value-weighted index (VWRET), the equal-weighted index (EWRET), CocaCola (CCRET), General Electric (GERET) and International Business Machines (IBMRET), from January 1963 to December 2011. Obtain the data from the sheet *stock_ monthly* of the file *data_ assignment1.xlsx*.

   value-weighted index is preferred

   (a) Import the data in R.

   (b) Transform the simple returns into log returns.

   (c) Plot the simple and log returns for all the series. **add comments with economic reasoning; 4 - 5 sentences**

   (d) Show the scatterplot of all the log return series. Interpret the graphs. **time series scatter plot**

   (e) Compute the sample mean, variance, skewness, excess kurtosis, minimum and maximum of log returns. Compare the two EWRET and VWRET series. **print every computed statistic**

   (f) Are the sample mean, skewness and excess kurtosis of log returns statistically different from zero? Use the 5% significance level to draw your conclusion and discuss their practical implications.

   by judgement or (better) with tests => last sentence

   (g) Obtain the histograms of the returns and compare them with normal and Student distributions that have the best fit of the empirical distribution. Do monthly log returns look approximately normal (Student)? Which looks closer to normal (Student): the VWRETD or EWRETD indexes or the individual companies? Confirm your observations with a Jarque-Bera test at the 5% significance level.

   also interpretation of the coefficients => economic interpretation

   (h) Estimate a CAPM regression for Coca-Cola, GE and IBM. You can assume the monthly risk free rate is 0% for this part. What are the alphas α? What are the betas β? Discuss the statistical significance of the estimates.

2. (6 points): The dataset pension contains information on net financial wealth (*nettfa*), age of the survey respondent (*age*), annual family income (*inc*), family size (*fsize*), and participation in certain pension plans for people in the U.S. The wealth and income variables are both in thousands USD (1,000USD).

   keep track of units

   (a) Import the data in R (Hint: use the library(haven))

   (b) How many single-person households are there in the dataset?

   (c) Use OLS to estimate the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u \tag{1}$$

   and report/interpret the results for the case of single-person households.

(d) Does the intercept from the regression in part (c) have an interesting meaning? Explain.

(e) Find the p-value for the test $H_0 : \beta_2 = 1$ against $H_1 : \beta_2 < 1$. Do you reject H_0 at the 1% significance level?

(f) Find the p-value for the test $H_0 : \beta_2 = 1$ against $H_1 : \beta_2 \neq 1$. Do you reject H_0 at the 1% significance level?

(g) If you run a simple regression of $nettfa$ on inc, is the estimated coefficient on inc much different from the estimate in part (c)? Why or why not?

(h) Estimate the following model with inc^2, age^2 and $fsize$ as additional regressors not conditioning on single households (e.g., using all values of fsize)

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 inc^2 + \beta_4 age^2 + \beta_5 fsize + u \tag{2}$$

What do you find? talk about economic intuition

(i) Calculate the F-test of the restriction $H_0 : \beta_3 = \beta_4 = 0$ in model (2). Do you reject the null hypothesis? Explain.

(j) Rescale the variable inc in (2) by dividing it by 10. Re-estimate (2) with the rescaled variable. What do you find? Is the statistical significance different?

(k) Test the original model in (1) for heteroskedasticity using the Breusch-Pagan test. What do you find?

(l) Estimate the original model in (1) with heteroskedasticity-robust (e.g., White) standard errors.

(m) Standardize all the variables in the model (1). Re-estimate (1) with the standardized variables. What do you find? How do you interpret the coefficients on inc and age?

3. (4 points): Use the data in ceo_salary.dta to answer this question

(a) Estimate the model

$$lsalary = \beta_0 + \beta_1 lsales + \beta_2 lmktval + \beta_3 ceoten + \beta_4 ceoten^2 + u \tag{3}$$

by OLS using all of the observations, where *lsalary*, *lsales*, *lmktvale* are the natural logarithms of the variables. Report results with "standard" (e.g., not adjusted) standard errors.

(b) Report the count, mean, standard deviation, min and max of the explanatory variables. Why are most of the explanatory variables in natural logs? Explain.

(c) Re-estimate the model with White standard errors and compare the t-stats.

(d) Obtain the residuals and standardize them. How many of them are above 1.96 in absolute value? If the standardized residuals were i.i.d draws from a standard normal distribution, about how many would you expect to be above 2 in absolute value with 177 draws?

(e) Add the dummy *college* and the interaction term *college·lsales* to the model in (3). Estimate this model. How you do interpret the coefficient on the interaction term? careful interpretation

(f) Let us suppose you found out that the variable *lsales* has measurement error because sales have been overstated, and the true value of the sales is 10% smaller (e.g., reduce each value of lsales by 10%). Call this variable *lsales_ adjusted*. Re-estimate the model in (3) with *lsales_ adjusted* as an explanatory variable instead of the original *lsales*. That is,

$$lsalary = \beta_0 + \gamma_1 lsales\_adjusted + \beta_2 lmktval + \beta_3 ceoten + \beta_4 ceoten^2 + u$$

What is the coefficient γ_1 on *lsales_ adjusted*? Compare it with the original one on *lsales*, β_1

(g) Drop all observations where the standardized residual in point (d) is greater than 2 in absolute value. Re-estimate the model in (3). Do you notice any difference in the OLS estimation?

4. (6 points): The dataset local_returns.dta contains panel information on the monthly stock returns of firms (variable: *ret*), returns of firms in the same area (variable: *city_ returns*), returns of firms in the same industry (variable: *indret*) and a few other risk factors (MRP, HML, MOM) from 2000 to 2010. The variable *permno* is the firm identifier, while *date* is the the time variable. Individual firms (e.g., permnos) are allocated to 20 cities and 12 industries.

   (a) Is it a balanced or unbalanced panel? Provide aggregate summary statistics.

   (b) Report average returns and volatilities of firms in the 20 cities (e.g., summary statistics by city). Do you see any differential returns amongst cities?

   (c) We are interested in estimating the following panel regression:

$$ret_{it} = \alpha + \beta_1 city\_returns_{it} + \beta_2 indret_{it} + \varepsilon_{it} \tag{4}$$

pooled OLS?

(d) Estimate (4) using pooled OLS without dummies.

(e) Test for heteroskedasticity in the residuals. What do you find?

(f) Estimate (4) using pooled OLS with time fixed effects (e.g., time dummies).

(g) Estimate (4) using pooled OLS with both time and firm fixed effects. Compare your results with those obtained in the previous point.

(h) Estimate (4) using pooled OLS with both time and firm fixed effects, and with double clustering on both firm and time dimensions. Compare your results with those obtained in (f) and (g).

(i) Estimate (4) using the Fixed Effect (FE) estimator (e.g., firm fixed effect).

(j) Estimate (4) using the First Difference (FD) estimator.

(k) Compare the FE, FD and pooled OLS with both time and fixed effects with double clustering (point (h)).

(l) Generate a dummy variable equal to 1 if the city=10, and equal to 0 otherwise. Let us assume that city=10 is San Francisco, and call this dummy $dummy_{SF}$. Create an interaction term between the $dummy_{SF}$ and the city_returns variable and add it to the model in (4). Re-estimate the model with the new interaction term

$$ret_{it} = \alpha + \beta_1 city\_returns_t + \beta_2 indret_t + \gamma_1 dummy_{SF} \cdot city\_returns_t + \varepsilon$$

Is it coefficient γ_1 statistically different from zero? How do you interpret it?