Part 0 – Open Stata, and make your own do-file

- Using windows explorer, make a new folder called `H:\rproject\clab\clab42`.
- Download the S&P500 description (stata) data, and the 2014 and 2015 stock price (excel) data from **Bb** and put it in this folder.
- Start Stata through the start menu button.
- In de white command window type `doedit` to start de do-file editor. Place the Stata screen on the left and the do file editor on the right such that you can easily switch between the two.
- In the first two lines of the do-file type

```
cls                              //this clears the screen
clear all                        //this clears the memory
cd "H:\rproject\clab\clab42"     //this is your path
```

- Save the do-file as `clab42.do`.
- In the following, as always, don't forget to apply Rules 1&2: paste the command in your `clab42.do` file, press control-s to save and control-d to run.

Part 1 – Appending and preparing public data

In the first part of this exercise you are going to import and append publicly available data of firms from the S&P500, the leading American stock market index. The aim is to create a dataset that contains information about the sector that a given firm operates in, and observe its growth(decline) in stock market value in two consecutive years. This will then be used in Part 2 to determine the (relative) cyclicality of the energy sector compared to other S&P500 firms in a differences-in-differences design applied to the Capital Asset Pricing Model (CAPM, Berk& DeMarzo, 2011, chapter 10-12).

1. Open the data `sp500desc2018.dta` and use `br` and `desc` to learn about it. Use Google to find out what the `tickersymbol` is. Finally type

   ```
   tab gicssector
   ```

   to see the various sectors that the firms are operating in.
2. The data that contains the starting and ending stock prices of the firms are in two excel files. They need to be imported before they can be used in Stata.

   ```
   import excel "sp500_2014.xls", firstrow case(lower) clear
   ```

   Then, you want to have a look at is using `br`, and save the file as `sp500_2014` so you can use it again later. Do the same for the 2015 data.
3. Now you have all the data saved in Stata format. To compare the 2014 stock market return to that of 2015, we need to have them both in one single dataset. There are two ways of doing this: `merge` or `append`. The `merge` command connects data horizontally by matching records on some variable. In this case, we need the `append` command to connect both datasets vertically. We want the stock market value of the firms of both years under one another such that they can both be used in the same regression as different units of observation. You achieve this by typing

   ```
   append using sp500_2014
   ```

   Then type `sort tick year` to sort the data and use `br` to see it. Do you understand the new data structure?

4. The open and close values are the opening and closing values of the stock within the year. To generate the return within the year, type
```
gen R = (close - open)/open
```
Type `br` to see the data again. Making a histogram is also insightful (`hist R`). Were these years any good for investors? Type `tabstat R, by(year) s(mean sd)` to see.
5. To make a distinction between firms from different sectors, we need to connect the returns data to S&P500 description data from **q1**.
```
merge m:1 tickersymbol using sp500desc2018.dta
```
Then type
```
tab _merge year, m
drop if _merge==2
gen msmpl = (_merge==3)
```
to understand how many records we could match. Why could we not match all records do you think? And why do we `drop if _merge==2`? The final command generates a variable that marks all the firms in the data that we have a description ("matched sample"); of the other firms we only observe the stock market prices, not the type of sector that the firm operates in.
6. Before turning to the analysis, it is useful to create a variable that marks all firms that we observe both in 2014 and in 2015. You do this by looking at each firm separately (`bys`) and testing if there are `_N==2` observations. This is the "balanced sample". All other firms that only appear once (in 2014 or 2015) get value 0:
```
bys tick: gen bsmpl=(_N==2)
```
Finally, type `tab bsmpl msmpl` to see how many firm-years we have with description, balance, and both.

Part 2 – The DiD

In the second part we are going to use the data to see the (relative) cyclicality of stocks using the CAPM. In section 12.3 of Berk & DeMarzo (2011), it is explained how the CAPM can be used to estimate the "Beta" of a stock. The Beta of a stock measures its degree of cyclicality. This measure has great interest in Finance, because it directly relates to the return that investors require (in market equilibrium) to hold it. The higher Beta, the more cyclical a firm is said to be, and the higher the return required to compensate (the "risk premium").

To find the Beta of stock i, you need to perform the following regression

$$R_{it} - r_f = Alpha_i + Beta_i(R_{Mkt,t} - r_f) + \varepsilon_{it}, \tag{1}$$

where R_{it} is the return of stock i in period t, r_f is the risk-free interest rate, and $R_{Mkt,t}$ is the return on the optimal market portfolio in period t, and ε_{it} a random noise term. The $Alpha_i$ is the structural over(under) performance of stock i (should be 0 in theory), and its $Beta_i$ is the degree if cyclicality of stock i.

Figure 12.2 from Berk & DeMarzo (2011) shows how the $Alpha_i$ and $Beta_i$ can be estimated using this model, by fitting a line through the data. There are three problems with this: *i*) what is $R_{Mkt,t}$? *ii*) what is r_f? *iii*) what if we only have a few periods? If we are really interested in the value of $Beta_i$, we need to solve all these problems. But what if we only want to *compare* the Beta of groups of stocks and test if they are the same? Then, we can use a differences-in-differences design, which only needs two periods of data, and no information on $R_{Mkt,t}$ or r_f.

Say we would like to compare two groups of stocks, $G_i = 0$ and $G_i = 1$. And we have only two periods of data $D_t = 0$ and $D_t = 1$. Then, we can re-write the model as (don't try it, its complicated):

$$R_{it} = Alpha_0 + (Beta_1 - Beta_0)(R_{Mkt,1} - R_{Mkt,0})X_{it} + \left(Alpha_1 - Alpha_0 + (Beta_1 - Beta_0)ER_{Mkt,0}\right)G_i + Beta_0\left(R_{Mkt,1} - R_{Mkt,0}\right)D_t + \epsilon_{it} \quad (2)$$

Again ϵ_{it} is a noise term, and $X_{it} = G_i \cdot D_t$. This model can be written shortly as

$$R_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + \epsilon_{it}, \quad (3)$$

where $\beta_1 \equiv (Beta_1 - Beta_0)\left(R_{Mkt,1} - R_{Mkt,0}\right)$, and the others defined similarly. The point from this derivation is that, if a change of X_{it} from 0 to 1 has an effect on returns ($\beta_1 \neq 0$), that must mean that $(Beta_1 - Beta_0) \neq 0$ and $\left(R_{Mkt,1} - R_{Mkt,0}\right) \neq 0$. Because market returns usually vary over time, it means $\beta_1 \neq 0 \Leftrightarrow Beta_0 \neq Beta_1$. So using this model we can test if the Beta's of both groups of stocks are the same using only two time periods, and without knowing the exact values of $R_{Mkt,t}$ and r_f.

7. Type
```
tab gicssector
gen util = (gicssector=="Utilities")
gen energy = (gicssector=="Energy")
```
to see what sectors there are in the data and make dummy variables for two sectors in particular: "Utilities" and "Energy". The first deals with electricity, while the second deals with oil (mainly).

8. To generate variables following the differences-in-differences design, we type
```
gen G = (util | energy)
gen D = (year==2015)
gen X = G*D
```
Explain what each of these commands does.

9. With all variables in place we can run the differences-in-differences regression
```
eststo did: reg R X G D, cluster(tickersymbol)
```
We use the `cluster()` option because many firms (`tickersymbol`) appear twice in our dataset. Can you explain what the coefficients are? Is the Beta of the "Utilities"/"Energy" similar to, or significantly different from, the rest?

10. Let's put what you have done into a graph (copy-paste into do-file):
```
predict Rcfact
replace Rcfact = Rcfact - _b[X]*X      //remove the effect
twoway (lfit R D if G==0, lc(blue)) (lfit R D if G==1, lc(black)) ///
       (line Rcfact D if G==1, lp(dash) lc(black) sort), ///
       xlabel(0 `""Before" "(2014)""' 1 `""After" "(2015)""') ///
       legend(order(1 "non-Energy" 2 "Energy" 3 ///
       "Energy if same cyclicality")) ytitle(R) xtitle("")
```
Interpret the graph; what are all lines?

11. Let's do the same analysis also on the balanced sample, and also those that were matched, and put everything into one table:
```
qui eststo didb: reg R X G D if bsmpl, cluster(tickersymbol)
qui eststo didbm: reg R X G D if bsmpl & msmpl, cluster(tickersymbol)
esttab did didb didbm, b(a2) se nogap star(* 0.10 ** 0.05 *** 0.01)
```

Do your conclusions change after restricting the sample?

12. The test $\beta_1 = 0$ tests $(Beta_1 - Beta_0)(R_{Mkt,1} - R_{Mkt,0}) = 0$. But what about β_1/β_3? This gives $\frac{\beta_1}{\beta_3} = \frac{(Beta_1 - Beta_0)(R_{Mkt,1} - R_{Mkt,0})}{Beta_0(R_{Mkt,1} - R_{Mkt,0})} = \frac{Beta_1 - Beta_0}{Beta_0}$. So by taking the ratio of β_1/β_3, we get the relative difference in Beta's; the market returns drop out completely. In Stata you get it by typing
```
nlcom _b[X]/_b[D]
```
So, are the "Utilities" and "Energy" sectors more or less cyclical that the rest? By what percentage? Is it significant?

Reference

Berk J., P. DeMarzo (2011). *Corporate Finance*. Pearson.