# 6th set SAS assignments

**[Task A]**

Data description:

| Variable | Variable Description |
|---|---|
| Stock/Ticker | Ticker Symbol |
| PIN | Probability of Informed Trading |
| mktcap | Market Capitalization |
| Company | Company Name |

**[1]** Create a library `saskurs` to access the files `mkt.sas7bdat` and `pins.sas7bdat`. Merge the two data sets with `PROC SQL` and keep only those stocks for which you have an observation for the PIN <u>and</u> the market capitalization. Compute the log of the market capitalization and name it `lmktcap`.

**[2]** Estimate the following linear regression model:

$$PIN_i = \alpha + \beta \cdot ln(mktcap_i) + \varepsilon_i$$

Rename the estimated parameters as `alpha` (intercept) and `beta` (lmktcap). Write the estimated parameters into an output data set `param`.

**[3]** Merge the estimated parameters to your data and compute the expected PIN (name it `pinhat`) as:

$$\widehat{PIN}_i = \hat{\alpha} + \hat{\beta} \cdot ln(mktcap_i)$$

**[4]** Use `PROC GPLOT` to create a plot of the regression line overlaying a scatter plot of the original data.

<u>Short</u>: Plot `pinhat` and `pin` versus `lmktcap`.

**[Task B]**

Data description:

| Variable | Variable Description |
| --- | --- |
| Obs | Index number of observation |
| Date | Date of Trading |
| Time | Time of Trading (5 minute intervals) |
| Buy | Number of buys in a certain 5 minute interval |
| Sell | Number of sells in a certain 5 minute interval |

**[1]** Access the data set `Buysell.sas7bdat` from the library `saskurs`. Compute the correlation coefficient between the number of buys and the number of sells.

**[2]** Write a macro which selects a subset of the data set from a flexible start date to a flexible end date. Use this macro to select all observations from January into a data set `bsjan` and all observations from February into a data set `bsfeb`.

**[3]** Write a macro which computes the mean, the 0.25 percentile, the 0.75 percentile and the sum of buys <u>and</u> sells for each five minute interval. Write an output data set containing the results. Call this macro for the complete data set, the January data set and the February data set resulting in three different output data sets.