QMM: Exercise Sheet 1 - Simple and Multiple Linear Regressions

Fabien Baeriswyl, Jérôme Reboulleau, Tom Ruszkiewicz

Exercise 1. For this exercise, use the `USComp.csv` dataset. This dataset is made of the following financial data related to 50 US companies: their assets, earnings per shape, number of employees, market value, their profits, revenues, stockholders equity and finally their total return to investors in 1997. We first want to study the relationship between the number of employees and the profits.

a) Plot the number of employees of a company against its profits (in what is called a scatterplot). Describe the relationship between the two.

b) Determine whether the correlation between the number of employees and the profits is zero or not by the use of a correlation test at $\alpha = 0.05$.

c) Fit the following linear model:

$$p_i = b_0 + b_1 e_i + \epsilon_i$$

where $i \in \{1, \dots, 50\}$, p_i is the profit of the ith company and e_i is the number of employees of the ith company. What do you expect the intercept b_0 to be? How do you interpret b_1 in this case? Report the regression equation. Is the slope coefficient significantly different from 0 at $\alpha = 0.01$?

d) Compute the SST, the SSR and the SSE. Using these figures, comment on the predictions of the model in c).

e) Compute the average number of employees within a company. Then, compute the 95% confidence interval for the average profits given the average number of employees.

f) Is the number of employees a good predictor of profits? If not, staying in the settings of a simple linear regression, what could you do in this case? (*Hint: try to find a better predictor of the profits and re-fit a new (simple) linear model.*)

Exercise 2. Using again the `USComp` dataset, suppose we now want to fit a multiple linear regression model and in particular, we want to regress the profits on all the other variables available to us.

a) Which variables are likely to be significant? Which one are likely to be eliminated from the linear fit?

b) Fit the (complete) multiple linear model

$$p_i = b_0 + b_1 e_i + b_2 a_i + b_3 mv_i + b_4 r_i + b_5 stoeq_i + trti_i + \epsilon_i$$

for $i \in \{1, \dots, 50\}$ where e_i is the number of employees, a_i the assets, mv_i the market value, r_i the revenues, $stoeq_i$ the stockholders equity and $trti_i$ the total return to investors in 1997 of the ith company. Is your intuition from points a) and b) confirmed?

c) Is there anything suspicious with one of the estimated coefficient of the regression? If so, what? Additionally, compute the Variance Inflation Factor (VIF) of the model in b) and describe its purpose.

d) Identify a simpler model than the one fitted in part b).