Statistical Data Analysis
Exercise Session 5

Exercise 1: Fish size dataset

Import the "River Fish" dataset. It contains measurements of 46 fish: "Length", "Breadth" and a dummy variable "Site"' indicating whether the fish was caught upstream or downstream. For this exercise, we will use "Length" and "Site" as the independent variables and "Breadth" as the dependent variable.

1. Start with an exploratory data analysis. Make boxplots and pairwise plots of your variables. Does a linear model seem appropriate?
2. Perform a regression analysis of this dataset. What are the parameter estimates and the error scale estimate?
3. Are the assumptions of the linear model satisfied? Make the appropriate plots (QQ-plot of the standardized residuals, index plot of the residuals, residuals versus fitted values, residuals versus each independent variable).
4. Plot the data together with the fitted regression lines. Color the observations according to the "Site" variable.
5. Perform a PCA analysis of (Breadth,Length), and retain two components. Add the first principal component on the previous plot. Also make a PCA scores plot and color the observations according to the "Site" variable.
6. Do you think a regression model is appropriate here? Explain why (not).

Exercise 2 : Chicago data.

Import the "Chicago" dataset. The data stems from a 1970s study on the fire: fires per 100 housing units relationship between insurance redlining (canceling policies or refusing to insure or renew) in Chicago and racial composition, fire and theft rates, age of housing and income in 47 zip codes. The variables are:

- race: racial composition in percent minority
- theft: theft per 1000 population
- fire: fires per 100 housing units
- age: percent of housing units built before 1939
- involact: new homeowner plan policies and renewals per 100 housing units
- income: median family income in thousands of dollars

For this exercise, we will use involact as our dependent variable and race, theft, fire, age and income as our independent variables.

1. Start with an exploratory data analysis. Make boxplots and pairwise plots of your variables. Does a linear model seem appropriate?
2. What are the parameter estimates and the error scale estimate?
3. Are the assumptions of the linear model satisfied? Make the appropriate plots.
4. Make a plot of fitted response values versus observed response values.

5. What is the design matrix X? What is the correlation matrix of the predictor variables, excluding the intercept term? Does multicollinearity seem to be an issue here?
6. Is at least one of the regression slopes different from zero?
7. Identify which of the predictors do seem to influence the dependent variable. Refit the model with only these predictor variables and check again the model assumptions.
8. Give a 95% confidence interval for β_1 (β_0 is the intercept).
9. Make the anova table and derive the R-squared value from it.
10. What is SSR($X_3|X_1$, X_2)? And SSR($X_1|X_2$, X_3)?
11. Test whether $(\beta_1, \beta_2)^t = (0, 0)^t$ based on a partial F-test, and on the Bonferonni method. Give both tests the same result?
12. Construct a 95% confidence interval for the mean response at $x_0 = (50, 11, 32, 60, 11000)^t$.
13. Construct a 95% confidence interval for the mean response at $x_0 = (100, 11, 32, 60, 21000)^t$.
14. What is $cor(\hat{\beta}_1, \hat{\beta}_2)$?
15. Would you include interaction terms and/or quadratic terms in your model? Explain why (not)

Exercise 3 : Firm data set.
This is the firm dataset used in the course notes (p.56 deel II).

1. Perform a regression analysis on the data. What is the estimated fit for stock firms and for mutual firms? Which coding scheme is used by R to re-parametrize the binary variable?
2. Now, change the parametrization such that stock firms are coded as 1 and mutual firms are coded as 0. Redo the analysis and check that the fitted lines are the same.
3. Add an interaction term. Is it significant?

Some useful functions:

- predict.lm, summary.lm, plot.lm, coef and other methods for lm objects.
- cov2cor
- anova function and associated methods.
- pairs (or car::scatterplotMatrix)
- boxplot
- levels