Univariate Data and Modelling – Exercises

Session 4 – Multiple Linear Regression and Selection of Variables

Exercise 1

Load the body.txt dataset from Toledo. This set contains body girth measurements as well as age, weight, height and gender for 507 physically active individuals - 247 men and 260 women. The data set has the following variables:

ID	Patient ID
Weight	Weight (kg)
Height	Height (cm)
Shoulder	Shoulder girth (cm)
Chest	Chest girth (cm)
Waist	Waist girth (cm)
Abdo	Abdominal girth (cm)
Hip	Hip girth (cm)
Thigh	Thigh girth (cm)
Bicep	Averaged Bicep girth (cm)
Forearm	Averaged Forearm girth (cm)
Knee	Averaged Knee girth (cm)
Calf	Averaged Calf maximum girth (cm)
Ankle	Averaged Ankle minimum girth (cm)
Wrist	Averaged Wrist minimum girth (cm)
Age	Age (years)
Sex	Male, Female

a) Explore the dataset by means of descriptive statistics. Are there any unexpected values? If yes, try to correct them and rerun the exploration. Draw appropriate plots, perform a correlation analysis, ... Try to get as much knowledge on the dataset before starting with model building.

b) Add a new dummy variable X1 to the dataset:

Sex	X1
Male	0
Female	1

- Finish the theoretic interaction model below using Y as response variable and X and X1 as predictor variables.

$Y = \beta_0 + \beta_1 X + \ldots$

- Explain mathematically the function of this dummy variable;
- Given this last answer, should X1 be quantitative or qualitative?

c) Make a subset of 50 random selected observations to test the final model in d) (hint: try the `?sample` command in R). The remaining observations will be used to build the model.

d) Follow the "Global structure for regression analysis" (P 8-38) as seen in theoretical class on this subset.
   - Use "Weight" as response variable and exclude the "Sex" and X1 variables for now;
   - Draw scatterplots of all the variables in function of the response variable and look for (possible polynomial) effects;
   - Test the full model without interaction, are all predictor variables significant? Remove the predictor variable with the highest P-value and rerun the regression. Keep doing this until all variables are significant. Use Akaike information criterion to reduce the model even further;
   - Use appropriate methods to check the proposed model and underlying assumptions (model misspecification and non-linearity, normality and independence of the errors, influential observations);
   - Discuss the R^2 and R_a^2 value. Can we use this model for predictions?
   - Calculate the R^2, R_a^2, and Cp value for all possible subsets. What do you think of your proposed model?
   - EXTRA: use the "scatterplot3d" package to make a 3D scatterplot of the Height, the Waist, and the response variable. Try to include the response plane.

e) Use the observations that were left out of the model-building process to make predictions on the weight. What is the standard deviation on these predictions (hint: use `se.fit = TRUE` in the `predict` command)? Discuss the 95% confidence interval;

f) Rebuild the model as in d), but include the X1 variable and all the interactions of X1 with the other variables (still exclude "Sex"!). Do not test assumptions or calculate R^2, R_a^2, and Cp value for all possible subsets;
   - Is X1 significant? What does this mean? Is any interaction significant? What does this mean?
   - Look at the R^2 and R_a^2 values. Would this model perform better in prediction than the model without X1?
   - Draw a scatterplot of the Height and the response variable. Include the regression line(s). Explain what this mean.

g) Rerun the predictions with the left-out observations. Is the prediction better than without the X1 variable?