Univariate Data and Modelling

Exercise Session 5 : Multiple Linear Regression.

Exercise 1

After graduation you decided to work as a statistical consultant. Your first assignment is to find the relation between subjects FEV and their current smoking status. Forced expiratory volume (FEV) is the amount of air an individual can exhale in the first second of a forceful breath and can be used as a model for pulmonary function. The primary question of interest for your client is whether the children of smokers suffer from reduced pulmonary function (measured as reduced FEV). The client gives you the dataset he wants you to investigate (FEV.DAT) and tells you that the data were recorded during surveys on primary and high schools. It contains the following variables:

Age	Subjects age (years)
Fev	Subjects FEV (liters) - Measured
Height	Subjects height (cm) - Measured
Sex	Subjects sex ("Male" / "Female")
Smoke	Subjects parents current smoking status ("Yes" / "No") - Self-reported

a) Look at this dataset with descriptive statistics.

b) Add a new dummy variable X1 to the dataset such that X1=0 if Sex ="Male" and X1=1 if Sex="Female". Also add a dummy variable X2 such that X2=0 if smoke="Yes" and X2=1 if smoke="NO" (hint: use the function **ifelse**).

c) During the discussion with the client, he mentioned that he ran some statistical methods on the data himself, yielding unexpected results.

- Compare the boxplots of FEV in function of the smoking status and explain; (hint: use the function **boxplot**)
- Run a simple linear regression with the dummy variable for smoking status (X2) as predictor variable and explain the outcome;(hint: use the function **lm**)
- Explain, using regression diagnostics, why we should not use this model (hint: look at the residuals plots);

d) Now you need to work on the report for the client. Run a multiple linear regression on the data with FEV as predicted variable. Use the global structure for regression analysis (Part II of Chapter 8 – page 38) as guidance. When you start model building, do not forget to include all interaction terms with the dummy variables. The original variables ("Sex" and "Smoke") should not be used in the model, only their dummies. Try to answer all the questions below:

- Should we use all the data or is it better to use only a subset (It is not needed to start subsetting straight away, but this is a good extra exercise)?
- Fit the full model on the complete dataset and look at the residual plots. (hint: use the function **lm**) Do you think you have to include polynomial terms? (hint: look at the residuals plots); Redraw the scatterplots of FEV in function of height and age using the **scatter.smooth(x, y)** function. What do you think now? Include all possible useful polynomial terms and re-examine the model.

- Explain all the terms of the final model and the corresponding estimated parameters (magnitude, sign, ...). What is the largest driving force in determining the FEV?
- Remember the primary question of interest. Is there a difference in FEV between smokers and non-smokers?(hint:look at the regression coefficients and p-values)

e) About a week after you handed the results of your analysis to the client, you get a very upset e-mail. The e-mail tells you that the client did a similar survey and used your model to predict the FEV of the subjects. He was surprised that the outcome did not corresponded at all with the true measured FEV and now he is questioning your statistical capabilities.

- What is wrong with this reasoning? Why should the proposed model not be used to make predictions in this case?(hint: Think about the primary question of interest during when the model was built).

Remark

When a function is mentioned in the hints, it is useful to read on the input arguments and output values of the function, by using the keyword "?**function**". For example, executing **?lm** will give you information on the **lm** function.