Regression Modelling

(STAT2008/STAT4038/STAT6038)

Tutorial 4 – More Multiple Linear Regression

Question One

The data file **brains.csv** (available on Wattle) contains data on the average brain and body weights of various species of mammals. These data were discussed in lectures as an additional example of simple linear regression, however, the data were actually collected as part of a larger study: Allison, T. & Cicchetti, D. (1976) "Sleep in Mammals: Ecological and Constitutional Correlates", Science, November 12, vol. 194, pp. 732-734.

Data from the full study are available in the data frame mammalsleep in the **R** library associated with the recommended text by Julian J. Faraway (Linear Models with *R*, Chapman & Hall/CRC, 2005). Instructions for accessing the faraway library are included in the sample assignments, (available in the "Assessment" topic on Wattle) or you could download the data as a .csv file (also available on Wattle) and then use read.csv() to read in the data. Don't forget to attach() the data once you have successfully accessed or input it.

(a) Plot brain weight (Y variable) against body weight (X variable). Use the command identify(body, brain, labels=row.names(mammalsleep)) to identify mammals that appear to be far from the general data crowd. You might like to read help(identify) before attempting this.

(b) Fit a simple linear regression of brain weight on body weight. Use your fitted model and the functions rstandard() and rstudent() to calculate the internally and externally Studentized residuals, respectively, for each data point. These standardised residuals should follow a Student's t distribution (and therefore be approximately normal for a large enough sample size). Which mammals have the largest (in absolute value) standardised residual of each type?

(c) Standardised residuals can be used to test whether there was a location shift in the model (i.e. whether or not an observation is a potential outlier that affected the calculation of the model). The externally Studentized residuals are calculated by excluding each observation from the data before calculating the residual for that observation, so the degrees of freedom associated with the error variance estimate used in the calculation of these residuals is 1 less than the error degrees of freedom for the overall model. Use the externally Studentized residuals from part (b) to assess whether there was a location shift in the model for humans.

(d) Clearly the data has an obvious skew in the values. Take the natural logarithm of both variables and refit the regression. Now which of the mammals have large internally and externally Studentized residuals?

(e) If time allows, you could try using the data from the full study for the purpose for which they were originally collected; to assess sleep in varying species of mammals. Note a number of the other variables have missing values (lm() will exclude any observation which has a missing value on any of the variables in the model) and there are potential problems with including some of the variables in a multiple regression model. Start by examining a scatterplot matrix for the data (using pairs()) and then try an initial regression of log(sleep) on log(body) and log(brain). Is log(brain) a significant addition to a model that already includes log(body)? Would log(body) be a significant addition to a model that already includes log(brain)?

Question Two

The data file **forbes.csv** (available on Wattle) contains Forbes' data on atmospheric pressure and the boiling point of water.

(a) As you might recall from when this example was covered in lectures, there was an obvious outlier in this data set. Fit a simple linear regression of the logarithm of the pressure on the boiling point and calculate the internally and externally Studentized residuals. What are the standardised residual values for the outlier?

(b) Produce a standardised residual plot by plotting the externally Studentized residuals against the fitted values and also examine both a Normal quantile (q-q) plot of the internally Studentized residuals and a bar plot of Cook's distances for this model.
[Note: these last two plots are ones that can be produced using the standard plot() function in **R**, where they correspond to using options: which=2 and which=4, respectively.
For further details see the help file for help(plot.lm), which is the function called by the generic plot() function to deal with linear model objects].

(c) Remove the outlier and refit the regression. Do you notice a dramatic change in the estimated coefficients? What about in the regression *MSE*?

(d) Produce a plot of the externally Studentized residuals against the fitted values and also examine the Normal qq plot and a bar plot of Cook's distances for this new model, excluding the outlier. Do you see any reason to doubt the underlying regression assumptions?

Question Three

The data file **addvar.csv** (available on Wattle) contains a dataset specifically constructed to illustrate some of the potential advantages of added variable plots in the investigation of non-linearity. It contains three columns, the first corresponding to a response variable and the second two to predictors.

(a) Plot the response versus each of the predictor variables. Do you notice any potential non-linearity?

(b) Fit a multiple regression of the response on both of the predictors and plot the residuals versus the fitted values. Now do you see any evidence of non-linearity?

(c) Plot the residuals from the regression in part (b) versus each of the predictors. Now do you see any evidence of non-linearity?

(d) Finally, construct the added variable plots for each of the predictors. Now do you see any evidence of non-linearity?

Question Four

The data file **savings.csv** (available on Wattle) contains information collected for 50 different countries world-wide. The dataset contains a measurement of the average aggregate personal savings rate (SavingsRate) for the populations of each country over the decade of 1960-1970. In addition, measurements of the average percentage of the population under 15 years of age (Pop15) and over 75 years of age (Pop75) during this period, as well as the countries average level of per-capita disposable income (DPI, measured in US dollars) and average percentage growth rate in personal disposable income (DPIgrowth) over the same decade.

(a) Fit a multiple regression of the savings rate on the other four predictor variables. Examine the internally Studentized residuals versus fitted values as well as versus each of the predictors individually and comment on what you see. Also, construct a normal q-q plot and comment on its appearance.

(b) Construct the four possible added variable plots and comment of their structure. Which variables appear to be significant even when the other predictors are already in the model? Do any of the predictors appear to be non-linearly related to the savings rate? Construct the appropriate partial *F*-tests to confirm your visual findings.

(c) Plot each of the predictors versus one another and calculate the correlation matrix for the four predictors. Do you think that multicollinearity is a problem?

(d) Which variable, if any, seems the most likely candidate for removal from the model? Refit the model without this predictor and compare the resulting coefficients, R^2 and *MSE* to those of the full regression.

(e) For this reduced regression, compute the leverages and plot them against the fitted values. Do any of the data points seem to have a potential for high influence? If so, draw the added variable plots and identify the high influence points on them.

(f) Again for this reduced model, calculate the leverages, *DFFITS*, *DFBETAS* and Cook's Distances for each of the data points. Which data points, if any, appear to be having a strong influence based on these measures?
