Regression Modelling

(STAT2008/STAT4038/STAT6038)

Tutorial 5 – Even More Multiple Linear Regression

Question One (this is Question 1 of Sample Assignment 2)

The data for this question are available in library(faraway). You can either follow the instructions for accessing this library 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.

The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy (a surgical procedure). In this assignment we are going to fit an appropriate multiple linear regression model to examine factors affecting lcavol (log of the cancer volume), which is a measure of the size of the cancer tumour (measured in ml).

(a) All of the other variables in the prostate dataset could potentially be included as predictors (explanatory variables) in a multiple regression model with lcavol as the response variable. Produce suitable plots and/or summary R output to investigate the contents of the variables svi, gleason and pgg45. How are these variables distributed? Discuss any potential problems with including these variables in a multiple regression model.

(b) Find an appropriate multiple linear regression model with lcavol as the response variable and lweight, age, lbph, lcp and lpsa as possible predictors. To simplify this exercise, exclude the variables mentioned in part (a) from consideration, assume that all the other variables are already measured on an appropriate scale (i.e. no further transformations are necessary), that an additive model is appropriate (i.e. no interaction terms or quadratic/higher order terms are needed), but do NOT exclude any potential outliers. Do NOT present output for multiple models, choose just ONE model!
Produce the ANOVA (Analysis of Variance) table for your chosen model and summary output showing the estimated coefficients and use these to justify your choice of model. Why have you included the explanatory variables that are included in your model and why have you chosen to exclude other possible predictors?

(c) For your chosen multiple regression model, construct a plot of the externally Studentised residuals against the fitted values and a normal Q-Q plot of the internally standardised residuals and use these plots to comment on the model assumptions.
Also produce selected statistics and/or a plot to investigate and discuss possible outliers and influential observations. Do NOT try to present a table of various statistics showing all 97 observations (though you could select just one statistic and present a relevant plot which shows all 97 observations).

(d) Perform a "nested model" F test to see whether or not any of the subset of possible predictors you have excluded from your chosen model would be a significant addition to your chosen model. If your chosen model includes 4 or 5 of the possible predictors (lweight, age, lbph, lcp and lpsa) then perform a test of the last two or three predictors as an addition to a model that already contains the other variables.
Review the above test results and the output in parts (b), (c) and (d) and also compare your chosen model with the simple linear regression model shown in the models solutions to Question 1 of Sample Assignment 1. Is your chosen a model an improvement in terms of reliably predicting the size of a prostate cancer tumour?

(e) Add an interaction term between lpsa and lcp to your chosen model (if your chosen model does not already include linear terms in lpsa and lcp, also add those terms to the model). Is this term a significant addition to the model? Interpret the coefficients of the terms involving both lpsa and lcp (and their interaction) in this expanded model and in your chosen model.

Question Two

The data file **clouds.csv** (available on Wattle) contains data concerning a cloud seeding experiment. On 24 "suitable" days during the summer of 1975 (a "suitable" day was determined to be any day having a "suitability measure", S of greater than 1.5, see below), a random decision was made whether to seed the clouds with an injection of silver iodide from a small aircraft and the subsequent rainfall was measured in a target area of 10^7 cubic meters. The data set contains 7 columns:

A – An indicator of whether the cloud was seeded or not (0 = not seeded; 1 = seeded);
D – Days after the first day of the experiment (June 16, 1975);
S – Suitability for seeding;
C – Percent cloud cover in the target area (measure using radar);
P – Rainfall in the target area during the hour preceding the seeding decision;
E – Echo motion category (either 0 or 1), a measure of the type of cloud; and
Y – Rainfall in the target area following the decision to seed or not.

We are interested in whether seeding has any effect on rainfall, and thus we wish to fit a regression of Y (or some transformation of it) on A, but we must potentially account for the effects of the other candidate predictor variables.

(a) Fit a regression of Y on all the predictors. Plot the residuals versus the fitted values and a normal q-q plot. Do you think that the response needs to be transformed? If so, what transformation would you suggest? (Recall that Y is a measure of rainfall within a fixed volume.)

(b) Re-fit a regression of your chosen transformation of Y on each of the predictors. Calculate the diagnostic measures t_i (the externally Studentized residuals), D_i (Cook's Distance), and h_{ii} (the leverages). Do any data points appear to be having an undue influence over the regression? Remove no more than of these 3 points and re-fit the regression. How does the removal of these points affect the *MSE* value? What about the coefficient values? (NOTE: The second day data point was noted by the investigators as a "disturbed day".)

(c) Construct an added variable plot for the predictor variable P, correcting for all the other predictors. Do you think that this predictor should be transformed? If so, by what function? (Again, recall that P is a measure of rainfall within a fixed volume.)

(d) Using the transformed predictor noted above as well as the other predictors, calculate the model selection criteria: s_p^2; R_a^2; $PRESS_p$; and C_p, for all the relevant models. (NOTE: We are trying to determine the effect of the predictor A, so any model without this predictor is not relevant to the current analysis.) Create plots of each of these criteria versus the number of parameters in the associated model. Which model(s) do you think fit the data the best? Do you think that cloud seeding has any discernible effect on rainfall?

(e) Redo the above calculations, but excluding the data points which were noted in part (b) as requiring investigation. Do the preferred models change at all?

Question Three

It can be shown that, if we fit the model: $Y = X\beta + \epsilon$, with independent and homoscedastic normally distributed errors; to a dataset from a population in which the all predictors are actually unrelated to the response variable (i.e. $\beta = 0$), then the expectation of the coefficient of determination is:

$$E(R^2) = 1 - E\left(\frac{SS_{Error}}{SS_{Total}}\right) = \frac{p-1}{n-1}$$

Comment on how this reinforces the fact that the coefficient of determination is not a good measure to use for model selection.

Question Four

The data file **highway.csv** (available on Wattle) contains data regarding automobile accident rates on 39 sections of large highways in a particular region of the United States during 1973. For each of the 39 highway sections, the following information was collected:

accdnt – Accidents per million vehicle-miles;
lngth – Length of the section in miles;
ADT – Average daily traffic in thousands of vehicles;
trcks – Truck traffic as a percent of total traffic;
spdlim – Maximum speed limit;
lwdth – Width of lanes in feet;
shld – Width of outer shoulder in feet;
intrchngs– Number of freeway-style interchanges per mile;
sgnls – Number of signaled crossroads per mile;
accsspt – Number of access points per mile; and
lns – Number of lanes of traffic (in both directions).

We are interested in finding a model for the accident rates, and in particular, determining what, if any, effect lowering speed limits would have.

(a) Fit a regression of accident rates on all the predictors. Construct a normal q-q plot. Do you think that the underlying errors may be reasonably assumed to be normal? If not, can you suggest a transformation of the response to a scale where normality may hold? Construct a residual versus fitted value plot. Do you think that the linearity assumption of the regression is reasonable? Do you notice anything else worthy of further investigation?

(b) For the regression fit in part (a), calculate the leverages (h_{ii}), Cook's distances (D_i), and $DFFITS_i$ for each of the data points. Do any of the points seem to be having an undue influence on the regression?

(c) Remove the two most influential points noted in part (b) and refit the regression. Construct the residual and normal q-q plots and comment on their appearance.

(d) For the regression with the two influential points removed, construct an added variable plot for lngth. Would you recommend including a squared term in this predictor?

(e) Now, using all the data points, and including a squared term in the predictor lngth, perform automatic model selection using the step() function in R and apply forward selection, backward elimination, and stepwise refinement. Which models are selected?

(f) Using the additional code from the surgery example discussed in lectures, find the best models of each different size (as measured by p, the number of parameters) For each of these models, calculate the Mallows' C_p, $PRESS_p$ and R_a^2 values. Which model or models do you prefer now?

(g) Repeat parts (e) and (f) without up to 3 influential points. Do the "good" models change?

(h) Based on your analysis, do you think that lowering the speed limit will have an impact on accident rates?