Regression Modelling

(STAT2008/STAT4038/STAT6038)

Tutorial 3 – Multiple Linear Regression

Question One

The data file **auscars.csv** (available on Wattle) contains data gathered by the NRMA on 62 different makes and models of automobiles selling in Australia in 1991. We have already examined the relationship between fuel efficiency (`L.100k`) and unladen weight of the vehicle (`Weight`) in the previous tutorial.

(a) Fit the simple linear regression of fuel efficiency on unladen weight and create the associated analysis of variance table. What is the fitted regression line? What is the interpretation of the slope parameter? What are the sums of squares for regression and error and the total sum of squares? Save the residuals from this regression.

(b) Fit a multiple regression of fuel efficiency on weight and engine capacity (`Capacity`) and create the associated analysis of variance table. What is the fitted regression equation? Compare the ANOVA table and fitted equation to the ones for the simple regression. What is the interpretation of the "slope" parameter associated with unladen weight? What are the sums of squares for regression and error and the total sum of squares?

(c) What is the predicted fuel efficiency for a car weighing 1500 kilograms and having an engine capacity of 1500 cubic centimetres? What is the standard error of the fit for this predicted value?

(d) Fit a linear regression with `Capacity` as the response variable and unladen weight as the predictor. Do you think that there is a relationship between these two variables? (Use plots as well as numerical output to answer this question.) Save the residuals from this regression.

(e) Fit a simple linear regression using the residuals from part (a) as the response variable and the residuals from part (d) as the predictor? Is there a relationship between these two sets of residuals? What is the slope estimate for the fitted line? Do you recognise this number? Plot the relationship between the two sets of residuals (this is called a partial regression or added variable plot – there will be more on how to interpret these plots later in the course).

Question Two

Suppose that we were examining the relationship between a response variable Y and two predictors x_1 and x_2 and scientific theory suggested that the functional form of the relationship was:

(a) $Y = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$

(b) $Y = \frac{x_1 x_2}{\beta_0+\beta_1 x_1+\beta_2 x_2}$

Rewrite each of these formulae in the form of a linear model on transformed variables.

Question Three (this is Question 2 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 assignment (which is available in the "Assessment" topic on Wattle) or you could download the data as a .csv files (also available in the "Assessment" topic on Wattle) and then use read.csv() to read in the data.

The dataset teengamb concerns a study of teenage gambling in Britain. In this assignment we are going to fit an appropriate multiple linear regression model to examine factors affecting the amount that teenagers will gamble (gambling expenditure measured in UK £ per year), including both teenagers who do and who do not regularly gamble.

(a) Transform gamble by creating a new variable trans.gamble <- log(gamble + 1). Compare histograms of gamble and trans.gamble and comment on which is more likely to be suitable for inclusion in a multiple regression model.
Assume that the researchers who collected the data believe that gambling expenditure differs by sex and is also strongly affected by factors such as education and socio-economic status. This is why they collected the variables verbal and status (as measures of education and socio-economic status respectively) and any multiple regression model will include status, verbal and sex as predictors so we can test these assertions (and control for the effects of these factors).
This leaves income as the only remaining observed variable (covariate). Construct an added variable plot to assess income as a possible addition to a multiple regression model for trans.gamble that already includes sex, verbal and status as predictors. Does this added variable plot suggest a transformation is required for income? The transformation we used in Question 2 of Assignment 1 was log(income). Construct a different added variable plot for log(income). Is this an improvement?

(b) Fit the multiple linear regression model with trans.gamble as the response variable and sex, verbal, status and log(income) as predictors. Construct a plot of the externally studentised residuals against the fitted values, a normal Q-Q plot of the internally studentised residuals and a bar plot of Cook's Distances for each observation. Comment on the model assumptions and on any unusual data points. Calculate appropriate influence statistics for the most unusual data point and comment on these statistics, but do NOT refine the model by removing this observation as a possible outlier.

(c) Produce the ANOVA table and the summary table of estimated coefficients for the multiple linear regression model in part (b). Interpret the overall and sequential F tests and the t-tests and the values of the estimated coefficients of the model. Are the earlier assertions in part (a) about sex, education and socio-economic status supported in the context of this model?

(d) To help the researchers interpret the model, plot gamble against income, with different plotting symbols for the two values of sex. Include your model on this plot by calculating predicted values for trans.gamble, for the full range of income values and for both values of sex, holding verbal and status at their mean values. Suitably back-transform the predictions and include them on the plot separately for both males and females. Also include point-wise 95% confidence intervals (but not 95% prediction intervals) on the plot.