Regression Modelling

(STAT2008/STAT4038/STAT6038)

Tutorial 2 – More Simple 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. Nine different variables were measured for each type of car, and the data set contains the names of these variables.

(a) Download and view the data file (once you have downloaded it; open the csv file in **Excel**). Use the `read.csv()` command to read the data into a data frame called `auscars` in `R`. Type the name of this new data frame `auscars` in `R` and check that the data has been read in correctly. Attach the data frame for use in the rest of this tutorial and at the end of today remember to save the `R` workspace, as we will be using the data again in Tutorial 3. If you don't save it now, you will have to do all this again. Even when you do use the same workspace, you will need to attach the data frame for each new `R` session.

(b) Fit a simple linear regression with the variable `L.100k` (which measures the fuel efficiency of the car in litres needed to travel 100 kilometres) as the response and `Weight` (which measures the unladen weight of the car in kilograms) as the predictor variable. What is the equation of the fitted regression line? What are the standard errors of the parameter estimates? Plot the data and superimpose the regression line, making sure that the limits on the *y*-axis range from 5 up to 20.

(c) Do you think that there is a significant relationship between the weight of a car and its fuel efficiency? If so, what is the nature of that relationship?

(d) Calculate the coefficient of determination, R^2. What is its interpretation?

(e) Calculate a 95% confidence interval for the parameter β_0. Is this a useful interval?

(f) Calculate a 95% confidence interval for the expected fuel efficiency of cars weighing 1800 kilograms. Calculate a corresponding 95% prediction interval for a single such car.

(g) Create a vector containing values spanning the range of `Weight` in the dataset using `seq(min(Weight), max(Weight), 10)` and calculate the predicted values of fuel efficiency and their standard errors for each of the values in this vector. Create two vectors one containing all the upper endpoints of 95% confidence intervals for each of these expected responses and the other containing all the lower endpoints. Draw both of these vectors as lines on the scatterplot; in other words, connect all the upper endpoints together and all the lower endpoints together. Repeat this procedure for 95% prediction intervals. What do you notice about these curves?

Question Two (this is Question 1 of Sample Assignment 1)

The data for this question and the next 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 datasets as .csv files (which are also available on Wattle) and use the same approach as in Question One of this tutorial 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). Of the variables included in this dataset, lcavol (log of the cancer volume) is a measure of the size of the cancer tumour and lpsa (log of the prostate specific antigen measure) is the result of a diagnostic blood test for prostate cancer.

(a) Plot lpsa against lcavol. Is there a significant correlation between lpsa and lcavol? Use R to conduct a suitable hypothesis test and present and interpret the results.

(b) Fit a simple linear regression model with lcavol as the response variable and lpsa as the predictor. Construct a plot of the residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of the leverages for each observation. Comment on the model assumptions and on any unusual data points.

(c) Produce the ANOVA (Analysis of Variance) table for the SLR model in part (b) and interpret the results of the F test. Are these results consistent with the hypothesis test you conducted in part (a)?

(d) What are the estimated coefficients of the SLR model in part (b) and the standard errors associated with these coefficients? Interpret the values of these estimated coefficients and perform t-tests to test whether or not these coefficients differ significantly from zero. What do you conclude as a result of these t-tests?

(e) Plot lcavol against lpsa. Include the fitted SLR model from part (b) as a line on the plot and also show 95% confidence intervals for the mean or expected value of lcavol (do NOT plot the 95% prediction intervals). Do the results of a PSA test appear to be a reliable predictor of the size of the prostate cancer tumour?

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

The dataset `teengamb` concerns a study of teenage gambling in Britain. For this assignment, we are interested in whether a teenager's `income` (measured in UK £ per week) can be used to predict the amount they will `gamble` (gambling expenditure measured in UK £ per year), at least for the teenagers who do regularly gamble.

The dataset teengamb concerns a study of teenage gambling in Britain. For this assignment, we are interested in whether a teenager's income (measured in UK £ per week) can be used to predict the amount they will gamble (gambling expenditure measured in UK £ per year), at least for the teenagers who do regularly gamble.

(a) Plot gamble against income. Describe the correlation shown in the plot. Would you expect a simple linear regression model to be a reasonable model for the relationship shown in the plot?

(b) Fit a simple linear regression model with gamble as the response variable and income as the predictor. Construct a plot of the residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of Cook's Distances for each observation. Comment on the model assumptions and on any unusual data points.

(c) In question 1, a natural log (to the base *e*) transformation had already been applied to both the response and predictor variables and appeared to produce reasonable results.
In this example, there are a number of teenagers who are not regular gamblers (their annual expenditure on gambling is very small or even zero). What is the problem with applying a log transformation in this situation? Exclude any teenager who spends less than £1 per year on gambling and fit another simple linear regression model with log(gamble) as the response variable and log(income) as the predictor. Check the same plots you produced for the earlier model in part (b). Are the same problems still apparent?

(d) Produce the ANOVA table and the table of the estimated coefficients for the revised SLR model in part (c). Interpret the values of the estimated coefficients for this SLR model and the results of the overall F test and the t-tests on the estimated coefficients.

(e) Use the revised SLR model from part (c) to predict the annual expenditure on gambling for three British teenagers, who were not included in the original study, but who have weekly incomes of £1, £5 and £20, respectively. Find 95% prediction intervals for these predictions. Do you think this revised SLR model is a good model for making all three of these predictions?

Question Four (optional extra – there will be no question like this one on the final exam)

In this question, we will use R to generate regression datasets and examine the distributions of the least-squares parameter estimates.

(a) Create a vector x containing the integers from –3 to 6.

(b) Create 1000 regression datasets and store the resulting least-squares parameter estimates using the following R commands:

```
> ests <- matrix(0,1000,2)
> for(i in 1:1000) {
+ y <- 2 - 3*x + rnorm(length(x),0,2)
+ ests[i,] <- lm(y~x)$coef}
```

(c) Plot the histograms of the slopes and intercepts.

(d) Calculate the means and variances of the 1000 parameter estimates using:

```
> apply(ests,2,mean)
> apply(ests,2,var)
```

Compare these values to those that normal theory would predict for them.

(e) Calculate the correlation of the slopes and intercepts. Plot a scatterplot of the slopes against the intercepts. Do these estimators appear to be uncorrelated? What does normal theory predict for this correlation? [Hint: Recall that $Var(b) = \sigma^2(X^TX)^{-1}$.]
