LINEAR MODELS AND GLM: ASSIGNMENT 3

Deadline: Monday March 22

Please write your answers in the form of a report, which should be submitted by email in the form of a PDF file. Be concise. You may include R output when appropriate, but you will be penalised for including boring and irrelevant output (one of the key skills in report-writing is to decide what to include and what to exclude).

You may be asked to present your report in class. This will be informal; you do not need to prepare slides for that.

Any R functions you write should be submitted in separate files (one file for each question).

Your entire submission should be a zip file with a top-level folder whose name identifies yourself.

Exercise 1: Give a justification for the alternative hypothesis in Tukey's one degree of freedom test for non-additivity. Reference: `http://www.jstor.org/stable/3001938`

Exercise 2: Implement an R function to perform Tukey's one degree of freedom test for non-additivity. Your function should accept a 2-dimensional table (matrix) as its argument. How would you use your function if the data is supplied as a data frame? [Hint: `?xtabs`]

Exercise 3: Use your function to perform the test for non-additivity in the `VADeaths` dataset.

Exercise 4: Implement an R function (or a suitable collection of R functions) to obtain an optimum Box-Cox transformation, including a profiled log-likelihood plot and a confidence interval. The profiled log-likelihood should look roughly like Figure 1 of the reference (without the Bayesian curve $p_u(\lambda)$); that is, it should include an indication of the confidence interval. It should be possible to supply the confidence level as an argument to your function.

Reference: `http://www.jstor.org/stable/2984418`. You need only consider the simple power transformation; equation (1).

Implementaton details. You might consider structuring your code as follows: The main function should return a list of class *"boxcox"*, with functions `plot.boxcox` and `print.boxcox` for plotting the profiled likelihood and printing the estimate and confidence bounds respectively.

Exercise 5: Use your function to find an optimal power transformation in the `VADeaths` dataset. Plot the transformed values as before (see Assignment 2). Does the plot "look" more additive? Perform Tukey's test for non-additivity on the transformed data.

Date: 10 March, 2010.

Exercise 6: Write an R function to fit a linear model minimising the sum of absolute errors, using the technique of iteratively reweighted least squares (IRLS). You can use the `lm()` function with weights.

You will need to choose a convergence criterion. What criterion did you choose?

Note that this is not the only solution for this problem (or even the first, or the best):
`http://www.jstor.org/stable/2285306`

Exercise 7: Use the function written in the previous question to fit regression lines to Anscombe's four regression datasets (`?anscombe` in R). Produce a plot similar to what you get from `example(anscombe)`, but with regression lines fitted using your code instead of the usual least squares regression lines.

Exercise 8: See the examples section in `help(esoph)` in R and re-fit the Binomial model there using the probit link function. Obtain an ANOVA table with suitable p-values by supplying the appropriate arguments to the `anova()` call (note that the default output of `anova()` for GLM fits does not include p-values).

Exercise 9: (You need to install the `alr3` package for this exercise.) To estimate the number of snow geese over Hudson Bay, small aircraft flew over the Bay and an "experienced person" estimated the number of geese in any flocks that were found. To test the reliability of this method, two experienced observers estimated the number of geese per flock, and a photograph of the flock was also taken. The number of geese were later counted from the photograph. The results are recorded in the `snowgeese` dataset in the alr3 package.
Fit a suitable model to the data. Which observer seems to give more accurate estimates?

Exercise 10: Consider the `highway` data, also in package alr3, which comes from an analysis relating the automobile accident rate, in accidents per million vehicle miles, to several potential terms. The response here is a rate, not a count. Is a Poisson regression model still appropriate? Fit a suitable model to the data.