MAST90104: Introduction to Statistical Learning

Assignment 3

Late assignments will be penalised two marks for each day overdue.
Extensions should be supported by a medical certificate. Requests for extensions must come on or before the due date for the assignment.

Please submit a scanned or other electronic copy of your work via the Learning Management System in one file - see this link for instructions

The .pdf must have in **one file**:

- handwritten or typed answers to the questions
- handwritten or typed R code used to produce your answers
- graphics required to answer the questions

If you have more than one file submitted, *only the last LMS .pdf file with your name on it will be marked.*

1. Suppose that there is one factor with 4 levels. Starting with the usual less than full rank model, find the matrices D and E and verify that $I_4 + DE$ is rank 4 for C_4 from the `contr.helmert(4)` matrix in R. Find the resulting reparameterisation and interpret it in terms of the mean responses for each of the 4 levels.

2. **You may not use theR `glm` command for this question.** Fit a binomial regression model to the O-rings data from the Challenger disaster, available in "orings.csv", using a *complementary log-log* link.

   Your solution should include the following:

   (a) parameter estimates

   (b) 95% CIs for the parameter estimates

   (c) a likelihood ratio test for the significance of the temperature coefficient

   (d) an estimate of the probability of damage when the temperature equals 29 Fahrenheit together with a 95% CI

   (e) a plot comparing the fitted c-log-log model to the fitted logit model.

3. Suppose the Y comes from an exponential family with pdf or pmf f of the form

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

   (a) Show that $\mathbb{E}Y = b'(\theta)$

   (b) Show that $\operatorname{Var} Y = b''(\theta)a(\phi)$.

4. Steve Sahyun of the University of Winsconsin has a website http://sahyun.net/neutron.php in which he explains neutron activation as a means of producing radioactive metals from standard metals. It includes data on samples of Silver (Ag), Aluminium (Al) and Copper (Cu) which have been subject to neutron activation. For each of the elements, two samples have been tested. The data file `Radioactive.csv` contains radioactive counts at various times after the initial measurement (with the square of time also recorded) as well as the name of the element and whether the count is for the first or second sample.

(a) Use `qplot` (in package `ggplot2`) to plot the log of counts versus time with different colours for the different elements and different shapes for samples one and two. Your answer should include the plot and your command to get it.

(b) Standard theory on radioactive emissions suggests that counts of them should be Poisson distributed with a varying rate as the radioactivity decays. This suggests a Poisson model for the counts with a log link. Comment on this in the light of the plot in (a). Fit the Poisson model with the log the mean having a linear dependence on time and including factors for both Material and Sample. Do diagnostic plots using `plot` for the model. Comment and re-fit, if needed, omitting some observations. Use stepwise AIC selection to see if some variables can be omitted. What does the final residual deviance indicate?

(c) Try re-fitting the model including TimeSquare as an extra variable to Time but still including Material and Sample. Carry out the additional steps in part (b). Comment on the contrast and comparison with (b).

(d) An alternative approach would be to run the same analyses just using the data for one material at a time. Do this. Comment on the comparison with (b) and (c).