M3/4S2 Spring 2019 - Assessed Coursework

Deadline: 16:00 Friday 8th March 2019

Recall that Imperial takes plagiarism very seriously.

For this coursework you are required to download a dataset personal to you. Your dataset is available on Blackboard. The filename is your cid.

The final report must be typed up and should be a properly structured document. Where appropriate, phrase your answer using proper sentences. `R` should be used for all coding and figure-plotting.

A printed copy must be submitted to the UG office. It should not exceed 10 A4 sides, not including code.

You must submit an electronic copy of your report, along with all code used. A good way to do this would be to produce your report as an RMarkdown document. Alternatively, you can supply code in a separate script. Files should be uploaded on Blackboard. Your code must be well-commented, such that is it clear what is being computed at every stage. Your code should be able to run through without errors; **CHECK YOUR CODE BEFORE SENDING.**

When reporting values in your report, round figures to 4 decimal points — using the `round` command. Do not round in your code.

You may assume throughout this coursework that any design matrix used has full rank.

Normal Linear Modelling (12 marks)

1. A clinical collaborator, who is not a statistician, has data from a trial into the effect of a treatment to reduce blood pressure and has enlisted your help with the statistical analysis. The variables available for the analysis are as follows:

   `female` - binary indicator: whether or not the subject is female.
   `dose` - total weight-standardized dose of the drug received (in mg/kg).
   `response` - decrease in systolic blood pressure, from baseline (mmHg).

   They asked for help because, after doing some preliminary analysis, they found a confusing effect. They fit the model `response ~ dose`, and found that each additional unit of the drug has a negative effect: the higher the dose, the smaller the decrease. Carry out an analysis of this data, with the aim of understanding the effects of the drug more precisely. You may use inbuilt R functions as needed. *(Hint: you may wish to fit more than one model.* `z~x*y` *is the syntax for a model with an interaction term, which allows the relationship between* `z` *and* `x` *to be different for different values of* `y`*.)*

   You should produce a clearly written account of your analysis, including

   - Details of exploratory analysis, with suitable plots.
   - Models that you have fit, and comparison of these using suitable tests.
   - Diagnostic plots for models, and discussion of any choices you have made (such as whether to exclude any outliers).

   Your account should be written so that a statistician with access to the data could reproduce your analysis. You should also give a brief plain language summary, accessible to the clinician who collected the data.

2. When fitting the Normal linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the Cook's distance C_i for observation i is defined to be

$$C_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T X^T X (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\hat{\sigma}^2},$$

   where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the estimator of $\widehat{\boldsymbol{\beta}}$ when the ith observation is excluded, and p is the rank of X.

   Use the fact that for an $n \times n$ matrix A, and row vectors u and v for which all terms below are well-defined,

$$(A + u^T v)^{-1} = A^{-1} - \frac{A^{-1}u^T v A^{-1}}{1 + vA^{-1}u^T},$$

to show that

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{(X^T X)^{-1}\boldsymbol{x}_i^T(y_i - \hat{y}_i)}{1 - h_{ii}},$$

where $h_{ii} = \boldsymbol{x}_i(X^T X)^{-1}\boldsymbol{x}_i^T$.

Hence deduce that Cook's distance can be expressed in terms of the standardized residual r_i as

$$C_i = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}.$$

3. Show that in a Normal linear model with an intercept, if the residuals $e_i = y_i - \widehat{y}_i$ satisfy

$$e_i = \alpha + \beta x_i,$$

for $i = 1, \ldots, n$, where $\boldsymbol{x}$ is a covariate of the model, then each residual is identically zero.

Generalized Linear Modelling (18 marks)

The remaining questions concern the results of a psychological experiment into reading accuracy in children. The data recorded is the count of the number of words correctly pronounced in a reading test before three errors were made. The scientist who collected the data is interested in the variables that best predict reading ability. An initial analysis used a Poisson GLM with a log link.

The variables available are as follows:

`x` - score out of 25 on a standardized inventory for attention (higher is better).
`q` - standardized measure of verbal fluency (higher is better).
`count` - count of words correctly pronounced before the third error.

4. Repeat the initial analysis by fitting a Poisson GLM, using `glm(, family="Poisson")`. Comment on goodness of fit, displaying appropriate plots to support your conclusions.

5. An alternative model for count data is the negative binomial distribution with mass function

$$\Pr(Y_i = y; p, r) = \binom{y + r - 1}{y}(1 - p)^r p^y, \qquad y \in \{0, 1, \ldots\}.$$

If r is taken to be a constant parameter, show that the negative binomial distribution is a member of the exponential family.

6. Derive the IWLS algorithm for the negative binomial GLM with the log link function. Your answer should state the form of the adjusted response variable z and the weights w clearly.

7. Write code to implement IWLS for a negative binomial GLM with log link. Be sure to choose a sensible initial estimate, and an efficient stopping rule.

8. Give a plain language summary of your results. This should include

   - Estimates and confidence intervals for all parameters of the model, on a suitable scale.
   - A comparison of the original Poisson model with the negative binomial: suitable diagnostic plots, together with reasoning about which model might be expected to fit better.
   - Note of any assumptions made.

Mastery Material - for M4 students only (5 marks)

9. Confidence intervals in the previous part make use of results that show the maximum likelihood estimator of $\widehat{\boldsymbol{\beta}}$ is approximately normally distributed around $\boldsymbol{\beta}$ with variance given by the Fisher information. Conduct simulations to determine the accuracy of this approximation in the regime given here, i.e. for Poisson and negative binomial responses, with the appropriate number of observations. State your conclusions clearly, accompanied by suitable plots.