Project 3
Computational Numerical Statistics
DM, FCT-UNL

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Deadline: 17/12/19

Generalized linear models

Select a dataset from the literature whose analysis via either a Logistic or Poisson regression model is adequate. The dataset should comprehend at least three covariates. All groups should have distinct data and problems to solve.

1. Begin by describing the data/problem and the question that needs to be answered. Produce a brief descriptive analysis of the data together with adequate graphical display.

2. Identify the exponential family to which your response variable belongs. Write the exponential canonical form and indicate all the relevant information that one can extract from it, all the way till Fisher information.

3. Describe in detail the GLM that is going to be used. Research in the literature about variable selection in the GLM. Identify one or two criteria for variable selection.

4. Begin by fitting your GLM using the `R` routine `glm()` and use it to perform variable selection according to your selected criteria. You are welcome to improvise using other possible `R` functions as long as you provide full detail on what you are doing and why. Provide all summary details from your best fitted model (estimates, standard errors, confidence intervals, etc). Discuss the results of your best fitted model and answer the main question of your problem.

5. Identify all the relevant information (link, inverse link, weight, etc) needed for the implementation of the IRWLS estimation procedure. Briefly describe the IRWLS steps for your particular GLM and implement it to fit the best model from 4.. Validate all your results from 4.. Avoid the use of the `lm()` in your code.

Bayesian inference

1. In a 2006 study published in The New England Journal of Medicine, 78 pairs of patients with Parkinson's disease were randomly assigned to receive treatment (which consisted of deep brain stimulation of a region of the brain affected by the disease) or control (which consisted of taking a prescription drug). The researchers found that in 50 of 78 pairs, the patients who received deep-brain stimulation had improved more than their partner in the control group. The parameter of interest is θ, the probability of doing better on treatment than control.

   (a) Explain why the $Beta(1, 1)$ is an adequate *prior* for this study. Say which values of θ are more plausible under this prior assumption.

   (b) Using the *prior* above, derive the *posterior* distribution of θ. Report the posterior mean, standard deviation, a 95% symmetric credible interval, an HPD interval, and plot both the *prior* and *posterior* distributions against the likelihood.

   (c) What is the posterior probability that $\theta > 0.5$?

   (d) Re-analyse the data using a Jeffreys prior; do your conclusions from (b) or (c) change?

   (e) Suppose 10 new pairs of patients are enrolled in the study. Based on your posterior, what is the probability that 6 or more will do better on treatment than control? And what is the expected number of patients that will do better on treatment than control?

   (f) Does your posterior represent strong belief that the treatment works? In the paper, the authors claim that deep-brain stimulation is "more effective than medical management". Based on your posterior, do you agree?