MAST30027: Modern Applied Statistics

Assignment 1, 2021.

Due: 5pm Monday August 16th

- This assignment is worth 12% of your total mark.
- To get full marks, show your working including 1) R commands and outputs you use, 2) mathematics derivation, and 3) rigorous explanation why you reach conclusions or answers. If you just provide final answers, you will get zero mark.
- The assignment you hand in must be typed (except for math formulas), and be submitted using LMS as a single PDF document only (no other formats allowed). For math formulas, you can take a picture of them. Your answers must be clearly numbered and in the same order as the assignment questions.
- The LMS will not accept late submissions. It is your responsibility to ensure that your assignments are submitted correctly and on time, and problems with online submissions are not a valid excuse for submitting a late or incorrect version of an assignment.
- We will mark a selected set of problems. We will select problems worth $\geq$ 50% of the full marks listed ($\geq$ 17 out of 34 for this assignment). For example, if we select 1-(b), (c), (e), 2-(a), and 3-(b) for marking, they will contribute $35(=\frac{7}{20}\times 100)$, $15(=\frac{3}{20}\times 100)$, $10(=\frac{2}{20}\times 100)$, $15(=\frac{3}{20}\times 100)$, $25(=\frac{5}{20}\times 100)$ to the full marks of 100 for the assignment 1.
- Also, please read the "Assessments" section in "Subject Overview" page of the LMS.

---

1. Fit a binomial regression model to the O-rings data from the Challenger disaster, using a *probit* link. You must use R (but without using the glm function); I want you to work from first principles.

   (a) (3 marks) Compute MLEs (maximum likelihood estimates) of the parameters in the model.

   (b) (7 marks) Compute 95% CIs for the estimates of the parameters. You should show how you derived the Fisher information.

   (c) (3 marks) Perform a likelihood ratio test for the significance of the temperature coefficient.

   (d) (3 marks) Compute an estimate of the probability of damage when the temperature equals 31 Fahrenheit (your estimate should come with a 95% CI, as all good estimates do).

   (e) (2 marks) Make a plot comparing the fitted probit model to the fitted logit model. To obtain the fitted logit model, you are allowed to use the glm function.

2. The data frame 'pima_subset' contains a subset of the pima data set. For details of the pima data set, please see the practical problem 2 for the week 2. You can obtain 'pima_subset' using the commands:

```
> library(faraway)
> missing <- with(pima, missing <- glucose==0 | diastolic==0 | triceps==0 | bmi == 0)
> pima_subset = pima[!missing, c(6,9)]
> str(pima_subset)
'data.frame': 532 obs. of  2 variables:
 $ bmi : num  33.6 26.6 28.1 43.1 31 30.5 30.1 25.8 45.8 43.3 ...
 $ test: int  1 0 0 1 1 1 1 1 1 0 ...
```

Using the 'pima_subset' data set, we will fit a binomial regression with a logit link with `test` as a response and `bmi` as a predictor to see the relationship between the odds of a patient showing signs of diabetes and his/her bmi. The odds o and probability p are related by

$$o = \frac{p}{1-p} \quad p = \frac{o}{1+o}.$$

(a) (3 marks) Please estimate the amount of increase in the log(odds) when the bmi increases by 7.

(b) (3 marks) Compute a 95% CI for the estimate.

You are allowed to use the `glm` function.

3. The gamma distribution with shape $\nu > 0$ and rate $\lambda > 0$ has p.d.f.

$$f(x;\nu,\lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}$$

for $x > 0$.

(a) (5 marks) Show that the gamma distribution is an exponential family.

(b) (5 marks) Obtain the canonical link and the variance function.