# HW Question Week 8

STA2101F 2021

## Due November 18 2021 11.59 pm

**Homework to be submitted through Quercus**

*Part 1. Project*

(a) Create a new `R project` for your final project. Create a new `R markdown` file to start recording the steps in your analysis. Write some code that reads your data into `R` from the original website where you obtained it, or from your own website that you create. (This is so I will be able to run your `.Rmd` file without actually storing your data on my computer.)
(b) Load your data and do some quick quality checks – are there any missing values? If so, how many? How will you handle them in the analysis?
(c) Construct some preliminary plots of the data, for example histograms, boxplots, and/or scatterplots, and comment on any anomalies.

*Part 2. Question for this week*

The `cloth` data in the library `SMPracticals` gives the number of faults, $y$, in each of $n = 32$ rolls of textile fabric of different lengths, $x$. Assume that the number of faults in roll $i$, say $y_i$, follows a Poisson distribution with rate $\lambda x_i$, where $x_i$ is known, and is the length of roll $i$.

(a) Show that the maximum likelihood estimate of $\lambda$ is given by $\hat{\lambda} = \bar{y}/\bar{x}$ and find an expression for the variance of $\hat{\lambda}$.

(b) Use `glm` to fit this Poisson model to the data, and give an approximate 95% confidence interval for $\lambda$.

(c) Carry out a goodness-of-fit test of the Poisson model using either the residual deviance, or Pearson's $\chi^2$, or both, and state your conclusions.

(d) Show that if it is assumed that $\lambda$ follows a Gamma distribution with shape and *rate* parameters $\alpha$ and $\beta$ respectively (i.e. with density $\{1/\Gamma(\alpha)\}\lambda^{\alpha-1}\beta^{\alpha}e^{-\lambda\beta}$), that the distribution of $y_i$ given $x_i$ is negative binomial, with

$$E(y_i \mid x_i) = \alpha\frac{1-\pi_i}{\pi_i} = \frac{\alpha}{\beta}x_i, \quad \mathrm{Var}(y_i) = \frac{\alpha}{\beta}\frac{x_i(\beta + x_i)}{\beta}.$$

(e) Fit this model using `glm.nb` and compare the confidence interval for $\lambda$ obtained from this model to that in (b). Assess the fit of the model using residual plots and summarize your conclusions.

(f) *PhD/Bonus (SM Exercise 10.5.1)*

Consider a set of $2n$ Poisson random variables $y_{11}\ldots,y_{1n}$ and $y_{21},\ldots,y_{2n}$, in a $2\times n$ contingency table:

$$\begin{array}{ccccc} y_{11} & \cdots & y_{1i} & \cdots & y_{1n} \\ y_{21} & \cdots & y_{2i} & \cdots & y_{2n} \end{array}$$

where $\eta_{1i}=\log\{E(y_{1i})\}=x_{1i}^T\beta$ and $\eta_{2i}=\log\{E(y_{2i})\}=x_{2i}^T\beta$, and $x_{ij}$ are fixed. Show that the conditional density of $y_{1i}$, given $y_{1i}+y_{2i}=m_i$ is distributed as a $Binom(m_i,p_i)$, where

$$\log\{\frac{p_i}{1-p_i}\}=(x_{1i}-x_{2i})^T\beta.$$

As noted in SM (not needed to prove), this implies that a contingency table in which a single, binary classification is regarded as the response can be analyzed using logistic regression. In this table, $y_{1i}$ represents the count associated with $x_{1i}$ for class “1”, say, and $y_{2i}$ represents the count associated with $x_{2i}$ for class “2”.