APPLIED STATISTICS, MATH10096

Workshop 3

1. Consider a survey in which a random sample of vegetarians and non-vegetarians was taken, and their weight y_i, height x_i, and gender were recorded. A model for the resulting data might be

$$\mathbb{E}(y_i) = \alpha + \nu_k + \gamma_j + \beta x_i \text{ if } y_i \text{ is for diet } k \text{ and gender } j, \quad j \in \{1, 2\}, \quad k \in \{1, 2\},$$

where $k = 1$ for vegetarian, $k = 2$ for non-vegetarian, $j = 1$ for female, $j = 2$ for male, $y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, and all Greek letters refer to model parameters.

   (a) Write out the model matrix (design matrix) and parameter vector for this model, ignoring identifiability problems. For the sake of concreteness assume a random sample of $n = 8$ subjects, the first 4 vegetarian, next 4 not, with the first two of each set of 4 being female and the next 2 male.

   (b) Re-write the model matrix and parameter vector so that the model is identifiable.

2. The "High School and Beyond" data is found in `hsb` in the `R` package `faraway` (which you need to install on your computer). Data was collected as a subset of the "High School and Beyond" study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. `hsb` is a data frame with 200 observations and 11 variables. To know more about the variables, please type `help(hsb)`. One purpose of the study was to determine which factors are related to the choice of the type of program, academic, vocational or general, that the students pursue in high school. Because so far we have not learned the necessary tools to fit a model that permits investigating this, we will follow Faraway (2014, Chapter 16) and change focus a little bit.

   (a) Model the math score in terms of the following five factors: gender, race, socioeconomic class, school type, and choice of high school program. Include all second-order interactions but no higher order interactions. Implement your model in `R` and check its underlying assumptions. **Hint**: A model including all second order interactions between predictor variables, say, `u`, `v`, `w`, and `z`, can simply be fitted as

```
lm(y ~ (u + v + w + z)^2)
```

   (b) Perform backwards model selection (you can use, for instance, the `step` function) and write down your final model using the identifiability constraints used in `R` by default. Do not forget to also check whether your final model does not violate any of the underlying assumptions of the linear model.

   (c) Briefly report on your findings based on the model you have arrived at in (b).

3. Are either, both or neither of the following two models GLMs? Explain.

$$\log(y_i) = \beta_0 + \beta_1 x_i, \quad y_i \sim \text{gamma},$$

$$\log(y_i) = \beta_0 + \beta_1 x_i, \quad \log(y_i) \sim \text{gamma}.$$

4. The following table classifies a random sample of women and men according to their belief in the afterlife:

	Believer	Non-Believer
Female	435	147
Male	375	134

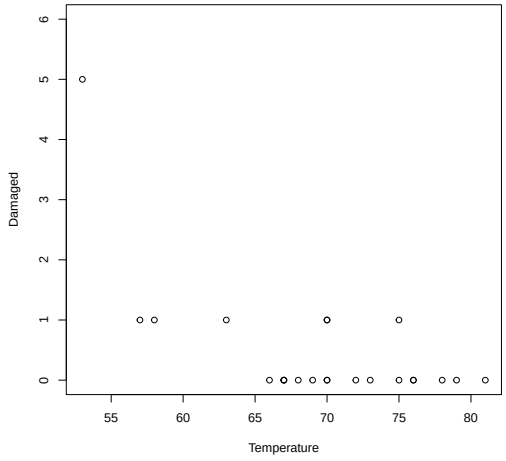
The data come from the US General Social Survey (1991), and the 'non-believer' category includes 'undecideds'. Are there differences between males and females in the holding of this belief? A null model for the data would be that belief does not depend on gender, in which case, if y_i is an observation of the counts in one of the cells of the table, then we could model the expected number of counts as

$$\mu_i \equiv \mathbb{E}(Y_i) = n\gamma_k\alpha_j \text{ if } y_i \text{ is data for gender } k, \text{ and faith } j.$$

Here n is the total number of people sampled, γ_1 and γ_2 are the proportion female and male in the population sampled, and α_1 and α_2 are the proportion of believers and non-believers in the population sampled.

For practical modelling purposes, it turns out that an appropriate response distribution to use here is the Poisson distribution. For the moment, show that this model can be turned into a GLM using the log link. Write out the full model matrix for the model, and then a version where the parameters are identifiable.

5. In January 1986 the space shuttle Challenger exploded shortly after take-off. Subsequent investigation eventually focused on the possibility of o-rings in the fuel tanks having failed as a result of the unusually low launch temperature (31 degrees F), hence allowing a fuel leak. It turned out that data on o-ring failure in previous launches was available for various launch temperatures. Here is a plot of the number of o-rings out of a total of 6 that failed in launches at different temperatures. The first few rows of the dataset are also shown.




```
##   temperature damaged undamaged
## 1          53       5         1
## 2          57       1         5
```

```
## 3          58        1         5
## 4          63        1         5
## 5          66        0         6
## 6          67        0         6
```

Write down a suitable GLM for these data. Don't forget to specify link, response distribution and the form of the model matrix.

6. Show that the exponential distribution with probability density function given by

$$f(y;\mu) = \frac{1}{\mu} e^{-\frac{1}{\mu}y}, \quad y > 0, \quad \mu > 0,$$

is in the exponential family of distributions. Further show, using the properties of the exponential family of distributions, that $\mathbb{E}[Y] = \mu$ and var$(Y) = \mu^2$.

7. A Bernoulli random variable, Y, takes the value 1 with probability p and 0 with probability $1 - p$, so that its probability mass function is

$$f(y;p) = p^y (1-p)^{1-y}, \quad y \in \{0,1\}, \quad p \in [0,1].$$

   (a) Find $\mu \equiv E(Y)$.

   (b) Show that the Bernoulli distribution can be written in general exponential family form.

8. Show that the Inverse Gaussian distribution with density

$$f(y;\mu,\gamma) = \sqrt{\frac{\gamma}{2\pi y^3}} \exp\left\{-\frac{\gamma(y-\mu)^2}{2\mu^2 y}\right\}, \quad y > 0, \mu > 0, \gamma > 0,$$

where μ is the mean and γ is a scaling parameter, can be written in general exponential family form.