Exercises 6 solutions

1. The design matrix for this model is given in §2.5 of the lecture notes as

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} .$$

It follows that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_I \end{pmatrix} \quad \text{and} \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \sum_{j=1}^{n_2} Y_{2j} \\ \vdots \\ \sum_{j=1}^{n_I} Y_{Ij} \end{pmatrix} ,$$

so that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y} = \begin{pmatrix} n_1^{-1} & 0 & \dots & 0 \\ 0 & n_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_I^{-1} \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \sum_{j=1}^{n_2} Y_{2j} \\ \vdots \\ \sum_{j=1}^{n_I} Y_{Ij} \end{pmatrix} = \begin{pmatrix} n_1^{-1}\sum_{j=1}^{n_1} Y_{1j} \\ n_2^{-1}\sum_{j=1}^{n_2} Y_{2j} \\ \vdots \\ n_I^{-1}\sum_{j=1}^{n_I} Y_{Ij} \end{pmatrix} ,$$

which is the vector of group means as required.

2. (a) Let:

Y_i denote the alkalinity for the ith case in the dataset
x_{1i} denote the value of `Qtemp` for the ith case in the dataset
x_{2i} denote the value of `Mtemp` for the ith case in the dataset
x_{3i} denote the value of `Mrain` for the ith case in the dataset

Then the fitted model is

$$Y_i = -12.372 + 4.158x_{i1} - 1.979x_{i2} + 0.213x_{i3} - 1.798x_{i2}x_{i3} + \varepsilon_i ,$$

where the (ε_i) are independent $N(0, 3.525^2)$ random variables.

(b) The interaction suggests that the effect of `Mtemp` is dependent on the value of `Mrain` (or vice versa). The negative coefficient suggests that the relationship between monthly temperature and alkalinity becomes more negative / less positive after a wet month than a dry month (equivalently that the relationship between monthly rainfall and alkalinity becomes more negative / less positive after a warm month than a cool month).

[In physical terms, the mechanism behind this is that rainfall affects the alkalinity by eroding and dissolving minerals from the soil and rocks in the surrounding region, and washing them into the lake: the temperature affects the speed of the chemical reactions involved in these weathering processes].

(c) The high p-value for `Mrain` suggests that given the other terms in the model, there is little evidence to reject the null hypothesis $H_0 : \beta_3 = 0$, where β_3 is the underlying coefficient of `Mrain`. The coefficient of `Mtemp` is borderline significant, with a p-value of 0.06. On the basis of this, we may be tempted to conclude that there is little evidence of a direct association between alkalinity and either `Mtemp` or `Mrain` taken individually, although the highly significant interaction suggests that there is strong evidence of an association when the two quantities are considered together.

(d) If temperatures were measured in degrees Fahrenheit rather than Celsius, then in place of x_{i1} we would have recorded $32+9x_{i1}/5 = \tilde{x}_{i1}$, say; and $32+9x_{i2}/5 = \tilde{x}_{i2}$, similarly. To figure out what would be the effect on the model, rearrange to obtain $x_{i1} = 5\,(\tilde{x}_{i1} - 32)/\,9$ and $x_{i2} = 5\,(\tilde{x}_{i2} - 32)/\,9$, and substitute back into the model equation from part (a):

$$\begin{aligned} Y_i &= -12.372 + [5 \times 4.158\,(\tilde{x}_{i1} - 32)\,/9] - [5 \times 1.979\,(\tilde{x}_{i2} - 32)\,/9] \\ &\quad +0.213x_{i3} - [5 \times 1.798\,(\tilde{x}_{i2} - 32)\,x_{i3}/9] + \varepsilon_i \\ &= \left[-12.372 - \frac{5 \times 4.158 \times 32}{9} + \frac{5 \times 1.979 \times 32}{9}\right] + \frac{5 \times 4.158}{9}\tilde{x}_{i1} \\ &\quad -\frac{5 \times 1.979}{9}\tilde{x}_{i2} + \left[0.213 + \frac{5 \times 1.798 \times 32}{9}\right] x_{i3} - \frac{5 \times 1.798}{9}\tilde{x}_{i2}x_{i3} + \varepsilon_i \\ &= -51.11 + 2.31\tilde{x}_{i1} - 1.099\tilde{x}_{i2} + 32.18x_{i3} - 0.999\tilde{x}_{i2}x_{i3} + \varepsilon_i\ . \end{aligned}$$

The structure of this model is exactly the same as before, but the coefficients have changed: the coefficients of `Qtemp`, `Mtemp` and the interaction have merely been scaled by a factor of 5/9 (as expected, since a change of one degree Fahrenheit has the same effect as a change of 5/9 degrees Celsius), but the intercept and the coefficient of `Mrain` have changed dramatically. The significance of `Qtemp`, `Mtemp` and the interaction term would not be affected by this change: in each case, the null hypothesis being tested is the same as before. However, the significance of the `Mrain` term (i.e. the coefficient of x_{i3}) would be affected. This is because with a change in measurement units, the coefficient of `Mrain` represents the effect upon alkalinity of a unit increase in monthly rainfall *when* $\tilde{x}_{i2} = 0$ *i.e. at a temperature of zero degrees Fahrenheit*: in the original model, the coefficient of `Mrain` represents the effect of a unit increase in monthly rainfall at a temperature of zero degrees *Celsius*. In the presence of interaction, by definition these two effects are different and thus the hypotheses being tested by the t-statistics in the regression output are different.

The purpose of this example is to illustrate why, in general, it is important to retain all relevant main effects in a model containing interactions (this is sometimes called the *principle of marginality*): an individual main effect (here, `Mrain`) can appear insignificant in a model, but a change in the units of measurement of a *different* covariate (here `Mtemp`) can lead to a very different picture. The reason, of course, is that the interpretation of each main effect depends on the units of measurement of each covariate with which it interacts (in particular, on where the zero is defined, which is often arbitrary). Similar considerations apply to models involving higher-order interaction terms: in general, if a model contains a high-order interaction then all of the relevant lower-order interactions should be retained. For this reason, you will find that R commands such as `step()` (for carrying out backward stepwise regression) will not consider removing main effects from a model containing the corresponding interaction terms.

3. Verifying each distribution belongs to exponential family.

   (a) Pmf of Bin(n, π) is $f(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$ for $y = 0, 1, \ldots, n$.

   Easier to take logs: hence obtain

   $\log f(y) = y \log(\frac{\pi}{1-\pi}) + n \log(1-\pi) + \log \binom{n}{y}$

   Hence, by comparing with the equation in the notes,

   $\theta = \log(\frac{\pi}{1-\pi})$ so $\pi = \frac{1}{1+\mathrm{e}^{-\theta}}$

   $b(\theta) = -n \log(1-\pi) = n \log(1+\mathrm{e}^{\theta})$ and $c(y, \phi) = \log \binom{n}{y}$.

   (b) Pdf of $\Gamma(\lambda, \alpha)$ is $f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \mathrm{e}^{-\lambda y}$ for $y > 0$.

   In order to take into account a dispersion parameter, write

   $\log f(y) = (-\frac{\lambda}{\alpha} y + \log \lambda)\alpha + (\alpha - 1) \log y - \log \Gamma(\alpha)$.

   Put $\theta = -\frac{\lambda}{\alpha}$, $\phi = \frac{1}{\alpha}$, $b(\theta) = -\log(-\theta)$

   and $c(y, \phi) = \frac{1}{\phi} \log \frac{1}{\phi} + (\frac{1}{\phi} - 1) \log y - \log \Gamma(\frac{1}{\phi})$.

   Verify mean and variance from

   $\mathrm{E}(Y) = b'(\theta)$ and $\mathrm{var}(Y) = b''(\theta) a(\phi)$

   where in these cases $a(\phi) = \phi$.

   Should obtain, after doing the necessary differentiation with respect to θ and then expessing in terms of the original parameters:

   for (a), $\mathrm{E}(Y) = n\pi$ and $\mathrm{var}(Y) = n\pi(1-\pi)$

   for (b), $\mathrm{E}(Y) = \frac{\alpha}{\lambda}$ (this is μ below) and $\mathrm{var}(Y) = \frac{\alpha}{\lambda^2}$.

   For (b), canonical link $g(\mu) = -1/\mu$ and variance function $V(\mu) = \mu^2$.

4. In all four cases, the distribution of the response belongs to the exponential family, but in case (d) the predictor is non-linear. More specifically:

   (a) Yes, have log link $\log \mu_i = -\beta x_i$. (Note that as $0 < \mu_i < 1$ we have that $\log \mu_i < 0$ and hence β is restricted to be positive if x_i is positive.)

   (b) Yes, have log link $\log \mu_i = \beta_0 + \beta_1 x_i$.

   (c) Yes, have log link $\log \mu_i = \log n_i + \beta x_i$

   (note that the coefficient of $\log n_i$ is known; this term is called an **offset**).

   (d) No, no monotonic function of μ_i that does not depend on an unknown parameter will give a linear predictor. However, if α were known then we have a glm, with predictor linear in the β's.