1. (40 points) Consider the Galapagos islands data with different geological variables measured from 30 islands. For example, the row for the island Baltra is as follows:

<table>
<thead>
<tr>
<th>#</th>
<th>Species</th>
<th>Area</th>
<th>Elevation</th>
<th>Nearest</th>
<th>Scruz</th>
<th>Adjacent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baltra</td>
<td>58</td>
<td>25.09</td>
<td>346</td>
<td>0.6</td>
<td>0.6</td>
<td>1.84</td>
</tr>
</tbody>
</table>

Consider the relationship between the number of species (counts) and these geological variables. The Poisson regression with the log link yields the following output.

Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Deviance Residuals:
Min 1Q Median 3Q Max
-8.2752 -4.4966 -0.9443 1.9168 10.1849

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.155e+00 5.175e-02 60.963 < 2e-16 ***
Area -5.799e-04 2.627e-05 -22.074 < 2e-16 ***
Elevation 3.541e-03 8.741e-05 40.507 < 2e-16 ***
Nearest 8.826e-03 1.821e-03 4.846 1.26e-06 ***
Scruz -5.709e-03 6.256e-04 -9.126 < 2e-16 ***
Adjacent -6.630e-04 2.933e-05 -22.608 < 2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3510.73 on 29 degrees of freedom
Residual deviance: 716.85 on 24 degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5

Based on the output, answer the following questions.

(a) What is the interpretation of the slope for “Area”?
(b) What is the fitted value $\hat{\mu}_{Baltra}$ for the island Baltra?
(c) What are the Pearson residual and deviance residual for the island Baltra?
(d) What is the sum of all response residuals? Why?
2. (28 points) A random vector $Y_{J \times 1} = (Y_1, \ldots, Y_J)$ is said to follow a multinomial distribution, denoted by $\text{Multinomial}(n, \pi)$, where $\pi = (\pi_1, \ldots, \pi_J)$ with $\sum_{j=1}^J \pi_j = 1$, if

$$f(y) = \frac{n!}{\prod_{j=1}^J y_j!} \prod_{j=1}^J \pi_j^{y_j}.$$ 

Show the following properties of $\text{Multinomial}(n, \pi)$.

(a) $\text{Multinomial}(n, \pi)$ is a $(J-1)$-dimensional exponential family.

(b) $\forall j \neq k$, $\text{Cov}(Y_j, Y_k) = -n\pi_j\pi_k$.

(c) $\forall j$, $(Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_J) | Y_j = y_j \sim \text{Multinomial}(n - y_j, (\frac{\pi_1}{1-\pi_j}, \ldots, \frac{\pi_J}{1-\pi_j}))$.

(d) Suppose $W_1, \ldots, W_J$ are independent with $W_j \sim \text{Poisson}(\lambda_j)$, where $\lambda_j > 0$ and $j = 1, \ldots, J$. Then

$$(W_1, \ldots, W_J) | \sum_{j=1}^J W_j \sim \text{Multinomial}(\sum_{j=1}^J W_j, (\frac{\lambda_1}{\sum_{j=1}^J \lambda_j}, \ldots, \frac{\lambda_J}{\sum_{j=1}^J \lambda_j})).$$

3. (32 points) Consider the linear mixed model $y = X\beta + Z\alpha + \epsilon$ where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ with $\text{rank}(X) = p < n$, $Z \in \mathbb{R}^{n \times q}$, $\beta \in \mathbb{R}^p$, $\alpha \sim N_q(0, \sigma^2_\alpha I)$ is a random vector of i.i.d. coefficients, and $\epsilon \sim N_n(0, \sigma^2_\epsilon I)$ is a vector of i.i.d. random errors and is independent of $\alpha$.

Consider the REML with $K_{(n-p)\times n}$ matrix s.t. $\text{rank}(K) = n - p$ and $KK = 0_{(n-p)\times p}$. We mentioned in class that the choice of $K$ is not unique, and different choices of $K$ lead to the same estimate. Let’s work out why here.

(a) Show that for any positive definite matrix $V$,

$$K^T(KVK^T)^{-1}K = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}.$$ 

(Hint: Note that $KK = KV^{1/2}V^{-1/2}X = 0_{(n-p)\times p}$.)

(b) Write the likelihood $\mathcal{L}(\beta, \sigma^2_\epsilon, \sigma^2_\alpha | Ky)$ in terms of $X$, $Z$, and $y$ (but without $K$!).

(Hint: You may assume (a) here. Also use Sylvester’s determinant theorem: For matrices $A_{a \times t}$ and $B_{t \times s}$, $\text{det}(I_a + AB) = \text{det}(I_t + BA)$)