COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

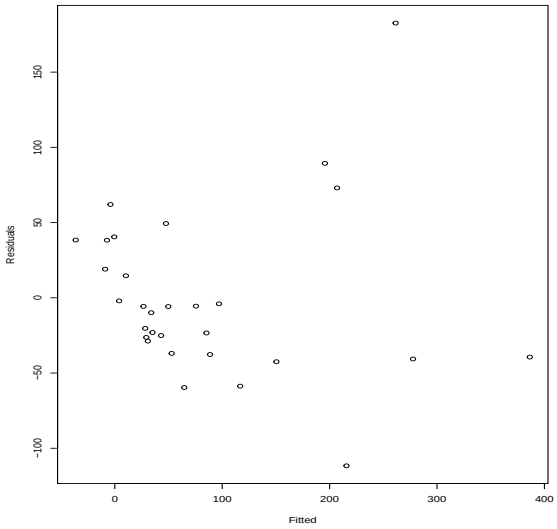
FRIDAY AUGUST 13, 2021 9:00 A.M. – 1:00 P.M.

STOR 665 Questions (100 points in total)

1. (50 points) Johnson and Raven (1973) studied tortoise species in Galapapos Islands. The dataset includes a count of the number of species of tortoise found on each island. In addition, there are five geographic variables for each island.

   Some R analysis results are included below.

```
> dim(gala2)
[1] 30  6
> gala2[1:5,]
          Species  Area Elevation Nearest Scruz Adjacent
Baltra         58 25.09       346     0.6   0.6     1.84
Bartolome      31  1.24       109     0.6  26.3   572.33
Caldwell        3  0.21       114     2.8  58.7     0.78
Champion       25  0.10        46     1.9  47.4     0.18
Coamano         2  0.05        77     1.9   1.9   903.82
> mod1<-lm(Species ~., gala2)
> plot(predict(mod1), residuals(mod1), xlab="Fitted", ylab="Residuals")
```



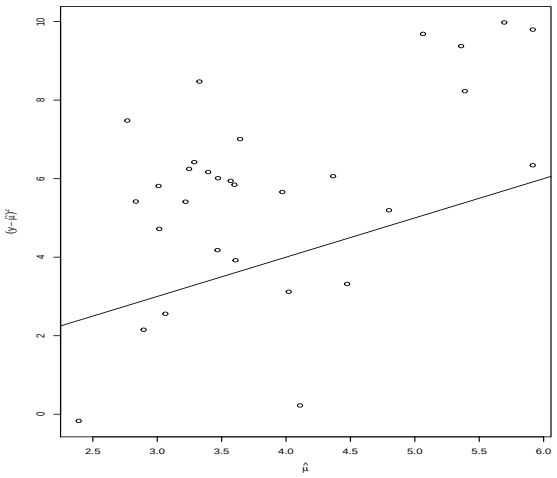
   (1a) (10 points) Write down the model fit in the above R analysis with clear notations. State the assumptions used for the model. Comment on the residual plot above.

```
> modp<-glm(Species ~ ., family=poisson, gala2)
> summary(modp)
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 ??  degrees of freedom
Residual deviance:  716.85  on ??  degrees of freedom
```

(1b) (18 points) Write down the model fit above with clear assumptions and notations. State the corresponding optimization problem and explain potential algorithms for solving it. Provide the missing numbers (marked as ??) in the output above.

(1c) (10 points) For the model in (1b), derive the relationship between the variance function and the mean of the response using standard GLM notations.

```
plot(log(fitted(modp)), log((gala2$Species-fitted(modp))^2),
       xlab=expression(hat(mu)),  ylab=expression((y-hat(mu))^2))
abline(0,1)
```



(1d) (12 points) Comment on the plot above and connect with (1c). What can you conclude? Explain what can be done to improve the analysis.

2. (50 points) Consider the following model,

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}; \quad i = 1, \ldots, a, j = 1, \ldots, n_i,$$

where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

(2a) (10 points) Assume μ and α_i's are fixed parameters. Derive $\hat{\mu}$ and $\hat{\alpha}_i$, and compute the variance of $\hat{\mu}$ and $\hat{\alpha}_i$ if the constraint $\sum_{i=1}^{I} n_i \alpha_i = 0$ is used. Explain why the constraint is needed.

(2b) (13 points) Assume μ is fixed and $\alpha_i \sim N(0, \sigma_a^2)$. In addition, α_i and ϵ_{ij} are independent for different j. Express the model in the matrix form and derive the variance matrix for the response vector. In addition, explain situations that model assumptions in (2a) and (2b) are suitable.

(2c) (14 points) Under the assumptions in (2b), derive the ANOVA estimator for σ_a^2. Explain the advantages and drawbacks of this estimator. Discuss alternative estimation methods as well and briefly compare them without the need of writing out the explicit form for alternative estimators.

(2d) (13 points) Under the assumptions in (2b), suppose $n_i = n$ for $\forall i$. Derive the EM algorithm for obtaining the maximum likelihood estimators of μ, σ_a^2, and σ_ϵ^2.