COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 20, 2010 9:00 A.M. – 1:00 P.M.

STOR 665 Questions

1. (40 points) The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Aylmer Fisher (1936) as an example of discriminant analysis. The dataset consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample, they are the length and the width of sepal and petal, in centimeters. Based on the combination of the four features, the goal is to distinguish the species from each other. Some R analysis results are included below.

```
> attach(iris)
> library(nnet)
> options(contrasts=c("contr.treatment", "contr.poly"))
> iris.mult <- multinom(Species~Sepal.Length+Sepal.Width
                        +Petal.Length+Petal.Width, data=iris)
> summary(iris.mult)
Call:
multinom(formula = Species ~ Sepal.Length + Sepal.Width + Petal.Length +
    Petal.Width, data = iris)
Coefficients:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    18.69037    -5.458424    -8.70740     14.24477   -3.097684
virginica    -23.83628    -7.923634   -15.37077     23.65978   15.135301
Std. Errors:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    34.97116     89.89215    157.0415     60.19170    45.48852
virginica     35.76649     89.91153    157.1196     60.46753    45.93406
Residual Deviance: 11.89973
AIC: 31.89973
```

(2a) (15 points) Explain the model fitted in the above R analysis. Derive the corresponding maximum likelihood function with clear notations.

(2b) (15 points) Explain briefly the meaning of the two estimated coefficients for Sepal.Length. Suppose we have the measurements for a particular iris flower, (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)=(4.9, 2.5, 4.5, 1.7). Calculate the estimated probabilities of this flower belonging to each of the three species.

(2c) (10 points) Which one of the three species was used as the reference category in the above analysis? What is the effect on the results if a different choice of the reference category is used?

2. (40 points) Plant leaves have tiny holes, called "stomata", through which they take up air, but also lose water. The problem for a plant is that if its stomata are too

small, it will not be able to get enough CO_2, and if they are too large, it will lose too much water. One hypothesis is that stomatal size will depend on the concentration of CO_2. Consider an experiment in which tree seedlings are grown under two levels of CO_2 concentration, with 3 trees assigned to each treatment, and after six months, stomatal size is measured at each of 4 random locations on each plant. The dataset is listed below

```
> stomata <- read.table("../datasets/stomata.txt", header=T, sep=" ")
> stomata
        area CO2 tree
1  1.6055739   1    1
2  1.6300711   1    1
3  1.5391189   1    1
4  1.7187315   1    1
5  1.3896163   1    2
6  1.5858805   1    2
7  1.4697276   1    2
8  1.9493473   1    2
9  1.5397020   1    3
10 1.2436558   1    3
11 0.8752505   1    3
12 0.9932352   1    3
13 3.1149370   2    4
14 2.7402102   2    4
15 2.4825228   2    4
16 2.8192831   2    4
17 2.8924475   2    5
18 2.8622759   2    5
19 2.8410755   2    5
20 3.0183753   2    5
21 2.6576575   2    6
22 2.0839150   2    6
23 2.2310707   2    6
24 2.3464027   2    6
```

(3a) (9 points) For the following R output, write out the model m0 and explain whether it is reasonable for this problem.

```
> m0= lm(area~CO2 + tree, stomata)
> m0
Call:
lm(formula = area ~ CO2 + tree, data = stomata)
Coefficients:
(Intercept)          CO2         tree
    0.01916      1.90244     -0.22997
> anova(m0)
```

```
Analysis of Variance Table
Response: area
          Df Sum Sq Mean Sq F value    Pr(>F)
CO2        1 8.8213  8.8213  143.75 7.397e-11 ***
tree       1 0.8462  0.8462   13.79  0.001286 **
Residuals 21 1.2886  0.0614
```

(3b) (9 points) A different model m1 was used with the output given below. Write out the model m1 with clear notations. Explain why the coefficient for tree6 is not available.

```
> stomata$CO2<-as.factor(stomata$CO2)
> stomata$tree<-as.factor(stomata$tree)
> m1 = lm(area~CO2 + tree, stomata)
> m1
Call:
lm(formula = area ~ CO2 + tree, data = stomata)
Coefficients:
(Intercept)         CO22        tree2        tree3        tree4        tree5
    1.62337      0.70639     -0.02473     -0.46041      0.45948      0.57378
      tree6
         NA
> anova(m1)
Analysis of Variance Table
Response: area
          Df Sum Sq Mean Sq  F value    Pr(>F)
CO2        1 8.8213  8.8213 184.5452 6.686e-11 ***
tree       4 1.2744  0.3186   6.6654  0.001788 **
Residuals 18 0.8604  0.0478
```

(3c) (8 points) Write out the model m2 in the following output and explain the relationship between models m1 and m2.

```
> m2 = lm(area~tree, stomata)
> anova(m2, m1)
Analysis of Variance Table
Model 1: area ~ tree
Model 2: area ~ CO2 + tree
  Res.Df    RSS Df   Sum of Sq F Pr(>F)
1     18 0.8604
2     18 0.8604  0 -2.2204e-16
```

(3d) (14 points) The final model used, m3, is listed below. Write out the model clearly. Express the model in the matrix form and give the covariance matrix of the response vector. Explain whether m3 is sensible compared with the other three models.

```
> library(nlme)
```

```
> m3=lme(area ~ CO2, stomata, ~1|tree)
> summary(m3)
Linear mixed-effects model fit by REML
 Data: stomata
Random effects:
 Formula: ~1 | tree
        (Intercept) Residual
StdDev:   0.2601957 0.218632
Fixed effects: area ~ CO2
               Value Std.Error DF  t-value p-value
(Intercept) 1.461659 0.1629435 18 8.970341  0.0000
CO22        1.212522 0.2304370  4 5.261837  0.0062
 Correlation:
     (Intr)
CO22 -0.707
Number of Groups: 6
```

3. (20 points) Consider a Poisson-Gamma model with $Y|U = u \sim Possion(u)$ and U follows a Gamma distribution with mean θ and variance θ^2/ν, where $\nu > 0$ is known. Suppose we have n observations $y_1, \ldots, y_n$. Derive an EM algorithm to estimate the parameter θ.
   Hint: If $Z \sim Gamma(\alpha, \beta)$, *then* $E[Z] = \alpha\beta$, $Var[Z] = \alpha\beta^2$, *and*

$$f(z|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} z^{\alpha-1} e^{-z/\beta}, \quad 0 \le z < \infty, \quad \alpha, \beta > 0.$$