STAT 510 **Final Exam** **Spring 2015**

Instructions: The is a closed-notes, closed-book exam. No calculator or electronic device of any kind may be used. Use nothing but a pen or pencil. Please write your name and answers on blank paper. Please do NOT write your answers on the pages with the questions. For questions that require extensive numerical calculations that you should not be expected to do without a calculator, simply set up the calculation and leave it at that. For example, $(3.45 - 1.67)/\sqrt{2.34}$ would be an acceptable answer. On the other hand, some quantities that are very difficult to compute one way may be relatively easy to compute another way. Part of this exam tests your ability to figure out the easiest way to compute things, based on the information provided and the relationships between various quantities. In general, I expect you to be able to invert diagonal matrices of any dimensions and non-diagonal 2×2 matrices. If you find yourself trying to deal with something more complicated, there is probably a better way to solve the problem.

1. An experiment was conducted to assess the effect of a drug on weight gain in mice. A total of 20 mice were housed in individual cages. A balanced and completely randomized design was used to assign the 20 mice to five doses of the drug (0, 5, 10, 15, or 20 units). The weight gained by each mouse during the two-week period immediately following administration of the drug was recorded. Assume the weight gains are independent and normally distributed with a constant variance $\sigma^2 > 0$ and expected values that may depend on the dose of the drug.

   In an R workspace, the weight gains of each mouse are stored in a vector y whose entries are ordered by the dose received (i.e., weight gains for mice that received 0 units of the drug are first, followed by weight gains of mice that received 5 units of the drug, etc.). Use the following code and output to complete parts (a) and (b) below.

```
> dose
 [1]  0  0  0  0  5  5  5  5 10 10 10 10 15 15 15 15 20 20 20 20
>
> dosef=factor(dose)
>
> anova(lm(y˜dose+dosef))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
dose       1   57.6  57.600 12.8955 0.002674 **
dosef      3   15.6   5.200  1.1642 0.356112
Residuals 15   67.0   4.467
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

   (a) Does a simple linear regression model with weight gain as the response and dose of the drug as the explanatory variable fit the data adequately relative to a cell-means model that allows one unrestricted mean weight gain for each dose? Support your answer with appropriate statistical analysis.

   (b) Suppose a simple linear regression model with weight gain as the response and dose of the drug as the explanatory variable was fit to the data. R (or almost any another statistical package) would give the value of a t statistic that could be used to test whether the slope coefficient on dose is zero. Determine the absolute value of that t statistic.

2. Suppose

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

   (a) Find a fully simplified expression for $\boldsymbol{P_X}$.

   (b) Suppose

$$\boldsymbol{W} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

   Find a fully simplified expression for $\boldsymbol{P_X W}$.

3. A study was conducted to compare activity levels of pigs. Two breeds of pigs (labeled 1 and 2) were studied. Pigs were grouped in pens, with four pigs in each pen. Five pens, randomly chosen from ten, were used to hold pigs of breed 1, and the other five pens were used to hold pigs of breed 2. A motion tracker was attached to each pig, and the total distance traveled over a one-week period was recorded for each pig. Let y_{ijk} be the total distance traveled by the kth pig in the jth pen for the ith breed ($i = 1, 2;\ j = 1, 2, 3, 4, 5;\ k = 1, 2, 3, 4$). Assume the following model holds.

$$y_{ijk} = \mu_i + p_{ij} + e_{ijk}, \tag{1}$$

where μ_1 and μ_2 are unknown constants in $\mathbb{R}$; the p_{ij} terms are normal random variables with mean 0 and unknown variance $\sigma_p^2 > 0$; the e_{ijk} terms are normal random variables with mean 0 and unknown variance $\sigma_e^2 > 0$; and all the p_{ij} and e_{ijk} terms are mutually independent. Suppose the y_{ijk} values are stored in a vector $\boldsymbol{y}$ that is ordered first by breed, then by pen, and then by pig; i.e.,

$$\boldsymbol{y} = (y_{111}, y_{112}, y_{113}, y_{114}, y_{121}, y_{122}, y_{123}, y_{124}, \dots, y_{251}, y_{252}, y_{253}, y_{254})'.$$

   (a) The model in (1) can be written in the form $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$ using the notational conventions we have discussed in class. Using Kronecker product notation where helpful, provide specific expressions for $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{Z}$, and $\boldsymbol{u}$.

   (b) Provide a specific and fully simplified formula for the F statistic you would use to test

$$H_0 : \mu_1 = \mu_2.$$

   Express your answer using summation notation rather than notation involving matrices and vectors.

   (c) Suppose, for each pig, the total distance traveled was separately recorded for each day of the one-week period. Let y_{ijkl} be the total distance traveled on the lth day by the kth pig in the jth pen for the ith breed ($i = 1, 2;\ j = 1, 2, 3, 4, 5;\ k = 1, 2, 3, 4;\ l = 1, \dots, 7$). Fully specify the one model you suspect would be most appropriate for this y_{ijkl} data. There is no one right answer, but some answers are better than others. To get full credit, the model you specify for the y_{ijkl} data needs to be consistent with model (1) for the y_{ijk} data because model (1) is assumed to hold and $y_{ijk} = \sum_{l=1}^{7} y_{ijkl}$.

4. A total of 30 plants were assigned to treatment with 0, 1, 2, 3, 4, or 5 units of an anti-fungal chemical using a balanced and completely randomized design. After treatment with the assigned amount of chemical, each plant was dusted with fungus spores and grown in an environment favorable to fungus development. After two weeks, a leaf was randomly selected from each plant, and imaging software was used to estimate the proportion of each leaf's surface that was covered with the fungus. The data are provided below, where `x` is the number of units of anti-fungal chemical applied to a plant and `y` is the estimated proportion of a leaf's surface covered with the fungus.

```
> rbind(x,y)
  [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8] [,9] [,10]
x 0.000 0.000 0.000 0.000 0.000 1.000 1.000 1.000 1.00 1.000
y 0.300 0.829 0.621 0.764 0.846 0.224 0.631 0.756 0.86 0.685

  [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20]
x 2.000 2.000 2.000 2.000  2.00 3.000 3.000 3.000 3.000 3.000
y 0.345 0.802 0.623 0.516  0.36 0.412 0.406 0.505 0.333 0.381

  [,21] [,22] [,23] [,24] [,25] [,26] [,27] [,28] [,29] [,30]
x 4.000 4.000 4.000 4.000 4.000 5.000 5.000 5.000 5.000 5.000
y 0.272 0.377 0.445 0.246 0.176 0.223 0.343 0.409 0.327 0.368
```

Note that the response is a continuous random variable that takes values in the interval $(0, 1)$. One way to model such data is to assume a beta-distributed response. The density of a beta-distributed random variable can be written as

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi}(1-y)^{(1-\mu)\phi}, \quad 0 < y < 1,$$

where $\mu \in (0, 1)$ and $\phi > 0$ are the parameters of the beta distribution and Γ is the gamma function. Under this parameterization, a random variable with density $f(y; \mu, \phi)$ has mean μ and variance $\mu(1-\mu)/(1+\phi)$. Let $\text{beta}(\mu, \phi)$ denote the beta distribution whose density is $f(y; \mu, \phi)$. Using the notational conventions of 510, we can write a generalized linear model for a beta response as follows:

$$y_i \overset{ind}{\sim} \text{beta}(\mu_i, \phi), \text{ where } \log\left(\frac{\mu_i}{1-\mu_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

An `R` package called `betareg` can be used to fit this generalized linear model. Different choices for the linear predictor $(\boldsymbol{x}_i'\boldsymbol{\beta})$ can be specified using the usual `R` syntax. Use the `R` code and output on page 4 to complete parts (a), (b), and (c) at the bottom of page 4.

```
> o1=betareg(y~1)
> o2=betareg(y~x)
> o3=betareg(y~factor(x))
>
> #Maximum likelihood estimates of the beta vector and the phi parameter
> #for each model fit to the data.
>
> coef(o1)
(Intercept)       (phi)
-0.05304535  5.31605651
> coef(o2)
(Intercept)           x       (phi)
  0.7485279  -0.3237699   9.3154437
> coef(o3)
(Intercept) factor(x)1 factor(x)2 factor(x)3 factor(x)4 factor(x)5      (phi)
  0.7332242 -0.1916561 -0.5981563 -1.0746948 -1.5133018 -1.3654358  9.8548887
>
> #AIC for each model fit to the data.
>
> AIC(o1)
[1] -9.359327
> AIC(o2)
[1] -23.97535
> AIC(o3)
[1] -17.68737
```

(a) Suppose a leaf is treated with 2 units of the anti-fungal chemical, dusted with fungus spores, and grown for two weeks in the environment favorable to fungus development used in this experiment. Using the output from the model preferred by AIC, estimate the expected value of the proportion of the leaf's surface that will be covered with the fungus.

(b) Compute the value of the likelihood ratio test statistic ($-2 \log \Lambda$) that could be used to test whether the model corresponding to `o2` fits the data adequately compared to the model corresponding to `o3`.

(c) The coefficient on `x` in the model corresponding to `o2` is estimated to be -0.3237699. Use the available output to compute a standard error for this estimate. (Do NOT attempt to derive the inverse Fisher information matrix. There is an easier way to obtain a reasonable standard error value from the available output.)

5. Imagine an online social network company where each user is "friends" with one or more other users. The company has two algorithms (labeled 1 and 2) for predicting the three friends each user is most interested in. Algorithms 1 and 2 are each used to find these top three friends for each user. For the users where the two algorithms return different results, the following process is carried out. A total of 40 e-mail messages are sent to each user over the course of several weeks. In 20 of the e-mail messages, the user is provided a link for visiting the social network site to see information about the three friends identified by algorithm 1. In the other 20 e-mail messages, the user is provided a link for visiting the social network site to see information about the three friends identified by algorithm 2. For $i = 1, \ldots, 1000$ users and $j = 1, 2$ algorithms, let y_{ij} be the number of times out of 20 that user i visited the social network site by clicking a link in an e-mail generated by algorithm j. Researchers at the company decide to fit the model

$$y_{ij} \overset{ind}{\sim} \text{binomial}(20, \pi_{ij}), \text{ where } \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_i + \alpha_j$$

and $\beta_1, \ldots, \beta_{1000}$ and α_1, α_2 are unknown parameters in $\mathbb{R}$. Below is a portion of the dataset assembled by the company, along with R code and a portion of some R output.

```
> head(d)
  customerID algorithm  y
1          1         1  9
2          1         2 17
3          2         1  6
4          2         2  4
5          3         1 18
6          3         2 13
> tail(d)
     customerID algorithm y
1995        998         1 8
1996        998         2 8
1997        999         1 2
1998        999         2 5
1999       1000         1 0
2000       1000         2 3

> o=glm(cbind(y,20-y)~customerID+algorithm,
        family=binomial(link=logit),data=d)

> summary(o)
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      4.086e-01  3.335e-01   1.225 0.220513
customerID2     -1.737e+00  4.956e-01  -3.504 0.000458 ***
customerID3      6.236e-01  5.056e-01   1.233 0.217402
.
.
.
[OUTPUT FOR CUSTOMER IDS 4 THROUGH 998 HAS BEEN OMITTED TO SAVE SPACE]
```

```
.
.
.
customerID999  -2.192e+00  5.342e-01  -4.103 4.07e-05 ***
customerID1000 -3.159e+00  6.875e-01  -4.594 4.34e-06 ***
algorithm2      4.352e-01  2.425e-02  17.944  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17595.5  on 1999  degrees of freedom
Residual deviance:  3109.6  on  999  degrees of freedom
AIC: 10647

Number of Fisher Scoring iterations: 15
```

(a) Determine the likelihood ratio statistic for testing

$$H_0 : \text{For all } i \text{ and } j,\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma \text{ for some } \gamma \in \mathbb{R}$$

versus

$$H_A : \text{For all } i \text{ and } j,\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_i + \alpha_j$$

for some $\beta_1, \ldots, \beta_{1000} \in \mathbb{R}$ and $\alpha_1, \alpha_2 \in \mathbb{R}$ such that $\beta_i + \alpha_j$ is not the same for all i and j.

(b) Doing the best you can with the code and output provided, fill in values or calculations for (i), (ii), (iii), (iv), and (v) in the following sentences.

> The odds of a given user following the link in an e-mail generated by algorithm ____(i)____ are estimated to be ____(ii)____ times the odds of that same user following the link in an e-mail generated by algorithm ____(iii)____. An approximate 95% confidence interval for this multiplicative effect on the odds is ____(iv)____ to ____(v)____.

[Do NOT write your answers on this sheet. Instead, write (i), (ii), (iii), (iv), and (v) on your answer sheet and provide a value for each.]