**STAT 511 Final Exam Spring 2010**

1. An experiment was conducted to study the effects of a drug and three diets on low-density lipoprotein (LDL) blood cholesterol levels. A total of 60 mice with high levels of LDL blood cholesterol were used in the experiment. Each mouse was housed in its own cage. The 60 mice were randomly assigned to the following six treatment groups using a balanced and completely randomized design with 10 mice in each treatment group.

| Treatment | Diet | Drug |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | 2 | 1 |
| 4 | 2 | 2 |
| 5 | 3 | 1 |
| 6 | 3 | 2 |

In the table above, Drug = 1 represents a placebo and Drug = 2 represents the drug of interest.

At the conclusion of the study, the improvement in LDL blood cholesterol level was recorded for each mouse. Positive values of the response indicate improvement (lowering) of LDL blood cholesterol level while negative values indicate an increase in LDL blood cholesterol. The response data was stored in a vector $\boldsymbol{y}$ and sorted by treatment group. Use the code and partial output below to help answer the following questions.

```
> diet=as.factor(rep(1:3,each=20))
> drug=as.factor(rep(rep(1:2,each=10),3))
> o=lm(y~diet+drug)
> summary(o)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.000      1.662   0.602  0.54974
diet2          4.000      2.036   1.965  0.05440 .
diet3          6.000      2.036   2.948  0.00466 **
drug2          2.500      1.662   1.504  0.13818
```

(a) According to the model fit by the researchers, what is the best linear unbiased estimate of the mean for Treatment 6?

(b) According to the model fit by the researchers, was the mean response for Treatment 1 significantly different from zero? Provide a test statistic, its degrees of freedom, a $p$-value, and a conclusion for a significance level 0.05 test.

(c) The researchers would like to know if there are any significant differences among the three diets with regard to lowering mean LDL blood cholesterol level. As far as possible based on the output provided, complete the entries in the table below. (Do not worry about making any adjustments for multiple testing in this case.)

| Effect Difference | Estimate | SE | $t$-statistic | $p$-value$\leq 0.05$? (yes or no) |
|---|---|---|---|---|
| Diet 1 − Diet 2 | | | | |
| Diet 1 − Diet 3 | | | | |
| Diet 2 − Diet 3 | | | | |

(d) Based on information in the output, estimate the error variance $\sigma^2$ for the Gauss-Markov linear model that the researchers fit to the data.

(e) State the degrees of freedom associated with the estimate in part (d).

2. Researchers were interested in studying the effects of three types of surgically implanted hearing aids (labeled 1, 2, and 3) on subjects with a certain type of hearing loss in both ears. For the purposes of this question, we will pretend that the experiment involved only a total of six subjects (even though many more subjects would likely be needed to obtain useful results). Two of the three hearing aid types were randomly assigned to each subject. The two hearing aid types randomly assigned to a subject were also randomly assigned to ears within subject. Prior to surgical implantation, each subject received a separate hearing test for each ear. Following surgical implantation, each ear of each subject was tested again. Thus, a pre-test and post-test score were available for each subject and each ear. In the code below, the variable *test* is coded as 1 for the pre-test and 2 for the post-test, and the variable *y* contains the test score. The variables *subj*, *ear*, and *hearaid* contain the subject ID, the ear ID, and the hearing aid type, respectively. All the variables except for *y* are factors in R.

```
> d
```

| | subj | ear | hearaid | test | y |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 54.8 |
| 2 | 1 | 1 | 1 | 2 | 66.9 |
| 3 | 1 | 2 | 2 | 1 | 56.6 |
| 4 | 1 | 2 | 2 | 2 | 73.5 |
| 5 | 2 | 3 | 1 | 1 | 43.9 |
| 6 | 2 | 3 | 1 | 2 | 56.4 |
| 7 | 2 | 4 | 2 | 1 | 44.8 |
| 8 | 2 | 4 | 2 | 2 | 61.7 |
| 9 | 3 | 5 | 1 | 1 | 40.8 |
| 10 | 3 | 5 | 1 | 2 | 51.6 |
| 11 | 3 | 6 | 3 | 1 | 38.7 |
| 12 | 3 | 6 | 3 | 2 | 74.5 |
| 13 | 4 | 7 | 1 | 1 | 52.4 |
| 14 | 4 | 7 | 1 | 2 | 63.8 |
| 15 | 4 | 8 | 3 | 1 | 48.5 |
| 16 | 4 | 8 | 3 | 2 | 81.2 |
| 17 | 5 | 9 | 2 | 1 | 32.9 |
| 18 | 5 | 9 | 2 | 2 | 49.6 |
| 19 | 5 | 10 | 3 | 1 | 39.3 |
| 20 | 5 | 10 | 3 | 2 | 73.6 |
| 21 | 6 | 11 | 2 | 1 | 69.8 |
| 22 | 6 | 11 | 2 | 2 | 84.6 |
| 23 | 6 | 12 | 3 | 1 | 60.5 |
| 24 | 6 | 12 | 3 | 2 | 96.3 |

```
> o=lmer(y~hearaid*test+(1|subj)+(1|ear),data=d)
```

Note that the last line of the code fits a linear mixed effects model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$.

(a) Determine the first and last row of the design matrix $\boldsymbol{X}$.

(b) Determine the first and last row of the matrix $\boldsymbol{Z}$.

(c) Specify the matrix $\boldsymbol{Z}$ using Kronecker product notation.

(d) Provide notation for the variance components of the linear mixed effects model fit by the R code.

(e) According to the model fit by the researchers, what is the variance of a single response in terms of the variance components in part (d)?

(f) According to the model fit by the researchers, what is the correlation between the pre-test and post-test scores for the left ear of subject 1 in terms of the variance components defined in part (d)?

(g) According to the model fit by the researchers, what is the correlation between the pre-test score for the left ear of subject 1 and the post-test score for the right ear of subject 1?

(h) According to the model fit by the researchers, how many different values can occur in the vector $\mathrm{E}(\boldsymbol{y})$? Does this seem appropriate based on the design of the experiment? Explain why or why not.

3. Consider an experiment designed to compare the resistance of two plant genotypes (1 and 2) to a fungal pathogen. Ten pots, each containing one plant of each genotype, were used in the experiment. The two plants in each pot were infected with the pathogen. 24 hours later, a leaf from each plant was sampled and examined under a microscope. The number of infected plant cells was recorded for each leaf. The presence of many infected plant cells suggests a high level of susceptibility to the fungus and a low level of resistance. The presence of few infected plant cells suggests that the plant is resistant to the fungus. Use relevant portions of the following code and output to answer the following questions.

```
> ### The data with pot and geno as factors.
>
> d
   pot geno  y
1    1    1  8
2    1    2 20
3    2    1 13
4    2    2 17
5    3    1 13
6    3    2 12
7    4    1 26
8    4    2 24
9    5    1 15
10   5    2 26
11   6    1 23
12   6    2 17
13   7    1 13
14   7    2 17
15   8    1  9
16   8    2  9
17   9    1  7
18   9    2 22
19  10    1 20
20  10    2 20
>
> ###  The 0.95 quantile of chi-square distributions
> ###  with degrees of freedom 1 through 20.  These
```

```
> ###  quantiles may be useful for answering some
> ###  questions.
>
> qchisq(0.95,1:20)
 [1]  3.841459  5.991465  7.814728  9.487729 11.070498 12.591587 14.067140
 [8] 15.507313 16.918978 18.307038 19.675138 21.026070 22.362032 23.684791
[15] 24.995790 26.296228 27.587112 28.869299 30.143527 31.410433
>
> ### Model 1 ###
>
> o1=glm(y~geno,family=poisson(link=log),data=d)
> summary(o1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4347  -0.7378  -0.3307   0.9221   2.6557

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.68785    0.08248  32.589   <2e-16 ***
geno2        0.22450    0.11062   2.029   0.0424 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 42.834
Residual deviance: 38.690
AIC: 134.40
>
> ### Model 2 ###
>
> o2=glm(y~pot+geno,family=poisson(link=log),data=d)
> summary(o2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.79513  -0.57521  -0.00501   0.57997   1.38657

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.52052    0.19874  12.683   <2e-16 ***
pot2         0.06899    0.26277   0.263   0.7929
pot3        -0.11333    0.27516  -0.412   0.6804
pot4         0.57982    0.23604   2.456   0.0140 *
pot5         0.38137    0.24516   1.556   0.1198
pot6         0.35667    0.24640   1.448   0.1477
pot7         0.06899    0.26277   0.263   0.7929
```

```
pot8         -0.44183    0.30211  -1.462   0.1436
pot9          0.03509    0.26495   0.132   0.8946
pot10         0.35667    0.24640   1.448   0.1477
geno2         0.22450    0.11062   2.029   0.0424 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 42.834
Residual deviance: 14.397
AIC: 128.10
```

(a) Which of the two models is preferred based on AIC?

(b) Compute BIC for model 2.

(c) Does model 2 fit significantly better than a model that assumes that all 20 counts are independent and identically distributed draws from a single Poisson distribution?

i. Provide a test statistic that can be used to answer this question.

ii. State the degrees of freedom for the test statistic.

iii. What conclusion do you reach if you conduct the test at the .05 significance level?

(d) Does model 2 fit significantly better than model 1?

i. Provide a test statistic that can be used to answer this question.

ii. State the degrees of freedom for the test statistic.

iii. What conclusion do you reach if you conduct the test at the .05 significance level?

(e) Based on the fit of model 1, write one sentence that will help the researchers understand the meaning of the estimated parameter labeled *geno2* in the R output.

(f) Based on the fit of model 1, estimate the mean number of infected cells for plants of genotype 2.

4. Consider the model

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}.) + \varepsilon_i \quad (i = 1, \ldots, n);$$

where $\beta_0$ and $\beta_1$ are unknown real-valued parameters, $y_i$ is the response for observation $i$, $x_i$ is the value of an explanatory variable for observation $i$, and $\varepsilon_1, \ldots, \varepsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$ with $\sigma^2 > 0$ an unknown variance component.

Find penalized least squares estimates of $\beta_0$ and $\beta_1$ using the penalty $\lambda^2\beta_1^2$. Treat $\lambda^2$ as a known positive number.

5. Consider the model

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}.) + \beta_2|x_i - \bar{x}.| + \varepsilon_i \quad (i = 1, \ldots, n);$$

where $\beta_0$, $\beta_1$, and $\beta_2$ are unknown real-valued parameters, $y_i$ is the response for observation $i$, $x_i$ is the value of an explanatory variable for observation $i$, and $\varepsilon_1, \ldots, \varepsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$ with $\sigma^2 > 0$ an unknown variance component.

Describe the model for the response mean in simple terms. Explain how to interpret the parameters $\beta_0$, $\beta_1$, and $\beta_2$. You may wish to sketch the mean function to help make it clear that you understand the model for the mean.

6. The coefficient of variation of a random variable is defined as the standard deviation of the random variable divided by its mean. Suppose we are interested in the coefficient of variation of soybean yields in the population of all Iowa farms that grew soybeans in 2009. Suppose the variable $y$ in the following code and output contains the 2009 soybean yields (in bushels per acre) for 30 Iowa farms randomly selected from the population of interest. Use the code and output to answer the questions below.

```
> y
 [1] 51.9 57.4 47.0 58.6 46.0 63.0 64.7 52.3 54.1 52.9 54.0 42.2 55.6 56.3 58.5
[16] 49.8 47.4 54.3 45.9 54.0 52.0 52.0 54.1 52.6 41.0 45.1 45.4 44.1 45.0 48.2
> mean(y)
[1] 51.51333
> sd(y)/mean(y)
[1] 0.1141792
> theta.hat=function(y,i)
+ {
+   sd(y[i])/mean(y[i])
+ }
> o=boot(y,theta.hat,5000)
> theta.hat.stars=o$t[,1]
> summary(theta.hat.stars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06438 0.10280 0.11140 0.11130 0.11990 0.15400
> sd(theta.hat.stars)
[1] 0.01279081
> quantile(theta.hat.stars,seq(0.05,0.95,by=0.05))
        5%        10%        15%        20%        25%        30%        35%
0.08976578 0.09476124 0.09819774 0.10060491 0.10280312 0.10455682 0.10633138
       40%        45%        50%        55%        60%        65%        70%
0.10792183 0.10971629 0.11143942 0.11321008 0.11485838 0.11650101 0.11802092
       75%        80%        85%        90%        95%
0.11991253 0.12212301 0.12441699 0.12743493 0.13232325
```

(a) Estimate the population coefficient of variation based on the sample of 30 Iowa farms.

(b) Use the bootstrap to estimate the standard error of the estimator of the population coefficient of variation.

(c) Use the bootstrap to estimate the bias of the estimator of the population coefficient of variation.

(d) Use the bootstrap to obtain a bias-corrected estimate of the population coefficient of variation.

(e) Provide an approximate 90% percentile bootstrap confidence interval for the population coefficient of variation.