**STAT 510** **Final Exam** **Spring 2016**

**Instructions**: This 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. If you find yourself trying to do exceedingly complex or tedious calculations, there is probably a better way to solve the problem.

1. Suppose $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is a known $n \times p$ matrix of rank $r$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector, $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, and $\sigma^2$ is an unknown, positive variance parameter. Suppose $\boldsymbol{c}$ is a $p$-dimensional vector of constants such that $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable. Let $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ be the best linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$. Let $\hat{\sigma}^2$ be the REML estimator of $\sigma^2$.

   (a) Explain what it means for $\boldsymbol{c}'\boldsymbol{\beta}$ to be estimable.

   (b) Give an expression for $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ in terms of $\boldsymbol{c}$, $\boldsymbol{X}$, and $\boldsymbol{y}$.

   (c) Give an expression for $\hat{\sigma}^2$ in terms of $\boldsymbol{X}$, $\boldsymbol{y}$, $n$, and $r$.

   (d) Show that $(\boldsymbol{c}'\hat{\boldsymbol{\beta}} - \boldsymbol{c}'\boldsymbol{\beta})/\sqrt{\hat{\sigma}^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}$ has a $t$-distribution with $n - r$ degrees of freedom. You may use, without proof, any facts in our course notes (other than the fact you have been asked to show here, of course).

2. An experiment was conducted to assess the effects of two different bacterial strains on the quantity of a particular protein in the blood of pigs. The two strains were randomly assigned to 8 pigs using a completely randomized design with 4 pigs per strain. At each of times 1, 2, 3, and 4 hours after infection with the strains, one blood sample was collected from each of the 8 pigs. The protein quantity in each blood sample was measured. A linear model of the form $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{e}$ was fit to the data using `proc mixed`. The code and output are provided below. Based on the fit of the model, estimate the covariance between the measurements of protein quantity in the blood samples taken from a single pig at times 1 and 3 hours after infection.

```
proc mixed;
  class pig strain time;
  model proteinQuantity=strain time strain*time / ddfm=satterthwaite;
  random pig(strain);
  repeated time / subject=pig type=ar(1);
run;
```

```
  Covariance Parameter Estimates

Cov Parm        Subject    Estimate

pig(strain)                  6.5047
AR(1)           pig          0.6772
Residual                     8.3603
```

```
         Type 3 Tests of Fixed Effects

                   Num      Den
Effect              DF       DF     F Value    Pr > F

strain               1     6.19        1.56    0.2573
time                 3     11.6       74.32    <.0001
strain*time          3     11.6        2.56    0.1053
```

3. An experiment was conducted to assess the effect of a virus infection on two plant genotypes (labeled $G1$ and $G2$). Plants were grown in a growth chamber with one plant per pot. A total of 18 pots – 6 containing plants of genotype $G1$ and 12 containing plants of genotype $G2$ – were arranged in the growth chamber using a completely randomized design. On each plant, one leaf was randomly selected for infection with the virus, and another leaf was randomly selected for infection with a control substance. One week after infection, a device was used to measure the color of each leaf. The measurement device returned a continuous score, where high values of the score are associated with healthy, dark green leaves and low values are associated with pale, unhealthy leaves. Let $y_{ijk}$ be the score for genotype $Gi$ ($i = 1, 2$), infection $j$ ($j = 1$ for control and $j = 2$ for virus), and plant $k$ ($k = 1, \ldots, 6$ for genotype $G1$ and $k = 7, \ldots, 18$ for genotype $G2$). Suppose

$$y_{ijk} = \mu_{ij} + p_k + e_{ijk},$$

where $\mu_{11}, \mu_{12}, \mu_{21}$, and $\mu_{22}$ are unknown parameters, $p_k \sim N(0, \sigma_p^2)$ for all $k$, $e_{ijk} \sim N(0, \sigma_e^2)$ for all $i, j, k$, all random effects and errors are mutually independent, and $\sigma_p^2$ and $\sigma_e^2$ are unknown variance components. R code and output are provided after parts (a) through (e) below. Answer parts (a) through (e) using whatever parts of the R code and output you judge to be useful.

(a) Provide the value of a test statistic that can be used to test for a genotype main effect.

(b) Provide the value of a test statsitic that can be used to test for an infection main effect.

(c) Provide the value of a test statistic that can be used to test for genotype × infection interaction.

(d) Estimate $\sigma_e^2$.

(e) Estimate $\sigma_p^2$.

```
> #The data are stored in a data frame d.
> #The columns labeled Control and Virus give the response
> #for the leaf infected with the control substance and
> #virus, respectively.
>
> d
  Plant Genotype Control Virus
1     1       G1    96.7  88.8
2     2       G1    90.6  79.1
3     3       G1    84.7  75.8
4     4       G1    92.7  81.0
5     5       G1    91.1  83.2
6     6       G1    78.3  76.7
```

```
7      7       G2    81.6  76.6
8      8       G2    77.8  87.0
9      9       G2    89.6  81.5
10    10       G2    93.8  85.5
11    11       G2    84.7  87.4
12    12       G2    87.1  77.7
13    13       G2    72.7  68.6
14    14       G2    79.1  80.2
15    15       G2    77.6  81.7
16    16       G2    72.2  74.9
17    17       G2    74.8  81.9
18    18       G2    83.4  73.5
> y=as.vector(t(cbind(d$Control,d$Virus)))
> geno=factor(rep(1:2,c(12,24)))
> infection=factor(rep(1:2,18))
> y
 [1] 96.7 88.8 90.6 79.1 84.7 75.8 92.7 81.0 91.1 83.2 78.3 76.7 81.6 76.6 77.8
[16] 87.0 89.6 81.5 93.8 85.5 84.7 87.4 87.1 77.7 72.7 68.6 79.1 80.2 77.6 81.7
[31] 72.2 74.9 74.8 81.9 83.4 73.5
> geno
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Levels: 1 2
> infection
 [1] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
Levels: 1 2
>
> anova(lm(y˜geno+infection+geno:infection))
Analysis of Variance Table

Response: y
               Df  Sum Sq Mean Sq F value  Pr(>F)
geno            1  157.53  157.53  4.2404 0.04769 *
infection       1  126.19  126.19  3.3967 0.07461 .
geno:infection  1   91.35   91.35  2.4590 0.12669
Residuals      32 1188.79   37.15
>
> avg=(d$Control+d$Virus)/2
> summary(lm(avg˜0+d$Genotype))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
d$GenotypeG1   84.892      2.169   39.14   <2e-16 ***
d$GenotypeG2   80.454      1.534   52.46   <2e-16 ***

Residual standard error: 5.313 on 16 degrees of freedom
Multiple R-squared:  0.9963,    Adjusted R-squared:  0.9958
F-statistic:  2142 on 2 and 16 DF,  p-value: < 2.2e-16
```

```
> diff=d$Control-d$Virus
> summary(lm(diff˜1))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.744      1.569   2.386   0.0289 *

Residual standard error: 6.658 on 17 degrees of freedom

> summary(lm(diff˜0+d$Genotype))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
d$GenotypeG1    8.250      2.439   3.383   0.0038 **
d$GenotypeG2    1.492      1.724   0.865   0.3998

Residual standard error: 5.974 on 16 degrees of freedom
Multiple R-squared:  0.4325,    Adjusted R-squared:  0.3615
F-statistic: 6.096 on 2 and 16 DF,  p-value: 0.01076
```

4. A company trains workers to perform a task that is repeated many times. Each time the task is performed, the outcome of the task can be classified as a success or a failure. The company assigns each of 50 trainees to one of five training times (0, 4, 8, 12, or 16 hours) using a balanced and completely randomized design. At the conclusion of the assigned training, each trainee independently performs the task 100 times, and the number of successes and the number of failures are recorded. The resulting data are plotted in Figure 1 on page 5. Two models (labeled MODEL 1 and MODEL 2) were fit to the data in `R` code provided on pages 5 and 6. The estimated values for mean number of successes according to these models are plotted (smooth curve for MODEL 1, dashes for MODEL 2) along with the data (circles) in Figure 1. Use the `R` code and output to complete parts (a) through (f) below.

   (a) Find $BIC$ for MODEL 1.

   (b) Using the code and output for MODEL 1, estimate the amount of training required to attain a success probability of 0.25.

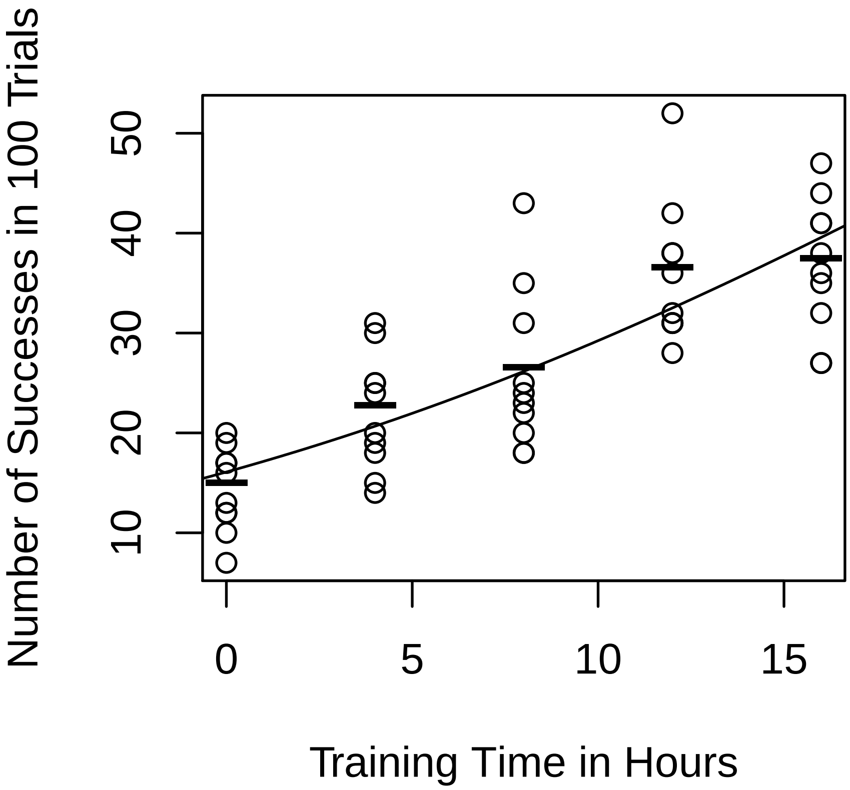
   (c) Using the code and output for MODEL 1, determine an appropriate standard error for the estimated amount of training required to attain a success probability of 0.25.

   (d) Provide a statement that shows you understand how to interpret the MODEL 1 coefficient whose estimate in the `R` output is 0.077.

   (e) Determine the value of a test statistic you would use to decide whether MODEL 1 or MODEL 2 is more appropriate for this dataset.

   (f) Give the approximate distribution of the test statistic in part (e) under the null hypothesis.

**Figure 1**



```
> #x is training time in hours.
> #y is the number of successes in 100 trials.
>
> rbind(x,y)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
x    0    0    0    0    0    0    0    0    0     0     4     4     4     4
y   10   16   12   13   20   17   12   19   17     7    31    24    30    14
  [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,27]
x     4     4     4     4     4     4     8     8     8     8     8     8     8
y    19    20    15    25    18    25    22    24    18    31    18    20    35
  [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37] [,38] [,39] [,40]
x     8     8     8    12    12    12    12    12    12    12    12    12    12
y    25    23    43    31    38    52    28    31    38    36    31    32    42
  [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49] [,50]
x    16    16    16    16    16    16    16    16    16    16
y    27    47    27    38    41    35    32    36    44    41
```

```
>
> #MODEL 1
>
> o1=glm(cbind(y,100-y)~x,family=binomial(link=logit))
>
> summary(o1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.652     0.0626  -26.37   <2e-16 ***
x               0.077     0.0059   13.05   <2e-16 ***

    Null deviance: 288.25  on 49  degrees of freedom
Residual deviance: 110.30  on 48  degrees of freedom
AIC: 350.8

> coef(o1)
(Intercept)           x
     -1.652       0.077
> vcov(o1)
            (Intercept)        x
(Intercept)     0.00392 -0.00032
x              -0.00032  0.00003
>
> #MODEL 2
>
> o2=glm(cbind(y,100-y)~factor(x),family=binomial(link=logit))
>
> summary(o2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.791    0.09033 -19.822  < 2e-16 ***
factor(x)4     0.531    0.11819   4.491 7.10e-06 ***
factor(x)8     0.739    0.11563   6.395 1.61e-10 ***
factor(x)12    1.211    0.11183  10.828  < 2e-16 ***
factor(x)16    1.250    0.11162  11.197  < 2e-16 ***

    Null deviance: 288.248  on 49  degrees of freedom
Residual deviance:  98.209  on 45  degrees of freedom
AIC: 344.71
```