STAT 510 **Final Exam** **Spring 2018**

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

Provided below are 0.975 quantiles from t distributions with degrees of freedom 5 through 25. One or more or these quantiles may be needed for answering some questions on this exam.

```
> rbind(5:25, round(qt(0.975, 5:25), 3))
      [,1]  [,2]  [,3]  [,4]  [,5]   [,6]   [,7]   [,8]  [,9]  [,10]  [,11] [,12]
[1,] 5.000 6.000 7.000 8.000 9.000 10.000 11.000 12.000 13.00 14.000 15.000 16.00
[2,] 2.571 2.447 2.365 2.306 2.262  2.228  2.201  2.179  2.16  2.145  2.131  2.12
     [,13]  [,14]  [,15]  [,16] [,17]  [,18]  [,19]  [,20] [,21]
[1,] 17.00 18.000 19.000 20.000 21.00 22.000 23.000 24.000 25.00
[2,]  2.11  2.101  2.093  2.086  2.08  2.074  2.069  2.064  2.06
```

1. Consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is an observed $n \times 1$ vector of response values, $\boldsymbol{X}$ is a known $n \times p$ matrix of rank r, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n\times n})$ for some unknown positive variance parameter σ^2.

   (a) Provide a definition for *column space* of $\boldsymbol{X}$.

   (b) Suppose $\boldsymbol{C}$ is a $q \times p$ matrix. What must be true about $\boldsymbol{C}$ in order for $\boldsymbol{C}\boldsymbol{\beta}$ to be estimable?

   (c) Suppose $\boldsymbol{C}\boldsymbol{\beta}$ is estimable. Provide an expression for the variance of the best linear unbiased estimator of $\boldsymbol{C}\boldsymbol{\beta}$.

   (d) What is the distribution of the $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}/\sigma^2$?

2. Researchers created a device to test the effectiveness of helmets at reducing the stress caused by head impacts. The device includes a head-shaped sensor on which a helmet can be placed, as well as a striking weight that can produce impacts to the front or side of a helmet placed on the sensor. The intensity of each impact can be controlled by the researchers. When an impact is delivered, a measurement of the amount of stress experienced by the head-shaped sensor is recorded. A measurement of 0 indicates no stress, while a measurement of 100 indicates stress high enough to cause serious brain injury.

   The researchers used the device to test a total of 10 helmets consisting of 5 helmets of type 1 and 5 helmets of type 2. The 10 helmets were tested in random order. When each helmet was tested, it was struck a total of 4 times: once with low impact to the front, once with high impact to the front, once with low impact to the side, and once with high impact to the side. The order of the 4 impacts was determined separately for each helmet using the following procedure. A fair coin was flipped. If the result of the flip was heads, the first two impacts were front impacts and the last two impacts were side impacts. If the result of the flip was tails, the first two impacts were side impacts and the last two impacts were front impacts. For the first two impacts, the coin was flipped again. If the result of the flip was heads, the first impact was at low intensity and the second was at high intensity. If the result of the flip was tails, the first impact was at high intensity and the second at low intensity. A coin was flipped a third time to determine the order of the impact intensities for the third and fourth impacts so that each order (low and then high vs. high and then low) was equally likely.

   Let $i = 1, 2$ index helmet types 1 and 2. Let $j = 1, \ldots, 5$ index helmets nested within helmet types. Let $k = 1, 2$ index the direction of impact, with $k = 1$ for front and $k = 2$ for side. Let $\ell = 1, 2$ index the intensity of the impact, with $\ell = 1$ for low and $\ell = 2$ for high. Let $y_{ijk\ell}$ be the stress measurement for the corresponding values of i, j, k, and ℓ. For $i = 1, 2$, $j = 1, \ldots, 5$, $k = 1, 2$, and $\ell = 1, 2$, consider the model

$$y_{ijk\ell} = \mu_{ik\ell} + a_{ij} + b_{ijk} + e_{ijk\ell}, \tag{1}$$

   where the $\mu_{ik\ell}$ values are unknown parameters, $a_{ij} \sim N(0, \sigma_a^2)$, $b_{ijk} \sim N(0, \sigma_b^2)$, $e_{ijk\ell} \sim N(0, \sigma_e^2)$, and all random terms are independent. Model (1) was fit to the dataset, and the following ANOVA table was obtained. Because we have a balanced experimental design, the type I and type III sums of squares are the same, and the lines of the ANOVA table can be reordered in a variety of ways without changing the results.

Source	Sum of Squares	Expected Mean Square
Type	226	
Direction	255	
Intensity	8910	
Type×Direction	207	
Type×Intensity	2	
Direction× Intensity	7	
Type×Direction×Intensity	9	
Helmet(Type)	254	$4\sigma_a^2 + 2\sigma_b^2 + \sigma_e^2$
Direction×Helmet(Type)	114	$2\sigma_b^2 + \sigma_e^2$
Error	59	σ_e^2
C. Total	10043	

   (a) We learned a shortcut for expressing sums of squares in summation notation that works for balanced designs like the one considered here. Use that shortcut to express the sum of squares for Direction × Intensity using summation notation.

(b) Compute a t statistic that can be used to test $H_0 : \bar{\mu}_{1\cdot\cdot} = \bar{\mu}_{2\cdot\cdot}$

(c) Compute the value of an unbiased estimator for σ_a^2

(d) The best linear unbiased estimator of $\bar{\mu}_{12\cdot} - \bar{\mu}_{11\cdot}$ is equal to 0.5 for this dataset. Provide a 95% confidence interval for $\bar{\mu}_{12\cdot} - \bar{\mu}_{11\cdot}$

(e) Compute a standard error for the best linear unbiased estimator of $\bar{\mu}_{121} - \bar{\mu}_{111}$

3. An experiment was conducted to assess the effect of a chemical seed treatment on the germination probability for plants of two genotypes. For each genotype, 120 seeds were divided into 10 batches, with 12 seeds per batch. The 10 batches for each genotype were treated with a randomly assigned chemical amount (0, 5, 10, 15, or 20 units), with 2 batches for each chemical amount. After chemical treatment, the 20 seed batches (10 for genotype 1 and 10 for genotype 2) were randomly assigned to 20 trays of soil, and the seeds from each batch were planted in their randomly assigned tray. Two weeks after planting, the number of seeds that germinated was recorded for each tray. Complete parts (a) through (f) below as well as possible, given the following code and partial output.

```
> #In the following data.frame d, genotype is a factor.
> #The variables chemical and y are numeric.
> #y is the number of seeds that germinated out of the 12 seeds in each tray.
> d
   genotype chemical  y
1         1        0  1
2         1        0  3
3         1        5  5
4         1        5  7
5         1       10  9
6         1       10  3
7         1       15  8
8         1       15 10
9         1       20 12
10        1       20 12
11        2        0  9
12        2        0  6
13        2        5  6
14        2        5 10
15        2       10  8
16        2       10  7
17        2       15 12
18        2       15  6
19        2       20 10
20        2       20 12
```

```
> o = glm(cbind(y, 12-y) ~ genotype + chemical + genotype:chemical,
+                          family = binomial(link = logit), data = d)
> summary(o)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6045  -1.0014   0.2257   1.3878   2.3864

Coefficients:
                   Estimate
(Intercept)           -1.50
genotype2              1.80
chemical               0.20
genotype2:chemical    -0.10

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 88.105  on 19  degrees of freedom
Residual deviance: 37.674  on 16  degrees of freedom

> vcov(o)
                  (Intercept) genotype2 chemical genotype2:chemical
(Intercept)             0.151    -0.151   -0.012              0.012
genotype2              -0.151     0.261    0.012             -0.020
chemical               -0.012     0.012    0.001             -0.001
genotype2:chemical      0.012    -0.020   -0.001              0.002
```

(a) Is the germination probability constant, or does it depend on the genotype and/or the amount of chemical? Compute the test statistic that you would use to answer this question. (You don't need to answer the question. Just compute the test statistic.)

(b) To find a p-value for the test statistic computed in part (a), what distribution would you use? In other words, state the null distribution of the test statistic you computed in part (a).

(c) Estimate the germination probability for seeds of genotype 1 that were treated with 0 units of the chemical.

(d) For the probability estimated in part (c), determine a confidence interval that has confidence level approximately equal to 95%.

(e) Estimate the amount of chemical that should be applied to the seeds to make the germination probability the same for both genotypes.

(f) Provide a standard error for the estimate computed in part (e).

4. Researchers investigated the effects of 20 different chemical compounds on the level of a protein in the blood of mice. On each of 5 days, 20 mice were randomly assigned to the 20 chemical compounds with one mouse for each compound. Each mouse was injected with its assigned compound, and then blood samples were taken from each mouse at 6 time points: 1, 2, 3, 4, 5 and 6 hours after injection. The same process was repeated each day with 20 different mice, so a total of 100 mice were used in the experiment. The level of the protein of interest was measured in each of the 600 blood samples.

   For $i = 1, \ldots, 5$, $j = 1, \ldots, 20$, and $k = 1, \ldots, 6$, let y_{ijk} be the protein level measurement on day i for chemical compound j at time k. For $i = 1, \ldots, 5$, $j = 1, \ldots, 20$, and $k = 1, \ldots, 6$, consider the model

$$y_{ijk} = \beta_0 + k\beta_1 + k^2\beta_2 + d_i + b_{0j} + kb_{1j} + k^2 b_{2j} + e_{ijk},$$

   where β_0, β_1, and β_2 are unknown fixed parameters and the other terms (aside from k, which is already defined as hours after injection) are random effects defined as follows. Let $\boldsymbol{d} = [d_1, \ldots, d_5]'$. For $j = 1, \ldots, 20$, let $\boldsymbol{b}_j = [b_{0j}, b_{1j}, b_{2j}]'$. For $i = 1, \ldots, 5$ and $j = 1, \ldots, 20$, let $\boldsymbol{e}_{ij} = [e_{ij1}, \ldots, e_{ij6}]'$. Suppose

$$\boldsymbol{d} \sim N(\boldsymbol{0}, \sigma_d^2 \boldsymbol{I}_{5\times 5}), \quad \boldsymbol{b}_j \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_b) \text{ for } j = 1, \ldots, 20, \text{ and}$$

$$\boldsymbol{e}_{ij} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_e) \text{ for } i = 1, \ldots, 5 \text{ and } j = 1, \ldots, 20,$$

   where σ_d^2 is an unknown positive variance parameter, $\boldsymbol{\Sigma}_b$ is an unknown 3×3 positive definite matrix, and $\boldsymbol{\Sigma}_e$ is an unknown 6×6 positive definite matrix. Finally, suppose that $\boldsymbol{d}$, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{20}$, and $\boldsymbol{e}_{11}, \ldots, \boldsymbol{e}_{520}$ are all independent.

   The researchers fit four versions of the model described above. The four versions – labeled A, B, C, and D – are identical except for the assumptions regarding the structure of $\boldsymbol{\Sigma}_e$. The four structures and their corresponding maximized REML log likelihoods are provided in the table below.

Version	Structure for $\boldsymbol{\Sigma}_e$	Maximized REML Log Likelihood
A	Diagonal constant variance ($\sigma^2 \boldsymbol{I}_{6\times 6}$ for some $\sigma^2 > 0$)	-1317
B	Compound Symmetry	-1235
C	AR(1)	-1158
D	Unstructured	-1064

   (a) Determine the dimension of the model parameter space for version D of the model.

   (b) Compute the likelihood ratio test statistic for testing the goodness of fit of version B of the model relative to version D of the model.

   (c) State the degrees of freedom associated with the test statistic in part (b).

   (d) Which of the four versions of the model is preferred according to AIC?