**STAT 510** **Exam 1** **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 the answer sheets provided. 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 8 mice were assigned to two treatment groups (1 and 2) using a completely randomized design with 4 mice per treatment. For $i = 1, 2$ and $j = 1, 2, 3, 4$, let $y_{ij}$ be the value of a response variable for the $j$th mouse that received treatment $i$. Suppose response values from all mice are mutually independent, and suppose $y_{ij} \sim N(\mu_i, \sigma^2)$ for all $i = 1, 2$ and $j = 1, 2, 3, 4$. Use the following R code and output to find a 95% confidence interval for $\mu_1 - \mu_2$.

```
> #y is the vector of responses.
> #The first four entries correspond to treatment 1.
> #The last four entries correspond to treatment 2.
> X
     [,1] [,2]
[1,]    1    1
[2,]    1    1
[3,]    1    1
[4,]    1    1
[5,]    1   -1
[6,]    1   -1
[7,]    1   -1
[8,]    1   -1
> t(X) %*% y
     [,1]
[1,] 52.4
[2,]  7.6
> Px = X %*% solve(t(X) %*% X) %*% t(X)
> t(y) %*% y - t(y) %*% Px %*% y
     [,1]
[1,] 2.52
> round(qt(.975, df = 1:10), 2)
[1] 12.71  4.30  3.18  2.78  2.57  2.45  2.36  2.31  2.26  2.23
```

2. Consider an experiment with two factors: $A$ with two levels (1 and 2) and $B$ with two levels (1 and 2). Suppose a completely randomized design is used to obtain values of a response variable for a total of 20 experimental units, with 5 experimental units for each of the 4 treatments formed by combining one level of factor $A$ with one level of factor $B$. For $i = 1, 2$, $j = 1, 2$, and $k = 1, \ldots, 5$, let $y_{ijk}$ be the value of a response variable for the $k$th experimental unit treated with level $i$ of factor $A$ and level $j$ of factor $B$. Suppose the response averages for the four treatment groups are

$$\bar{y}_{11\cdot} = 5.9 \quad \bar{y}_{12\cdot} = 3.4 \quad \bar{y}_{21\cdot} = 2.2 \quad \bar{y}_{22\cdot} = 2.5.$$

Consider the following `R` code and output.

```
> A = factor(rep(1:2, each = 10))
> A
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
Levels: 1 2
> B = factor(rep(rep(1:2, each = 5), 2))
> B
 [1] 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2
Levels: 1 2
> #y is the vector of responses, ordered appropriately relative to A and B;
> #i.e., the first 5 entries correspond to the treatment (A=1,B=1),
> #the next 5 to the treatment (A=1,B=2), etc.
> o = lm(y ~ A + B + A:B)
> yhat = fitted(o)
> sum((y - yhat)^2)
[1] 15.0
```

(a) Provide the vector that would be produced by the `R` command `coef(o)`.

(b) Provide the entry in the third row and fourth column of the matrix that would be produced by the command `vcov(o)`.

3. Suppose

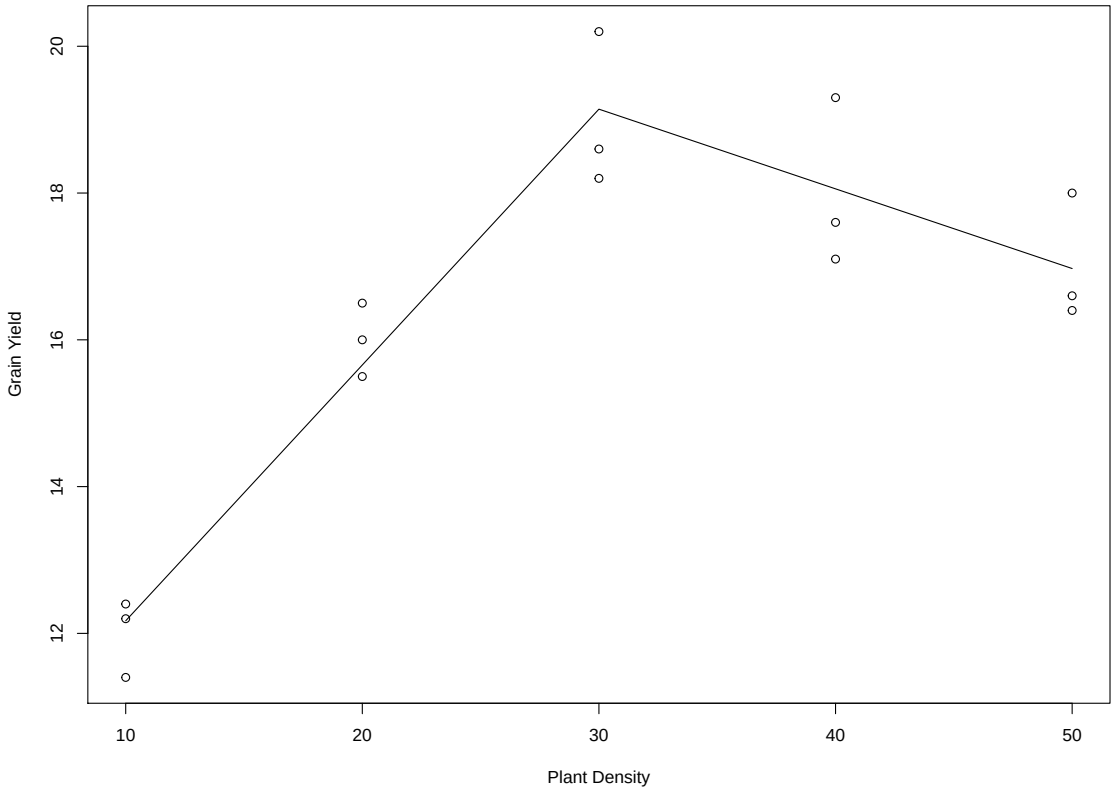
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left( \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix} \right).$$

Prove that $\bar{y}_{\cdot} = (y_1 + y_2)/2$ is NOT the BLUE of $\mu$.

4. Reconsider the example discussed in class where 5 different plant densities (10, 20, 30, 40, and 50) were assigned to 15 plots of land using a balanced and completely randomized design. At the end of the study, grain yield was recorded for each plot. Suppose the plant density data are available in an R workspace as a numeric vector x.

```
> x
 [1] 10 10 10 20 20 20 30 30 30 40 40 40 50 50 50
```

Suppose the yield data are available in the same R workspace as a numeric vector y. The figure below shows a scatterplot of the data, along with the fit of a continuous piecewise linear function.




The continuous piecewise linear function is linear for $x \in (-\infty, 30]$ and also linear for $x \in [30, \infty)$. The two linear pieces of the function are allowed to have different slopes, but because the function is continuous, both linear pieces must have the same value at $x = 30$. Let $\beta_0$ be the intercept (i.e., the value of the piecewise linear function at $x = 0$). Let $\beta_1$ be the slope for $x \in (-\infty, 30)$, and let $\beta_2$ be the slope for $x \in (30, \infty)$. For appropriate choices of the R vectors x1 and x2, ordinary least squares estimates of $\beta_0$, $\beta_1$, and $\beta_2$ can be obtained from the following R code.

```
piecewiseLinear = lm(y ~ x1 + x2)
coef(piecewiseLinear)
```

(a) Define the vectors x1 and x2 so that the code above will provide the ordinary least squares estimates of $\beta_0$, $\beta_1$, and $\beta_2$.

Now let $\boldsymbol{X}_1$ be the $15 \times 1$ matrix with all entries equal to 1. Let $\boldsymbol{X}_2$ be the model matrix for the simple linear regression model that can be fit with the code

```
simpleLinear = lm(y ~ x)
```

Let $\boldsymbol{X}_3$ be the model matrix corresponding to the piecewise linear model that was the subject of part (a). Let $\boldsymbol{X}_4$ be the model matrix corresponding to the cell means model that can be fit with the code

```
cellMeans = lm(y ~ 0 + factor(x))
```

Consider the following code and output.

```
> anova(simpleLinear, cellMeans)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ 0 + factor(x)
  Res.Df   RSS   Df   Sum of Sq        F    Pr(>F)
1     13  51.9
2     10   7.5    3        44.4   19.733   0.00016

> coef(cellMeans)
factor(x)10 factor(x)20 factor(x)30 factor(x)40 factor(x)50
         12          16          19          18          17

> sum((y - mean(y))^2)
[1] 95.1
```

(b) Provide the sequential sums of squares and degrees of freedom for the ANOVA table associated with the column space sequence $\mathcal{C}(\boldsymbol{X}_1) \subset \mathcal{C}(\boldsymbol{X}_2) \subset \mathcal{C}(\boldsymbol{X}_3) \subset \mathcal{C}(\boldsymbol{X}_4)$. Your table should be formatted as follows.

| Source | Sum of Squares | Degrees of Freedom |
|---|---|---|
| Linear | | |
| Piecewise Linear | | |
| Cell Means | | |
| Error | | |
| Corrected Total | | |