STAT 510 **Exam 1** **Spring 2019**

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 $\boldsymbol{X}$ and $\boldsymbol{W}$ are any two matrices with n rows for which $\mathcal{C}(\boldsymbol{X}) = \mathcal{C}(\boldsymbol{W})$. Show that $\boldsymbol{P_X} = \boldsymbol{P_W}$.

2. Suppose $\hat{\theta}$ is an estimator of a parameter θ. The bias of $\hat{\theta}$ as an estimator of θ is defined to be $E(\hat{\theta}) - \theta$. The mean squared error of $\hat{\theta}$ as an estimator of θ is defined to be $E\left[(\hat{\theta} - \theta)^2\right]$. Show that $E\left[(\hat{\theta} - \theta)^2\right]$ is equal to the variance of $\hat{\theta}$ plus the square of the bias of $\hat{\theta}$ as an estimator θ.

3. An experiment was conducted to study the effect of a fertilizer on corn yield. Four fertilizer amounts (0, 2, 4, and 10 pounds per acre) were assigned to 20 plots of land using a balanced and completely randomized design with 5 plots per fertilizer amount. Let $x_1 = 0$, $x_2 = 2$, $x_3 = 4$, and $x_4 = 10$. Let y_{ij} be the yield for the jth plot that received x_i pounds of fertilizer per acre ($i = 1, 2, 3, 4$, $j = 1, \ldots, 5$). The researchers fit to the data the following model:

$$y_{ij} = \beta_1 + \beta_2(x_i - \bar{x}.) + \epsilon_{ij}, \text{ where } \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \tag{1}$$

$\bar{x}. = 4$ (the average fertilizer amount), and $\beta_1, \beta_2 \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$ are unknown parameters.

   (a) Let θ be the expected value of yield for plots receiving 4 pounds of fertilizer per acre; i.e., $\theta = E(y_{3j})$ for all $j = 1, \ldots, 5$. Find a fully simplified expression for the estimator of θ obtained from the fit of model (1) to the data.

   Suppose that, unknown to the researchers, the true model for the data is actually

$$y_{ij} \sim N(\mu_i, 36), \text{ where } \mu_1 = 160, \mu_2 = 180, \mu_3 = 200, \mu_4 = 252, \tag{2}$$

   and all yields are independent. Note that this model (2) is a special case of the cell-means model discussed in class. The remainder of this problem is concerned with the performance of the researchers' analysis based on the fit of model (1), considering that the data actually follow the cell-means model (2). Note that model (2) implies that θ in part (a) has true value $\mu_3 = 200$.

   (b) Find the variance of the researchers' estimator of θ in part (a) considering that model (2) is the true model.

   (c) Find the bias of the researchers' estimator of θ in part (a) considering that model (2) is the true model.

   (d) Find the mean squared error of the researchers' estimator of θ in part (a) considering that model (2) is the true model.

(e) If the researchers were to fit a cell-means model to their data, would the mean squared error of the cell-means-model estimator of θ be less than, equal to, or greater than the mean squared error computed in part (d)? Explain. (Again, assume that model (2) is the true model.)

(f) Suppose the researchers construct a test for lack of fit of their model (1) against a cell-means model. Once again assuming that model (2) is the true model, determine the distribution of the lack-of-fit test statistic. Be as specific as possible.

4. A company that manages a regional chain of pizza restaurants decides to launch an advertising campaign to increase delivery sales on Valentines Day. The company's marketing department produces two versions (say, 1 and 2) of an advertisement that can be mailed to potential customers. For future reference, the company would like to know which version of the advertisement is better at generating sales. The company has 10 stores in each of 3 regions. Within each region, 5 stores (randomly selected from the 10) were assigned advertisement 1. The other five stores in each region were assigned advertisement 2. For each store, 100 copies of the assigned advertisement were mailed to 100 randomly selected households within five miles of the store. The total of pizza delivery sales on Valentines Day to the 100 households that received a mail advertisement were recorded for each store. Let y_{ijk} be the total dollar value of pizza delivery sales to the 100 households corresponding to region i, advertisement j, and store k ($i = 1, 2, 3$, $j = 1, 2$, and $k = 1, \ldots, 5$). The researchers assume the cell-means model

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \text{ where } \epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \tag{3}$$

and $\mu_{ij} \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. Use the R code and partial output at the end of this exam to complete parts (a) through (d) below.

(a) Determine the Least Squares Mean (LSMEAN) for advertisement 1 and the LSMEAN for advertisement 2.

(b) Provide a 95% confidence interval for $\mu_{11} - \mu_{12}$.

(c) Suppose the researchers fit to the data the additive model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \text{ where } \varepsilon_{ijk} \overset{iid}{\sim} N(0, \eta^2) \tag{4}$$

and μ, α_i, $\beta_j \in \mathbb{R}$ and $\eta^2 > 0$ are unknown parameters. Determine the F statistic for testing the lack of fit of model (4) relative to model (3).

(d) Suppose an executive for the company asks for your interpretation of the results of this experiment. Please write a few sentences summarizing the results for the executive.

R Code and Output for Problem 4

```
> head(d)
  region ad store   y
1      1  1     1 180
2      1  1     2 100
3      1  1     3 106
4      1  1     4  91
5      1  1     5 138
6      1  2     6 103
> d$region
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
Levels: 1 2 3
> d$ad
 [1] 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2
Levels: 1 2
> d$store
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
[25] 25 26 27 28 29 30
> o = lm(y ~ region + ad + region:ad, data = d)
> summary(o)

Coefficients:
            Estimate Std. Error
(Intercept)   123.00      ?????  (Note that all but the last standard error
region2        56.00      ?????   in the column has been intentionally
region3        -5.00      ?????   blocked out with question marks.)
ad2           -28.00      ?????
region2:ad2   -11.00      ?????
region3:ad2   108.00      36.00

> rbind(10:30, round(qt(0.975, 10:30), 3))
       [,1]   [,2]   [,3]  [,4]   [,5]   [,6]  [,7]  [,8]   [,9]  [,10]
[1,] 10.000 11.000 12.000 13.00 14.000 15.000 16.00 17.00 18.000 19.000
[2,]  2.228  2.201  2.179  2.16  2.145  2.131  2.12  2.11  2.101  2.093
      [,11] [,12]  [,13]  [,14]  [,15] [,16]  [,17]  [,18]  [,19]  [,20]
[1,] 20.000 21.00 22.000 23.000 24.000 25.00 26.000 27.000 28.000 29.000
[2,]  2.086  2.08  2.074  2.069  2.064  2.06  2.056  2.052  2.048  2.045
      [,21]
[1,] 30.000
[2,]  2.042
```