STAT 510 Homework 4

Due Date: 11:00 A.M., Wednesday, February 7

1. Suppose $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ for some unknown $\sigma^2 > 0$. Let $\hat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$.

   (a) Determine the distribution of
   $$\begin{bmatrix} \hat{\boldsymbol{y}} \\ \boldsymbol{y} - \hat{\boldsymbol{y}} \end{bmatrix}.$$

   (b) Determine the distribution of $\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}$.

2. Consider an experiment with four treatments and a completely randomized design. Suppose y_{ij} is the measurement of the response for the jth experimental unit treated with the ith treatment. Suppose n_i is the number of experimental units that provided an observation of the response for treatment i ($i = 1, 2, 3, 4$). Suppose the sample mean of the response observations, the sample variance of the response observations, and number of response observations for each treatment are as follows:

Treatment i	Sample Mean $\bar{y}_{i\cdot}$	Sample Variance $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$	Number of Observations n_i
1	31.2	4.1	3
2	39.5	3.4	2
3	22.8	2.8	2
4	26.3	3.2	4

Suppose all response observations are independent and normally distributed with variance σ^2 and expected value that depends on the treatment according to the following table:

Treatment	Expected Value of the Response
1	β_1
2	$\beta_1 + \beta_2$
3	$\beta_1 + \beta_2 + \beta_3$
4	$\beta_1 + \beta_2 + \beta_3 + \beta_4$

Let $\boldsymbol{y} = [y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{31}, y_{32}, y_{41}, y_{42}, y_{43}, y_{44}]'$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4]'$.

   (a) The stated assumptions about the distribution of the response values can be summarized by writing $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{X}$ is an appropriately chosen model matrix. Write the appropriate matrix $\boldsymbol{X}$.

   (b) Provide a numerical value for the BLUE of β_4.

   (c) Provide a numerical value for the standard error of the estimate computed in part (b).

3. Consider the simple linear regression model
   $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n$$
   where $\epsilon_1, \ldots, \epsilon_n$ are iid and have a normal distribution $N\left(0, \sigma^2\right)$ and β_0, β_1, and $\sigma^2 > 0$ are unknown parameters. The model matrix for this linear model can be written as $\boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{x}]$, where $\boldsymbol{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{x} = [x_1, \ldots, x_n]'$.

(a) We have learned that the least squares estimator of $\boldsymbol{\beta}$ in a general linear model with a full-rank matrix $\boldsymbol{X}$ is given by $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Simplify this expression for the special case of simple linear regression to obtain an expression for the least squares estimators of β_0 and β_1. Express your final answers using summation notation.

(b) There are some computational advantages to working with a model matrix whose columns are orthogonal. For the simple linear regression problem consider the model matrix

$$\boldsymbol{W} = [\mathbf{1}, \boldsymbol{x} - \bar{x}_\cdot \mathbf{1}]$$

This design matrix is obtained by "centering" the explanatory variable around its mean $\bar{x}_\cdot = \sum_{i=1}^{n} x_i/n$. Find a matrix $\boldsymbol{B}$ so that $\boldsymbol{X}\boldsymbol{B}^{-1} = \boldsymbol{W}$. It will follow that

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{\alpha}$$

where $\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{\alpha}$.

(c) Derive expressions for the least squares estimators of α_0 and α_1 (where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)'$ from part (b)) using $\hat{\boldsymbol{\alpha}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{y}$.

(d) Multiply $\hat{\boldsymbol{\alpha}}$ from part (c) by $\boldsymbol{B}^{-1}$ from part (b) to obtain expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$.

(e) Show that your answer to part (a) matches your answer to part (d).

4. An experiment was conducted to assess whether students rate their instructors differently depending on the perceived gender of their instructors. One female instructor (say, Jane) taught two online sections of a class entirely through an online course management system, where the only contact bewteen the instructor and students was through email or online discussion board comments. Likewise, one male instructor (say, John) taught two other online sections of the same class through the same online course management system. In one section actually taught by Jane, students were given Jane's identity as the instructor, but in the other section actually taught by Jane, students were told that their instructor was John. Similarly, in one section actually taught by John, students were given John's identity as the instructor, but in the other section actually taught by John, students were told their instructor was Jane.

   At the end of the semester, students completed a course evaluation that included scores on 12 questions about the quality of their instructors. For each student, the 12 scores provided by that student were combined together to obtain a single numerical instructor rating. This numerical rating serves as the response variable in this experiment. Larger values of the single numerical rating correspond to higher (i.e., better) instructor ratings. Please study the following R code and output and use it to answer parts (a) and (b) on page 4.

```
> #y is a vector.  Each entry in y contains the numerical instructor
> #rating provided by a student in one of the four online sections of the
> #course that were involved in this experiment.
>
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.100   2.750   3.300   3.228   3.700   5.000
```

```
> #Actual Instructor is represented by a factor in R called ai.
> #The ith entry in ai is the name of the instructor (either Jane or John)
> #who actually taught the ith student.
>
> ai
 [1] Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane
[16] Jane Jane Jane Jane Jane John John John John John John John John John John John
[31] John John John John John John John John John John John John John John
Levels: Jane John
>
> #Perceived Instructor is represented by a factor in R called pi.
> #The ith entry in pi is the name of the instructor (either Jane or John)
> #whom the ith student perceives to be his/her instructor.
>
> pi
 [1] Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane John John John John John
[16] John John John John John Jane Jane Jane Jane Jane Jane Jane Jane Jane Jane
[31] John John John John John John John John John John John John John John
Levels: Jane John
>
> #The number of students for each combination of ai and pi is
> #given in the following table.
>
> table(ai,pi)
      pi
ai     Jane John
  Jane   10   10
  John   10   13
>
> #Some results from fitting a particular linear model to
> #the instructor ratings are as follows:


> o=lm(y~ai+pi+ai:pi)
> summary(o)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.8500     0.2252  12.657 2.18e-15 ***
aiJohn         -0.0600     0.3184  -0.188  0.85152
piJohn          0.8700     0.3184   2.732  0.00941 **
aiJohn:piJohn  -0.1831     0.4371  -0.419  0.67766
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) The researchers conducting this study are not sure how to interpret the output produced by R. In particular, they wonder how to interpret the output line

```
piJohn          0.8700     0.3184   2.732  0.00941 **
```

that shows up in the table produced by the `summary(o)` command. They understand that 0.00941 is a small p-value that indicates statistical significance at the 0.01 level, but they aren't sure what hypothesis is being tested. Provide a brief explanation to help the researchers interpret this result in the context of their study. Do NOT assume that the researchers know the meaning of statistical terms like *simple effect*, *main effect*, or *interaction*. If you use those words in your answer, you'll need to explain their meaning for the researchers.

(b) Provide an approximate 95% confidence interval for the main effect of the factor *Perceived Instructor*.