STAT 510 **Exam 1** **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. 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). (You don't need to actually do the calculation, but set it up so the answer could be obtained easily by typing what you write into a simple calculator.)

2. Let A be a factor with two levels labeled $A1$ and $A2$. Let B be a factor with two levels labeled $B1$ and $B2$. Suppose the four treatment combinations obtained by combining one level of A with one level of B were randomly assigned to a total of 6 experimental units. After treatment, the value of a response variable was measured for each experimental unit. The observed response values are provided in the following table.

Experimental Unit	Level of Factor A	Level of Factor B	Observed Response
1	$A1$	$B1$	3
2	$A1$	$B1$	5
3	$A1$	$B2$	10
4	$A2$	$B1$	2
5	$A2$	$B2$	10
6	$A2$	$B2$	12

(a) Consider the cell means Gauss-Markov model that allows the expected value of the response to vary with treatment in an unconstrained manner (i.e., the four treatment means can be any real numbers). Compute LSMEANS for factor A.

(b) Fill in the missing entries in the following ANOVA table. To receive full credit, each missing entry must be calculated exactly or written down so that the answer could be obtained easily by entering what is written into a simple calculator. (Recopy the table onto your answer sheet; don't fill in the entries on this page.)

Source	Degrees of Freedom	Type I Sum of Squares
A	?	?
B	?	?
$A \times B$	?	?
Error	?	?
C. Total	?	?

3. 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*.

4. Researchers were interested in comparing the effects of three diets (labeled 1, 2, and 3) on weight gained by pigs. Within each of two research farms, 15 pigs were randomly assigned to the three diets using a completely randomized design with 5 pigs per diet. Thus, the experiment involved a total of 30 pigs. Immediately before placing each pig on its assigned diet, the initial weight of each pig was recorded. For $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, \ldots, 5$, let x_{ijk} be the initial weight of the kth pig assigned diet j on farm i, and let y_{ijk} be the weight gained by the kth pig while being fed diet j at farm i. Suppose for $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, \ldots, 5$,

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + \phi_i + \delta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{1}$$

where $\beta_0, \beta_1, \phi_1, \phi_2, \delta_1, \delta_2, \delta_3, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}$,and γ_{23} are unknown real-valued parameters and all the ϵ_{ijk} terms are independent normal random variables with mean 0 and some unknown positive variance σ^2. Use the SAS code and output on page 5 to complete parts (a) through (c) on page 6.

```
proc glm;
  class farm diet;
  model y=x farm diet farm*diet / solution;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	288.18	48.03	2.00	0.1068
Error	23	551.68	23.99		
Corrected Total	29	839.86			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	64.58	64.58	2.69	0.1144
farm	1	11.25	11.25	0.47	0.5002
diet	2	191.60	95.80	3.99	0.0324
farm*diet	2	20.75	10.37	0.43	0.6540

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	55.92	55.92	2.33	0.1404
farm	1	10.37	10.37	0.43	0.5174
diet	2	190.12	95.06	3.96	0.0332
farm*diet	2	20.75	10.37	0.43	0.6540

Parameter			Estimate	Standard Error	t Value	Pr > |t|
Intercept			56.79 B	11.87	4.79	<.0001
x			0.57	0.38	1.53	0.1404
farm	1		−2.50 B	3.21	−0.78	0.4435
farm	2		0.00 B	.	.	.
diet	1		2.91 B	3.13	0.93	0.3624
diet	2		−1.18 B	3.10	−0.38	0.7065
diet	3		0.00 B	.	.	.
farm*diet	1	1	3.66 B	4.44	0.82	0.4179
farm*diet	1	2	0.22 B	4.41	0.05	0.9606
farm*diet	1	3	0.00 B	.	.	.
farm*diet	2	1	0.00 B	.	.	.
farm*diet	2	2	0.00 B	.	.	.
farm*diet	2	3	0.00 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

(a) The solution to the normal equations provided by SAS for this problem is in the last section of the output on page 5. Provide a solution to the normal equations for this problem that is different than the solution provided by SAS.

(b) Consider the simple linear regression model

$$y_{ijk} = \alpha_0 + \alpha_1 x_{ijk} + \varepsilon_{ijk}, \tag{2}$$

where α_0 and α_1 are unknown real-valued parameters and all the ε_{ijk} terms are independent normal random variables with mean 0 and some unknown positive variance η^2. Determine the value of an F statistic that can be used to test whether the simple linear regression model (2) fits the data adequately relative to the full model given in expression (1) on page 4.

(c) If the simple linear regression model (2) from part (b) was fit to the data, almost any statistical package would provide the value of a t statistic for testing the null hypothesis $H_0 : \alpha_1 = 0$ based on the fit of model (2). Determine the absolute value of that t statistic.