STAT 640: Homework 10

Due **Wednesday, April 20, 11:59pm MT** on the course Canvas webpage. Please follow the homework guidelines on the syllabus.

Name: Hannah Butler

Problem 1

Consider the data in the file `coagulation.csv` on Canvas, which comes from an experiment of diet on blood coagulation time. Animals were independently randomized to four diets and the time for blood coagulation was measured.

a. Write the one-way ANOVA model that can be used to analyze this data.

Answer:

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \boldsymbol{Y}_3 \\ \boldsymbol{Y}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_4 & \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{0}_4 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 \\ \mathbf{1}_8 & \mathbf{0}_8 & \mathbf{0}_8 & \mathbf{0}_8 & \mathbf{1}_8 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \boldsymbol{\epsilon}$$

---

b. Is this a balanced design?

Answer: No, this is not a balanced design because we have a different number of replicates in each treatment group.

---

c. Is there a relationship between diet and coagulation time? Conduct your test at level $\alpha = 0.05$. In your response, provide the null hypothesis, alternative hypothesis, test statistic, the reference distribution of the test statistic under the null, and a brief conclusion.

Answer: Here, we want to test the null hypothesis that all of the $\alpha_i = 0$, or that none of the diets have an effect on blood coagulation time. In symbols, $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$. The alternative hypothesis would be that at least one of the diets has an effect on blood coagulation time, or in symbols, $H_1 : \alpha_1 \neq 0$ or $\alpha_2 \neq 0$ or $\alpha_3 \neq 0$ or $\alpha_4 \neq 0$ or any combination of those. In order to test this, we need to impose the distributional assumption on the model:

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}).$$

Without this assumption, we can not perform inference as usual.

We test the null hypothesis using an ANOVA F-test. The F-test statistic is

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{\boldsymbol{Y}^T(\boldsymbol{P_X} - \boldsymbol{J}_N)\boldsymbol{Y}/(t-1)}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}/(N-t)} = 13.57143 \quad \textit{(See code below).}$$

```
N <- nrow(coag)                          # Total number of observations
t <- length(unique(coag$diet))           # Number of treatment levels
Y <- coag$time                           # Response vector
X <- coag %>%                            # ANOVA Design matrix
  pivot_wider(names_from = "diet", values_from = 1) %>%
  mutate_all(function(x) ifelse(is.na(x), 0, 1)) %>%
  as.matrix()

Px <- X[,-1] %*% solve(t(X[,-1])%*%X[,-1]) %*% t(X[,-1])  # Proj. onto space of X
Jn <- matrix(1, nrow = N, ncol = N)                       # Matrix of overall means

F_num <- (t(Y) %*% (Px - Jn/N) %*% Y)/(t-1)
F_den <- (t(Y) %*% (diag(1, nrow = N) - Px) %*% Y)/(N-t)
F_stat <- F_num/F_den
```

```
## F-test statistic: 76 / 5.6 = 13.57143
```

Under H_0, this statistic is distributed as a central $\mathcal{F}$ random variable with $4-1=3$ and $24-4=20$ degrees of freedom. It can also be noted that this a GLRT, since we utilize all of the available information in the design matrix to test the hypothesis.

```
## P-Value:  4.658471e-05  - Reject
```

```
# check
coag_mod <- lm(time ~ 0 + . - seq, coag)
anova(coag_mod)                          # My SSR is off. Jn incorrect?
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       4  98532 24633.0  4398.8 < 2.2e-16 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under the null hypothesis, we obtain an F-statistic of 13.57 **(I think this is wrong)** with a p-value of 0.000047 **(This would also be wrong)**. According to these values, the experimental data provides evidence that *at least* one of the diets has an effect on coagulation time, so we reject the null hypothesis that none of the diets has an effect on blood coagulation time. Further testing would be needed to see which diet(s) have a *statistically* significant effect.

---

d. Is there a difference in coagulation times between diets B and D? Conduct your test at level $\alpha = 0.05$. In your response, provide the null hypothesis, alternative hypothesis, test statistic, the reference distribution of the test statistic under the null, and a brief conclusion.

Answer: Here, we can use the contrast $\boldsymbol{g}^T\boldsymbol{\beta} = 0$ with $\boldsymbol{g}^T = \begin{bmatrix} 0 & 0 & 1 & 0 & -1 \end{bmatrix}$ to test the null hypothesis $H_0 : \alpha_2 = \alpha_4$ against the alternative $H_1 : \alpha_2 \neq \alpha_4$. The unique BLUE estimate would be computed as

$$\boldsymbol{g}^T\hat{\boldsymbol{\beta}} = \widehat{\alpha_2 - \alpha_4} = 66 - 61 = 5.$$

This was found using the LSE $\hat{\boldsymbol{\beta}}^T = \begin{bmatrix} 0 & 61 & 66 & 68 & 61 \end{bmatrix}$, although it does not matter which LSE we use, since all will produce the same BLUE estimate for $\alpha_2 - \alpha_4$, according to the Gauss-Markov Theorem.

```
bbh <- c(0, tapply(coag$time, coag$diet, mean))  # LSE (0, group means)
g <- c(0, 0, 1, 0, -1)                           # contrast
n2 <- sum(coag$diet == "B")
n4 <- sum(coag$diet == "D")
```

```
## F-test statistic: 85.71429 / 5.6 = 15.30612
```

```
## P-Value: 0.0008635834 - Reject
```

We get an F-statistic of 15.306 and under the null hypothesis, $H_0 : \alpha_2 = \alpha_4$, we get a p-value of 0.00086, which is suggests that there is a difference in average blood coagulation times between animals on diet B and animals on diet A. Therefore, our decision would be to reject the null hypothesis.

---

e. Find the contrast $\boldsymbol{g}^\mathsf{T}\boldsymbol{\beta}$ that results in the largest possible F statistic for $H_0 : \boldsymbol{g}^\mathsf{T}\boldsymbol{\beta} = 0$. Report the contrast ($\boldsymbol{g}^\mathsf{T}\boldsymbol{\beta}$), a corresponding $\boldsymbol{d}$, and the value of its F statistic.

Answer: Since the denominator does not involve any contrasts, we should focus our attention on maximizing the numerator of the F test statistic. If I were to guess, we should take the derivative of the numerator with respect to $\boldsymbol{g}$ and set this equal to zero. We then solve for the $\hat{\boldsymbol{g}}$ that satisfies this equation and check that it is a maximum. Since the numerator is a quadratic form, $\hat{\boldsymbol{g}}$ will be a maximum if the leading coefficient of the numerator is negative.

Problem 2

(Adapted from Casella, 2008, *Statistical Design*) The data in `ivd.csv` on Canvas contain measurements of the in vitro digestibility (IVD) of alfalfa grown at different temperatures. The variable `temp` has four levels: 17, 22, 27, and 32 degrees Celsius. Each level has four randomly-assigned replicates.

a. Write the one-way ANOVA model that can be used to analyze this data, with treatment levels $i = 1, 2, 3, 4$ corresponding to temperatures 17, 22, 27, 32.

Answer:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}, \quad E[\boldsymbol{\epsilon}] = \mathbf{0}, \quad \sigma^2[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I},$$

or, in matrix terms,

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \boldsymbol{Y}_3 \\ \boldsymbol{Y}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_4 & \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{0}_4 \\ \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 \\ \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{1}_4 & \mathbf{0}_4 \\ \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{1}_4 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \boldsymbol{\epsilon}_{ij}$$

---

b. Provide the ANOVA table and for the overall F-test give the following: the null hypothesis, test statistic, the distribution of the test statistic under the null, and a brief conclusion.

```
N <- nrow(ivd)                                          # Total number of observations
t <- length(unique(ivd$temp))                           # Number of treatment levels
Y <- ivd$ivd                                            # Response vector
X <- ivd %>%                                            # ANOVA design matrix
  mutate(T17 = temp == 17, T22 = temp == 22
         , T27 = temp == 27, T32 = temp == 32
         ) %>%
  mutate_all(function(x) as.numeric(x)) %>%
  select(-c(ivd, temp)) %>%
  as.matrix()

Px <- X %*% solve(t(X)%*%X) %*% t(X)
Jn <- Px
Jn[Px != 0] = 1/N

F_num3 <- (t(Y) %*% (Px - Jn) %*% Y)/(t-1)
F_den3 <- (t(Y) %*% (diag(1, nrow = N) - Px) %*% Y)/(N-t)
F_stat3 <- F_num3/F_den3
cat("F-test Statisic:", F_stat3)
```

```
## F-test Statisic: 238515.9
```

```
# check
mod_ivd <- lm(ivd ~ 0 + as.factor(temp), ivd)
anova(mod_ivd)
```

```
## Analysis of Variance Table
##
## Response: ivd
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(temp)   4 144501   36125  238516 < 2.2e-16 ***
## Residuals        12      2       0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

---

c. Give an interpretation of what testing the following contrasts would mean (assume $H_0 : \boldsymbol{g}^\mathsf{T}\boldsymbol{\beta} = 0$):

- $\boldsymbol{g}_1^\mathsf{T}\boldsymbol{\beta} = \begin{bmatrix} 0 & -3 & 1 & 1 & 1 \end{bmatrix}\boldsymbol{\beta}$
- $\boldsymbol{g}_2^\mathsf{T}\boldsymbol{\beta} = \begin{bmatrix} 0 & -3 & -1 & 1 & 3 \end{bmatrix}\boldsymbol{\beta}$

Answer: The first contrast tests if the effect on IVD at the higher 3 temperatures is the same as the effect at 17 degrees on IVD.

The second contrast tests whether the difference in effect on IVD between 22 degrees and 27 degrees is equal to three times the difference in effect on IVD between 32 degrees and 17 degrees.

---

d. Assuming the model from (a), what are the distributions of $\boldsymbol{g}^\mathsf{T}\hat{\boldsymbol{\beta}}$ for the two $\boldsymbol{g}^\mathsf{T}$ in (c)?

Answer:

Problem 3

Come up with two possible studies, **one a designed experiment and one an observational study**, that could be conducted related to the graduate student experience. The study topic (pedagogical, biomedical, socioeconomic, political, etc.) is up to you. You don't need to worry about cost, but the studies should be feasible and ethical. For each study:

1. Describe the study in no more than a few sentences
2. Is the study observational or a designed experiment?
3. What are the treatments?
4. What are the EUs and OUs?
5. For the designed experiment, is blocking necessary?
6. Are there other important factors? What are their levels?

Answer:

Designed Experiment

Recruit 100 graduate students and assign them to groups with different levels of mandatory mental health maintenance. Group 1 is allowed to maintain their mental health how they see fit, group 2 must attend a weekly 1-hour counseling session, group 3 must attend a weekly 1-hour yoga session, and group 4 must meditate 1 hour per week. Students who currently report none of the current treatments as part of their regular weekly routine and have no plans to incorporate them are eligible for the study. The study is conducted in the first 8 weeks of the spring semester and at the end, each student receives a score for their stress level determined by answering a questionnaire. This is a designed experiment, since the treatment levels are determined by the experimenters and the students are randomly assigned to be in one of the treatment groups. Blocking may be necessary, if we believe year in program or study discipline may have some effect on stress level. For this study, there are no other factors being considered.

Observational Experiment

Research Question: Do PhD students who have been in the program longer spend less time on campus? Asked 50 students in each of years 1-4 how many hours they spend on campus per week. This is an observational experiment, since we do not assign students to groups, but observe students that can be divided into groups. The "treatment" here is year in PhD program, with 4 levels: 1st, 2nd, 3rd, and 4th. The experimental units are students, the observational units are students. We might suspect that the program or discipline (art, STEM, Humanities, etc.) might have an effect on time spent on campus as well, so we would want to block in order to better isolate the effect of year on the time spent on campus. I don't think there are other important factors here.
