

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:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_4 & \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{0}_4 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 \\ \mathbf{1}_8 & \mathbf{0}_8 & \mathbf{0}_8 & \mathbf{0}_8 & \mathbf{1}_8 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \epsilon$$

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:

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

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

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

```
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 ~. - seq, coag)
anova(coag_mod)
```

```
## Analysis of Variance Table
##
## Response: time
##          Df Sum Sq Mean Sq F value    Pr(>F)
## diet      3     228    76.0   13.571 4.658e-05 ***
## 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 with a p -value of 0.000047. 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 $\mathbf{g}^T \boldsymbol{\beta} = 0$ with $\mathbf{g}^T = [0 \ 0 \ 1 \ 0 \ -1]$ 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

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

This was found using the LSE $\hat{\boldsymbol{\beta}}^T = [0 \ 61 \ 66 \ 68 \ 61]$, 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_num2 <- (g %*% bbh)^2 / (1/n2 + 1/n4)
F_den2 <- F_den # Same denominator as before
F_stat2 <- F_num2/F_den2
```

```
## 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, $H_0 : \alpha_2 = \alpha_4$, $F \sim \mathcal{F}(1, 20, \theta = 0)$ and we get a p -value of 0.00086. This 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 $\mathbf{g}^T \boldsymbol{\beta}$ that results in the largest possible F statistic for $H_0 : \mathbf{g}^T \boldsymbol{\beta} = 0$. Report the contrast $(\mathbf{g}^T \boldsymbol{\beta})$, a corresponding \mathbf{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 \mathbf{g} and set this equal to zero. We then solve for the $\hat{\mathbf{g}}$ that satisfies this equation and check that it is a maximum. Since the numerator is a quadratic form, $\hat{\mathbf{g}}$ 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:

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

or, in matrix terms,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_4 & \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{0}_4 \\ \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 \\ \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{1}_4 & \mathbf{0}_4 \\ \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{1}_4 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \epsilon_{ij}$$

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.

Answer: For an overall F -test, we assess the null hypothesis that there is no difference in average IVD of alpha grown at the 4 different temperature levels, or $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$, with the alternative hypothesis being that at least one of the tested temperature levels has alfalfa with a different average IVD than another temperature level. The F -test statistic is computed as

$$\frac{\mathbf{Y}^T(\mathbf{P}_X - \mathbf{J}_N)\mathbf{Y}/3}{\mathbf{Y}^T(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}/12} = 12.89 \quad (\text{See code below})$$

and under the null hypothesis, is distributed as a central \mathcal{F} random variable with $4 - 1 = 3$ and $16 - 4 = 12$ degrees of freedom.

```
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 <- matrix(1, nrow = N, ncol = N)/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

## F-test Statistic: 1.952292 / 0.1514583 = 12.88996

## P-Value: 0.0004618518 - Reject

## Analysis of Variance Table
##
## Response: ivd
##              Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(temp)  3  5.8569  1.95229    12.89 0.0004619 ***
## Residuals      12  1.8175  0.15146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

With an F -test statistic of 12.89 and a p -value of 0.00046 under H_0 , we reject the null hypothesis and say that the data provides evidence that the average IVD of alfalfa grown in at least one temperature differs from the average IVD of alfalfa at another temperature. Further testing would be required to identify the groups that differ. A summary of the results can be found below in the ANOVA table.

Source	d.f.	SS	MS	F-statistic	P-value
Treatment	3	5.856875	1.9522917	12.88996	0.0004619
Residuals	12	1.817500	0.1514583	NA	NA
Total	15	7.674375	NA	NA	NA

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

- $\mathbf{g}_1^T \boldsymbol{\beta} = [0 \quad -3 \quad 1 \quad 1 \quad 1] \boldsymbol{\beta}$
 - $\mathbf{g}_2^T \boldsymbol{\beta} = [0 \quad -3 \quad -1 \quad 1 \quad 3] \boldsymbol{\beta}$
-

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 there is a linear trend in the group effect of temperature on IVD. This is valid because the treatment levels are numerical and evenly spaced, and the design is balanced.

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

Answer: Assuming that the errors are iid $\sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, $\mathbf{g}^T \hat{\boldsymbol{\beta}}$ is also normally distributed, with mean $\mathbf{g}^T \boldsymbol{\beta}$ and variance $\sigma^2 \mathbf{g}^T (\mathbf{X}^T \mathbf{X})^{-} \mathbf{g}$.

Using the generalized inverse

$$(\mathbf{X}^T \mathbf{X})^{-} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} \end{bmatrix},$$

for the first contrast $[0 \quad -3 \quad 1 \quad 1 \quad 1] \boldsymbol{\beta}$,

$$\mathbf{g}^T \hat{\boldsymbol{\beta}} \sim N(\alpha_2 + \alpha_3 + \alpha_4 - 3\alpha_1, 3\sigma^2),$$

and for the second contrast $[0 \quad -3 \quad -1 \quad 1 \quad 3] \boldsymbol{\beta}$,

$$\mathbf{g}^T \hat{\boldsymbol{\beta}} \sim N(-3\alpha_1 - \alpha_2 + \alpha_3 + 3\alpha_4, 5\sigma^2).$$

```
Xinv <- rbind(0, cbind(0, solve(t(X) %*% X)))
g <- c(0, -3, 1, 1, 1)

g %*% Xinv %*% g
```

```
##      [,1]
## [1,]    3
```

```
g2 <- c(0, -3, -1, 1, 3)

g2 %*% Xinv %*% g2
```

```
##      [,1]
## [1,]    5
```

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

Description: An experiment is conducted at CSU to determine whether regular mental health maintenance strategies have any impact on the stress level of graduate students in PhD programs. Students will be assigned to one of four treatment groups over the course of 8 weeks during the semester and asked to complete a questionnaire to gauge their stress level at the end.

Treatment: The factor that will be considered is mental health maintenance strategy. For this experiment, there are 4 levels: “Control”, where students are not given a specific routine to follow, “Yoga,” where students are required to attend a 1-hour yoga class each week, “Counseling,” where students are required to attend a 1-hour group counseling session each week, and “Mediation,” where students are required to meditate for a total of one hour per week.

Experimental/Observational Units: The experimental and observational units are the students.

Blocking: Blocking will be required, since we expect that program and year in program will also have an effect on the student stress level. However, we are not interested in measuring these effects and would simply like to control for these factors.

Other Factors: I do not think there are other important factors that are being considered.

Observational Study

Description: An observational study will be conducted to see how long, on average, PhD students of different years spend on campus.

Treatment: The factor in question is the number of years the student has been in their PhD program. We question students in any of the first 4 years of their study how many hours they estimate that they spend on campus per week.

Experimental/Observational Units: The students are the experimental and observational units for this study

Blocking: Blocking by program is required because we believe that different programs have different requirements that will influence the amount of time spent on campus by students.

Other Factors: I don't think there are other factors to be considered.
