STAT 640: Homework 8

Due **Wednesday, March 30, 11:59pm MT** on the course Canvas webpage. Please follow the homework guidelines on the syllabus.

Name: Hannah Butler

Problem 1

Researchers wish to study the effectiveness of different building materials for pig shelters. They obtain data on 9 pig farms, three that have shelters of straw, three that have shelters of wood, and three that have shelters of brick. For each farm, they have a measure of wolf attack severity, ranging from 0 (no attacks) to 10 (all pigs killed by wolves). The observed data, in the order of `("straw","straw","straw","wood","wood","wood","brick","brick","brick")`, are:

```
Y <- cbind(9:1)
```

Define the categorical indicators

```
straw <- rep(c(1, 0), times=c(3, 6))
wood  <- rep(c(0, 1, 0), each=3)
brick <- c(rep(0, 6), rep(1, 3))
```

a. Fit a linear model with no intercept and the predictors `straw`, `wood`, and `brick`–in that order (i.e., `lm(Y ~ 0 + straw + wood + brick)`). Construct the sum of squares (SS) and mean squares (MS) for a Sequential (Type I) ANOVA table using only basic matrix operations in R. Verify your numbers against `anova()`.

Answer:

```
# design matrix
X <- cbind(straw, wood, brick)
# parameter estimates
bh <- solve( t(X)%*%X ) %*% (t(X)%*%Y); t(bh)
```

```
##      straw wood brick
## [1,]     8    5     2
```

```
RSS <- t(Y - X%*%bh) %*% (Y - X%*%bh); RSS
```

```
##      [,1]
## [1,]    6
```

```
# Type 1 SS: in order
P_0 <- diag(0, 9)
# straw | nothing else
P_x1 <- straw %*% solve(t(straw) %*% straw) %*% straw
SS_X1_0 <- t(Y) %*% (P_x1 - P_0) %*% Y

# wood | straw
X2 <- cbind(straw, wood)
P_x2 <- X2 %*% solve(t(X2)%*%X2) %*% t(X2)
SS_X2_X1 <- t(Y) %*% (P_x2 - P_x1) %*% Y
```

```r
# brick | straw, wood
P_x3 <- X %*% solve( t(X)%*%X ) %*% t(X)
SS_X3_X2 <- t(Y) %*% (P_x3 - P_x2) %*% Y
```

Because each group accounts for 1 degree of freedom, the MSE values will be the same: .

Source	Estimate	Degrees of Freedom	Sum of Squares (I)	Mean Squares
Straw	8	1	192	192
Wood	5	1	75	75
Brick	2	1	12	12
Residuals	NA	6	6	1

```r
#check
pig_df <- data.frame(Y, X)
pig_lm <- lm(Y ~ 0 + straw + wood + brick
             , pig_df
             )
anova(pig_lm)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## straw      1    192     192     192 8.796e-06 ***
## wood       1     75      75      75 0.0001307 ***
## brick      1     12      12      12 0.0134000 *
## Residuals  6      6       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b. Repeat (a) but now *with* an intercept and a different ordering: `lm(Y ~ wood + brick + straw)`. Construct the sum of squares (SS) and mean squares (MS) for a Sequential (Type I) ANOVA table using only basic matrix operations in R. Verify your numbers against `anova()`.

Answer: Since the design matrix $\boldsymbol{X}_{int}$ will no longer be one of full rank, we must use a generalized inverse to compute $\hat{\boldsymbol{\beta}}$. This can be found by partitioning $\boldsymbol{X}_{int} = \begin{bmatrix}\mathbf{1} & \boldsymbol{X}\end{bmatrix}$ and finding the generalized inverse as $\boldsymbol{G} = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{bmatrix}$

```
# design matrix with intercept
X_int <- cbind(intercept = 1, wood, brick, straw)
# inverse of rank = 3 block
XX_inv <- solve(t(X_int[,1:3])%*%X_int[,1:3])
# Generalized inverse
G <- bdiag(XX_inv, 0) %>%
  as.matrix()
bh_int <- G %*% (t(X_int)%*%Y)

# Check Generalized Inverse
#(t(X_int) %*% X_int) %*% G %*% (t(X_int) %*% X_int)

# Just the intercept
P_x0 <- X_int[,1] %*% solve(t(X_int[,1])%*%X_int[,1]) %*% t(X_int[,1])
# Wood | Intercept
X1 <- X_int[,1:2]
P_x1 <- X1 %*% solve(t(X1) %*% X1) %*% t(X1)
SS_X1_int <- t(Y) %*% (P_x1 - P_x0) %*% Y
# Brick | Intercept + Wood
X2 <- X_int[,1:3]
P_x2 <- X2 %*% solve(t(X2)%*%X2) %*% t(X2)
SS_X2_X1 <- t(Y) %*% (P_x2 - P_x1) %*% Y
# Straw | Intercept + Wood + Brick
P_xfull <- X_int %*% G %*% t(X_int)
SS_Xfull_X1X2 <- t(Y) %*% (P_xfull - P_x2) %*% Y
# Residual Sum of Squares
RSS <- t(Y) %*% (diag(1, 9) - P_xfull) %*% Y
```

Source	Estimate	Sum of Squares (I)	Mean Squares
Wood	-3	0	0
Brick	-6	54	54
Straw	0	0	0
Residuals	NA	6	1

```
#check
pig_lm_2 <- lm(Y~ wood + brick + straw, pig_df)
anova(pig_lm_2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## wood       1      0       0       0 1.000000
## brick      1     54      54      54 0.000325 ***
## Residuals  6      6       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

---

c. The ANOVA table in (b) looks strange. Explain why this happens.

Answer: I'm not entirely sure. However, the intercept term represents the average wolf attack severity when the pig shelter is none of the three tested. However, we did not have any observations which did not fall into one of the three shelter categories, so when considering the variables sequentially, the sum of squares for wood given the intercept

---

Problem 2

Consider the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I})$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix}^\mathsf{T}$ and design matrix:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Letting $\boldsymbol{J}_n = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ and $\mathbf{0}_{j,k}$ be a $j \times k$ matrix of zeros, suppose we have the following projection matrices:

$$\boldsymbol{P}_1 = \begin{bmatrix} \boldsymbol{J}_3 & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \end{bmatrix} \qquad \boldsymbol{P}_2 = \begin{bmatrix} \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \boldsymbol{J}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \end{bmatrix} \qquad \boldsymbol{P}_3 = \begin{bmatrix} \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \boldsymbol{J}_3 \end{bmatrix}$$

$$\boldsymbol{P}_4 = \begin{bmatrix} \boldsymbol{J}_3 & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \boldsymbol{J}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \end{bmatrix} \qquad \boldsymbol{P}_5 = \begin{bmatrix} \boldsymbol{J}_3 & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \boldsymbol{J}_3 \end{bmatrix} \qquad \boldsymbol{P}_6 = \begin{bmatrix} \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \boldsymbol{J}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \boldsymbol{J}_3 \end{bmatrix}$$

$$\boldsymbol{P}_7 = \begin{bmatrix} \boldsymbol{J}_3 & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \boldsymbol{J}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \boldsymbol{J}_3 \end{bmatrix} \qquad \boldsymbol{P}_8 = \begin{bmatrix} \mathbf{0}_{3,3} & \mathbf{0}_{3,6} \\ \mathbf{0}_{6,3} & \boldsymbol{J}_6 \end{bmatrix}$$

For each of the following quantities ((a) through (f)), provide the following information:

1. Does this represent a valid F-statistic for an F-test? If yes, answer remaining questions; if no, explain why not and then skip remaining questions.
2. What are the null and alternative hypotheses being tested?
3. Does this F-statistic correspond to a GLRT? If no, why not?
4. Would this F-statistic appear in a Type I ANOVA Table?
5. Would this F-statistic appear in a Type III ANOVA Table?

a. $\dfrac{\boldsymbol{Y}^\mathsf{T}(\boldsymbol{P}_4 - \boldsymbol{P}_1)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_4 - \boldsymbol{P}_1)}{\boldsymbol{Y}^\mathsf{T}(\boldsymbol{I} - \boldsymbol{P}_4)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_4)}$

b. $\dfrac{\boldsymbol{Y}^\top(\boldsymbol{P}_4 - \boldsymbol{P}_1)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_4 - \boldsymbol{P}_1)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_7)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_7)}$

c. $\dfrac{\boldsymbol{Y}^\top(\boldsymbol{P}_7 - \boldsymbol{P}_5)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_7 - \boldsymbol{P}_5)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_7)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_7)}$

d. $\dfrac{\boldsymbol{Y}^\top(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_3 - \boldsymbol{P}_2)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_6)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_6)}$

e. $\dfrac{\boldsymbol{Y}^\top(\boldsymbol{P}_6 - \boldsymbol{P}_8)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_6 - \boldsymbol{P}_8)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_7)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_7)}$

f. $\dfrac{\boldsymbol{Y}^\top(\boldsymbol{P}_7 - \boldsymbol{P}_8)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_7 - \boldsymbol{P}_8)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_7)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_7)}$

Answers:

$$\frac{\boldsymbol{Y}^\top(\boldsymbol{P}_4 - \boldsymbol{P}_1)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_4 - \boldsymbol{P}_1)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_4)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_4)}$$

1. Yes this represents a valid F test
2. $H_0 : \beta_2 = 0$, $H_a : \beta_2 \neq 0$
3. Yes, there is a difference of one degree of freedom between the models being compared
4. Yes
5. Not sure, look up

$$\frac{\boldsymbol{Y}^\top(\boldsymbol{P}_4 - \boldsymbol{P}_1)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_4 - \boldsymbol{P}_1)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_7)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_7)}$$

1. No, I don't think so

$$\frac{\boldsymbol{Y}^\top(\boldsymbol{P}_7 - \boldsymbol{P}_5)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_7 - \boldsymbol{P}_5)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_7)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_7)}$$

1. Yes, this represents a valid F test
2. $H_0 : \beta_2 = 0$, $H_a : \beta_2 \neq 0$
3. Yes, I think so
4. No. In a type I ANOVA table, and this design matrix, we would only see quantities for $X_2 \mid X_1$, but no $X_2 \mid X_1, X_3$.
5. Maybe, I'm not sure so I need to look it up

$$\frac{\boldsymbol{Y}^\top(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_3 - \boldsymbol{P}_2)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_6)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_6)}$$

1. No, these models are not nested

$$\frac{\boldsymbol{Y}^\top(\boldsymbol{P}_6 - \boldsymbol{P}_8)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_6 - \boldsymbol{P}_8)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_7)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_7)}$$

1. No

$$\frac{\boldsymbol{Y}^\top(\boldsymbol{P}_7 - \boldsymbol{P}_8)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{P}_7 - \boldsymbol{P}_8)}{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{P}_7)\boldsymbol{Y}/\mathsf{rank}(\boldsymbol{I} - \boldsymbol{P}_7)}$$

1. No?

Problem 3

From Montgomery, (1997): An experiment is conducted to assess the effect of cotton content (percent) on tensile strength of men's shirts. Five levels of cotton percentage are considered, with five shirts tested for strength at each level. The results are included in the following data:

```
strength <- c(7, 7, 15, 11, 9, 12, 17, 12, 18, 18, 14, 18, 18, 19, 19, 19, 25,
              22, 19, 23, 7, 10, 11, 15, 11)
cotton <- sort(rep(c(15, 20, 25, 30, 35), 5))
```

Using basic matrix operations in `R`, conduct a complete model utility test, which is the GLRT of the full model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, 5$ versus the intercept-only reduced model, $Y_{ij} = \mu + \epsilon_{ij}$. Your answer should reproduce each value in:

```
fit<- lm(strength ~ factor(cotton))
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: strength
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## factor(cotton)  4 475.76  118.94  14.757 9.128e-06 ***
## Residuals      20 161.20    8.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

```
# Design Matrix
X_c <- cbind(intercept = 1
             , c15 = rep(c(1,0), times = c(5, 20))
             , c20 = rep(c(0,1,0), times = c(5, 5, 15))
             , c25 = rep(c(0,1,0), times = c(10, 5, 10))
             , c30 = rep(c(0,1,0), times = c(15, 5, 5))
             , c35 = rep(c(0,1), times = c(20, 5))
             )

# Projection Matrix
H <- bdiag(solve( t(X_c[,-6])%*%X_c[,-6] ), 0) %>%
  as.matrix()
P_int <- X_c[,1] %*% solve( t(X_c[,1])%*%X_c[,1]) %*% t(X_c[,1])
P_X <- X_c %*% H %*% t(X_c)
RSS <- t(strength) %*% (diag(1, 25) - P_X) %*% strength
SS <- t(strength) %*% (P_X - P_int) %*% strength
```
