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
```

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
# 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

# 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: .

```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 $\boldsymbol{G} = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{bmatrix}$

```r
# design matrix with intercept
X_int <- cbind(1, X)
# inverse of full rank block
XX_inv <- solve(t(X)%*%X)
# Generalized inverse
G <- diag(c(0, diag(XX_inv)))
bh_int <- G %*% (t(X_int)%*%Y)

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

P_x0 <- X[,1] %*% solve(t(X[,1])%*%X[,1]) %*% t(X[,1])
X1 <- X_int[,1:2]
X2 <- X_int[,1:3]
P_x1 <- X1 %*% solve(t(X1) %*% X1) %*% t(X1)
P_x2 <- X2 %*% solve(t(X2)%*%X2) %*% t(X2)
P_xfull <- X_int %*% G %*% t(X_int)
SS_X1_int <- t(Y) %*% (P_x1 - P_x0) %*% Y
SS_X2_X1 <- t(Y) %*% (P_x2 - P_x1) %*% Y
SS_Xfull_X1X2 <- t(Y) %*% (P_xfull - P_x2) %*% Y

#check
pig_lm_2 <- lm(Y~.
               , pig_df
               )
anova(pig_lm_2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## straw      1   40.5    40.5    40.5 0.0007066 ***
## wood       1   13.5    13.5    13.5 0.0104017 *
## Residuals  6    6.0     1.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(pig_lm_2)
```

```
##
## Call:
## lm(formula = Y ~ ., data = pig_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0000     0.5774   3.464 0.013400 *
## straw         6.0000     0.8165   7.348 0.000325 ***
## wood          3.0000     0.8165   3.674 0.010402 *
```

```
## brick              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 6 degrees of freedom
## Multiple R-squared:    0.9,  Adjusted R-squared:  0.8667
## F-statistic:    27 on 2 and 6 DF,  p-value: 0.001
```

---

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

Answer:

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

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

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

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

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

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

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

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

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

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

Answers:

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:

```r
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:

```r
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:
