

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

```

# 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: .

```

#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    8.796e-06 ***
## wood       1     75       75    0.0001307 ***
## brick      1     12       12    0.0134000 *
## Residuals  6      6        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Source	Estimate	Sum of Squares (I)	Mean Squares
Wood	-3	0	0
Brick	-6	54	54
Straw	0	0	0
Residuals	NA	6	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 \mathbf{X}_{int} will no longer be one of full rank, we must use a generalized inverse to compute $\hat{\beta}$. This can be found by partitioning $\mathbf{X}_{int} = \begin{bmatrix} \mathbf{1} & \mathbf{X} \end{bmatrix}$ and finding the generalized inverse as

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

```
# 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
```

```
#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 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$, $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \beta_3]^\top$ and design matrix:

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

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

$$\begin{aligned} \mathbf{P}_1 &= \begin{bmatrix} \mathbf{J}_3 & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \end{bmatrix} & \mathbf{P}_2 &= \begin{bmatrix} \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{J}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \end{bmatrix} & \mathbf{P}_3 &= \begin{bmatrix} \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{J}_3 \end{bmatrix} \\ \mathbf{P}_4 &= \begin{bmatrix} \mathbf{J}_3 & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{J}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \end{bmatrix} & \mathbf{P}_5 &= \begin{bmatrix} \mathbf{J}_3 & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{J}_3 \end{bmatrix} & \mathbf{P}_6 &= \begin{bmatrix} \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{J}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{J}_3 \end{bmatrix} \\ \mathbf{P}_7 &= \begin{bmatrix} \mathbf{J}_3 & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{J}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{J}_3 \end{bmatrix} & \mathbf{P}_8 &= \begin{bmatrix} \mathbf{0}_{3,3} & \mathbf{0}_{3,6} \\ \mathbf{0}_{6,3} & \mathbf{J}_6 \end{bmatrix} \end{aligned}$$

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. $\frac{\mathbf{Y}^\top(\mathbf{P}_4 - \mathbf{P}_1)\mathbf{Y}/\text{rank}(\mathbf{P}_4 - \mathbf{P}_1)}{\mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_4)\mathbf{Y}/\text{rank}(\mathbf{I} - \mathbf{P}_4)}$
- b. $\frac{\mathbf{Y}^\top(\mathbf{P}_4 - \mathbf{P}_1)\mathbf{Y}/\text{rank}(\mathbf{P}_4 - \mathbf{P}_1)}{\mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_7)\mathbf{Y}/\text{rank}(\mathbf{I} - \mathbf{P}_7)}$
- c. $\frac{\mathbf{Y}^\top(\mathbf{P}_7 - \mathbf{P}_5)\mathbf{Y}/\text{rank}(\mathbf{P}_7 - \mathbf{P}_5)}{\mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_7)\mathbf{Y}/\text{rank}(\mathbf{I} - \mathbf{P}_7)}$
- d. $\frac{\mathbf{Y}^\top(\mathbf{P}_3 - \mathbf{P}_2)\mathbf{Y}/\text{rank}(\mathbf{P}_3 - \mathbf{P}_2)}{\mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_6)\mathbf{Y}/\text{rank}(\mathbf{I} - \mathbf{P}_6)}$
- e. $\frac{\mathbf{Y}^\top(\mathbf{P}_6 - \mathbf{P}_8)\mathbf{Y}/\text{rank}(\mathbf{P}_6 - \mathbf{P}_8)}{\mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_7)\mathbf{Y}/\text{rank}(\mathbf{I} - \mathbf{P}_7)}$
- f. $\frac{\mathbf{Y}^\top(\mathbf{P}_7 - \mathbf{P}_8)\mathbf{Y}/\text{rank}(\mathbf{P}_7 - \mathbf{P}_8)}{\mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_7)\mathbf{Y}/\text{rank}(\mathbf{I} - \mathbf{P}_7)}$

Answers: To have a valid F statistic, the three conditions below must be met:

1. The numerator must be distributed as a χ^2 random vector
2. The denominator must be distributed as a *central* χ^2 random vector
3. The numerator must be independent of the denominator.

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

1. This is not a valid F statistic, because the denominator is not a central χ^2 random variable. This has to do with \mathbf{P}_4 projecting onto a subspace of $\mathcal{C}(\mathbf{X})$, rather than $\mathcal{C}(\mathbf{X})$ itself.

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

1. This is a valid F statistic, according to the three conditions listed above. Since $(\mathbf{P}_4 - \mathbf{P}_1)$ is a projection matrix, we can use Proposition 3.18 to confirm that it is distributed as a χ^2 random vector. Since \mathbf{P}_7 projects onto the full column space of \mathbf{X} , the centrality parameter for $\frac{1}{\sigma^2}\mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_7)\mathbf{Y}$ will be zero. Finally, because \mathbf{P}_4 projects onto a subspace of \mathbf{P}_7 , the numerator and denominator will be independent (not sure why??)

2. $H_0 : \beta_2 = 0$, $H_a : \beta_2 \neq 0$.

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

1. Using the same justifications as we did for b. above, we can see that this is a valid F statistic.
2. $H_0 : \beta_2 = 0$, $H_a : \beta_2 \neq 0$.
3. This is a GLRT, since \mathbf{P}_7 is in both the numerator and denominator.

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

1. No, this is not a valid F statistic, since \mathbf{P}_8 does not project onto $\mathcal{C}(\mathbf{X})$. (Although I do think it projects onto a subspace of $\mathcal{C}(\mathbf{X})$, where the first three components are 0 and the last 6 components of the elements are the same).

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

1. Yes, this is a valid F statistic, since we have a central χ^2 random variable in the denominator, \mathbf{P}_8 projects onto a subspace of \mathbf{P}_6 ($\mathbf{P}_6 - \mathbf{P}_8$ is itself a projection matrix), and \mathbf{P}_6 projects onto a subspace of \mathbf{P}_7 (The numerator is independent of the denominator).
2. $H_0 : \beta_2 = \beta_3$, $H_a : \beta_2 \neq \beta_3$
3. This is not a GLRT, since we are not utilizing all of the information available to test the hypothesis.

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

1. This is also a valid F statistic.
2. $\mathbf{H}_0 : \beta_1 = 0$ and $\beta_2 = \beta_3$, $\mathbf{H}_a : \beta_1 \neq 0$ or $\beta_2 \neq \beta_3$.
3. This is a GLRT.

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 $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, \dots, 5$ 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
```
