HW3

Samuel Olson

Outline

- Q1: part f) through i) need re-review
- Q2

Problem 1

Case Study 5.1.1 from *The Statistical Sleuth* describes a dietary restriction study. Female mice were assigned to one of the following six treatment groups:

1. **NP:** unlimited, nonpurified, standard feed
2. **N/N85:** normal diet before weaning and normal diet (85 kcal/week) after weaning
3. **N/R50:** normal diet before weaning and reduced calorie (50 kcal/week) after weaning
4. **R/R50:** reduced calorie diet before and after weaning (50 kcal/week)
5. **N/R50lopro:** normal diet before weaning, reduced calorie (50 kcal/week) after weaning, and reduced protein
6. **N/R40:** normal diet before weaning and severely reduced calorie (40 kcal/week) after weaning

The response of interest was mouse lifetime in months.

Download the corresponding data file at http://www.statisticalsleuth.com/ or access it by installing and loading the R package `Sleuth3` and examining `case0501`. To do that latter, try the following R commands:

```
require(Sleuth3)
```

```
## Loading required package: Sleuth3
```

```
## Warning: package 'Sleuth3' was built under R version 4.4.2
```

```
# case0501
```

Complete the following parts under the assumption that a Gauss-Markov model with normal errors and an unrestricted mean for each of the six treatment groups is appropriate for these data.

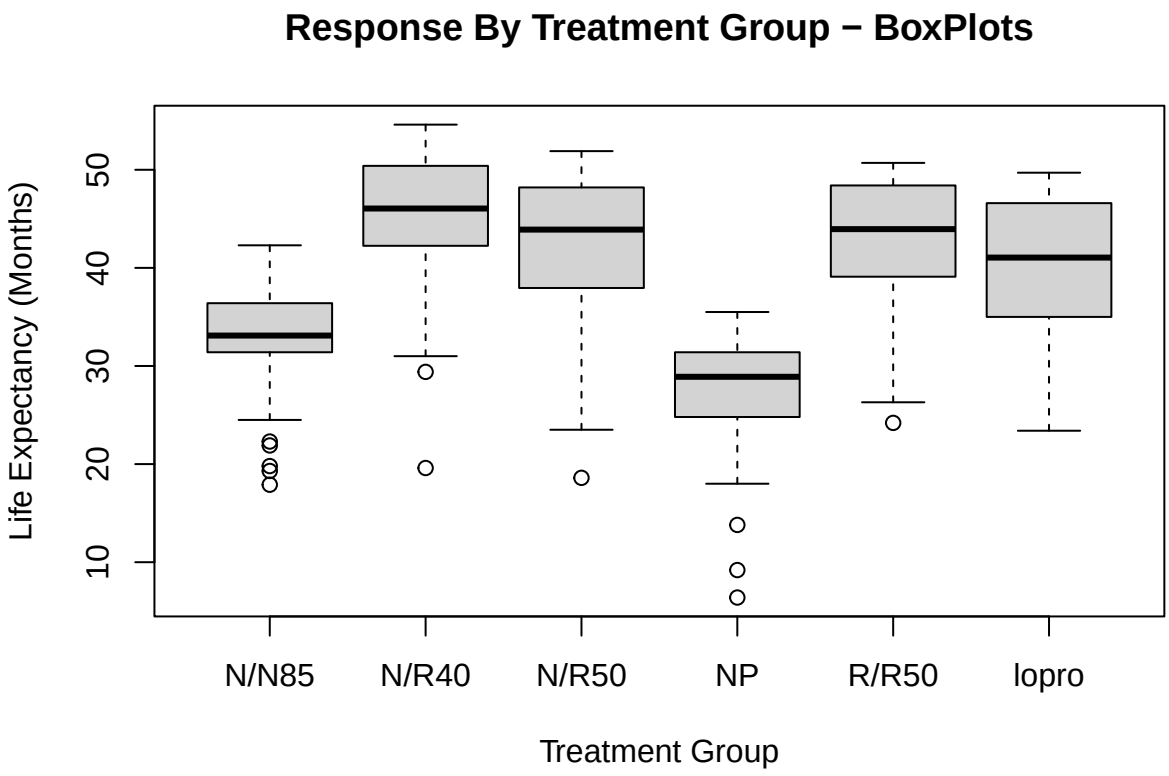
Note:

Doing this problem primarily in R.

a)

Create side-by-side boxplots of the response for this dataset, with one boxplot for each treatment group. Be sure to clearly label the axes of your plot.

```
boxplot(formula = Lifetime ~ Diet,
        data = case0501,
        main = "Response By Treatment Group - BoxPlots",
        xlab="Treatment Group",
        ylab = "Life Expectancy (Months)")
```



b)

Find the SSE (sum of squared errors) for the full model with one unrestricted mean for each of the six treatment groups.

```
lm(formula = Lifetime ~ Diet,
           data = case0501) |>
  deviance()
```

```
## [1] 15297.42
```

c)

Compute $\hat{\sigma}^2$ for the full model.

```
fullModel <- lm(formula = Lifetime ~ Diet,
          data = case0501)

numer <- lm(formula = Lifetime ~ Diet,
          data = case0501) |>
          deviance()
denom <- fullModel$df

denom
```

```
## [1] 343
```

```
numer/denom
```

```
## [1] 44.59888
```

d)

Find the SSE for a reduced model that has one common mean for the N/N85, N/R50, N/R50lopro, and N/R40 treatment groups and unrestricted means for the other two treatment groups.

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Modify
# levels(case0501$Diet)
# "N/N85" "N/R40" "N/R50" "NP"    "R/R50" "lopro"
mergedGroup <- levels(case0501$Diet)[c(1,3,6,2)]

reduced <- case0501 |>
  mutate(
    newDiet = case_when(
      Diet %in% mergedGroup ~ "N/N85+N/R50+N/R50lopro+N/R40",
      # only change mergedGroup matches
      TRUE ~ as.character(Diet)
    )
  ) |>
  mutate(newDiet = factor(newDiet))

redModel <- lm(Lifetime ~ newDiet,
           data = reduced)
deviance(redModel)
```

```
## [1] 20287.99
```

e)

Use the answers from parts (b) through (d) to compute an F-statistic for testing the null hypothesis that the mean of the response vector is in the column space associated with the reduced model vs. the alternative that the mean of the response vector is in the column space of the full model but not in the column space of the reduced model.

Explicitly, we're testing:

$$H_0 : E(\boldsymbol{y}) \in \mathcal{C}(\mathbf{X}_0)$$

$$H_a : E(\boldsymbol{y}) \in \mathcal{C}(\mathbf{X}) \setminus \mathcal{C}(\mathbf{X}_0)$$

Using the answers from the prior parts of the question, noting the difference in degrees of freedom between the full and reduced model is 3:

$$F = \frac{(SSE_{\text{Reduced}} - SSE_{\text{Full}})/(df_{\text{Reduced}} - df_{\text{Full}})}{SSE_{\text{Full}}/df_{\text{Full}}} = \frac{((20287.99 - 15297.42)/3)}{(15297.42/343)} = 37.3$$

Checking directly against the R output comparing the two models:

```
anova(redModel, fullModel)
```

```
## Analysis of Variance Table
##
## Model 1: Lifetime ~ newDiet
## Model 2: Lifetime ~ Diet
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1    346 20288
## 2    343 15297  3    4990.6 37.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

f)

Explain to the scientists conducting this study what the F-statistic in part (e) can be used to test. Consider the context of the study (i.e., pay attention to the description of the experiment and the descriptions of the treatments) and use terms non-statistician scientists will understand.

The partial F-test, comparing the full and reduced model, in part e) is evidence to to test whether the full model is significantly better than the reduced model, which goes hand-in-hand with testing whether there is significant difference among the N/N85, N/R50, N/R50lopro and N/R40 treatment groups, i.e. the groups being merged effectively into a single treatment group (having the same treatment means in life expectancy).

Specifically with regards to study interpretation, the partial F-statistic test from part e) can be used to test whether the expected lifetime (average life expectancy) is affected (or correlated, depending on treatment assignment and study design) by different diets among female mice who are treated with normal diet before weaning. Specifically, the calculated partial F-statistic is 37.3 with p-value near zero ($<$ 2.2e-16). This is evidence to support using the full model in lieu of the reduced model, at the $\alpha = 0.05$ level, and in support of the finding that there is significant difference among the N/N85, N/R50, N/R50lopro and N/R40 treatment group means.

g)

Consider an F-statistic of the form given on slide 20 of slide set 2. Provide the $\boldsymbol{C}$ matrix and $\boldsymbol{d}$ vector and compute the F-statistic corresponding to the test of the hypotheses in part (e).

```
# Touch this up
y <- case0501$Lifetime
I <- diag(1, length(y))
r <- length(levels(case0501$Diet))
xmat <- model.matrix(~0 + case0501$Diet)
proj <- function(x){x %*% MASS::ginv(t(x)%*%x) %*% t(x)}
hat.sig2 <- t(y) %*% (I-proj(xmat)) %*% y / (length(y)-r)
hat.b <- solve(t(xmat)%*%xmat) %*% t(xmat) %*% y
C <- matrix(c(1, -1, 0, 0, 0, 0,
              1, 0, -1, 0, 0, 0,
              1, 0, 0, 0, 0, -1),
            byrow = TRUE,
            nrow = 3)
Fstat <- t(C %*% hat.b) %*% solve(C %*% solve(t(xmat)%*%xmat) %*% t(C)) %*% (C %*% hat.b)/3/hat.sig2
Fstat[[1]]
```

```
## [1] 37.29968
```

$$H_0 : C\boldsymbol{\beta} = \mathbf{d}$$

$$H_a : C\boldsymbol{\beta} \neq \mathbf{d}$$

where,

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

according to the order of the treatments in the dataset, $q = 3$ and $\mathbf{d} = \mathbf{0}$.

$$F = \frac{(C\hat{\boldsymbol{\beta}} - \mathbf{d})'(C(\mathbf{X}'\mathbf{X})^{-1}C')^{-1}(C\hat{\boldsymbol{\beta}} - \mathbf{d})/q}{\hat{\sigma}^2}$$

and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad \hat{\sigma}^2 = \frac{\mathbf{y}'(\mathbf{I} - P_{\mathbf{X}})\mathbf{y}}{n - r}$$

From the code and output below, $F = 37.3$, the same result as in part e), and having similar interpretation within the context of the study.

h)

Use R to obtain the p-value associated with the F-statistic in part (g). Provide the interpretation of the p-value. That is, what probability does it reflect? If you are not sure what I am looking for, **pick up any undergraduate Statistics textbook for examples.** (Sick burn)

```
p_value <- 1 - pf(q = Fstat[[1]],
                  df1 = 3,
                  df2 = length(y) - r)
p_value
```

```
## [1] 0
```

Again, similar to part f), the p-value is practically zero, in fact it is rounded to 0 via the method used.

The p-value represents the probability of obtaining an F-statistic as extreme as 37.3 (or more extreme) under the null hypothesis being true, i.e. that contrasts among the diet groups have no effect on lifetime (average life expectancy).

i)

Evaluate the strength of evidence against the null hypothesis based on the p-value found in part (h). Do not use the p-value to make a decision about rejecting or failing to reject the null hypothesis - I am not interested in that. For more background reading, consider the following reference: https://www.amstat.org/asa/files/pdfs/p-valuesstatement.pdf.

The p-value from part (h) is extremely small ($p < 0.001$), indicating extremely strong evidence against $H_0 : C\boldsymbol{\beta} = \mathbf{0}$.

Problem 2

Consider a two-factor experiment with factors A and B. Factor A represents gender and has two levels (male coded as 1/female coded as 2). Factor B reflects a patient's smoking history and has four levels (never coded as 1, light coded as 2, median coded as 3, heavy coded as 4). The data set contains a third variable, `fat`, which we will ignore for this analysis. Let the response variable, `exercise`, denote the patient's achievement score in some exercise routine that can be used as a proxy for cardiovascular fitness. The higher the score, the better the patient's cardiovascular fitness. The data are saved in a text file `stress.txt`. You may use R or SAS to analyze these data, but you have to submit all your code and results. (I will present my solution using SAS.) We will fit a cell-means model to these data estimating a patient's achievement score based on gender and smoking history.

Note:

Doing this problem primarily in SAS.

a)

Set up a contingency table similar to the one on slide 7 of the lecture slides that reflects all possible factor level combinations. Use the parameterization introduced on slide 6 of the same set of slides and specify each cell mean.

b)

Specify the model matrix for this model. (I realize this is a big matrix and I will ask you to do this only once.)

c)

Specify the corresponding $\boldsymbol{\beta}$-vector and obtain its OLS estimate.

d)

Obtain the standard error associated with the OLS estimator of each cell mean. Start by specifying the relevant formula and show your calculations at least once, i.e., for at least one of the cell means.

e)

Specify the parameter representation reflecting the main effect of gender and also its point estimate.

f)

Is there an interaction between gender and smoking? Similarly to the previous parts, specify all relevant parameter representations.

g)

Specify $\boldsymbol{C}$ allowing you to test for a main effect of gender. State the appropriate null- and alternative hypothesis using parameter representation. Obtain the corresponding value of the test statistic, df and p-value and provide a conclusion in the context of the data.

h)

Specify $\boldsymbol{C}$ allowing you to test for a main effect of smoking. State the appropriate null- and alternative hypothesis using parameter representation. Obtain the corresponding value of the test statistic, df and p-value and provide a conclusion in the context of the data.

i)

Specify $\boldsymbol{C}$ allowing you to test for an interaction between gender and smoking. State the appropriate null- and alternative hypothesis using parameter representation. Obtain the value of the relevant test statistic, df and p-value. Provide an interpretation of the result that a scientist unfamiliar with technical statistical terms can understand. Would you argue that the interaction is of practical importance? Briefly explain.

j)

Provide a 95% confidence interval for the mean associated with male patients who never smoked. Show all your work.

k)

Provide a 95% confidence interval for the mean effect of gender. Show all your work.

l)

Obtain the residuals for the fitted models and use them to check the necessary assumptions that allow us to fit the proposed model. Please submit and explain any graphical displays that you might use.