HW3

Samuel Olson

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.

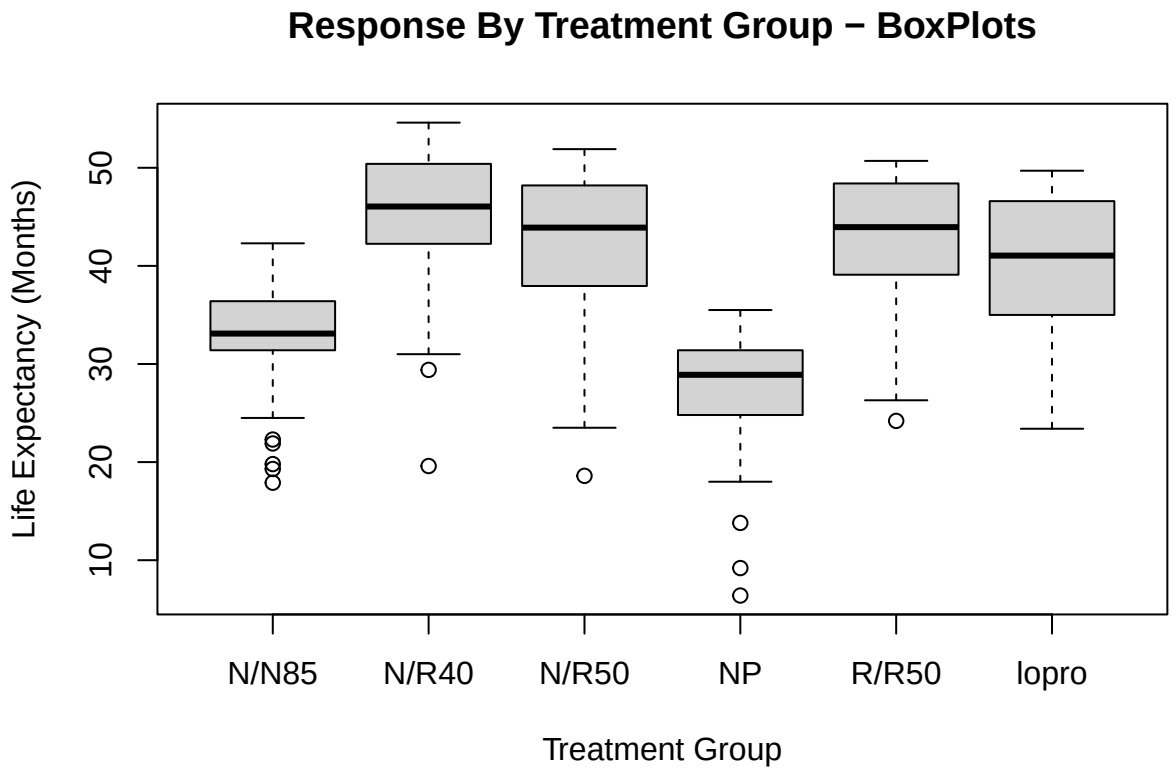
Note:

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.

```r
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

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.

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 (Combining 4 groups into 1 group effectively frees up 3 extra degrees of freedom):

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

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.

Explicitly, we're testing:

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

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

But we should not expect non-statisticians to understand that! I still check the slides to double check. Anyway:

What we're doing is a partial F-test, comparing the full and reduced model, from part e) to determine if there is evidence in support of the full model being significantly better than the reduced model. This is to say we're testing whether it is appropriate to group together the N/N85, N/R50, N/R50lopro, and N/R40 treatment groups. This is to test whether there is significant difference in the average life expectancy within the N/N85, N/R50, N/R50lopro and N/R40 treatment groups. As this value has been calculated, and has been provided, we may say: The calculated partial F-statistic is 37.3 with its corresponding p-value near zero ($<$ 2.2e-16) is overwhelming evidence in support of using the full model in lieu of the reduced model, at the $\alpha = 0.05$ level, such that we have overwhelming evidence that at least, one group mean life expectancy is different among the N/N85, N/R50, N/R50lopro and N/R40 treatment groups, in the linear model that includes NP and R/R50.

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).

Our Hypotheses are:

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

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

where:

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

And:

$\mathbf{d} = \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ (the zero vector).

Again, we have 3 degrees of freedom difference in the full vs. reduced model, meaning we use $q = 3$ in the following equation to calculate our F statistic:

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

where:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

And:

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

Calculating:

```
# 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
```

From the code and output, $F = 37.3$, 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. So we'd say there is overwhelming evidence that the F-statistic we observed, 37.3, is extremely unlikely to have been observed under the null hypothesis being true.

The interpretation in the context of the study is very similar to that provided previously: The p-value indicates overwhelming evidence that at least one group mean life expectancy is different among the N/N85, N/R50, N/R50lopro and N/R40 treatment groups, in the linear model that includes NP and R/R50.

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 the null hypothesis: $H_0 : C\boldsymbol{\beta} = \mathbf{0}$.

Apologies for repetition here, but I believe the study interpretation is warranted from this question: The p-value indicates overwhelming evidence that at least one group mean life expectancy is different among the N/N85, N/R50, N/R50lopro and N/R40 treatment groups, in the linear model that includes NP and R/R50.

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 R, again.

```
stress <- read.table(file = "stress.txt",
                     header = TRUE,
                     sep = ",")
d <- stress
```

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.

```
require(dplyr)
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
aggStress <- stress |>
  group_by(gender, smoking) |>
  summarise(meanScore = mean(Score), .groups = "drop")

stressTable <- aggStress |>
  pivot_wider(names_from = smoking, values_from = meanScore, names_prefix = "Smoking_")

round(stressTable, 2)
```

```
## # A tibble: 2 x 5
##   gender Smoking_1 Smoking_2 Smoking_3 Smoking_4
##    <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1      1      26.0      14.1      19.9      16.0
## 2      2      19.8      12.1      12.1      10.2
```

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.)

2 Genders, 4 levels of smoking, 3 reps of each gender-smoking combination.

This means our model matrix (design matrix) has 8 columns and 24 rows.

This looks like:

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

Weeeeee!

c)

Specify the corresponding β-vector and obtain its OLS estimate.

```
gender <- rep(c("Male", "Female"), each = 12)
smoking <- rep(c("Never", "Light", "Medium", "Heavy"), each = 3, times = 2)

stressDat <- data.frame(gender = factor(gender), smoking = factor(smoking))
X <- model.matrix(~ 0 + gender:smoking, data = stressDat)
X <- X[, order(colnames(X))]

y <- stress$Score

beta_hat <- solve(t(X) %*% X) %*% (t(X) %*% y)

# X[0,]
round(beta_hat,2)
```

```
##                              [,1]
## genderFemale:smokingHeavy  10.20
## genderFemale:smokingLight  12.13
## genderFemale:smokingMedium 16.03
## genderFemale:smokingNever  19.87
## genderMale:smokingHeavy    12.07
## genderMale:smokingLight    19.83
## genderMale:smokingMedium   14.07
## genderMale:smokingNever    25.97
```

```
o=lm(Score~0+smoking:gender, data =d)
coef(o)
```

```
## smoking:gender
##       2.801111
```

The primary representation I use throughout this problem is of the form:

$$\begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{41} \\ \beta_{21} \\ \beta_{22} \\ \beta_{23} \\ \beta_{24} \end{bmatrix} = \begin{bmatrix} 25.96667 \\ 14.06667 \\ 19.86667 \\ 16.03333 \\ 19.83333 \\ 12.06667 \\ 12.13333 \\ 10.20000 \end{bmatrix}$$

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.

$$\text{SE}(\hat{\beta}_j) = \sqrt{\sigma^2 \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$$

where $\hat{\beta}_j$ is the OLS estimate for cell mean j and σ^2 is the estimated residual variance, given by:

$$\hat{\sigma}^2 = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{y}}{n - p}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix, n is the number of observations, and p is the number of estimated parameters (cells).

Each diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ determines the variance of $\hat{\beta}_j$, meaning that:

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{jj}$$

For the first cell mean (e.g., Male, Never Smoked):

$$\text{SE}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2 \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{11}}$$

```r
require(dplyr)

n <- nrow(X)
p <- ncol(X)
I <- diag(1, n)

P_X <- X %*% solve(t(X) %*% X) %*% t(X)
sigma2_hat <- as.numeric(t(y) %*% (I - P_X) %*% y / (n - p))
var_beta_hat <- sigma2_hat * solve(t(X) %*% X)
SE_beta_hat <- sqrt(diag(var_beta_hat))
names(SE_beta_hat) <- colnames(X)

SE_beta_hat
```

```
##  genderFemale:smokingHeavy  genderFemale:smokingLight
##                   1.764031                   1.764031
## genderFemale:smokingMedium  genderFemale:smokingNever
##                   1.764031                   1.764031
##    genderMale:smokingHeavy    genderMale:smokingLight
##                   1.764031                   1.764031
##   genderMale:smokingMedium    genderMale:smokingNever
##                   1.764031                   1.764031
```

Note: Because we have a balanced design, the SE of each $\hat{\beta}$ is the same, the value being 1.764031.

e)

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

```
beta_hat_values <- as.numeric(beta_hat)

main_effect_gender <- mean(beta_hat_values[1:4]) - mean(beta_hat_values[5:8])

main_effect_gender
```

```
## [1] -3.425
```

Let $\boldsymbol{\beta}$ be the vector of cell means:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{\text{Female,Never}} \\ \beta_{\text{Female,Light}} \\ \beta_{\text{Female,Medium}} \\ \beta_{\text{Female,Heavy}} \\ \beta_{\text{Male,Never}} \\ \beta_{\text{Male,Light}} \\ \beta_{\text{Male,Medium}} \\ \beta_{\text{Male,Heavy}} \end{bmatrix}$$

We express the main effect of gender as a linear contrast:

$$\mathbf{C}\boldsymbol{\beta}$$

where the contrast matrix $\mathbf{C}$ is:

$$\mathbf{C} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{bmatrix}$$

Thus, the main effect of gender can be calculated by:

```
C <- matrix(c(1/4, 1/4, 1/4, 1/4, -1/4, -1/4, -1/4, -1/4), nrow = 1)
beta_hat
```

```
##                               [,1]
## genderFemale:smokingHeavy  10.20000
## genderFemale:smokingLight  12.13333
## genderFemale:smokingMedium 16.03333
## genderFemale:smokingNever  19.86667
## genderMale:smokingHeavy    12.06667
## genderMale:smokingLight    19.83333
## genderMale:smokingMedium   14.06667
## genderMale:smokingNever    25.96667
```

```
C %*% beta_hat
```

```
##        [,1]
## [1,] -3.425
```

```r
d=stress
d$gender <- factor(d$gender)
d$smoking <- factor(d$smoking)
d <- d[order(d$gender, d$smoking),]

o=lm(Score~0+smoking:gender, data =d)
o$df
```

```
## [1] 16
```

```r
estimate=function(lmout,C,a=0.05){
  b=coef(lmout)
  V=vcov(lmout)
  df=lmout$df
  Cb=C%*%b
  se=sqrt(diag(C%*%V%*%t(C)))
  tval=qt(1-a/2,df)
  low=Cb-tval*se
  up=Cb+tval*se
  m=cbind(C,Cb,se,low,up)
  dimnames(m)[[2]]=c(paste("c",1:ncol(C),sep=""),
              "estimate","se",
              paste(100*(1-a),"% Conf.",sep=""),
              "limits")
  return(m)
}

# manual calculation
mean(coef(o)[1:4])-mean(coef(o)[5:8])
```

```
## [1] 5.425
```

```r
C = matrix(c(1/4,1/4,1/4,1/4,-1/4,-1/4,-1/4,-1/4),nrow =1)

estimate(o,C)
```

```
##        c1   c2   c3   c4    c5    c6    c7    c8 estimate       se 95% Conf.
## [1,] 0.25 0.25 0.25 0.25 -0.25 -0.25 -0.25 -0.25    5.425 1.247358  2.780718
##        limits
## [1,] 8.069282
```

The estimated main effect of gender is 5.425 via the above parametrization, this is comparing $\hat{\beta}_{male} - \hat{\beta}_{female}$ (average male effects over smoking levels minus average female effects over smoking levels).

f)

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

H_0: No Interaction

$$(\mu_{\text{Male},s} - \mu_{\text{Female},s}) - (\mu_{\text{Male},s'} - \mu_{\text{Female},s'}) = 0 \quad \text{for all } s \neq s'$$

where $\mu_{\text{Male},s}$ is the mean exercise score for males at smoking level s, $\mu_{\text{Female},s}$ is the mean exercise score for females at smoking level s.

This hypothesis states that the difference between genders must be the same at all smoking levels.

We have our same vector of cell means:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{\text{Female,Never}} \\ \beta_{\text{Female,Light}} \\ \beta_{\text{Female,Medium}} \\ \beta_{\text{Female,Heavy}} \\ \beta_{\text{Male,Never}} \\ \beta_{\text{Male,Light}} \\ \beta_{\text{Male,Medium}} \\ \beta_{\text{Male,Heavy}} \end{bmatrix}$$

To test for no interaction, we use the contrast matrix:

$$\mathbf{C}_{\text{int}} = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

Thus, testing for interaction involves checking:

$$\mathbf{C}_{\text{int}}\boldsymbol{\beta} = \mathbf{0}$$

$$\begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{\text{Female,Never}} \\ \beta_{\text{Female,Light}} \\ \beta_{\text{Female,Medium}} \\ \beta_{\text{Female,Heavy}} \\ \beta_{\text{Male,Never}} \\ \beta_{\text{Male,Light}} \\ \beta_{\text{Male,Medium}} \\ \beta_{\text{Male,Heavy}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

```
C <- matrix(c(1, -1, 0, 0, -1, 1, 0, 0,
              1, 0, -1, 0, -1, 0, 1, 0,
              1, 0, 0, -1, -1, 0, 0, 1), nrow = 3)

C %*% beta_hat
```

```
##            [,1]
## [1,]  -1.866667
## [2,] -36.433333
## [3,]  52.133333
```

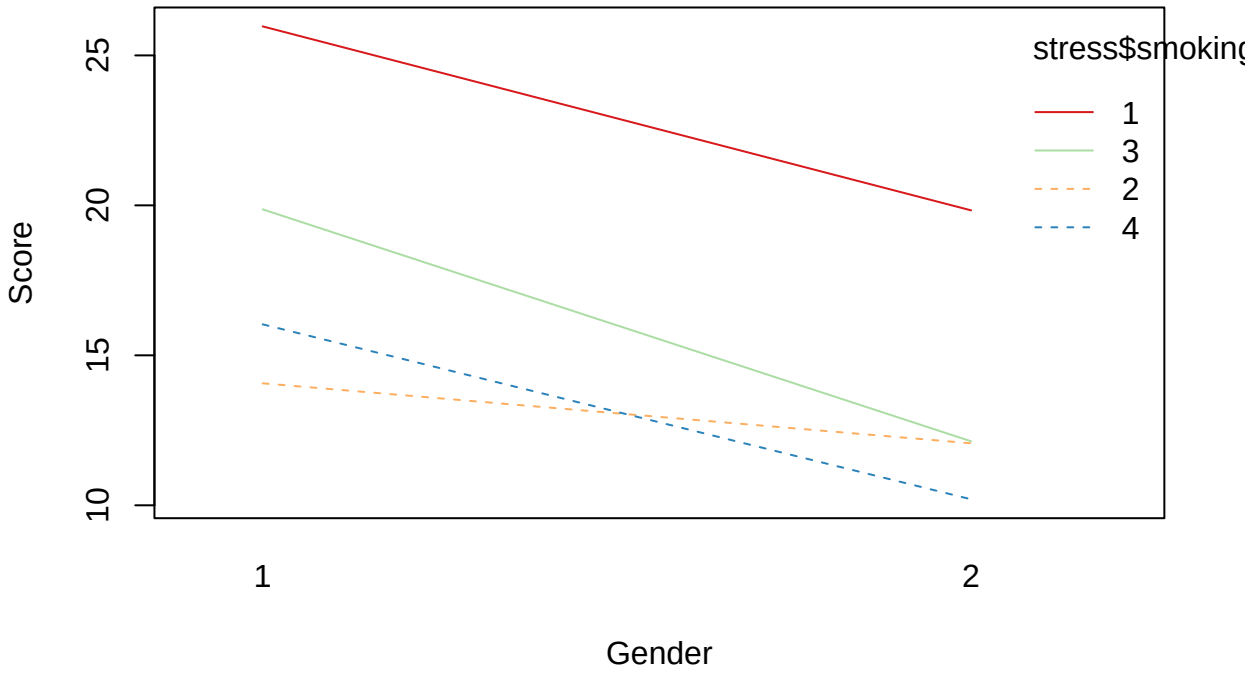
These values are not zero! So there is some evidence, no comment on strength, to suppose there are interaction effects. Leaving that for the F test and latter part of this problem.

There is also the graphical representation to consider, which similarly shows the potential for there being interaction effects by the non-parallel, intersecting lines.

```
require(RColorBrewer)
```

```
## Loading required package: RColorBrewer
```

```
interaction.plot(stress$gender, stress$smoking, stress$Score,
                 xlab = "Gender",
                 ylab = "Score",
                 col = brewer.pal(4, "Spectral"),
                 lty = 1:2)
```



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.

The hypotheses:

$$H_0 : \alpha_G = 0$$

$$H_A : \alpha_G \neq 0$$

Using our parameter representation from part e), the main effect of gender is:

$$\alpha_G = \frac{1}{4} \sum_s (\beta_{\text{Male},s} - \beta_{\text{Female},s})$$

Thus, the null hypothesis states that the mean exercise score for males and females is the same on average across smoking levels, while the alternative hypothesis states that there is a difference in mean exercise score between genders.

To test for the main effect of gender, we define the contrast matrix:

$$\mathbf{C} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{bmatrix}$$

Testing for the main effect of gender involves solving:

$$\mathbf{C}\boldsymbol{\beta} = 0$$

$$\begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{bmatrix} \begin{bmatrix} \beta_{\text{Female,Never}} \\ \beta_{\text{Female,Light}} \\ \beta_{\text{Female,Medium}} \\ \beta_{\text{Female,Heavy}} \\ \beta_{\text{Male,Never}} \\ \beta_{\text{Male,Light}} \\ \beta_{\text{Male,Medium}} \\ \beta_{\text{Male,Heavy}} \end{bmatrix} = 0$$

The test statistic for testing $H_0 : \alpha_G = 0$ is:

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

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated cell means, $(\mathbf{X}'\mathbf{X})^{-1}$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$, q = 1 (number of constraints tested), and $\hat{\sigma}^2$ is the residual variance estimate.

The numerator degrees of freedom: $df_{\text{num}} = 1$, since we are testing a single contrast.

The denominator degrees of freedom: $df_{\text{den}} = n - p$, where n is the total number of observations and p = 8 (one parameter for each gender-smoking combination), giving 24 - 8 = 16.

```r
# alt method, not used
C_gender <- matrix(c(1/4, 1/4, 1/4, 1/4, -1/4, -1/4, -1/4, -1/4), nrow = 1)

num <- t(C_gender %*% beta_hat) %*% solve(C_gender %*% solve(t(X) %*% X) %*% t(C_gender)) %*% (C_gender
den <- sigma2_hat * 1
F_stat <- num / den

df_num <- 1
df_den <- nrow(X) - ncol(X)

p_value <- 1 - pf(F_stat, df_num, df_den)

list(F_stat = F_stat, df_num = df_num, df_den = df_den, p_value = p_value)
```

```r
# alt method
test=function(lmout,C,d=0){
  b=coef(lmout)
  V=vcov(lmout)
  dfn=nrow(C)
  dfd=lmout$df
  Cb.d=C%*%b-d
  Fstat=drop(t(Cb.d)%*%solve(C%*%V%*%t(C))%*%Cb.d/dfn)
  pvalue=1-pf(Fstat,dfn,dfd)
  list(Fstat=Fstat,pvalue=pvalue)
}

C = matrix(c(1/4,1/4,1/4,1/4,-1/4,-1/4,-1/4,-1/4),nrow =1)
C
```

```
##      [,1] [,2] [,3] [,4]  [,5]  [,6]  [,7]  [,8]
## [1,] 0.25 0.25 0.25 0.25 -0.25 -0.25 -0.25 -0.25
```

```r
test(o,C)
```

```
## $Fstat
## [1] 18.91547
##
## $pvalue
## [1] 0.0004970517
```

Within the context of the study, we have overwhelmingly strong evidence (p: 0.00049, with F-statistic of 18.92) against the null hypothesis that there is no difference in average exercise scores between males and females when accounting for the potential effects of smoking, i.e. that averaged across all smoking levels, there is overwhelming evidence there are differences in the mean "exercise level" or fitness level between genders.

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.

To test the main effect of smoking, we have:

$$H_0 : \alpha_S = 0$$

$$H_A : \alpha_S \neq 0$$

where the main effect of smoking is:

$$\alpha_S(s) = \frac{1}{2} \sum_g \left(\beta_{g,s} - \bar{\beta}_S \right), \quad \text{for } s = \text{Never, Light, Medium, Heavy}$$

To test for the main effect of smoking, define:

$$\mathbf{C} = \frac{1}{2} \begin{bmatrix} 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 \end{bmatrix}$$

To be frank, I realized my initial test needed to be multiplied by the constant 1/2. My fingers hurt and I'm lazy, so I'm just tacking that on to the calculations and expressions here.

That being said,

We are testing:

$$\mathbf{C}\boldsymbol{\beta} = 0$$

Where:

$$\frac{1}{2} \begin{bmatrix} 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_{\text{Female,Never}} \\ \beta_{\text{Female,Light}} \\ \beta_{\text{Female,Medium}} \\ \beta_{\text{Female,Heavy}} \\ \beta_{\text{Male,Never}} \\ \beta_{\text{Male,Light}} \\ \beta_{\text{Male,Medium}} \\ \beta_{\text{Male,Heavy}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The corresponding F statistic takes the familiar form:

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

where q = 3 (since we test 3 independent constraints), and $\hat{\sigma}^2$ is the residual variance.

Since the F statistic uses to degrees of freedom, note:

Numerator: $df_{\text{num}} = 3$ (corresponding to the 3 constraints). Denominator: $df_{\text{den}} = n - p$, where n is total observations and p = 8 (one parameter per gender-smoking combination), giving 24 - 8 = 16.

```r
# Again, alt method
C_smoking <- matrix(c(1, -1,  0,  0, 1, -1,  0,  0,
                      1,  0, -1,  0, 1,  0, -1,  0,
                      1,  0,  0, -1, 1,  0,  0, -1),
                    nrow = 3, byrow = TRUE)
adjustedC <- (1/2) * C_smoking

num <- t(adjustedC %*% beta_hat) %*% solve(adjustedC %*% solve(t(X) %*% X) %*% t(adjustedC)) %*% (adjus
den <- sigma2_hat * 3
F_stat <- num / den

df_num <- 3
df_den <- nrow(X) - ncol(X)

p_value <- 1 - pf(F_stat, df_num, df_den)

list(F_stat = F_stat, df_num = df_num, df_den = df_den, p_value = p_value)
```

```r
C = matrix(c(c(1/2,-1/2,0,0,1/2,-1/2,0,0),
           c(1/2,0,-1/2,0,1/2,0,-1/2,0),
           c(1/2,0,0,-1/2,1/2,0,0,-1/2)),
           byrow = T,nrow =3)
test(o,C)
```

```
## $Fstat
## [1] 13.7615
##
## $pvalue
## [1] 0.0001071936
```

Within the context of the study, we have extremely strong evidence (p: 0.0001) against the null hypothesis that there is no difference in average exercise scores between smoking levels when accounting for (averaging across) the potential effects of gender. This is to say we have overwhelming evidence that the mean patient fitness levels averaged across genders is different for at least one smoking level compared to the mean fitness levels of the other smoking levels.

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.

Our initial hypotheses:

$$H_0 : \gamma_s = 0, \quad \forall s$$

$$H_A : \gamma_s \neq 0 \text{ for at least one } s$$

We define our C matrix as:

$$\mathbf{C}_{\text{int}} = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

Testing for an interaction involves solving:

$$\mathbf{C}_{\text{int}}\boldsymbol{\beta} = 0$$

Giving us:

$$\begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{\text{Female,Never}} \\ \beta_{\text{Female,Light}} \\ \beta_{\text{Female,Medium}} \\ \beta_{\text{Female,Heavy}} \\ \beta_{\text{Male,Never}} \\ \beta_{\text{Male,Light}} \\ \beta_{\text{Male,Medium}} \\ \beta_{\text{Male,Heavy}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Our F statistic is of the typical form:

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

where q = 3 (testing 3 independent constraints) and $\hat{\sigma}^2$ is the residual variance estimate.

Again, as we are working with an F statistic, we have two degrees of freedom to note:

Numerator: $df_{\text{num}} = 3$ (corresponding to the 3 constraints). Denominator: $df_{\text{den}} = n - p$, where: n is total observations and p = 8 (one parameter per gender-smoking combination), again totalling 16.

Calculating:

```
C = matrix(c(c(1,-1,0,0,-1,1,0,0),
             c(1,0,-1,0,-1,0,1,0),
             c(1,0,0,-1,-1,0,0,1)),byrow = T,nrow =3)
test(o,C)
```

```
## $Fstat
## [1] 0.9494756
##
## $pvalue
## [1] 0.4401953
```

Interpretation:

Within the context of the study, we have little to no evidence (p: 0.440) against the null hypothesis that there are no interaction effects, e.g. that there is little to no evidence that there is a difference in average exercise scores by gender that change/differ depending on smoking level.

The practical significance of this is that we may consider using an additive model appropriate. That point aside, this is important though. Namely, by having little to no evidence of interaction effects, we have more solid footing in interpreting the main effects such as those outlined previously such as gender or smoking effects. Were there interaction effects, these types of effects would require additional adjustments. This is helpful for practical reasons in that it would allow researchers or policymakers to better understand the impact of particular factors, e.g. targetted treatments, though those are not necessarily relevant or of note for the study in question.

j)

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

The formula we use for this problem is:

$$c^T \hat{\boldsymbol{\beta}} \pm t_{n-r,1-\alpha/2} \sqrt{\hat{\sigma}^2 c^T (\mathbf{X}^T \mathbf{X})^{-1} c}$$

Given the estimated mean $\hat{\beta}$ for the category Male, Never Smoked, (the row of the beta vector), our calculation is given by:

$$\hat{\beta} \pm t_{n-r,1-\alpha/2} \cdot \text{SE}$$

where: SE is the standard error of the estimated mean (Male, Never Smoked).

Calculating:

```
stress <- read.table(file = "stress.txt",
                     header = TRUE,
                     sep = ",")

stress_data <- stress

stress_data$gender <- as.factor(stress_data$gender)
stress_data$smoking <- as.factor(stress_data$smoking)

model <- lm(Score ~ gender:smoking - 1, data = stress_data)

beta_male_never <- coef(model)["gender1:smoking1"]
se_male_never <- summary(model)$coefficients["gender1:smoking1", "Std. Error"]

df <- df.residual(model)

t_value <- qt(0.975, df)

MOE <- t_value * se_male_never

CI_lower <- beta_male_never - MOE
CI_upper <- beta_male_never + MOE

CI_lower
```

```
## gender1:smoking1
##         22.22709
```

```
CI_upper
```

```
## gender1:smoking1
##         29.70625
```

This gives the interval (22.23, 29.71).

k)

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

We are again using the formula:

$$c^T \hat{\boldsymbol{\beta}} \pm t_{n-r,1-\alpha/2} \sqrt{\hat{\sigma}^2 c^T (\mathbf{X}^T \mathbf{X})^{-1} c}$$

The mean effect of gender can be computed as the difference between the mean achievement scores for male and female patients:

$$\hat{\delta} = \hat{\beta}_{\text{Male}} - \hat{\beta}_{\text{Female}}$$

The confidence interval is of the form:

$$\hat{\delta} \pm t_{n-r,1-\alpha/2} \cdot SE$$

where: SE is the standard error of the main effect of gender, and:

$$\hat{\delta} = \hat{\beta}_{\text{Male}} - \hat{\beta}_{\text{Female}}$$

Calculating:

```
# using the alternative method because my setup was all messed up previously for gender effects
C = matrix(c(1/4,1/4,1/4,1/4,-1/4,-1/4,-1/4,-1/4), nrow = 1)
estimate(o,C)
```

```
##        c1   c2   c3   c4    c5    c6    c7    c8 estimate       se 95% Conf.
## [1,] 0.25 0.25 0.25 0.25 -0.25 -0.25 -0.25 -0.25    5.425 1.247358  2.780718
##        limits
## [1,] 8.069282
```

This gives the interval (2.78, 8.07), bearing in mind that I'm using the parametrization consistent with the prior gender effect question, i.e. comparing average male to average female averaged over smoking levels respectively.

l)

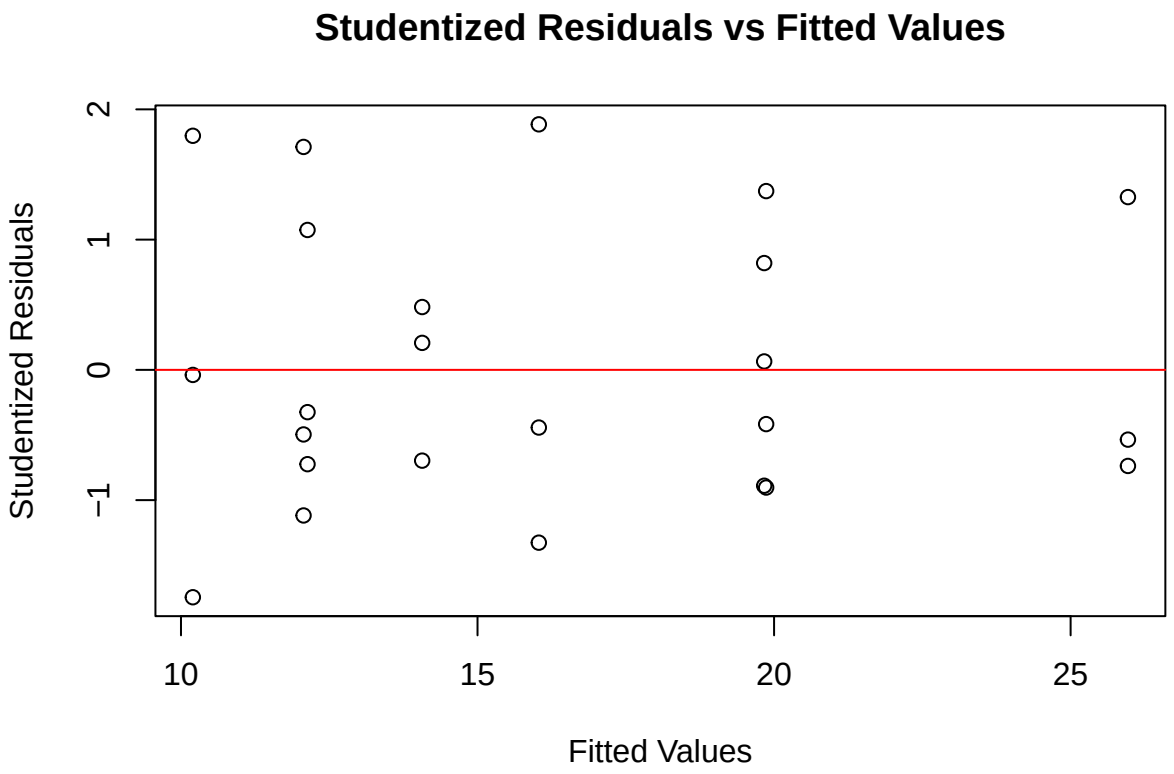
Obtain the residuals for the fitted models and use them to check the necessary assumptions that allow us to fit the proposed model. Please submit and explain any graphical displays that you might use.

```
stress_data <- stress

stress_data$gender <- as.factor(stress_data$gender)
stress_data$smoking <- as.factor(stress_data$smoking)

# Fit the cell-means model
model <- lm(Score ~ gender:smoking - 1, data = stress_data)

# Extract residuals and fitted values
residuals <- resid(model)
fitted_values <- fitted(model)
```

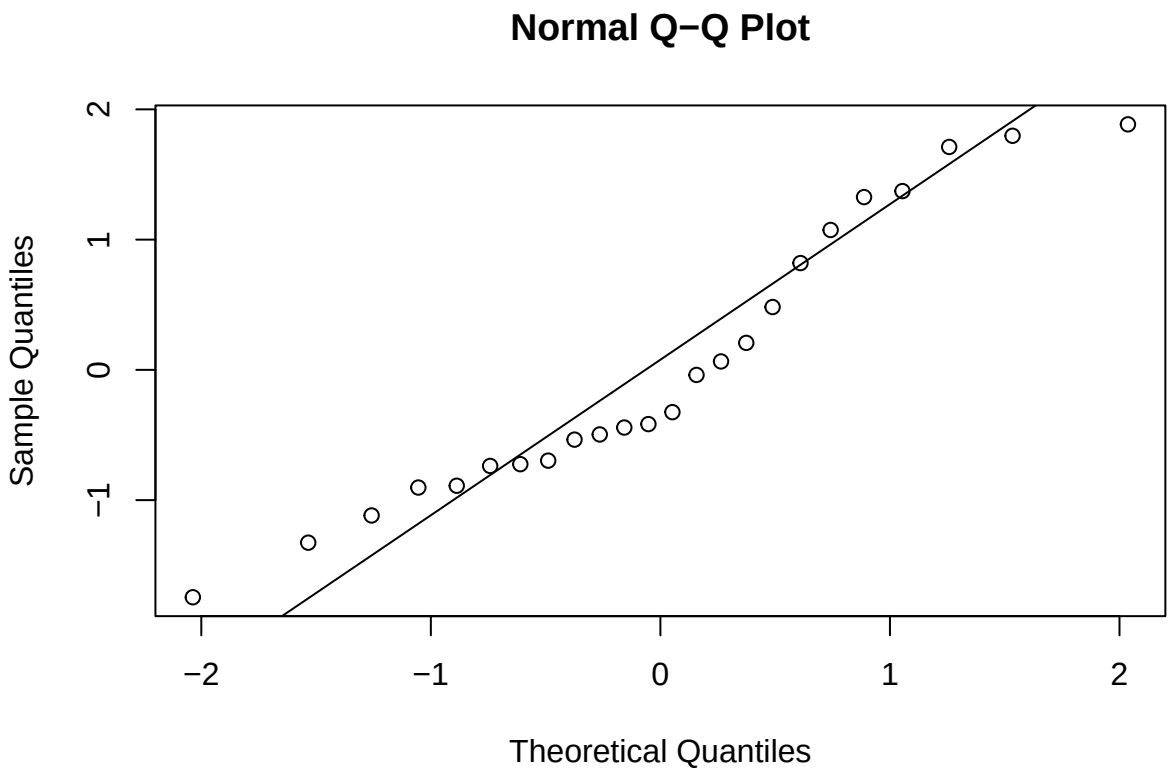
```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
stdresids <- studres(model)

plot(model$fitted.values, stdresids, main = "Studentized Residuals vs Fitted Values", xlab = "Fitted Va
abline(h = 0, col = "red")
```

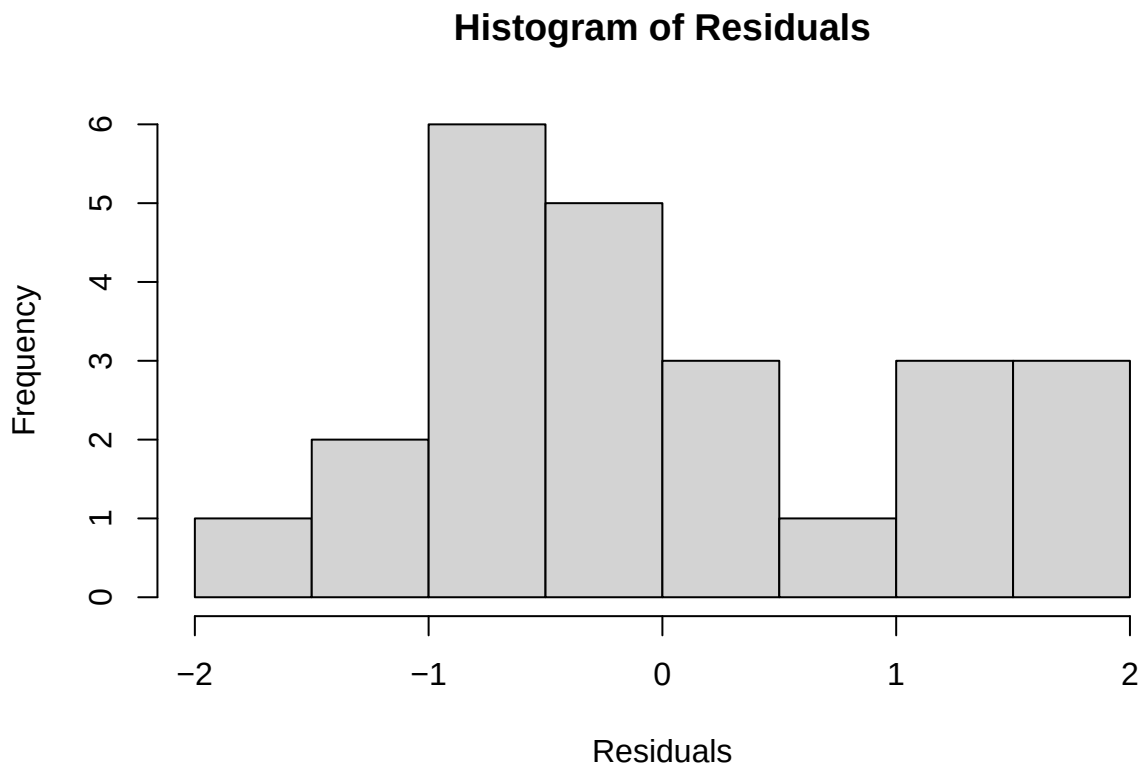



```
qqnorm(stdresids)
qqline(stdresids, col = "black")
```




```
# shapiro.test(stdresids)
```

```
hist(stdresids, main="Histogram of Residuals", xlab="Residuals", breaks=10)
```




```
require(moments)
```

```
## Loading required package: moments
```

```
require(lmtest)
```

```
## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 4.4.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.4.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
mean(stdresids)
```

```
## [1] 0.01444866
```

```r
median(stdresids)
```

```
## [1] -0.3704723
```

```r
skewness(x = stdresids)
```

```
## [1] 0.4199657
```

```r
kurtosis(x = stdresids) - 3
```

```
## [1] -0.9730195
```

```r
require(car)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
leveneTest(Score ~ factor(gender):factor(smoking), data = stress, center = mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value Pr(>F)
## group  7   0.651 0.7089
##       16
```

```r
leveneTest(Score ~ factor(gender):factor(smoking), data = stress, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  7  0.2226 0.9742
##       16
```

The first assumption to review is the independence assumption, which predominantly is determined by reviewing the study design. Though limited information is provided, we don't have reason to suspect this being violated, such as through clustering of experimental units in the study.

That being said:

The residual plot (studentized) looks good for the purposes of diagnosing a number of our the model assumptions. To begin with, we generally observe a random spread of residuals across the range of fitted values (there is not a clear trend present in the residual plot above). Because of this, We do not observe any trends or noticeable patterns in the above plots, such that we have reason to believe our linearity assumption is not being violated. Additionally, it the residuals do not appear clustered, such that they appear to be randomly spread (constant variance) and centered around zero.

However, there are some potential issues with the normality assumption. In particular, the QQ plot does not closely align with the reference line within the first theoretical quantile, so there are some concerns about the normality distribution, even with considering the studentized residuals. Furthermore, the QQ plot appears to have a slight "S" shape/curve, further suggesting deviation from normality. The histogram of residuals (studentized) provide a similar picture of not appearing especially normal, though these considerations are largely visual.

Overall, we have reason to suspect our normality assumption is being violated but the other assumptions appear appropriate.

Additionally, I've included some Levene Tests to test the equal variances assumption. Via these tests, we achieve relatively large p-values for both, such that we have evidence to support the null hypothesis of equal variances between factors, which is also an assumption in our model.

Also, in support of normality, we do also check the skewness and (excess) kurtosis of the (studentized) residuals. Though they do not match exactly with the values we'd expect for a normal distribution, the observed statistics as shown do not deviate substantially. However, there are the visual inconsistencies noted previously such that this assumption may still be violated.