Homework 4

Mary Silva, Wyara Moura Silva, and Peter Trubey

November 10, 2017

Chapter 8

Question 1

```
mileage = data.frame(
  model = c(rep('A',2),rep('B',3),rep('C',3),rep('D',2)),
  mileage = c(22,26,28,24,29,29,32,28,23,24)
)
mileage.model = lm(mileage ~ model, data = mileage)
anova(mileage.model)[["Pr(>F)"]][1] <= 0.05
```

```
## [1] FALSE
```

Question 3

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \qquad \varepsilon_{i,j} \sim N(0, \sigma^2) Y = X\beta + \varepsilon \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,n} \\ y_{2,1} \\ \vdots \\ y_{2,n} \\ y_{3,1} \\ \vdots \\ y_{3,n} \end{pmatrix} = \begin{pmatrix} 1,0,0 \\ \vdots \\ 1,0,0 \\ 1,1,0 \\ \vdots \\ 1,1,0 \\ 1,0,1 \\ \vdots \\ 1,0,1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,n} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,n} \\ \varepsilon_{3,1} \\ \vdots \\ \varepsilon_{3,n} \end{pmatrix}$$

To prevent singularity in the model matrix, we drop the α for Setosa, and μ, instead of representing the grand mean, now represents the average for Setosa.

The unknown parameters for our one-way ANOVA are contained in the *beta* vector, μ, α_2, and α_3, with μ representing Setosa's mean value, and α_2 and α_3 representing Versicolor's and Virginica's average deviations from that mean value.

```
data(iris)
sl.model = lm(Sepal.Length ~ Species, data = iris)
anova(sl.model)
```

```
## Analysis of Variance Table
##
## Response: Sepal.Length
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Species     2 63.212  31.606  119.26 < 2.2e-16 ***
## Residuals 147 38.956   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sl.model$coefficients
```

```
##       (Intercept) Speciesversicolor  Speciesvirginica
##             5.006             0.930             1.582
```

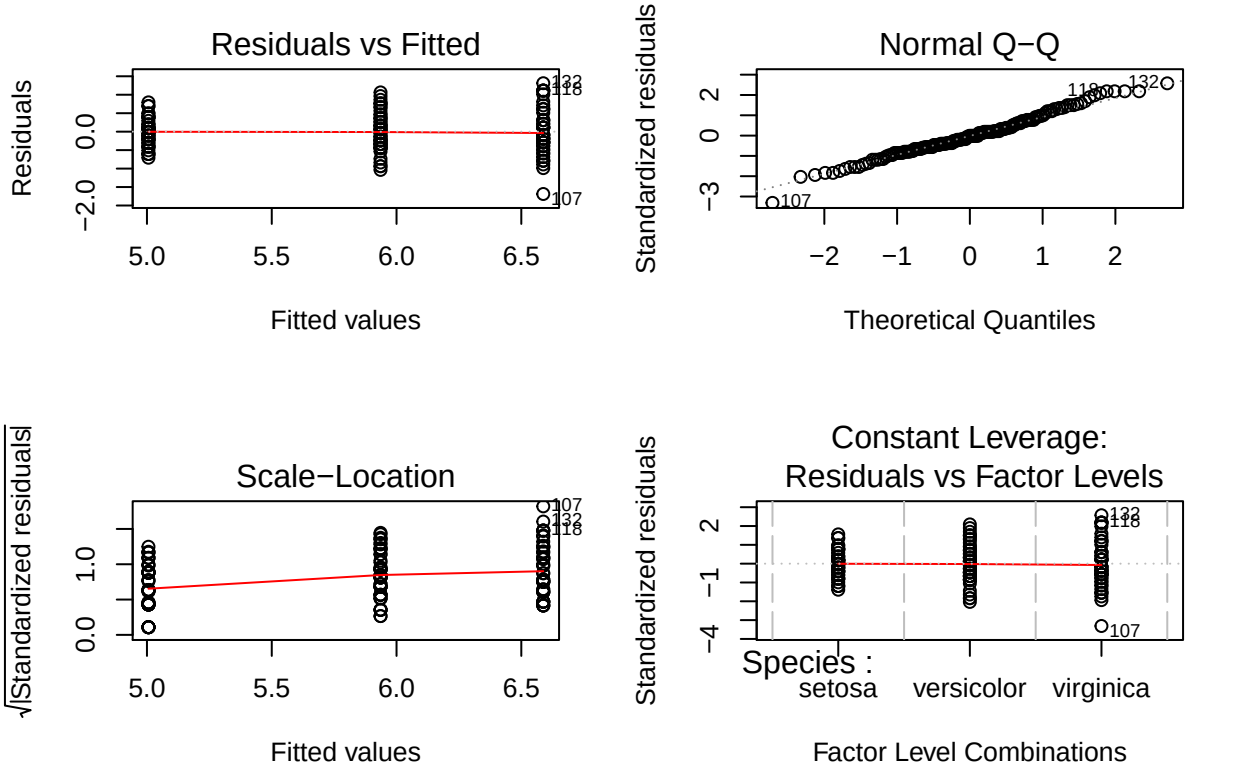
The parameter estimates are given as above, with Setosa having an average value of 5.006, Versicolor deviating from Setosa by 0.930, and Virginica deviating from Setosa by 1.582.

Question 4

The assumption required for model inference in QUestion 3 is that the errors, $\varepsilon_{i,j}$ are distributed normally, with mean 0 and variance σ^2.

To check if there is a problem with the residuals, we can observe the fit plots associated with the model.

```
par(mfrow=c(2,2))
plot(sl.model)
```



We don't see any serious issues, but there is some worrying straying from the diagonal on the QQ plot. Additionally, we can conduct a Shapiro Wilks test of residuals to test whether the residuals are not normally distributed.

```
shapiro.test(sl.model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sl.model$residuals
```

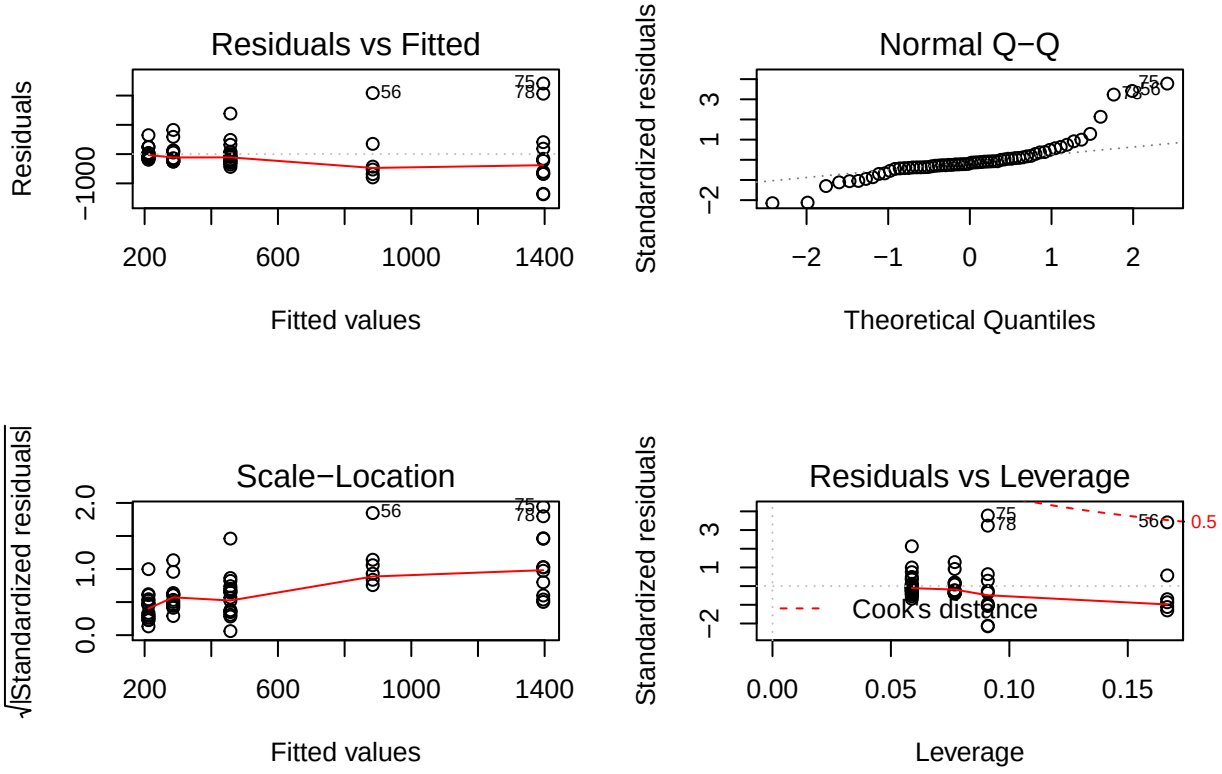
```
## W = 0.9879, p-value = 0.2189
```

With a p-value of 0.2189, we have no evidence that the residuals are not normally distributed.

Question 5

```
url = 'http://personal.bgsu.edu/~mrizzo/Rx/Rx-data/PATIENT.DAT'
patient = read.table(url, sep = '\t')
names(patient) = c('stomach','bronchus','colon','ovary','breast')
patient.long = na.omit(stack(patient))
names(patient.long) = c('time','organ')
```

```
patient.model = lm(time ~ organ, data = patient.long)
par(mfrow = c(2,2))
plot(patient.model)
```



The residuals are clearly not normal, there is marked deviance from the diagonal on the QQ plot, and strong heteroscedasticity on the residuals vs fitted values plot.

```
shapiro.test(patient.model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  patient.model$residuals
## W = 0.8068, p-value = 9.981e-08
```

with a p-value of less than 0.05, we have strong evidence that the errors are not normally distributed. Therefore, this model formulation is not valid. We can look at a log transformation of the Y variable, survival time.

```
patient.long$logtime = log(patient.long$time)
patient.logmodel = lm(logtime ~ organ, data = patient.long)
summary(patient.logmodel)
```

```
##
## Call:
## lm(formula = logtime ~ organ, data = patient.long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3805 -0.6607  0.1025  0.8207  2.0460
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.96792    0.33147  14.988  < 2e-16 ***
## organbronchus -0.01475    0.44033  -0.033  0.97339
## organcolon     0.78120    0.44033   1.774  0.08120 .
## organovary     1.18270    0.58985   2.005  0.04955 *
## organbreast    1.59068    0.48961   3.249  0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.195 on 59 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.1726
## F-statistic: 4.286 on 4 and 59 DF,  p-value: 0.004122
```

In this model, the intercept takes the mean value for stomach. It appears that Bronchus does not deviate significantly from stomach, while colon shows a weak significant difference from stomach with a p-value of 0.08120. Ovary and Breast are strongly significantly different from stomach.

```
patient.logaov = aov(logtime ~ organ, data = patient.long)
TukeyHSD(patient.logaov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = logtime ~ organ, data = patient.long)
##
## $organ
##                          diff        lwr      upr     p adj
## bronchus-stomach -0.01474955 -1.2537924 1.224293 0.9999997
## colon-stomach     0.78120255 -0.4578403 2.020245 0.3981146
## ovary-stomach     1.18269662 -0.4770864 2.842480 0.2763506
## breast-stomach    1.59068365  0.2129685 2.968399 0.0158132
## colon-bronchus    0.79595210 -0.3575340 1.949438 0.3072938
## ovary-bronchus    1.19744617 -0.3994830 2.794375 0.2296079
## breast-bronchus   1.60543320  0.3041254 2.906741 0.0083352
## ovary-colon       0.40149407 -1.1954351 1.998423 0.9540004
## breast-colon      0.80948110 -0.4918267 2.110789 0.4119156
## breast-ovary      0.40798703 -1.2987803 2.114754 0.9615409
```

Using Tukey's Honest Significant Difference (HSD) test, we see that breast is significantly different from

stomach and Bronchus, but no other significant pairwise differences, even at alpha values up to 0.2.

Chapter 9

Question 1

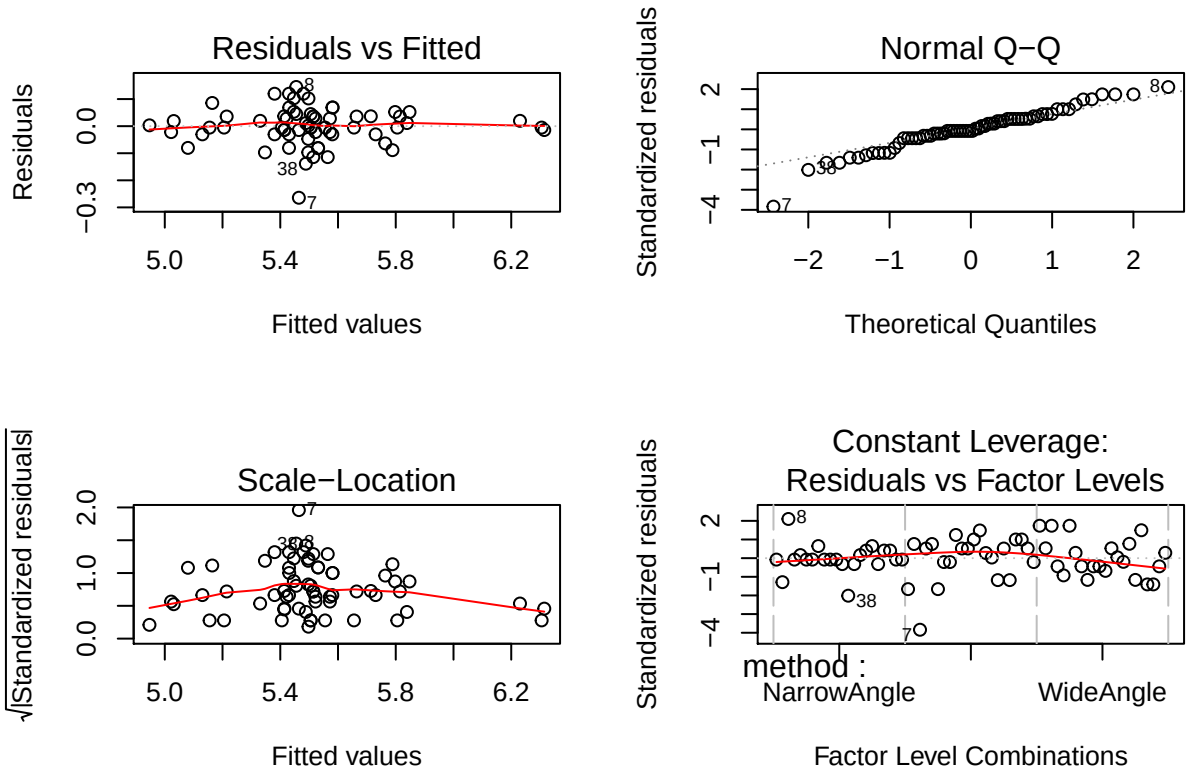
```
url = 'http://personal.bgsu.edu/~mrizzo/Rx/Rx-data/rounding.txt'
rounding = read.table(url, header = TRUE)
names(rounding) = c('time','method','player')
rounding$player = as.factor(rounding$player)
str(rounding)
```

```
## 'data.frame':    66 obs. of  3 variables:
##  $ time  : num  5.4 5.5 5.55 5.85 5.7 5.75 5.2 5.6 5.5 5.55 ...
##  $ method: Factor w/ 3 levels "NarrowAngle",..: 2 1 3 2 1 3 2 1 3 2 ...
##  $ player: Factor w/ 22 levels "1","2","3","4",..: 1 1 1 2 2 2 3 3 3 4 ...
```

```
rounding.model = lm(time ~ method + player, data = rounding)
```

Treating players as the relevant blocks, we can't look for interaction effects between blocks and players as we would have more than 66 terms to estimate, meaning we would run out of degrees of freedom.

```
par(mfrow=c(2,2))
plot(rounding.model)
```



There might be some heteroscedasticity in the residuals, with a bulge in variance right in the middle of the fitted values, but that might also be because there are more data there. We should check the Shapiro Wilks

test.

```
shapiro.test(rounding.model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rounding.model$residuals
## W = 0.95277, p-value = 0.01356
```

The shapiro test concludes that the fitted values are not normal. We should investigate how to fix that.