Homework 6

Hyunjoon Rhee

5/3/2021

Problem 1

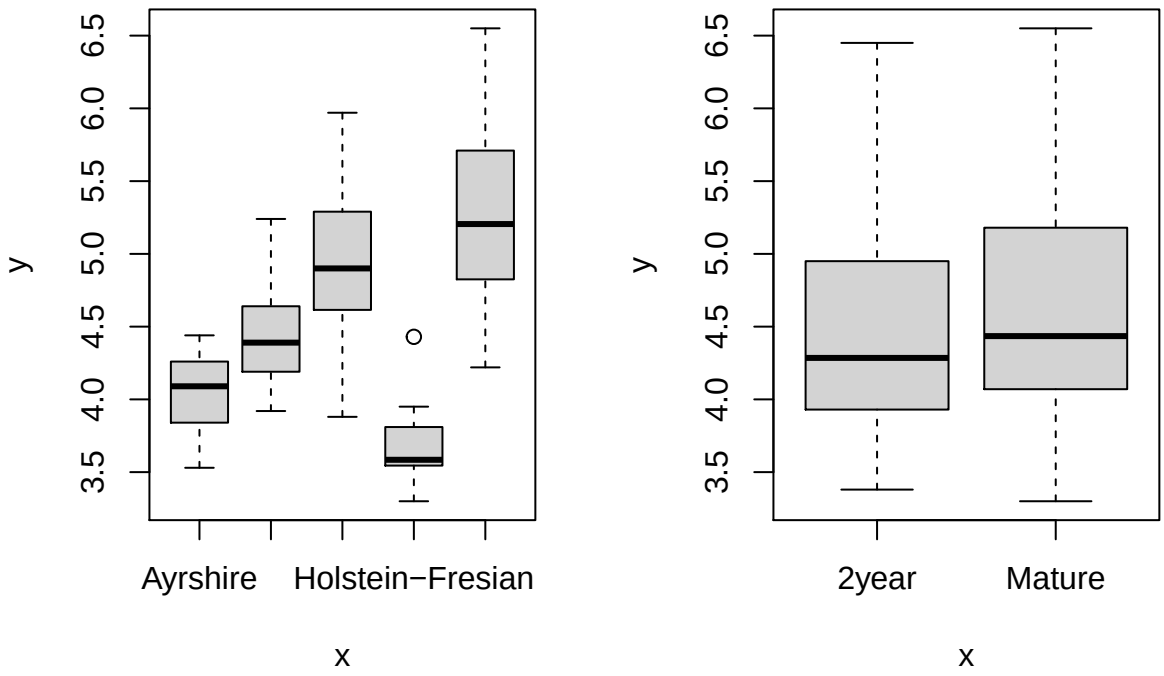
a)

```
library(faraway)
attach(butterfat)
```

```
head(butterfat)
```

```
##   Butterfat    Breed    Age
## 1      3.74 Ayrshire Mature
## 2      4.01 Ayrshire  2year
## 3      3.77 Ayrshire Mature
## 4      3.78 Ayrshire  2year
## 5      4.10 Ayrshire Mature
## 6      4.06 Ayrshire  2year
```

```
par(mfrow=c(1,2))
plot(Breed, Butterfat, data = butterfat)
plot(Age, Butterfat, data = butterfat)
```



b)

```
model=aov(Butterfat~Breed+Age+Breed*Age, data = butterfat)
summary(model)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Breed        4  34.32   8.580  49.565 <2e-16 ***
## Age          1   0.27   0.274   1.580  0.212
## Breed:Age    4   0.51   0.128   0.742  0.566
## Residuals   90  15.58   0.173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p value for breed and age interaction term is greater than $\alpha = 0.05$ which means that it fails to reject that there is an interaction term has zero effect on the response.

c)

```
maineffect = lm(Butterfat ~ Breed + Age, data = butterfat)
summary(maineffect)
```

```
##
## Call:
## lm(formula = Butterfat ~ Breed + Age, data = butterfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0202 -0.2373 -0.0640  0.2617  1.2098
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.00770    0.10135  39.541  < 2e-16 ***
## BreedCanadian           0.37850    0.13085   2.893  0.00475 **
## BreedGuernsey           0.89000    0.13085   6.802 9.48e-10 ***
## BreedHolstein-Fresian -0.39050    0.13085  -2.984  0.00362 **
## BreedJersey             1.23250    0.13085   9.419 3.16e-15 ***
## AgeMature               0.10460    0.08276   1.264  0.20937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4138 on 94 degrees of freedom
## Multiple R-squared:  0.6825, Adjusted R-squared:  0.6656
## F-statistic: 40.41 on 5 and 94 DF,  p-value: < 2.2e-16
```

From problem 1 b) it was concluded that the interaction term has no statistical meaning. Therefore, the additive model was introduced. Looking at the p value of each term, only the 'Age' term has a p value that is greater than $\alpha = 0.05$. This means that only age rejects the null hypothesis and conclude that age does not have statistical significance.

```
anova(lm(Butterfat ~ Breed, data=butterfat))
```

```
## Analysis of Variance Table
##
## Response: Butterfat
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Breed      4 34.321  8.5803  49.802 < 2.2e-16 ***
## Residuals 95 16.368  0.1723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(Butterfat ~ Age, data=butterfat))
```

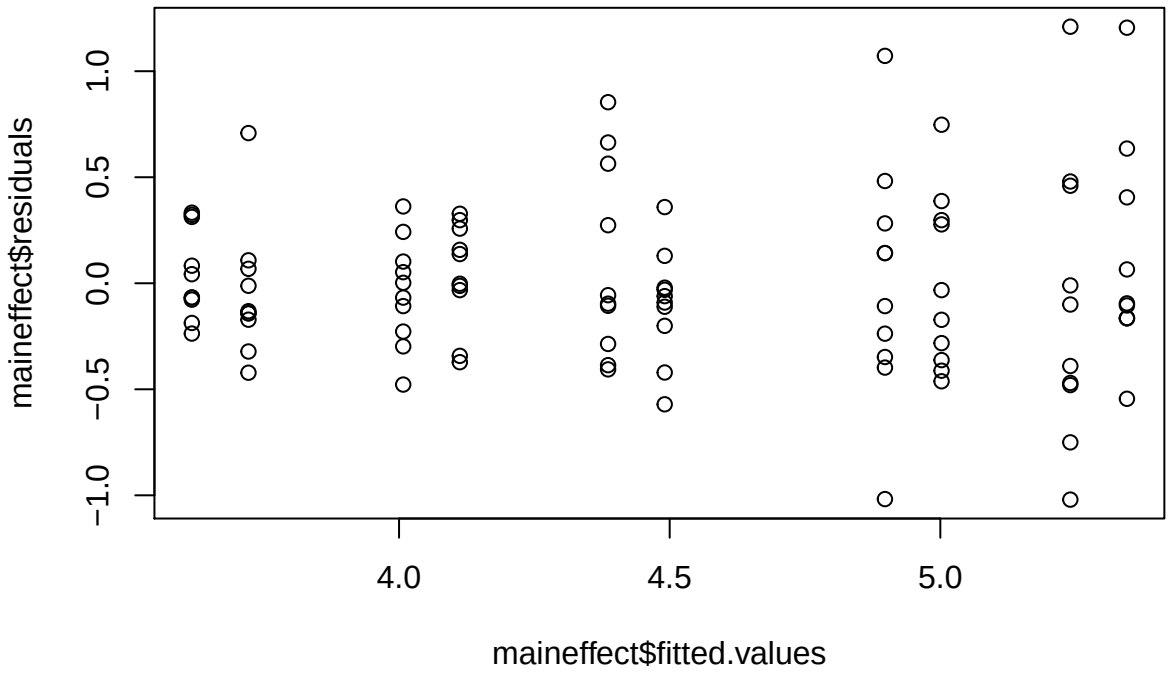
```
## Analysis of Variance Table
##
## Response: Butterfat
##           Df Sum Sq Mean Sq F value Pr(>F)
## Age        1  0.274 0.27353  0.5317 0.4676
## Residuals 98 50.415 0.51444
```

Looking at the anova table of the two of the p values, it is evident that there is a statistical difference between the breed and the butterfat, whereas age does not have statistical difference between age and butterfat. This is the same result as the above analysis.

d)

Constant variance check

```
plot(maineffect$fitted.values, maineffect$residuals)
```




```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.2
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

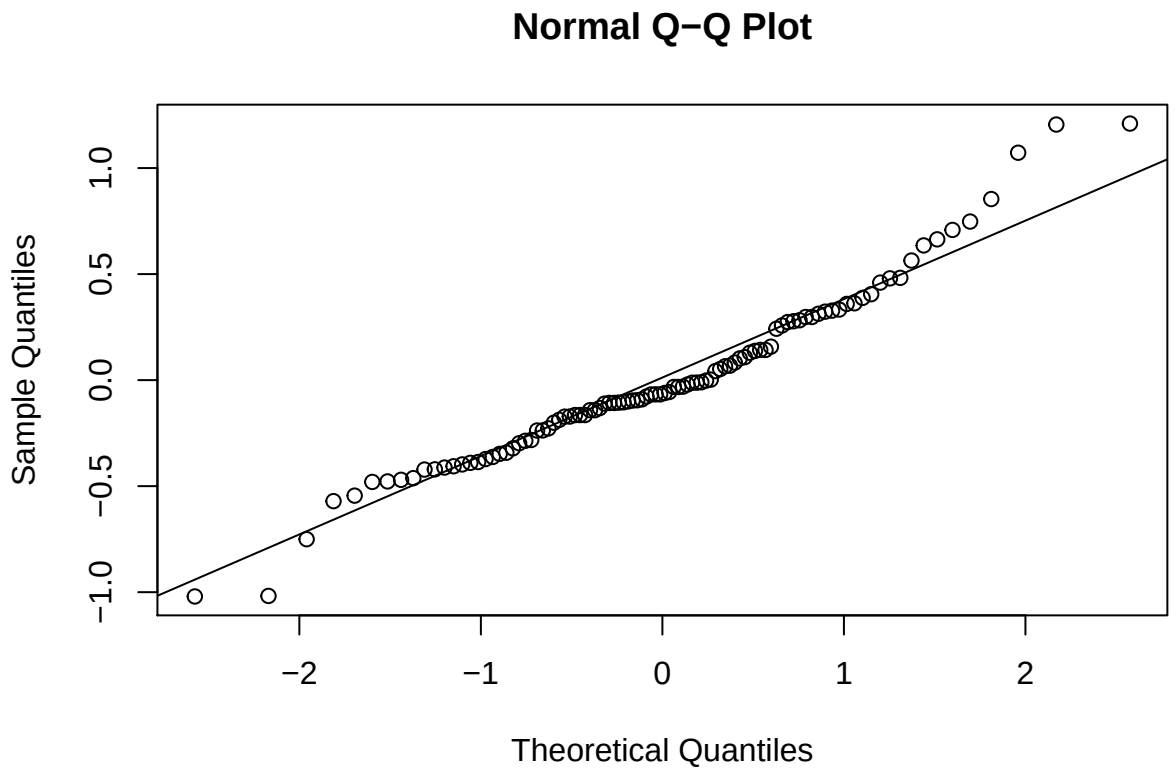
```
bptest(maineffect)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  maineffect
## BP = 14.739, df = 5, p-value = 0.01154
```

It is hard to see from the graph if there is a constant variance, so Breusch-Pagan test was performed. The bptest shows a p value that is less than $\alpha = 0.05$, which means it can reject the null hypothesis and conclude that it is not homoskedastic.

Normality check

```
qqnorm(maineffect$residuals)
qqline(maineffect$residuals)
```




```
shapiro.test(maineffect$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  maineffect$residuals
## W = 0.96347, p-value = 0.007168
```

The qqplot seems to show a straight line, but for further analysis, shapiro-wilk test was performed. Because the p value is smaller than $\alpha = 0.05$, it is evident that it can reject the null hypothesis and conclude that the residual is not normally distributed.

```
dwtest(maineffect)
```

```
##
```

```
##  Durbin-Watson test
##
## data:  maineffect
## DW = 2.0367, p-value = 0.4531
## alternative hypothesis: true autocorrelation is greater than 0
```

The Durbin-Watson test shows it has a p value that is greater than $\alpha = 0.05$, which means that it fails to reject the null hypothesis and conclude that the errors are statistically not correlated.

Constant variance ,normality, correlation assumptions were not met, meaning that them model itself is questionable.

e)

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
bestbreed = butterfat %>% filter(Breed == 'Jersey' | Breed == 'Guernsey') %>% select(Butterfat, Breed)
t.test(Butterfat ~ Breed, data = bestbreed)
```

```
##
##  Welch Two Sample t-test
##
## data:  Butterfat by Breed
## t = -1.9895, df = 36.367, p-value = 0.05421
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.691530171  0.006530171
## sample estimates:
## mean in group Guernsey   mean in group Jersey
##                 4.9500                 5.2925
```

Looking at the results in Problem 1 a), it is known that Jersey and Guernsey are the two breeds that has most butterfat content. Because the p value of the t test is greater than $\alpha = 0.05$, it fails to reject the null hypothesis and conclude that it is hard to say there is a statistical difference between the best and the second.

Problem 2

a)

```
anova(lm(Speed ~ Run + Expt, data = morley))
```

```
## Analysis of Variance Table
##
## Response: Speed
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Run        1    412     412  0.0733 0.7872081
## Expt       1  72581   72581 12.9172 0.0005138 ***
## Residuals 97 545032    5619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(Speed ~ Run, data = morley))
```

```
## Analysis of Variance Table
##
## Response: Speed
##           Df Sum Sq Mean Sq F value Pr(>F)
## Run        1    412   411.7  0.0653 0.7988
## Residuals 98 617612  6302.2
```

Looking at the p value of the anova table, it is much greater than $\alpha = 0.05$, which means that there is not a significant difference among the run groups.

b)

By using blocking factor, it can categorize the difference between the run groups and compare the response. Without a blocking factor, it would have not earned any meaningful data related to the mean of the speed.

Problem 3

a)

```
prob3 = lm(yield~ ., alfalfa)
anova(prob3)
```

```
## Analysis of Variance Table
##
## Response: yield
##            Df  Sum Sq Mean Sq F value   Pr(>F)
## shade       4  87.402  21.851  7.1254 0.003533 **
## irrigation  4  16.562   4.141  1.3502 0.307872
## inoculum    4 155.894  38.974 12.7091 0.000284 ***
## Residuals  12  36.799   3.067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova table shows that when the $\alpha = 0.05$, shade and inoculum show statistical significance whereas irrigation does not.

b)

```
TukeyHSD(aov(yield~., alfalfa), "inoculum")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = yield ~ ., data = alfalfa)
##
## $inoculum
##      diff        lwr       upr     p adj
## B-A -0.72  -4.250202  2.810202 0.9633433
## C-A -0.08  -3.610202  3.450202 0.9999928
## D-A -0.86  -4.390202  2.670202 0.9326392
## E-A -6.60 -10.130202 -3.069798 0.0005166
## C-B  0.64  -2.890202  4.170202 0.9759059
## D-B -0.14  -3.670202  3.390202 0.9999332
## E-B -5.88  -9.410202 -2.349798 0.0014163
## D-C -0.78  -4.310202  2.750202 0.9515868
## E-C -6.52 -10.050202 -2.989798 0.0005764
## E-D -5.74  -9.270202 -2.209798 0.0017334
```

The confidence interval shows that everything but E-A, E-B, E-C, E-D do not contain 0 in the confidence interval, which means that there is no significant difference between every pair but the ones that contains E in it. This shows that A-D show significant difference with E, but the rest do not.