HW6

Tianqi Wu

4/17/2020

```
library(faraway)
attach(butterfat)
library(lmtest)
attach(morley)
attach(alfalfa)
```

Problem 1

1(a)

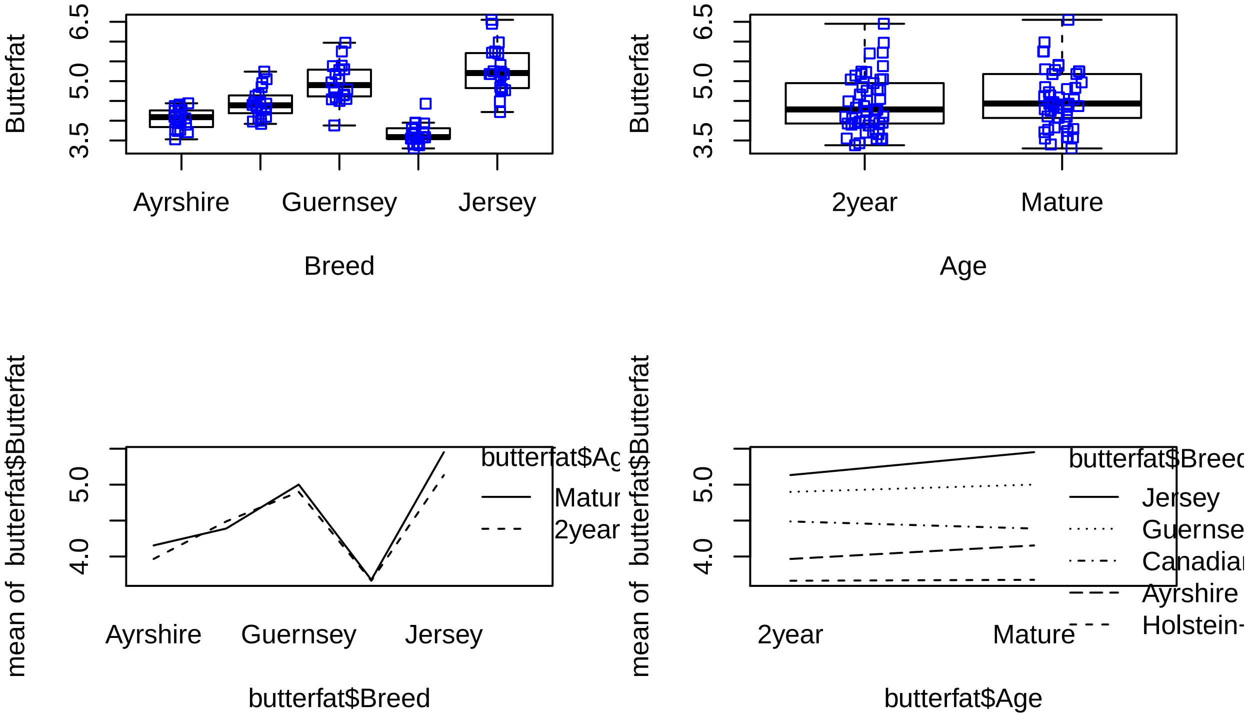
Since our predictors are two categorical variables. Boxplots are appropriate here. As we can see, Breed of Holstein-Fresian has smallest butterfat and Jersey has largest butterfat. There is not big difference between age of 2year and mature.

```
par(mfrow=c(2,2))

boxplot(Butterfat~Breed, data=butterfat, outline=FALSE)
stripchart(Butterfat~Breed, data=butterfat, method="jitter",
           col="blue", vertical=TRUE, add=TRUE)

boxplot(Butterfat~Age, data=butterfat, outline=FALSE)
stripchart(Butterfat~Age, data=butterfat, method="jitter",
           col="blue", vertical=TRUE, add=TRUE)

interaction.plot(butterfat$Breed, butterfat$Age, butterfat$Butterfat)
interaction.plot(butterfat$Age, butterfat$Breed, butterfat$Butterfat)
```



1(b)

Since the Breed:Age has p-value 0.5658 > 0.05, we fail to reject the null hypothesis and conclude that there is no interaction between breed and age.

```
model.1b=lm(Butterfat ~ Breed*Age, butterfat)
anova(model.1b)
```

```
## Analysis of Variance Table
##
## Response: Butterfat
##           Df Sum Sq Mean Sq F value Pr(>F)
## Breed      4 34.321  8.5803 49.5651 <2e-16 ***
## Age        1  0.274  0.2735  1.5801 0.2120
## Breed:Age  4  0.514  0.1285  0.7421 0.5658
## Residuals 90 15.580  0.1731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

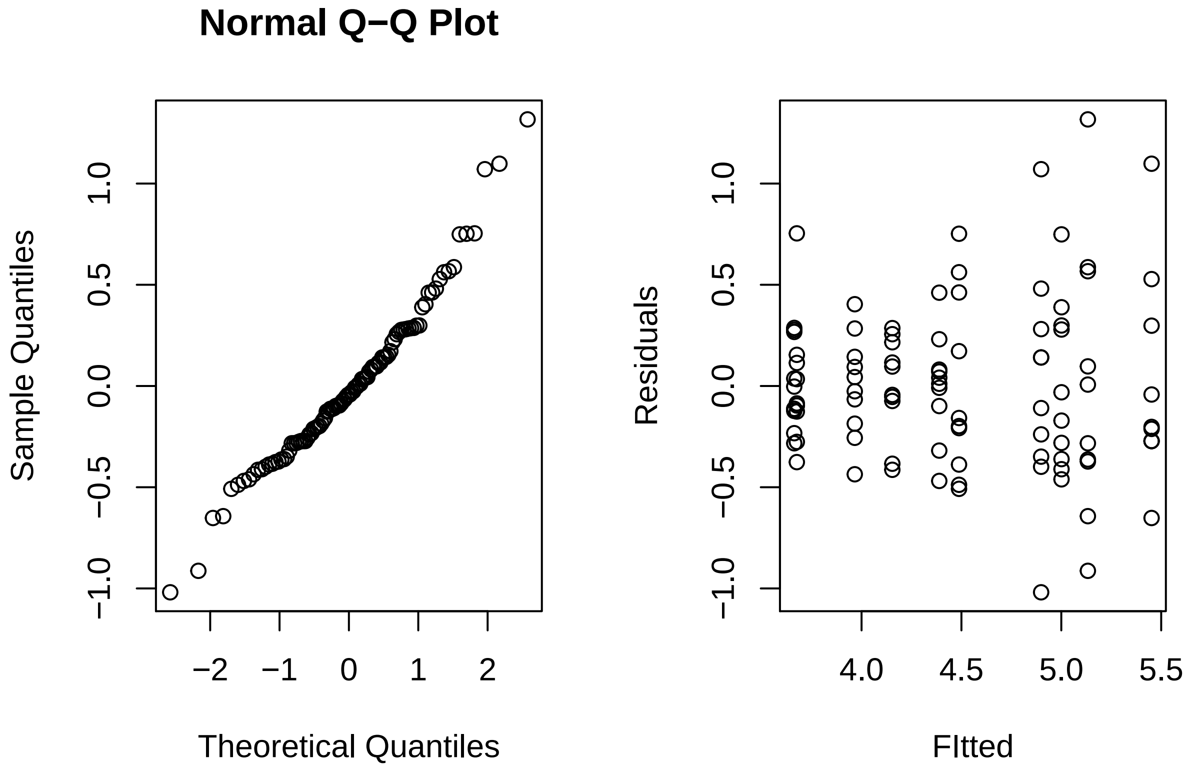
1(c)

Since the p-value for Breed < 0.05 and p-value for Age > 0.05, we conclude that there is statistically significant difference between breeds but there is no such difference between ages.

1(d)

Since The p-value for wilks-shapiro test is 0.01635 < 0.05, we reject the null hypothesis and conclude that the normality assumption is violated. However, from the qq plot, the data only shows light skewness. Since the p-value for Levene's test is 0.007686 < 0.01, we reject the null hypothesis and conclude that homocedasticity assumption is violated. Also, the residual plot shows increasing residual as fitted value increases. It indicates we may need some transformation. No outlier is detected.

```
par(mfrow=c(1,2))
qqnorm(model.1b$res)
plot(model.1b$fitted, model.1b$res, xlab="FItted", ylab="Residuals")
```



```
shapiro.test(residuals(model.1b))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.1b)
## W = 0.96828, p-value = 0.01635
```

```
summary(lm(abs(model.1b$res) ~ Breed*Age))
```

```
##
## Call:
## lm(formula = abs(model.1b$res) ~ Breed * Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50800 -0.14200 -0.06060  0.09965  0.80200
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.940e-01  7.558e-02   2.567  0.01191 *
## BreedCanadian                   1.956e-01  1.069e-01   1.830  0.07056 .
## BreedGuernsey                   2.290e-01  1.069e-01   2.143  0.03485 *
## BreedHolstein-Fresian          -2.040e-02  1.069e-01  -0.191  0.84906
## BreedJersey                      3.210e-01  1.069e-01   3.003  0.00346 **
## AgeMature                        1.388e-16  1.069e-01   0.000  1.00000
## BreedCanadian:AgeMature         -2.104e-01  1.512e-01  -1.392  0.16737
## BreedGuernsey:AgeMature         -7.980e-02  1.512e-01  -0.528  0.59885
## BreedHolstein-Fresian:AgeMature  3.760e-02  1.512e-01   0.249  0.80412
## BreedJersey:AgeMature           -1.302e-01  1.512e-01  -0.861  0.39133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.239 on 90 degrees of freedom
## Multiple R-squared:  0.2133, Adjusted R-squared:  0.1346
## F-statistic: 2.711 on 9 and 90 DF,  p-value: 0.007686
```

```
## outliers
n = nrow(butterfat)
p = ncol(butterfat)
jack=rstudent(model.1b)
qt(.05/(2*n), n-p-1)
```

```
## [1] -3.603392
```

```
sort(abs(jack), decreasing=TRUE)[1:5]
```

```
##       82       99       60       48       96
## 3.544433 2.893426 2.815850 2.667897 2.371741
```

1(e)

From previous boxplots, we can see that the best breed is Jersey and the second best is Guernsey. From the tukey pairwise CIs, since CI of Jersey-Guernsey includes 0, we do not enough evidence to conclude that the best breed is clearly superior to the second best.

```
TukeyHSD(aov(Butterfat ~ Breed + Age, data=butterfat), "Breed")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Butterfat ~ Breed + Age, data = butterfat)
##
## $Breed
##                               diff        lwr        upr     p adj
## Canadian-Ayrshire           0.3785  0.0145538  0.7424462 0.0373310
## Guernsey-Ayrshire           0.8900  0.5260538  1.2539462 0.0000000
## Holstein-Fresian-Ayrshire -0.3905 -0.7544462 -0.0265538 0.0290906
## Jersey-Ayrshire             1.2325  0.8685538  1.5964462 0.0000000
## Guernsey-Canadian           0.5115  0.1475538  0.8754462 0.0016067
## Holstein-Fresian-Canadian -0.7690 -1.1329462 -0.4050538 0.0000006
## Jersey-Canadian             0.8540  0.4900538  1.2179462 0.0000000
## Holstein-Fresian-Guernsey -1.2805 -1.6444462 -0.9165538 0.0000000
## Jersey-Guernsey             0.3425 -0.0214462  0.7064462 0.0752825
## Jersey-Holstein-Fresian     1.6230  1.2590538  1.9869462 0.0000000
```

Problem 2

Since p-value for Run > 0.05, we fail to reject the null hypothesis and conclude that there is no significantly difference between runs. The relative efficiency is 1.1689. It indicates that we would gain 16.89% efficiency with blocking factor.

```
morley$Expt = as.factor(morley$Expt)
morley$Run = as.factor(morley$Run)
anova(lm(Speed~Expt+Run, morley))
```

```
## Analysis of Variance Table
##
## Response: Speed
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Expt       4  94514 23628.5  4.3781 0.003071 **
## Run       19 113344  5965.5  1.1053 0.363209
## Residuals 76 410166  5396.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## Analysis of Variance Table
##
## Response: Speed
##           Df Sum Sq Mean Sq F value Pr(>F)
## Run       19 113344  5965.5  0.9456 0.5313
## Residuals 80 504680  6308.5
```

```
# efficiency
sigma_crd = 6308.5
sigma_rcbd = 5396.9
sigma_crd/sigma_rcbd
```

```
## [1] 1.168912
```

Problem 3

From the anova result, since p-value of inoculum < 0.05, we reject the null hypothesis and conclude that there a difference between inoculum. From the summary, the p-value of inoculumE < 0.05 and it is the only level that is significantly different. From the tukey pairwise CIs, E-A, E-B, E-C and E-D are significantly different since their CIs do not include 0. In conclusion, level E is significantly different from other four levels.

```
model.3 = lm(yield~inoculum+irrigation+shade, alfalfa)
anova(model.3)
```

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

```
TukeyHSD(aov(yield~inoculum+irrigation+shade, alfalfa), "inoculum")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = yield ~ inoculum + irrigation + shade, 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
```