STAT 425 - HW6

Josh Janda

20 April, 2020

```
library(tidyverse)
library(faraway)
library(lmtest)
```

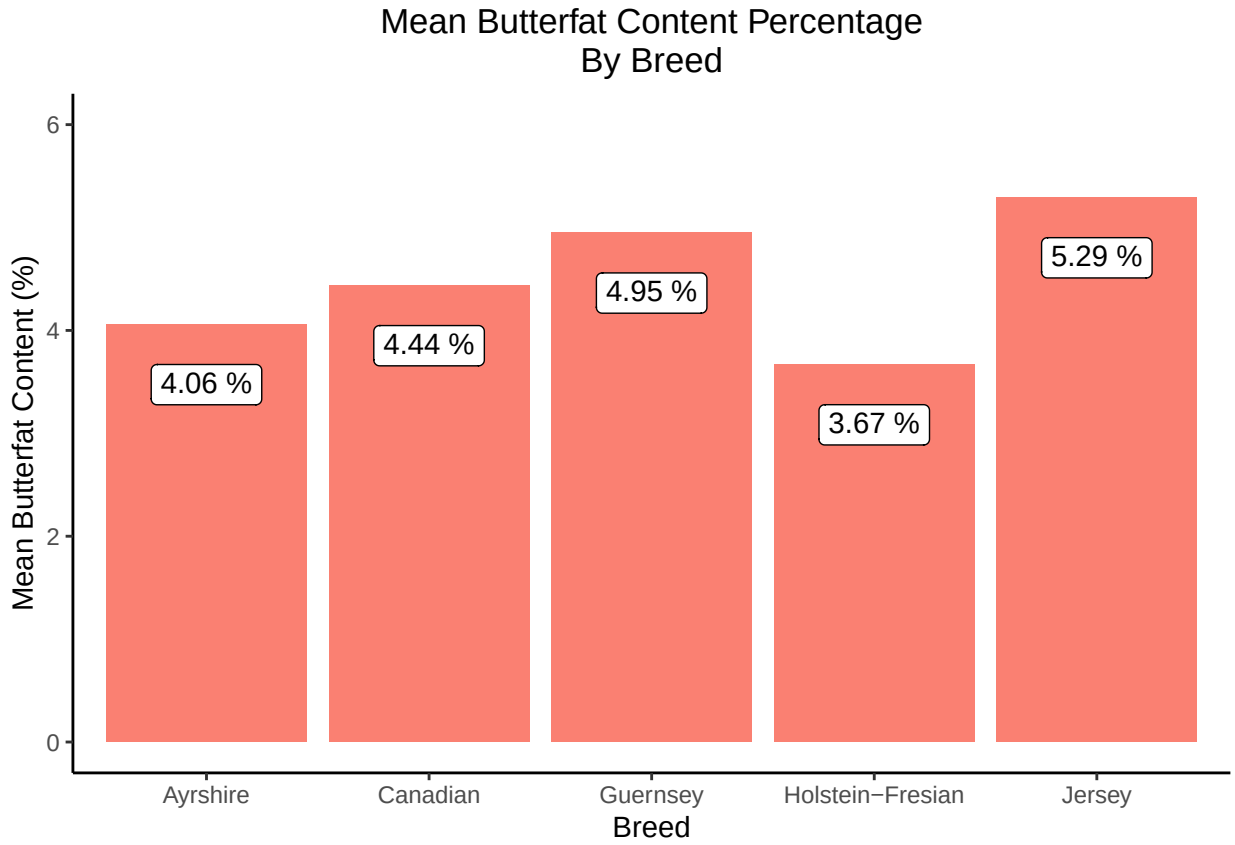
Problem 1.

Data on the content of milk from Canadian cows of five different breeds and two different ages can be found in the butterfat dataset.

A.

Make appropriate plots of the data

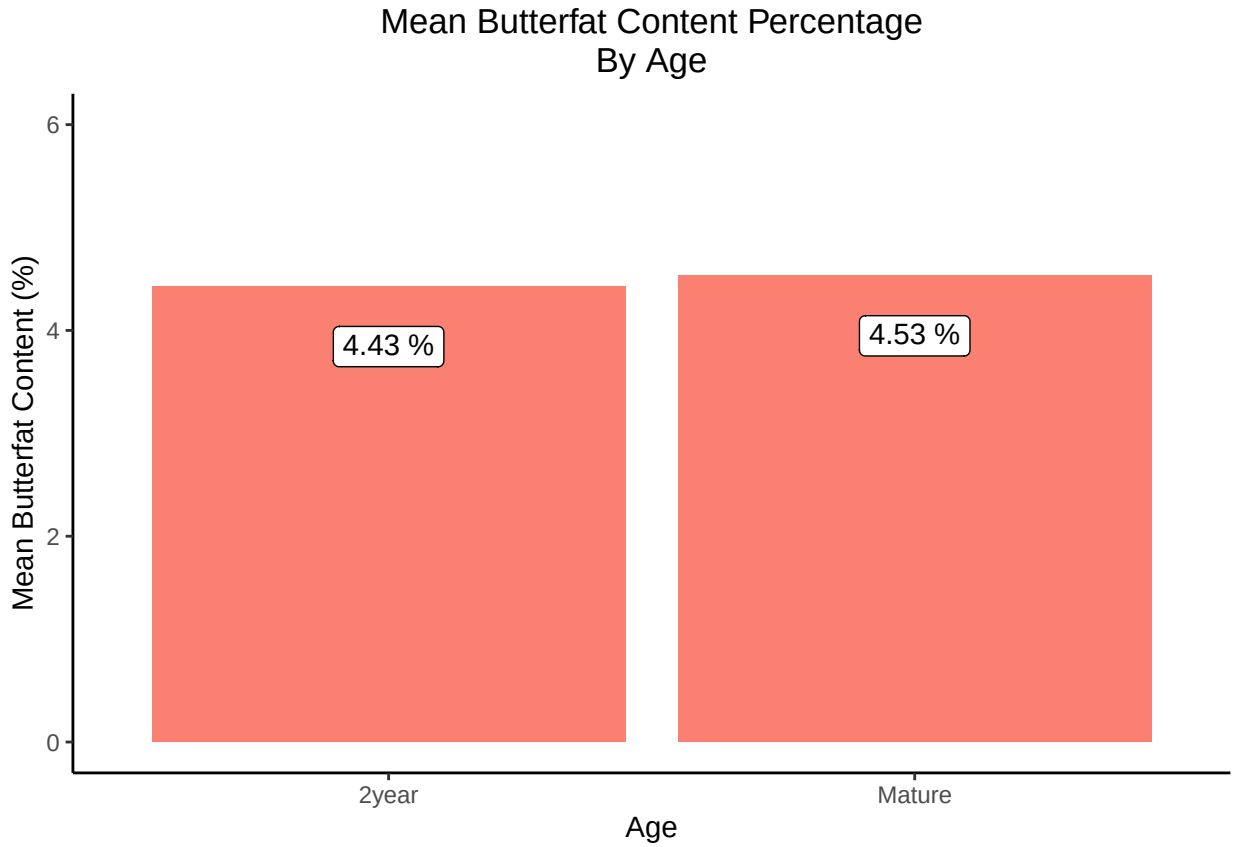
```
butterfat1 = butterfat %>% group_by(Breed) %>% summarise(MeanButterfat = mean(Butterfat))
ggplot(data = butterfat1, aes(x = Breed, y = MeanButterfat)) +
  geom_bar(stat = 'identity', fill = 'salmon') +
  geom_label(aes(label = paste(round(MeanButterfat, 2), '%')), vjust = 2) +
  labs(x = 'Breed', y = 'Mean Butterfat Content (%)', title = 'Mean Butterfat Content Percentage\nBy Br
  theme(panel.background = element_blank(),
        axis.line = element_line(color = 'black'),
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits = c(0, 6))
```



The plot above shows the mean butterfat content percentage between each breed of Canadian cows. We can see that the Jersey breed has the highest mean butterfat content percentage at 5.29%. The breed with the lowest mean butterfat content percentage is the Holstein-Fresian breed at 3.67%

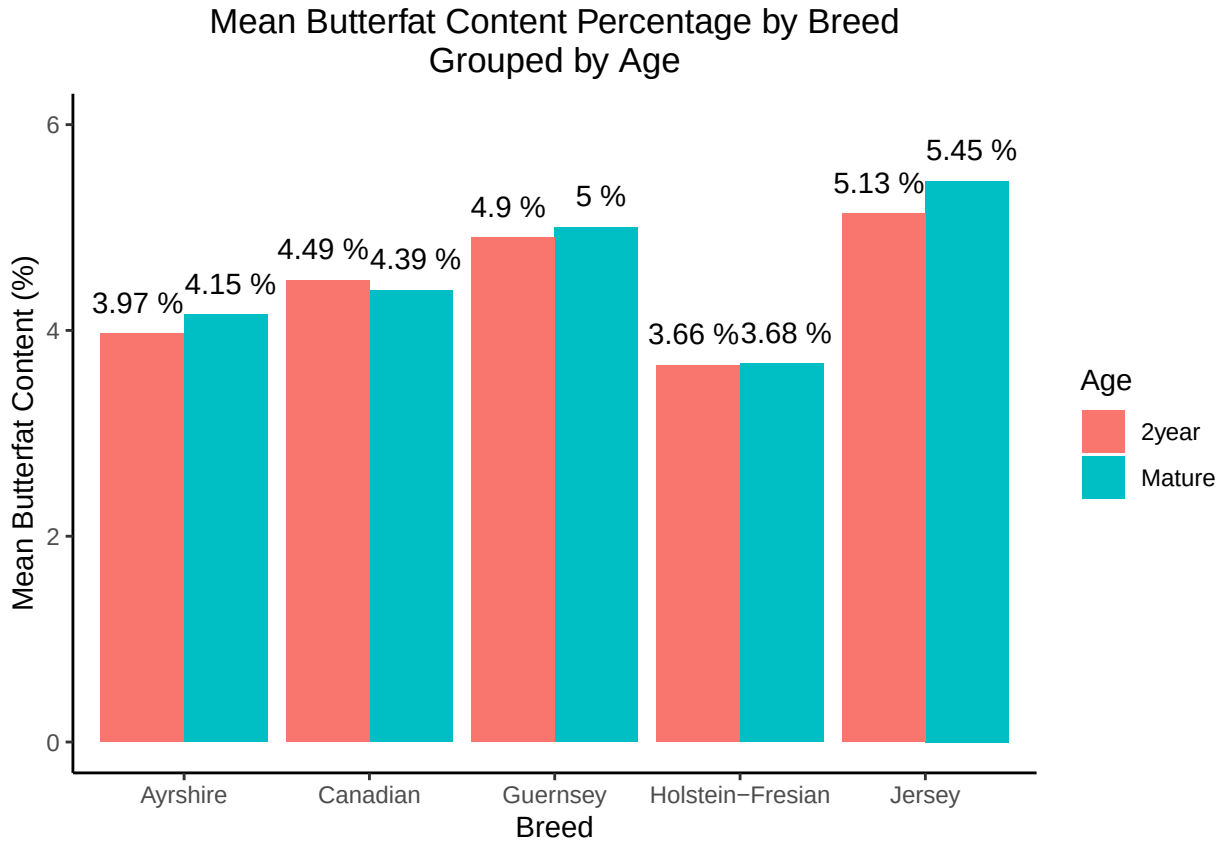
From this plot, we can gain information such as which breed outputs the highest amount of butterfat on average.

```
butterfat1 = butterfat %>% group_by(Age) %>% summarise(MeanButterfat = mean(Butterfat))
ggplot(data = butterfat1, aes(x = Age, y = MeanButterfat)) +
  geom_bar(stat = 'identity', fill = 'salmon') +
  geom_label(aes(label = paste(round(MeanButterfat, 2), '%')), vjust = 2) +
  labs(x = 'Age', y = 'Mean Butterfat Content (%)', title = 'Mean Butterfat Content Percentage\nBy Age'
  theme(panel.background = element_blank(),
        axis.line = element_line(color = 'black'),
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits = c(0, 6))
```



The plot above shows the mean butterfat content percentage between each breed of Canadian cows. We can see that Mature cows have the highest mean butterfat content percentage at 4.53%. Cows that are age 2year have the lowest mean butterfat content percentage at 4.43%. These content percentages do not differ by much, possibly telling us that this factor is not very useful for predicting butterfat content.

```
butterfat1 = butterfat %>% group_by(Breed, Age) %>% summarise(MeanButterfat = mean(Butterfat))
ggplot(data = butterfat1, aes(x = Breed, y = MeanButterfat, fill = Age)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  geom_text(aes(label = paste(round(MeanButterfat, 2), '%')),
            vjust = -1, position = position_dodge(width = 1)) +
  labs(x = 'Breed', y = 'Mean Butterfat Content (%)',
       title = 'Mean Butterfat Content Percentage by Breed\nGrouped by Age') +
  theme(panel.background = element_blank(),
        axis.line = element_line(color = 'black'),
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits = c(0, 6))
```



The plot above let's us see the mean butterfat content percentage between ages in each breed. The breed with the largest difference of mean butterfat content percentage between ages is the Jersey breed, which also was the breed with the highest mean butterfat content percentage. The breed with the smallest difference of mean butterfat content percentage between ages is the Holstein-Fresian breed, which also was the breed with the lowest mean butterfat content percentage.

From this plot we are able to gain insight on how the the variables **breed** and **age** interact between each other. While this plot does suggest differences between content percentages between each age for each age, the differences are not large possibly indicating that there is not an interaction between these variables.

B.

Determine whether there is an interaction between breed and age.

```
butter_no_int = lm(Butterfat ~ ., data = butterfat)
butter_int = lm(Butterfat ~ Breed*Age, data = butterfat)
anova(butter_no_int, butter_int)
```

```
## Analysis of Variance Table
##
## Model 1: Butterfat ~ Breed + Age
## Model 2: Butterfat ~ Breed * Age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     94 16.094
## 2     90 15.580  4   0.51387 0.7421 0.5658
```

The ANOVA table above gives us an F-Statistic of .7421 and an associated p-value of .5658. Therefore, we fail to reject the null hypothesis and conclude that the smaller model without the interaction terms is the better model.

I can conclude that there is not interaction between breed and age.

C.

Determine whether there is statistically significant difference between breeds and also ages.

```
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
```

From the ANOVA output above, we can reject H_0 at $\alpha = .05$ and conclude that there is a statistically significant difference between breeds and their Butterfat content percentage.

```
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
```

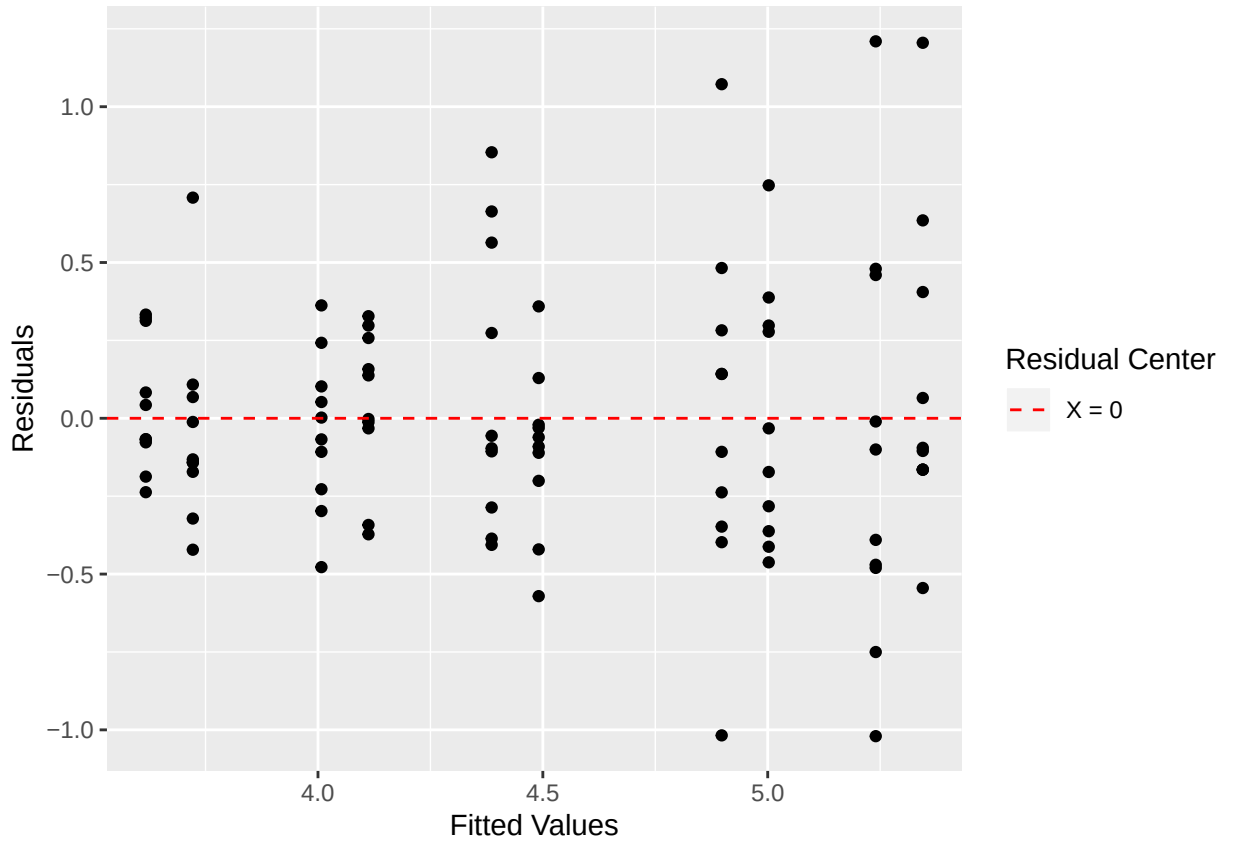
From the ANOVA output above, we fail to reject H_0 at $\alpha = .05$ and conclude that there is not a statistically significant difference between ages and their Butterfat content percentage.

D.

Present regression diagnostics for your chosen model and comment whether the assumptions have been met.

First, I will check the model for heteroskedasticity.

```
ggplot() +
  geom_point(aes(x = butter_no_int$fitted.values,
                 y = butter_no_int$residuals)) +
  geom_hline(aes(yintercept = 0, linetype = "X = 0"), color = 'red') +
  labs(linetype = "Residual Center",
       x = "Fitted Values", y = "Residuals") +
  scale_linetype_manual(values = c(2))
```



The residual plot above on my chosen model suggests that the model is not homoskedastic. The variance of the residuals seems to get larger as the fitted value increases.

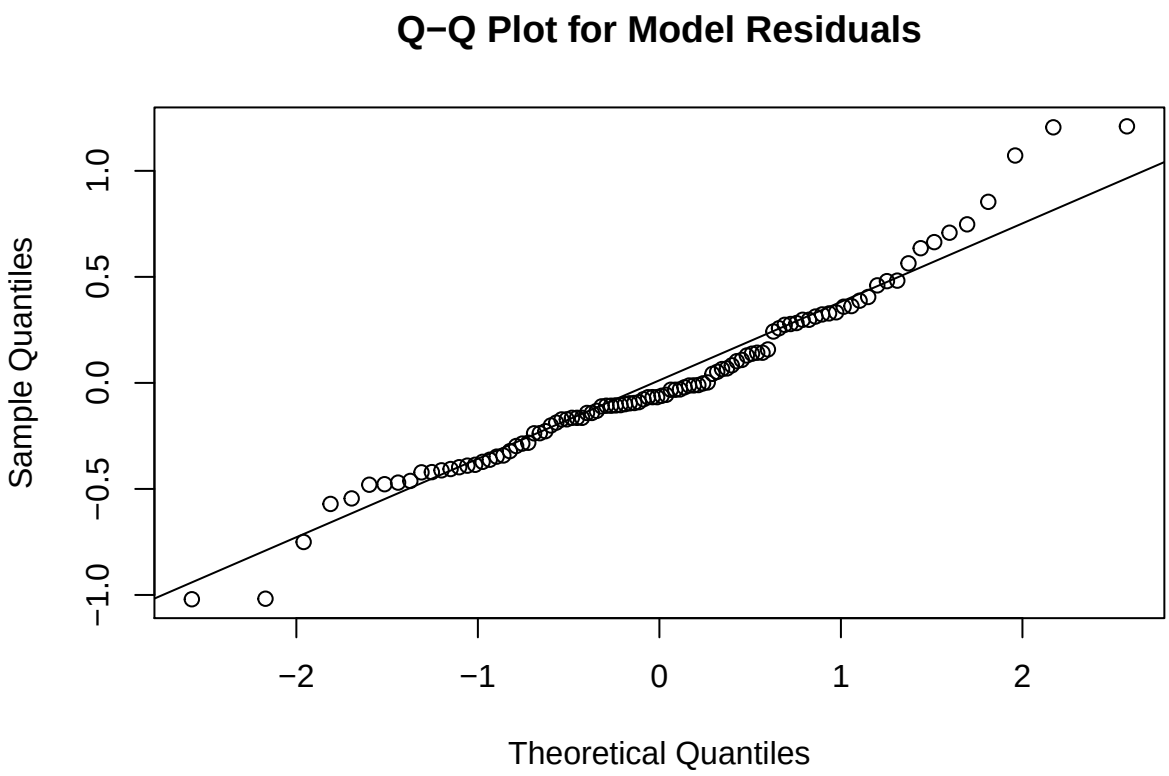
```
bptest(butter_no_int)
```

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

To further confirm heteroskedasticity, I performed the Breusch-Pagan test and got a p-value of .01154. Therefore, I can reject H_0 at $\alpha = .05$ and conclude that this model is not homoskedastic.

Next, I will check the model to see if the residuals are normally distributed.

```
qqnorm(butter_no_int$residuals, main = 'Q-Q Plot for Model Residuals')
qqline(butter_no_int$residuals)
```



Looking at the Q-Q Plot above, I believe that the residuals are mostly normally distributed. The residuals seem to follow a linear trend around the true Q-Q normal line, with some deviation at the lowest and highest theoretical quantiles. This may be due to the heteroskedasticity in the model.

To confirm normality in the residuals, I will perform a Shapiro-Wilk test. The null hypothesis for this test is that the data comes from a normal distribution.

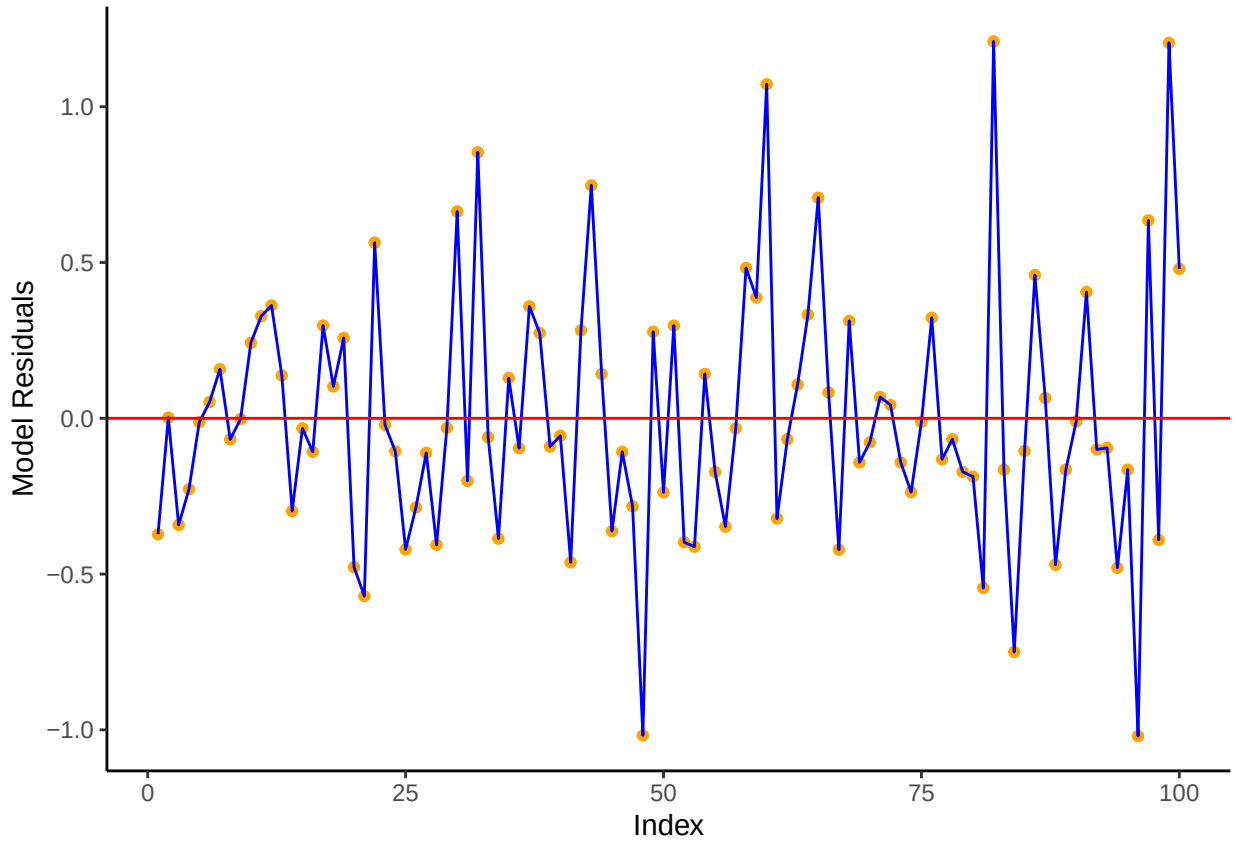
```
shapiro.test(butter_no_int$residuals)
```

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

With a p-value of .007168, I reject H_0 at $\alpha = .01$ and conclude that the residuals are not normally distributed.

The last model assumption I would like to check is if there is any correlation between the residuals.

```
res_df = data.frame(index = 1:100, res = butter_no_int$residuals)
ggplot(res_df, aes(x = index, y = res)) +
  geom_point(color = 'orange') +
  geom_line(color = 'blue') +
  geom_hline(yintercept = 0, color = 'red') +
  labs(x = 'Index', y = 'Model Residuals') +
  theme(panel.background = element_blank(),
        axis.line = element_line(color = 'black'))
```



Looking at the correlation plot above, there seems to be some autocorrelation between residuals and their indices. This is determined by looking for patterns of multiple positive residuals following one another or multiple negative residuals following one another.

To confirm this correlation, I will utilize the Durbin-Watson test. The null hypothesis for this test is that there is no correlation between the models residuals.

```
dwtest(butter_no_int)
```

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

With a p-value of .4531, I fail to reject H_0 and confirm that the errors are not correlated at a significant level.

Overall, I have checked the model assumptions of:

1. Constant Error Variance (Homoskedastity)
2. Residuals Being Normally Distributed
3. Errors Have a Correlation of 0

Of these assumptions, our model violates assumptions 1 and 2. Therefore, this model has not met all assumptions and should not be deemed as a useful model. Some steps may be taken to remedy these

assumption violations, such as a box-cox transformation to tend the residuals towards a normal distribution or by adding a weight to the model to create homoskedasticity.

E.

Is the best breed in terms of butterfat content clearly superior to the second best breed?

The best breed is the Jersey breed with a mean butterfat content by percentage of 5.29%. The second best breed is the Guernsey breed with a mean butterfat content by percentage of 4.95% (see plot in part a). In order to test if the Jersey breed is clearly superior, I will test to see if these means are significantly different.

```
top2_breeds = butterfat %>% filter(Breed == 'Jersey' | Breed == 'Guernsey') %>% select(Butterfat, Breed
t.test(Butterfat ~ Breed, data = top2_breeds)
```

```
##
##  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 output above on the Welch's Two Sample t-test, I fail to reject H_0 that the difference between these means is equal to zero. Therefore, I cannot say that the best breed in terms of butterfat content is clearly superior to the second best breed.

Problem 2.

The morley data can be seen as a randomized block experiment with Run as the treatment factor and Expt as the blocking factor. Is there a difference between runs and what efficiency is gained by blocking?

```
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
```

We can see that the run variable is not significant while the experiment number is. This makes sense as the run number contains no relevant information regarding the experiment, it is just the label of the experiment.

To test if there is a difference between runs, I will run an ANOVA test.

```r
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 output above, we fail to reject H_0 and conclude that there is not a difference between run groups.

Utilizing the experiment number as a blocking factor, we gain efficiency of this experiment as we are able to accurately test the difference between the run groups by comparing between experiment numbers. If we only performed one experiment with multiple more runs, we would not be able to accurately test difference between means of the speed.

Problem 3.

The alfalfa data arise from a Latin square design where the treatment factor is inculum and the blocking factors are shade and irrigation. Test the significance of the effects and determine which levels of the treatment factor are significantly different.

```r
anova(lm(yield ~ ., data = alfalfa))
```

```
## 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
```

Looking at the ANOVA plot above, we are able to test the significance of the effects on the yield of alfalfa. I can see that both shade and inoculum are statistically significant at $\alpha = .01$, while irrigation is not. This tells me that shade and seed incolum may be of greater importance to yield compared to the amount of irrigation.

Next, I want to determine which levels of the treatment factor are statistically significant.

```r
AB = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('A', 'B')))
AC = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('A', 'C')))
AD = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('A', 'D')))
AE = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('A', 'E')))
BC = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('B', 'C')))
BD = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('B', 'D')))
BE = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('B', 'E')))
CD = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('C', 'D')))
CE = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('C', 'E')))
```

```r
DE = t.test(yield ~ inoculum, data = alfalfa %>% filter(inoculum %in% c('D', 'E')))

results = data.frame(Groups = c('AB', 'AC', 'AD', 'AE',
                                'BC', 'BD', 'BE', 'CD',
                                'CE', 'DE'),
                     t.Stat = c(AB$statistic, AC$statistic, AD$statistic, AE$statistic,
                                BC$statistic, BD$statistic, BE$statistic, CD$statistic,
                                CE$statistic, DE$statistic),
                     p.value = c(AB$p.value, AC$p.value, AD$p.value, AE$p.value,
                                BC$p.value, BD$p.value, BE$p.value, CD$p.value,
                                CE$p.value, DE$p.value))
results
```

```
##    Groups      t.Stat    p.value
## 1      AB  0.68210451 0.51533799
## 2      AC  0.05847053 0.95520836
## 3      AD  0.60262283 0.56874433
## 4      AE  3.62265707 0.01428077
## 5      BC -0.43986514 0.67312908
## 6      BD  0.09268475 0.92879651
## 7      BE  3.11455707 0.02174571
## 8      CD  0.44740484 0.66648277
## 9      CE  3.13621057 0.01589613
## 10     DE  2.70989970 0.02867233
```

For treatment levels to be considered significantly different, I will use a significance level of $\alpha = .05$.

- For groups A and B, the yield **is not** significantly different between groups.
- For groups A and C, the yield **is not** significantly different between groups.
- For groups A and D, the yield **is not** significantly different between groups.
- For groups A and E, the yield **is**significantly different between groups.
- For groups B and C, the yield **is not** significantly different between groups.
- For groups B and D, the yield **is not** significantly different between groups.
- For groups B and E, the yield **is** significantly different between groups.
- For groups C and D, the yield **is not** significantly different between groups.
- For groups C and E, the yield **is** significantly different between groups.
- For groups D and E, the yield **is** significantly different between groups.

Overall, groups A and E, B and E, C and E, and D and E are significantly different from one another. Note that all four seed types A-D are significantly different than group E and that group E is a constant. This tells me that possibly only one seed type is needed compared to seed E, as there are no significant differences between groups A-D.