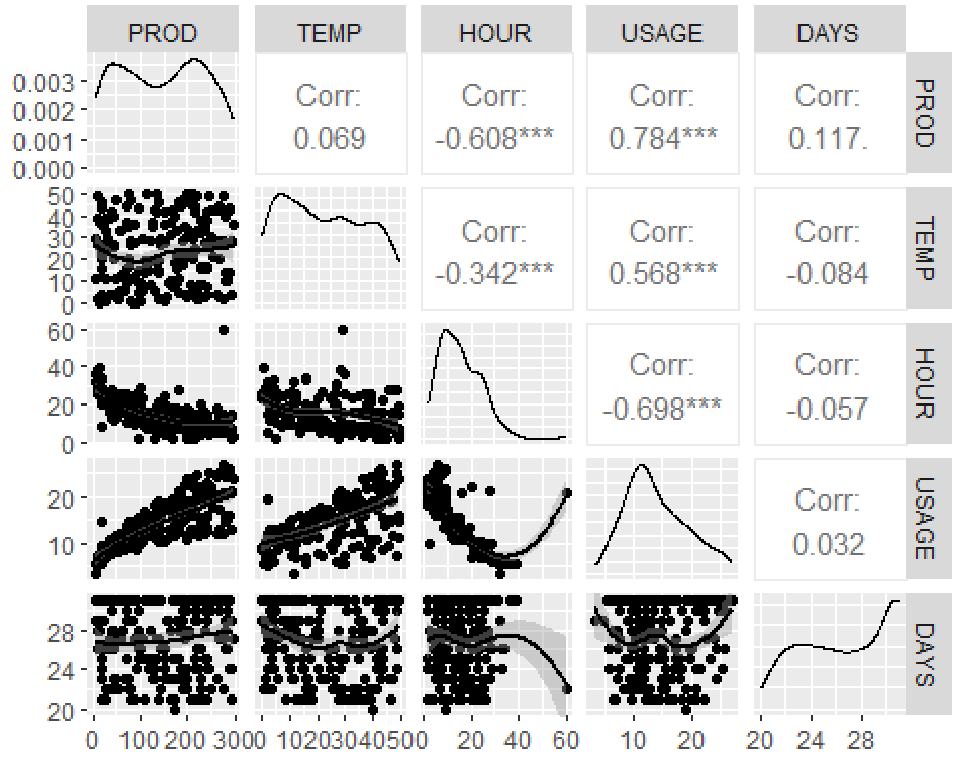
603 Assignment Three

Maria Delgado

2024-04-03

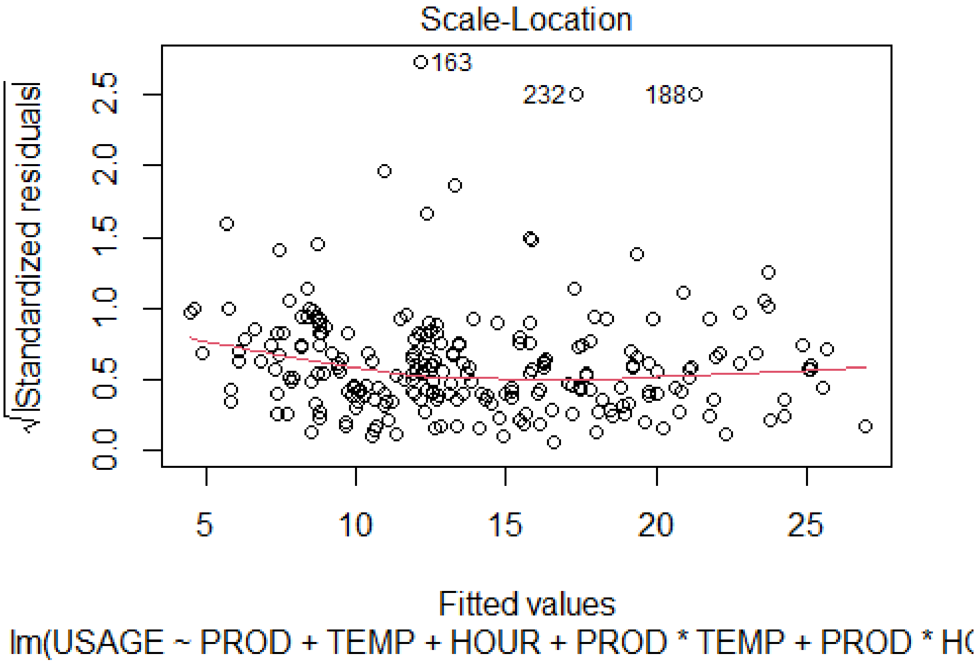
```
library(olsrr)

## Warning: package 'olsrr' was built under R version 4.3.3

##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers

library(ggplot2)
library(GGally)

## Warning: package 'GGally' was built under R version 4.3.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

library(mctest)
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.3.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.3.3

##
## Attaching package: 'zoo'

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

library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':
##
##     cement
```

```
library(cowplot)

## Warning: package 'cowplot' was built under R version 4.3.3

library(agricolae)

## Warning: package 'agricolae' was built under R version 4.3.3

library(FSA)

## Warning: package 'FSA' was built under R version 4.3.3

## Registered S3 methods overwritten by 'FSA':
##   method       from
##   confint.boot car
##   hist.boot    car

## ## FSA v0.9.5. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.

                            #QUESTION ONE PART A

water <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment Three/
water.csv")
head(water)

##    PROD TEMP HOUR USAGE DAYS
## 1 171.3 39.7  9.5  19.0   20
## 2  19.4 16.0 20.0   6.6   21
## 3  18.7 12.1 26.0   6.7   21
## 4  25.6 39.0 24.0   9.5   21
## 5  25.6 39.0 23.0   9.5   21
## 6 139.2 14.3 16.0  12.2   21

water_interaction_model = lm(USAGE~PROD+TEMP+HOUR+PROD*TEMP+PROD*HOUR, data=w
ater)

ggpairs(water,lower = list(continuous = "smooth_loess", combo = "facethist",
discrete = "facetbar", na = "na"))
```



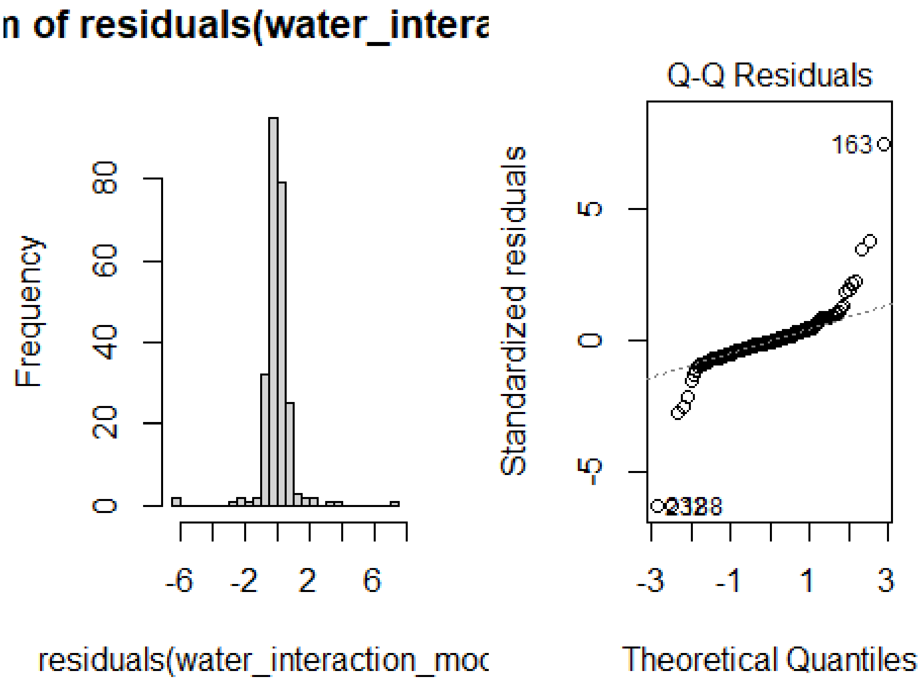
```
water_firstordermodel = lm(USAGE~PROD+TEMP+HOUR, data=water)
imcdiag(water_firstordermodel, method="VIF")

##
## Call:
## imcdiag(mod = water_firstordermodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##         VIF detection
## PROD 1.6452         0
## TEMP 1.1738         0
## HOUR 1.8548         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================

#Multicollinearity is a condition where predictor variables in a regression m
odel are highly correlated, leading to difficulties in interpretation, unstab
le estimates, and decreased model reliability. The VIF test is a commonly use
d method to detect multicollinearity by examining the inflation of variance i
```

```
n the coefficient estimates. From the output above: collinearity is not detec
ted as VIF is less than 5.

                               #QUESTION ONE PART B

plot(water_interaction_model, which = 3)
```



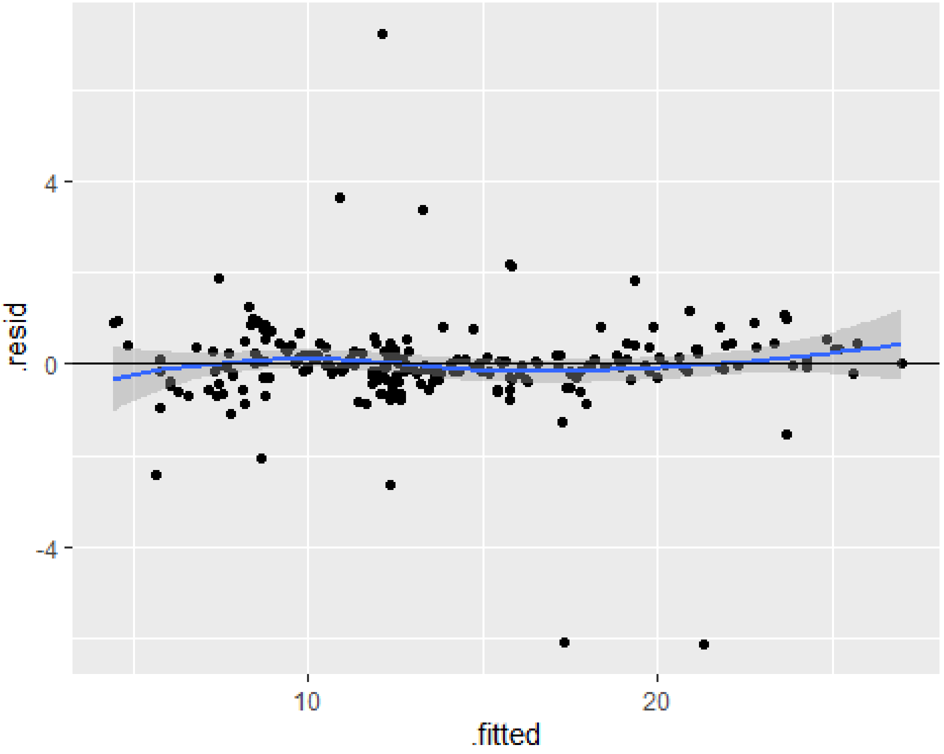
```
#Based on the residual plot, it appears that the residuals are evenly distrib
uted around zero, showing consistent variance and lacking any discernible pat
tern.This observation suggests that there may not be an issue with heterosced
asticity.

#H0: heteroscedasticity is not present (homoscedasticity) Ha: heteroscedastic
ity is present

bptest(water_interaction_model)

##
##  studentized Breusch-Pagan test
##
## data:  water_interaction_model
## BP = 2.0057, df = 5, p-value = 0.8484

#The Breusch-Pagan test is utilized to examine whether the variance of residu
als in a regression model exhibits homoscedasticity (constant variance) or he
teroscedasticity (varying variance) concerning the predictor variables.The re
sults present the outcome of the Breusch-Pagan test conducted on the interact
```

```
ion model. With a p-value of 0.8484, which exceeds the conventional significa
nce level of 0.05, we fail to reject the null hypothesis. Hence, the test off
ers evidence supporting the absence of heteroscedasticity, signifying constan
t variance (homoscedasticity).

                           #QUESTION ONE PART C

# (H0) states that the sample data are significantly normally distributed, an
d the alternative hypothesis (Ha) states that the sample data are not signifi
cantly normally distributed.
shapiro.test(residuals(water_interaction_model))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(water_interaction_model)
## W = 0.67655, p-value < 2.2e-16

#The p-value obtained (2.2e-16) is extremely small, indicating strong evidenc
e against the null hypothesis.
par(mfrow=c(1,2))
hist(residuals(water_interaction_model), breaks = 24)
plot(water_interaction_model, which=2)
```

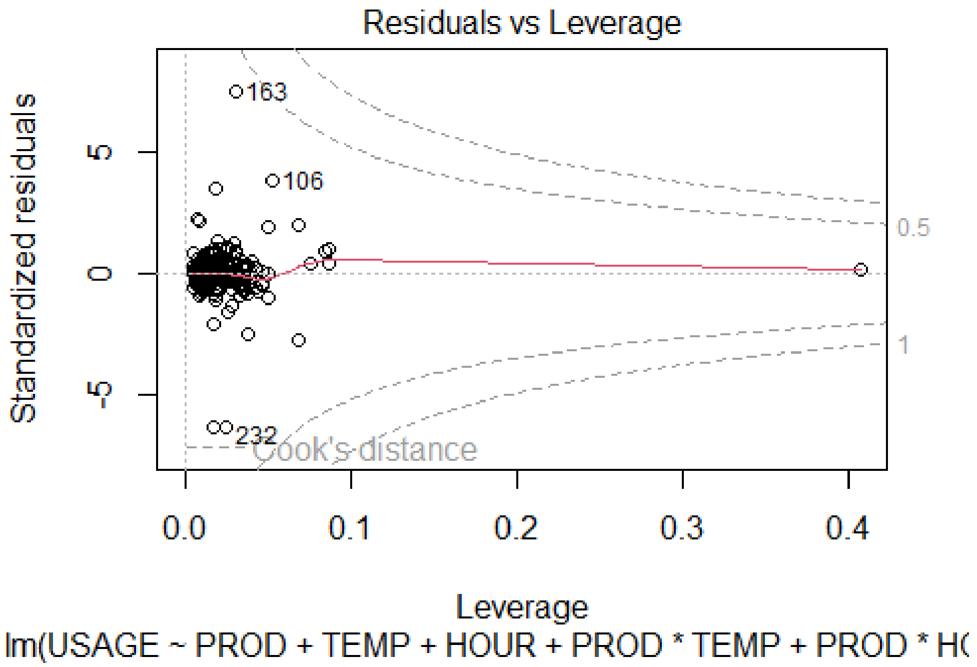


```
##The histogram displays a distribution that is more peaked (leptokurtic) com
pared to a normal distribution, with heavier tails. The Q-Q plot further conf
irms this observation, as it shows deviations from the straight line on both
```

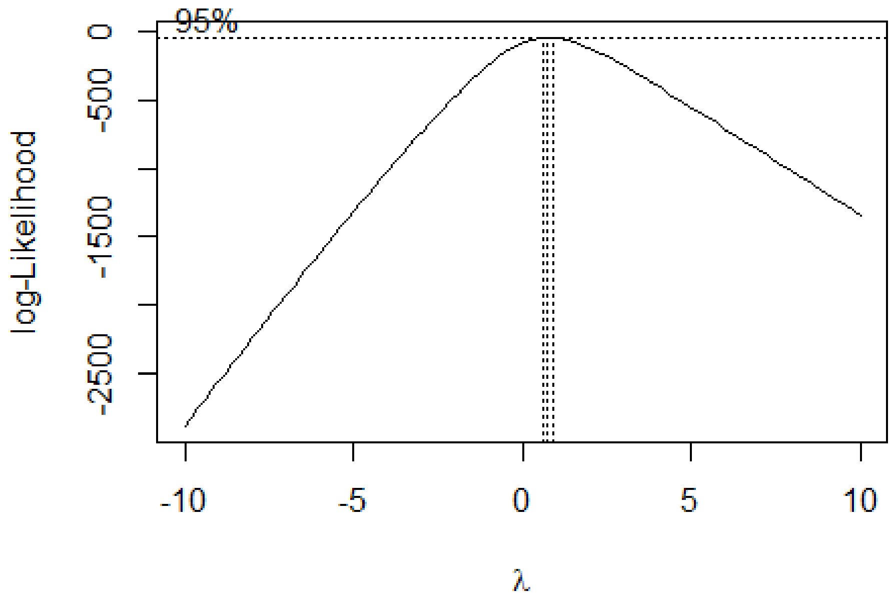
```
ends. These findings indicate that the residuals do not follow a normal distr
ibution.

                              #QUESTION ONE PART D

ggplot(water_interaction_model, aes(x=.fitted, y=.resid)) +
  geom_point() + geom_smooth()+
  geom_hline(yintercept = 0)

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



```
#There seems to be  a  consistent deviation from the zero line in one directi
on (e.g., the residuals consistently above or below zero), it might indicate
a problem with the linearity assumption.

                          #QUESTION ONE PART E

plot(water_interaction_model,which=c(5))
```



```
#From the plot, there does not appear to be influential outlirs.



                       #QUESTION ONE PART F

#One of the assumptions, namely homoscedasticity, is not met. To address this
, we can apply a Box-Cox transformation. If this proves ineffective, we may c
onsider introducing polynomial terms for the most highly correlated variable
with the response.

library(MASS)
bc = boxcox(water_interaction_model,lambda=seq(-10,10))
```



```
bestlambda = bc$x[which(bc$y==max(bc$y))]
water_bcmodel = lm((((USAGE^bestlambda)-1)/bestlambda) ~ PROD+TEMP+HOUR+PROD*
TEMP+PROD*HOUR, data=water)
summary(water_bcmodel)
```

```
##
## Call:
## lm(formula = (((USAGE^bestlambda) - 1)/bestlambda) ~ PROD + TEMP +
##     HOUR + PROD * TEMP + PROD * HOUR, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65119 -0.14141 -0.00348  0.12792  3.14516
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.275e+00  2.258e-01  32.224   <2e-16 ***
## PROD        -1.914e-03  1.075e-03  -1.780   0.0764 .
## TEMP         1.485e-03  4.133e-03   0.359   0.7198
## HOUR        -1.236e-01  7.577e-03 -16.306   <2e-16 ***
## PROD:TEMP    4.733e-04  2.337e-05  20.256   <2e-16 ***
## PROD:HOUR    4.611e-04  3.614e-05  12.760   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4603 on 243 degrees of freedom
## Multiple R-squared:  0.9645, Adjusted R-squared:  0.9638 
## F-statistic:  1322 on 5 and 243 DF,  p-value: < 2.2e-16
```

```
cat("The model after Box-Cox transformation has Breusch-Pagan test p-value ="
,
    bptest(water_bcmodel)$p.value,
    "\n",
    "The reduced model has Breusch-Pagan test p-value =",
    bptest(water_interaction_model)$p.value,
    "\n\n")
```

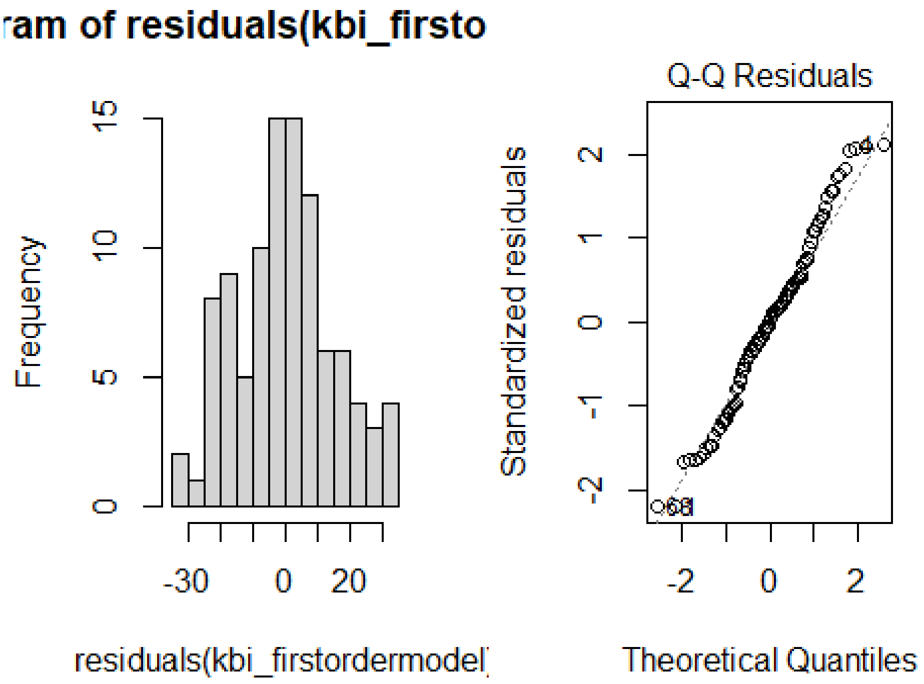
```
## The model after Box-Cox transformation has Breusch-Pagan test p-value = 0.
9155836 
##  The reduced model has Breusch-Pagan test p-value = 0.8483625
```

```
cat("The model after Box-Cox transformation has Shapiro-Wilk p-value =",
    shapiro.test(residuals(water_bcmodel))$p.value,
    "\n",
    "The reduced model has Shapiro-Wilk p-value =",
    shapiro.test(residuals(water_interaction_model))$p.value)
```

```
## The model after Box-Cox transformation has Shapiro-Wilk p-value = 1.027677
e-19 
##  The reduced model has Shapiro-Wilk p-value = 1.490894e-21
```

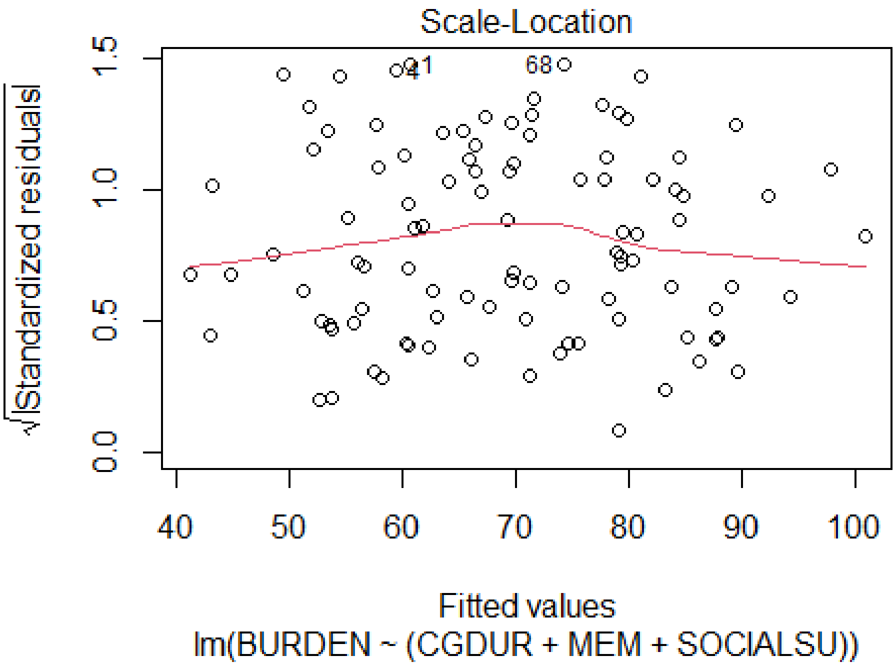
```
                              #QUESTION Two PART A

KBI <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment Three/KB
I.csv")
kbi_firstordermodel=lm(BURDEN~(CGDUR+ MEM +SOCIALSU) , data=KBI)
summary(kbi_firstordermodel)
```

```
## 
## Call:
## lm(formula = BURDEN ~ (CGDUR + MEM + SOCIALSU), data = KBI)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -32.672  -9.977   0.367   7.774  31.523 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 115.53922   12.36816   9.342 3.86e-15 ***
## CGDUR         0.12168    0.06486   1.876   0.0637 .  
## MEM           0.56612    0.10232   5.533 2.73e-07 ***
## SOCIALSU     -0.49237    0.08930  -5.514 2.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.25 on 96 degrees of freedom
## Multiple R-squared:  0.4397, Adjusted R-squared:  0.4222
## F-statistic: 25.12 on 3 and 96 DF,  p-value: 4.433e-12

##Check Normality
par(mfrow=c(1,2))
hist(residuals(kbi_firstordermodel), breaks = 12)
plot(kbi_firstordermodel, which=2)
```



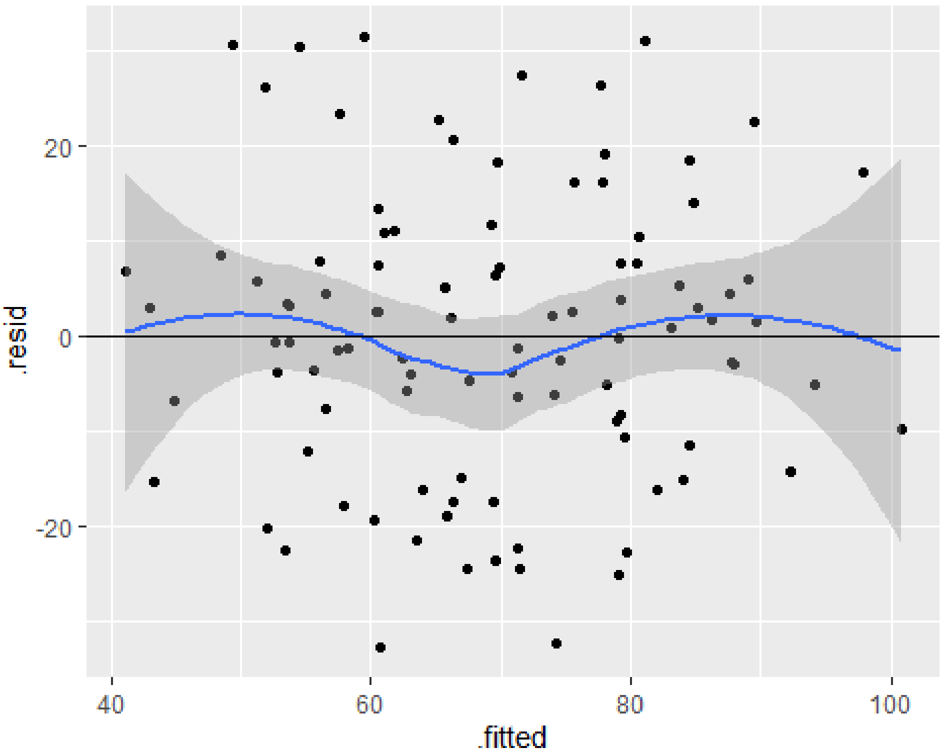
```
#The histogram displays a single peak, symmetrical distribution with no appar
ent skewness, suggesting the sample conforms to a normal distribution. Additi
onally, the qq-plot indicates that the residuals closely align with the refer
ence line, with negligible deviations in either tail.

#Shapiro-Wilk normality test: Null Hypothesis (H0): The sample data exhibit s
ignificant normal distribution.Alternative Hypothesis (Ha): The sample data d
o not exhibit significant normal distribution.

shapiro.test(residuals(kbi_firstordermodel))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(kbi_firstordermodel)
## W = 0.98407, p-value = 0.2716

#The Shapiro-Wilk normality test additionally verifies that the residuals fol
low a normal distribution, as indicated by the p-value of 0.2716, which excee
```

```
ds the significance level of 0.05, leading to acceptance of the null hypothes
is.

#Check Homoscedasticity
plot(kbi_firstordermodel, which = 3)
```

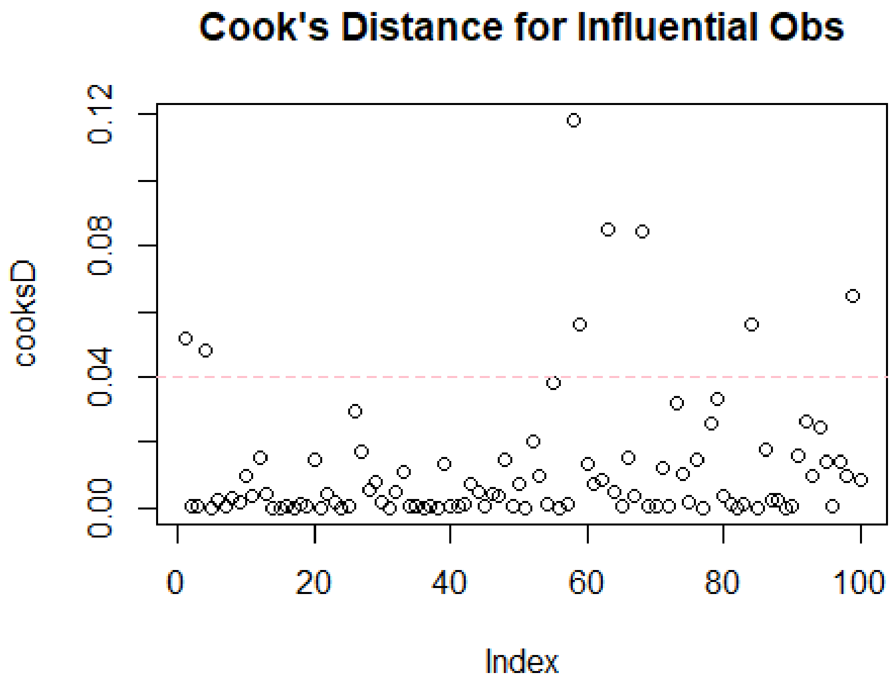


```
#It is evident that there is no discernible pattern in the variability of the
residuals across the observed range of values. The plot depicts an almost hor
izontal line, indicating that the spread of the residuals remains relatively
consistent regardless of the magnitude of the measured values. This observati
on aligns with the concept of homoscedasticity

#Breusch-Pagan test: Null Hypothesis (H0): Absence of heteroscedasticity (hom
oscedasticity). Alternative Hypothesis (Ha): Presence of heteroscedasticity.
bptest(kbi_firstordermodel)

##
##  studentized Breusch-Pagan test
##
## data:  kbi_firstordermodel
## BP = 2.0208, df = 3, p-value = 0.5681

#The output presents the results of the Breusch-Pagan test conducted for the
first-order model. With a p-value of 0.5681, surpassing the significance thre
shold of 0.05, we accept the null hypothesis. Thus, the test indicates no evi
dence of heteroscedasticity, suggesting constant variance (homoscedasticity).
```

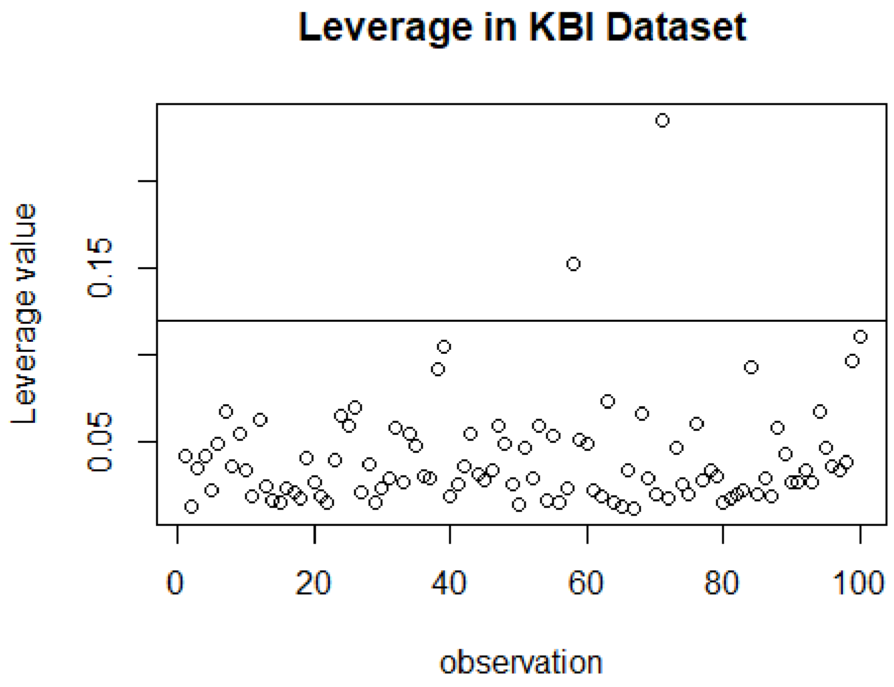
```
#Check Linearity
ggplot(kbi_firstordermodel, aes(x=.fitted, y=.resid)) +
  geom_point() + geom_smooth()+
  geom_hline(yintercept = 0)

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



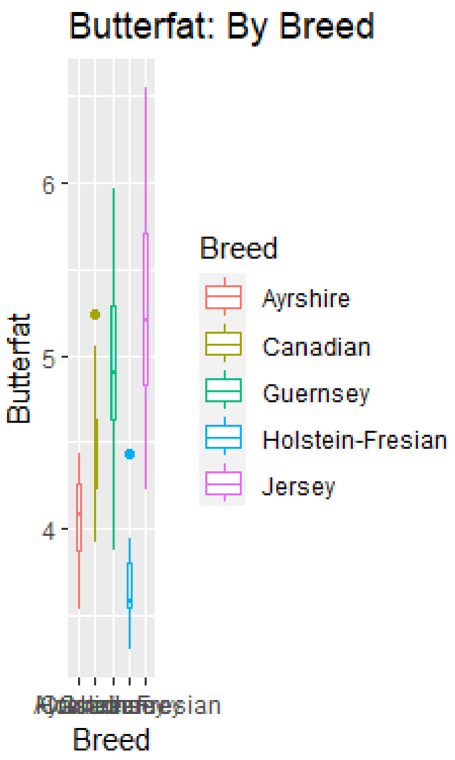
```
#Observing the residuals plot for the first-order regression, it is evident t
hat there is no discernible pattern or trend present in the residuals.

                                #QUESTION TWO PART B

# Detecting influential points with Cook’s distance

# find Cook's distance for each observation in the dataset
cooksD <- cooks.distance(kbi_firstordermodel)

# Plot Cook's Distance with a horizontal line at 4/n to see which observation
s
# exceed this threshold
n <- nrow(KBI)
plot(cooksD, main = "Cook's Distance for Influential Obs")
abline(h = 4/n, lty = 2, col = "pink") # add cutoff line
```

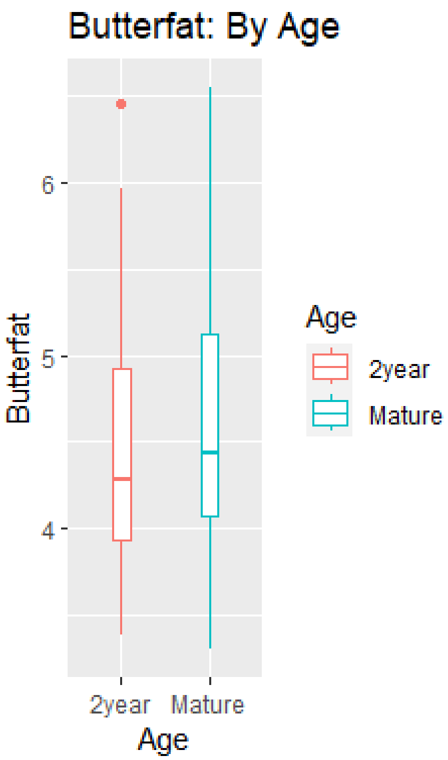


```
# identify influential points
influential_obs <- as.numeric(names(cooksD)[(cooksD > (8/n))])

# define new data frame with influential points removed
KBI_outliers_removed <- KBI[-influential_obs, ]

# Identify outliers with Leverage method
n <- nrow(KBI)
p <- length(coef(kbi_firstordermodel))
lev <- hatvalues(kbi_firstordermodel)
outlier3p <- lev[lev > (3 * p / n)]
print(outlier3p)

##        58        71
## 0.1527990 0.2352185

##        58        71
## 0.1527990 0.2352185
plot(rownames(KBI), lev, main = "Leverage in KBI Dataset", xlab = "observatio
n", ylab = "Leverage value")
abline(h = 3 * p / n)
```

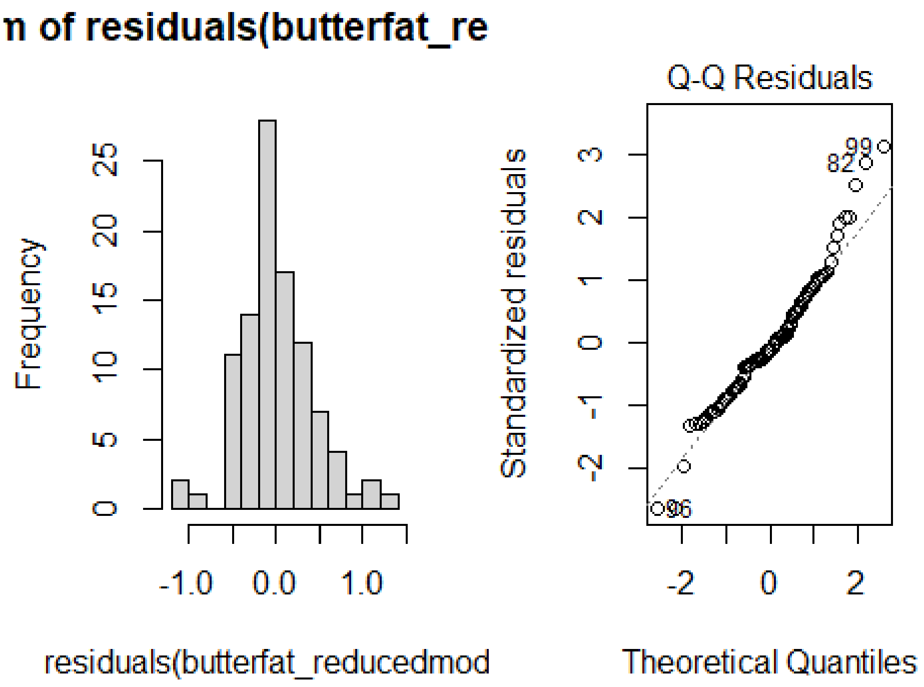


```
# remove outliers calculated in previous step
KBI_outliers_removed <- KBI[-c(58, 71), ]

#The identified outliers, based on leverage values exceeding the threshold, a
re:

#Observation 58 with a leverage value of approximately 0.1527990
#Observation 71 with a leverage value of approximately 0.2352185

                               #QUESTION TWO PART C
kbi_no_utliers_model=lm(BURDEN~(CGDUR+ MEM +SOCIALSU) , data=KBI_outliers_rem
oved)

#Comparing two models
cat("The model from Assignment 2 Problem 4(c) has adjusted r-squared =",
    summary(kbi_firstordermodel)$adj.r.squared,
    "\n",
    "The model with outliers removed has adjusted r-squared =",
    summary(kbi_no_utliers_model)$adj.r.squared,
    "\n\n")

## The model from Assignment 2 Problem 4(c) has adjusted r-squared = 0.422220
7
##  The model with outliers removed has adjusted r-squared = 0.4299905

cat("The model from Assignment 2 Problem 4(c) has RMSE =",
    sigma(kbi_firstordermodel),
    "\n",
    "The model with outliers removed has RMSE =",
    sigma(kbi_no_utliers_model))

## The model from Assignment 2 Problem 4(c) has RMSE = 15.24611
##  The model with outliers removed has RMSE = 15.19434

#After removing outliers from the dataset, the model exhibits an improved adj
usted R-squared value (0.43 compared to 0.422) and a lower root mean squared
error (RMSE). This indicates that eliminating just two influential points has
had a beneficial impact on the linear model's fit.

                       #QUESTION THREE PART A

butterfat <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment Th
ree/butterfat.csv")
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

library("cowplot")
p = ggplot(data=butterfat, aes(x=Breed, y=Butterfat, color = Breed)) +
        geom_boxplot(width=0.2) +
        ggtitle("Butterfat: By Breed")

q = ggplot(data=butterfat, aes(x=Age, y=Butterfat, color = Age)) +
        geom_boxplot(width=0.2) +
        ggtitle("Butterfat: By Age")

plot_grid(p, q, ncol = 2, nrow = 1)
```

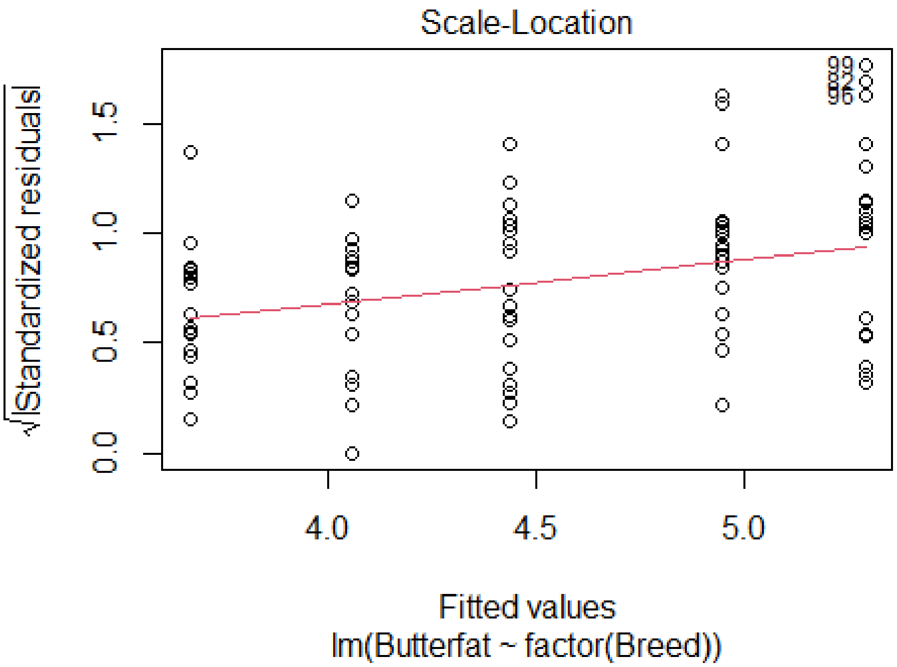




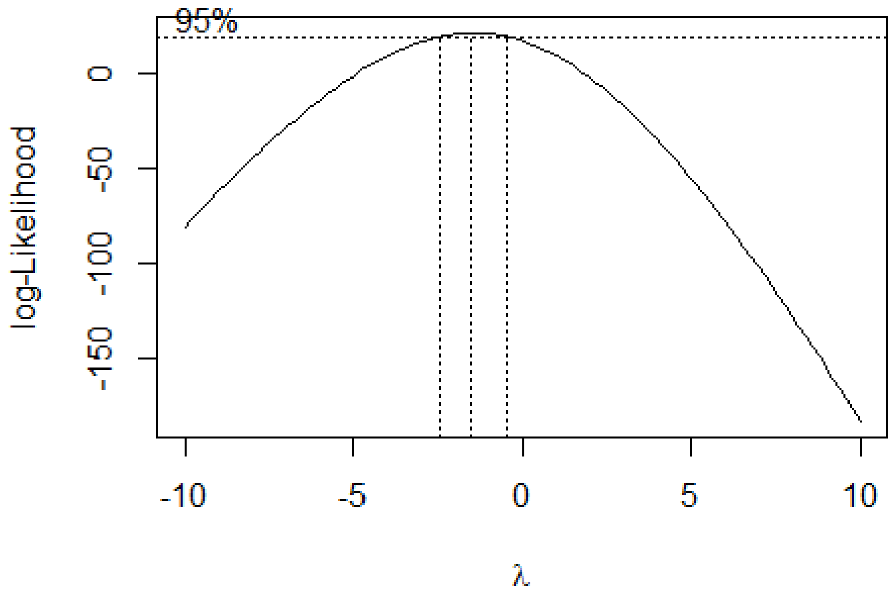
```
#Based on the boxplots displayed above, it's evident that the variability in
butterfat content remains consistently low across different ages, with the me
dians closely aligned. However, when considering the breed factor, the interq
uartile range (IQR) displays more variability. Upon visual examination, the H
olstein-Fresian breed exhibits the smallest IQR, whereas the Jersey breed dis
plays the greatest. Additionally, there appears to be notable separation betw
een the medians of the breeds.

                              #QUESTION THREE PART B
butterfat_firstordermodel = lm(Butterfat~factor(Age)+factor(Breed), data = bu
tterfat)
summary(butterfat_firstordermodel)

##
## Call:
## lm(formula = Butterfat ~ factor(Age) + factor(Breed), 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 ***
## factor(Age)Mature              0.10460    0.08276   1.264  0.20937
## factor(Breed)Canadian          0.37850    0.13085   2.893  0.00475 **
## factor(Breed)Guernsey          0.89000    0.13085   6.802 9.48e-10 ***
## factor(Breed)Holstein-Fresian -0.39050    0.13085  -2.984  0.00362 **
## factor(Breed)Jersey             1.23250    0.13085   9.419 3.16e-15 ***
## ---
## 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

#According to the summary of the first-order model, the Age factor does not s
how statistical significance, as indicated by a p-value of 0.21 (greater than
alpha = 0.05). Hence, it is advisable to retain only the breed variable in ou
r model.

butterfat_reducedmodel = lm(Butterfat~factor(Breed), data = butterfat)
summary(butterfat_reducedmodel)

##
## Call:
## lm(formula = Butterfat ~ factor(Breed), data = butterfat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07250 -0.27213 -0.05125  0.22363  1.25750
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.06000    0.09281  43.743  < 2e-16 ***
## factor(Breed)Canadian           0.37850    0.13126   2.884  0.00486 **
## factor(Breed)Guernsey           0.89000    0.13126   6.780 1.01e-09 ***
## factor(Breed)Holstein-Fresian -0.39050    0.13126  -2.975  0.00371 **
## factor(Breed)Jersey             1.23250    0.13126   9.390 3.33e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4151 on 95 degrees of freedom
## Multiple R-squared:  0.6771, Adjusted R-squared:  0.6635
## F-statistic:  49.8 on 4 and 95 DF,  p-value: < 2.2e-16

                         #QUESTION THREE PART C
#Check Normality
par(mfrow=c(1,2))
hist(residuals(butterfat_reducedmodel), breaks = 12)
plot(butterfat_reducedmodel, which=2)
```



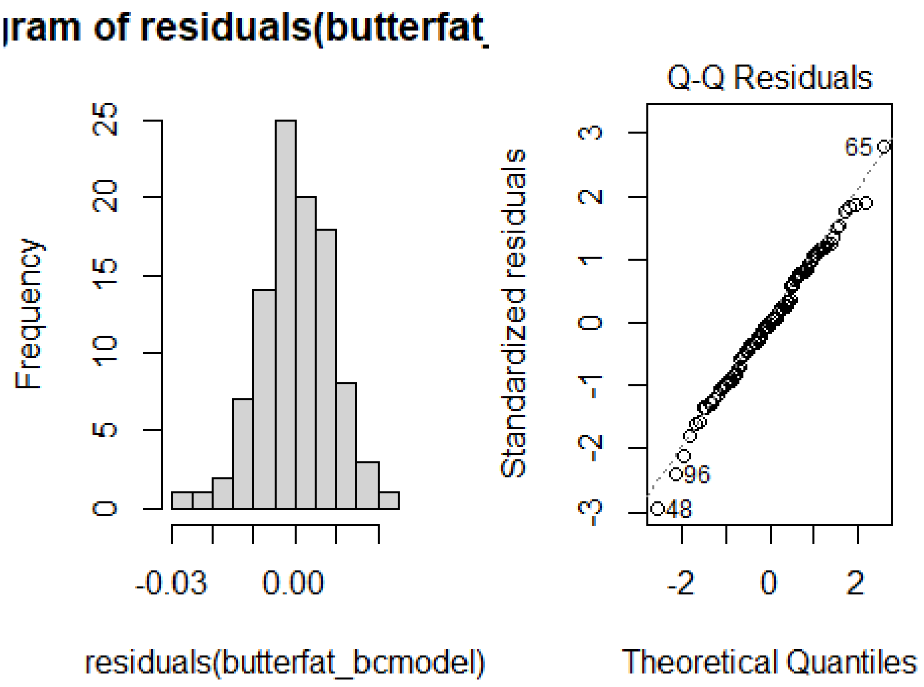
```
#The data appears to deviate from a normal distribution, as evidenced by the
histogram's lack of symmetry and right skew. This observation is further conf
irmed by the QQ-plot, where the residuals deviate noticeably from the middle
line.

#Shapiro-Wilk normality test:

#Null Hypothesis (H0): The sample data are normally distributed.
#Alternative Hypothesis (Ha): The sample data are not normally distributed.

shapiro.test(residuals(butterfat_reducedmodel))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(butterfat_reducedmodel)
## W = 0.96805, p-value = 0.01571

#The p-value of 0.01571 is below the significance level of 0.05, indicating t
hat the residuals do not follow a normal distribution (rejecting the null hyp
othesis).

#Check Homoscedasticity

plot(butterfat_reducedmodel, which = 3)
```



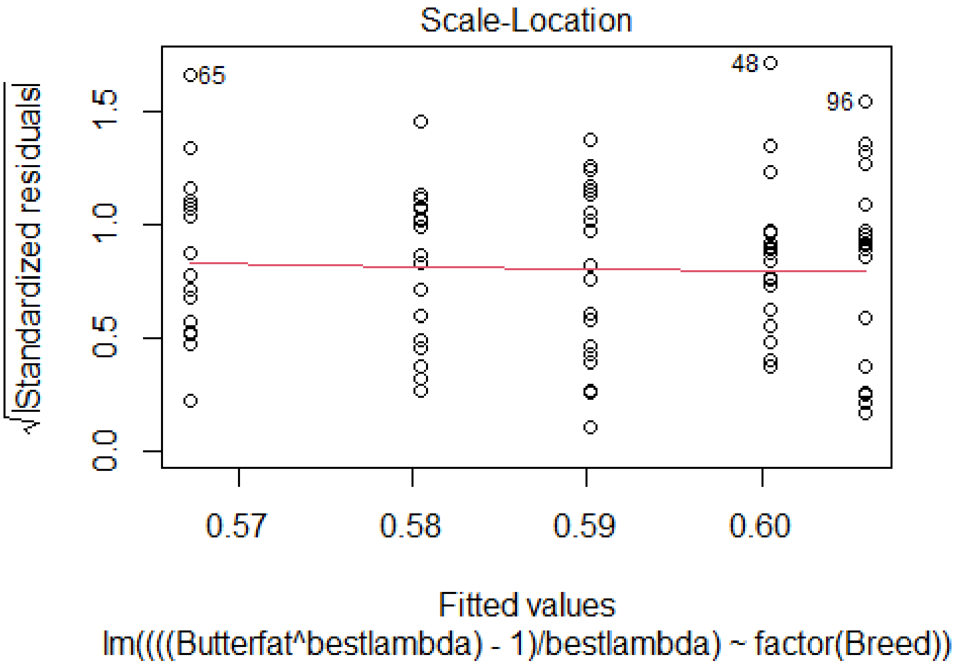
```
#In the Spread-Location plot, there seems to be a slight systematic variation
in the spread of the residuals across the measured values, noticeable by the
upward trend of the red line.

#Breusch-Pagan test:

#Null Hypothesis (H0): Homoscedasticity is present.
#Alternative Hypothesis (Ha): Heteroscedasticity is present.

bptest(butterfat_reducedmodel)

##
##  studentized Breusch-Pagan test
##
## data:  butterfat_reducedmodel
## BP = 13.389, df = 4, p-value = 0.009525

#Based on the Breusch-Pagan test result with a p-value of 0.0095, which is le
ss than 0.05, we reject the null hypothesis and conclude that heteroscedastic
ity is present.Furthermore, the diagnostic analysis indicates that our linear
model violates its assumptions. Specifically, the residuals do not exhibit no
rmal distribution and display varying variance.

                         #QUESTION THREE PART D

library("MASS")
bc = boxcox(butterfat_reducedmodel,lambda=seq(-10,10))
```



```
bestlambda = bc$x[which(bc$y==max(bc$y))]
butterfat_bcmodel = lm((((Butterfat^bestlambda)-1)/bestlambda) ~ factor(Breed
), data = butterfat)

cat("The model after Box-Cox transformation has adjusted r-squared =",
    summary(butterfat_bcmodel)$adj.r.squared,
    "\n",
    "The model with the interaction term has adjusted r-squared =",
    summary(butterfat_reducedmodel)$adj.r.squared,
    "\n\n")

## The model after Box-Cox transformation has adjusted r-squared = 0.7167454
##  The model with the interaction term has adjusted r-squared = 0.6635023

cat("The model after Box-Cox transformation has RMSE =",
    sigma(butterfat_bcmodel),
    "\n",
    "The reduced model has RMSE =",
    sigma(butterfat_reducedmodel))

## The model after Box-Cox transformation has RMSE = 0.008731683
##  The reduced model has RMSE = 0.415078

                        #QUESTION THREE PART E
par(mfrow=c(1,2))
hist(residuals(butterfat_bcmodel), breaks = 12)
plot(butterfat_bcmodel, which=2)
```



```
#Check Homoscedasticity
plot(butterfat_bcmodel, which = 3)
```

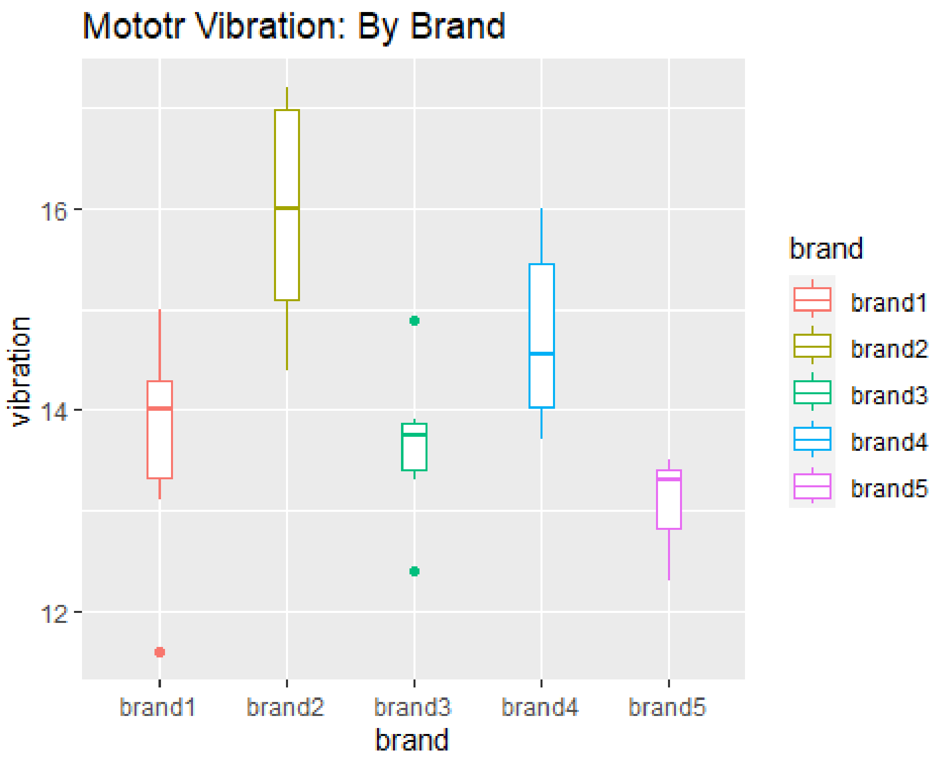


```
cat("The model after Box-Cox transformation has Breusch-Pagan test p-value ="
,
    bptest(butterfat_bcmodel)$p.value,
    "\n",
    "The other hand, the reduced model has Breusch-Pagan test p-value =",
    bptest(butterfat_reducedmodel)$p.value,
    "\n\n")

## The model after Box-Cox transformation has Breusch-Pagan test p-value = 0.
9729943
##  The other hand, the reduced model has Breusch-Pagan test p-value = 0.0095
24592

cat("The model after Box-Cox transformation has Shapiro-Wilk p-value =",
    shapiro.test(residuals(butterfat_bcmodel))$p.value,
    "\n",
    "The reduced model has Shapiro-Wilk p-value =",
    shapiro.test(residuals(butterfat_reducedmodel))$p.value)

## The model after Box-Cox transformation has Shapiro-Wilk p-value = 0.964305
5
##  The reduced model has Shapiro-Wilk p-value = 0.01570535

#Following the Box-Cox transformation, notable improvements are observed in t
he distribution of residuals. Specifically, the histogram displays symmetry,
and the residuals closely adhere to the theoretical line in the QQ-plot. Furt
hermore, the Shapiro-Wilk normality test confirms the normal distribution of
residuals, evidenced by a p-value of 0.96, which exceeds the threshold of 0.0
5. Regarding the Scale-Location plot, the red line maintains a horizontal tra
jectory, indicating consistent variance without discernible patterns. Additio
nally, the Breusch-Pagan test yields a p-value of 0.973, surpassing the signi
ficance level of 0.05. This implies that the null hypothesis is not rejected,
signifying the absence of heteroscedasticity. Consequently, the evidence sugg
ests the presence of homoscedasticity.

                                             #QUESTION FOUR PART A

vibration <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment Th
ree/vibration.csv")
head(vibration)

##   vibration  brand
## 1      13.1 brand1
## 2      15.0 brand1
## 3      14.0 brand1
## 4      14.4 brand1
## 5      14.0 brand1
## 6      11.6 brand1

##A
#The motor vibration, measured in microns, serves as the response variable, w
ith each motor representing an individual experimental unit.
```

```
                     #QUESTION FOUR PART B
##The treatment is the brand of bearing, there are five brands so there will
be five treatment.

                           #QUESTION FOUR PART C
#H0: Testing whether all treatment means are equal, implying that the impact
of the brand of bearing on motor vibration is zero.Ha: Asserting that at leas
t one effect is not zero.

                           #QUESTION FOUR PART D
anova(lm(vibration~brand, data=vibration))
```

```
## Analysis of Variance Table
##
## Response: vibration
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## brand      4 30.855  7.7138   8.444 0.0001871 ***
## Residuals 25 22.838  0.9135                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The ANOVA table shows that the calculated F-value (Fcal) is 8.44 with a corr
esponding p-value of 0.0001871, which is less than the chosen significance le
vel (alpha = 0.05). Therefore, we reject the null hypothesis. This indicates
strong evidence to suggest that the average motor vibration differs significa
ntly across the brands of bearings at the chosen significance level of 0.05.
Also: Total degrees of freedom (dfs): 29; Total sum of squares (SST): 53.693.

                                        #QUESTION FOUR PART E
ggplot(data=vibration, aes(x=brand, y=vibration, color = brand)) +
  geom_boxplot(width=0.2) +
  ggtitle("Mototr Vibration: By Brand")
```



```
#Based on the plot presented, it appears that the variability among groups re
presenting different brands exceeds the variability within each group. While
there are outliers identified for brand 3 (two outliers) and brand 1 (one out
lier), their presence does not appear to significantly distort the overall pa
ttern. Although outliers can elevate the variance within a group, the substan
tial variance observed between groups suggests that these outliers do not exe
rt a substantial influence on the overall analysis.

                                         #QUESTION FOUR PART F

mean(vibration$vibration[vibration$brand=="brand1"])

## [1] 13.68333

mean(vibration$vibration[vibration$brand=="brand2"])

## [1] 15.95

mean(vibration$vibration[vibration$brand=="brand3"])

## [1] 13.66667

mean(vibration$vibration[vibration$brand=="brand4"])

## [1] 14.73333

mean(vibration$vibration[vibration$brand=="brand5"])

## [1] 13.08333
```

```
#Unadjusted paired t-tests

pairwise.t.test(vibration$vibration, vibration$brand, p.adj = "none")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  vibration$vibration and vibration$brand
##
##        brand1  brand2  brand3  brand4
## brand2 0.00038 -       -       -
## brand3 0.97615 0.00035 -       -
## brand4 0.06865 0.03689 0.06464 -
## brand5 0.28728 2.3e-05 0.30058 0.00618
##
## P value adjustment method: none

#In conducting pairwise t-tests without adjustment for multiple comparisons, the analysis suggests three distinct groups among the brands based on vibration levels: (1, 3, 5), (3, 1, 4), and (2). Within each group, there are no significant differences in vibration levels (p > 0.05). However, a significant difference is observed between brands 4 and 5 (p < 0.05), indicating that they are not statistically indifferent

pairwise.t.test(vibration$vibration, vibration$brand, p.adj = "bonferroni")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  vibration$vibration and vibration$brand
##
##        brand1  brand2  brand3  brand4
## brand2 0.00376 -       -       -
## brand3 1.00000 0.00348 -       -
## brand4 0.68648 0.36891 0.64642 -
## brand5 1.00000 0.00023 1.00000 0.06184
##
## P value adjustment method: bonferroni

scheffe.test(aov(vibration~brand, data = vibration),"brand", group=TRUE,console=TRUE)

##
## Study: aov(vibration ~ brand, data = vibration) ~ "brand"
##
## Scheffe Test for vibration
##
## Mean Square Error  : 0.9135333
##
## brand,  means
##
```

```
##         vibration       std r         se  Min  Max    Q25   Q50    Q75
## brand1  13.68333 1.1940128 6 0.3901994 11.6 15.0 13.325 14.00 14.300
## brand2  15.95000 1.1674759 6 0.3901994 14.4 17.2 15.100 16.00 16.975
## brand3  13.66667 0.8164966 6 0.3901994 12.4 14.9 13.400 13.75 13.875
## brand4  14.73333 0.9395034 6 0.3901994 13.7 16.0 14.025 14.55 15.450
## brand5  13.08333 0.4792355 6 0.3901994 12.3 13.5 12.825 13.30 13.400
##
## Alpha: 0.05 ; DF Error: 25
## Critical Value of F: 2.75871
##
## Minimum Significant Difference: 1.833094
##
## Means with the same letter are not significantly different.
##
##         vibration groups
## brand2  15.95000      a
## brand4  14.73333     ab
## brand1  13.68333      b
## brand3  13.66667      b
## brand5  13.08333      b
```

```
CRD = aov(vibration~brand, data = vibration)
tvalue = qt(0.025, CRD$df.residual, lower.tail = F)
MSE = sum((CRD$residuals)^2/CRD$df.residual)
r = length(vibration$vibration[vibration$brand=="brand1"])
LSD = tvalue*sqrt(2*MSE/r)
LS = LSD.test(CRD, trt="brand")
LS
```

```
## $statistics
##     MSerror Df     Mean       CV  t.value      LSD
##   0.9135333 25 14.22333 6.719869 2.059539 1.136505
##
## $parameters
##          test p.ajusted name.t ntr alpha
##   Fisher-LSD      none  brand   5  0.05
##
## $means
##         vibration       std r         se      LCL      UCL  Min  Max    Q25
Q50
## brand1  13.68333 1.1940128 6 0.3901994 12.87970 14.48696 11.6 15.0 13.325
14.00
## brand2  15.95000 1.1674759 6 0.3901994 15.14637 16.75363 14.4 17.2 15.100
16.00
## brand3  13.66667 0.8164966 6 0.3901994 12.86304 14.47030 12.4 14.9 13.400
13.75
## brand4  14.73333 0.9395034 6 0.3901994 13.92970 15.53696 13.7 16.0 14.025
14.55
## brand5  13.08333 0.4792355 6 0.3901994 12.27970 13.88696 12.3 13.5 12.825
13.30
```

```
##           Q75
## brand1 14.300
## brand2 16.975
## brand3 13.875
## brand4 15.450
## brand5 13.400
##
## $comparison
## NULL
##
## $groups
##        vibration groups
## brand2  15.95000      a
## brand4  14.73333      b
## brand1  13.68333     bc
## brand3  13.66667     bc
## brand5  13.08333      c
##
## attr(,"class")
## [1] "group"
```

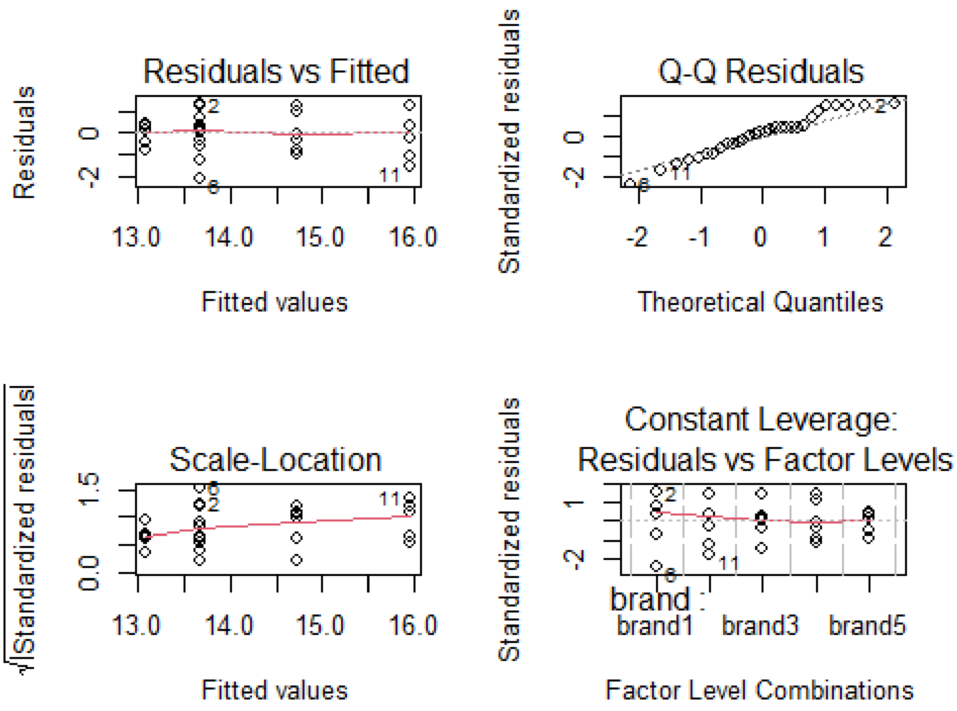
```
pairwise.t.test(vibration$vibration, vibration$brand, p.adj = "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  vibration$vibration and vibration$brand
##
##        brand1  brand2  brand3  brand4
## brand2 0.00313 -       -       -
## brand3 0.97615 0.00313 -       -
## brand4 0.32321 0.22134 0.32321 -
## brand5 0.86183 0.00023 0.86183 0.04329
##
## P value adjustment method: holm
```

```
TukeyHSD(aov(vibration~brand, data = vibration), conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = vibration ~ brand, data = vibration)
##
## $brand
##                      diff        lwr        upr     p adj
## brand2-brand1  2.26666667  0.6460270  3.8873064 0.0031588
## brand3-brand1 -0.01666667 -1.6373064  1.6039730 0.9999998
## brand4-brand1  1.05000000 -0.5706397  2.6706397 0.3418272
## brand5-brand1 -0.60000000 -2.2206397  1.0206397 0.8112981
## brand3-brand2 -2.28333333 -3.9039730 -0.6626936 0.0029299
## brand4-brand2 -1.21666667 -2.8373064  0.4039730 0.2106883
## brand5-brand2 -2.86666667 -4.4873064 -1.2460270 0.0002024
## brand4-brand3  1.06666667 -0.5539730  2.6873064 0.3268245
## brand5-brand3 -0.58333333 -2.2039730  1.0373064 0.8262091
## brand5-brand4 -1.65000000 -3.2706397 -0.0293603 0.0445279
```

```
#After conducting various pairwise t-tests with different adjustment methods,
the following summary of significant differences among the brands can be prov
ided:

#Scheffe and Bonferroni methods identify significant differences among the fo
llowing groups: (5, 3, 1, 4) and (4, 2).
#Newman-Keuls, Fisher, and unadjusted paired t-tests reveal significant diffe
rences between the following groups: (5, 3, 1) and (3, 1, 4), as well as (2).
#Tukey and Holm methods also indicate significant differences among the follo
wing groups: (5, 3, 1), (3, 1, 4), and (4, 2).

                        #QUESTION FOUR PART G

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



```
#Null hypothesis (H0): The data exhibits homoscedasticity, indicating the abs
ence of heteroscedasticity.
#Alternative hypothesis (Ha): The data displays heteroscedasticity.

bptest(CRD)
```

```
## 
##  studentized Breusch-Pagan test
## 
## data:  CRD
## BP = 4.5697, df = 4, p-value = 0.3344

bartlett.test(vibration~brand, data=vibration)

## 
##  Bartlett test of homogeneity of variances
## 
## data:  vibration by brand
## Bartlett's K-squared = 4.0967, df = 4, p-value = 0.3931

#Based on the results of the Breusch-Pagan test and the Bartlett test for hom
ogeneity of variances, we can conclude that there is no evidence to reject th
e null hypothesis (H0) that heteroscedasticity is not present. Therefore, the
data supports the assumption of homoscedasticity.

#Null hypothesis (H0): The sample data exhibit significant normal distributio
n. Alternative hypothesis (Ha): The sample data do not exhibit significant no
rmal distribution.
shapiro.test(residuals(CRD))

## 
##  Shapiro-Wilk normality test
## 
## data:  residuals(CRD)
## W = 0.95996, p-value = 0.3091

#In this case, with a p-value of 0.3091, which is greater than the significan
ce level of 0.05, we fail to reject the null hypothesis. There is no signific
ant evidence to suggest that the sample data are not normally distributed.

#A test for equality of treatment means
kruskal.test(vibration~brand, data=vibration)

## 
##  Kruskal-Wallis rank sum test
## 
## data:  vibration by brand
## Kruskal-Wallis chi-squared = 16.967, df = 4, p-value = 0.001961

#With a p-value of 0.001961, the Kruskal-Wallis test result is significant at
a significance level of 0.05, indicating that there are significant differenc
es in vibration levels among the brands.Following a significant Kruskal-Walli
s test, a post-hoc analysis, such as the Dunn test, can be performed to deter
mine which specific pairs of brands differ significantly from each other in t
erms of vibration levels.

dunnTest(vibration~brand,data=vibration,method="none")
```

```
## Warning: brand was coerced to a factor.

## Dunn (1964) Kruskal-Wallis multiple comparison

##   with no adjustment for p-values.

##         Comparison          Z      P.unadj        P.adj
## 1  brand1 - brand2 -2.3638672 0.0180852958 0.0180852958
## 2  brand1 - brand3  0.3447306 0.7302968900 0.7302968900
## 3  brand2 - brand3  2.7085978 0.0067568196 0.0067568196
## 4  brand1 - brand4 -1.1983493 0.2307810509 0.2307810509
## 5  brand2 - brand4  1.1655178 0.2438094438 0.2438094438
## 6  brand3 - brand4 -1.5430800 0.1228113776 0.1228113776
## 7  brand1 - brand5  1.4117540 0.1580224068 0.1580224068
## 8  brand2 - brand5  3.7756212 0.0001596094 0.0001596094
## 9  brand3 - brand5  1.0670234 0.2859612815 0.2859612815
## 10 brand4 - brand5  2.6101033 0.0090514878 0.0090514878

#Conclusion: All three assumptions (normality, constant variance, and indepen
dence) have been satisfied. The ANOVA F-statistics indicate rejection of the
null hypothesis (H0), suggesting that the means of the groups or the treatmen
t effects of different brand bearings are not equal.
```