Stat631 HW 3

Brandon Keck

Exercise 1.

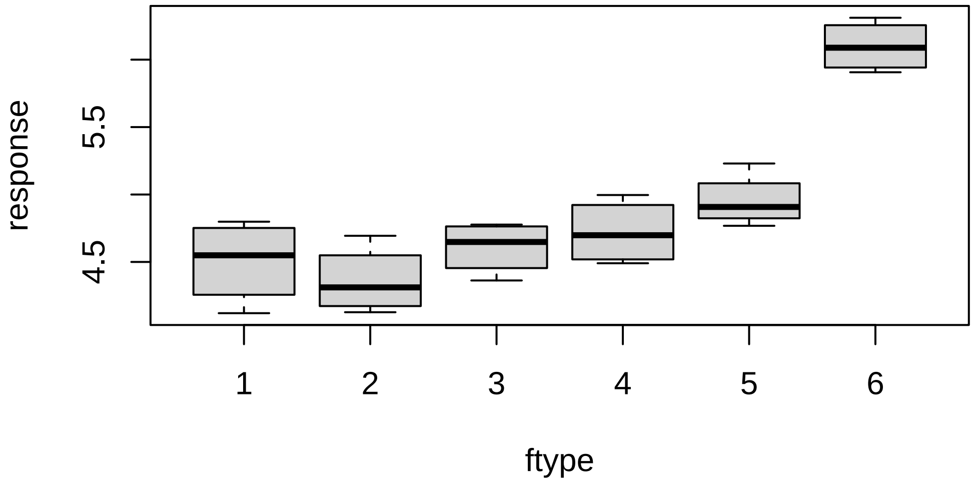
```
library(tidyverse)
library(car)
library(readr)

# Load in the dataset
CTAcars <- read.table("car.dat", header = TRUE)
head(CTAcars)
```

```
  type response
1    1 4.705398
2    1 4.120209
3    1 4.798508
4    1 4.393436
5    2 4.127314
6    2 4.404476
```

```
# Convert categorical data to a factor variable
CTAcars$ftype <- as.factor(CTAcars$type)
# head(CTAcars)
```

```
boxplot(response ~ ftype, data = CTAcars)
```



(a)

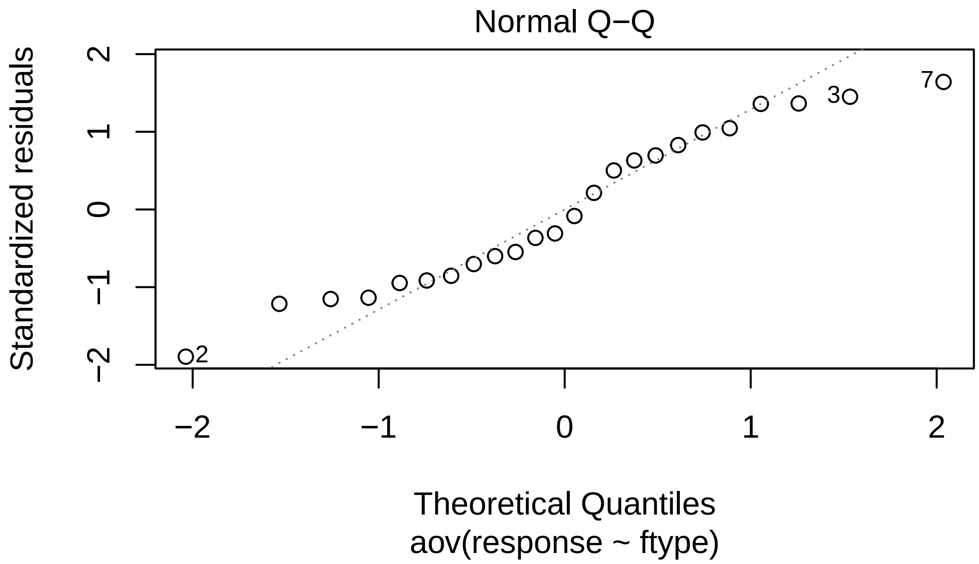
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

vs

H_A : At least one mean is different

- **Independence**: We can assume that the car manufacturers were selected at random and that the oil consumption of one car does not influence that of another. Therefore indepdence holds both within and between the groups.
- **Normality**: The residuals deviate from the QQ line, suggesting the presence of heavy tails. However, the Shapiro-Wilk test indicates that normality can still be assumed
- **Equal variance**: Both the residuals vs fitted values plot and the results of Levene's test suggest that the assumption of homoscadasticity is met.

```
full.model <- aov(response ~ ftype, data = CTAcars)
plot(full.model, which = 2)
```



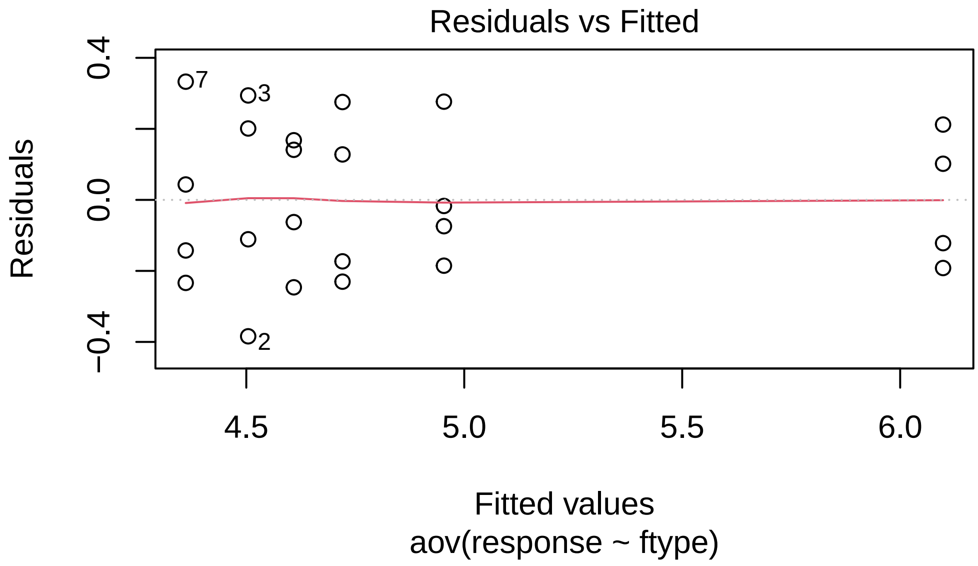
```
shapiro.test(resid(full.model))
```

```

	Shapiro-Wilk normality test

data:  resid(full.model)
W = 0.94535, p-value = 0.2144
```

```
plot(full.model, which = 1)
```



```
leveneTest(full.model)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5  0.6041 0.6977
      18
```

```
anova(full.model)
```

```
Analysis of Variance Table

Response: response
          Df Sum Sq Mean Sq F value    Pr(>F)
ftype      5 7.9958 1.59916  29.193 4.887e-08 ***
Residuals 18 0.9860 0.05478
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under the null hypothesis, we would expect the F value to be approximately 1. However, the observed F value is 29.193 with a p-value of 4.887e-08. Since the p-value is well below the significance level of $\alpha = 0.05$ we reject H_0 and conclude that there is evidence to suggest that at least one car manufacturer differs in oil consumption compared to the others.

(b)

(i)

Imported cars: Nissan (3) and Mercedes (6)

Domestic cars: Ford (1), Chevrolet (2), Lincoln (4), and Cadillac (5)

$C(-\frac{1}{4}, -\frac{1}{4}, \frac{1}{2}, -\frac{1}{4}, -\frac{1}{4}, \frac{1}{2})$

```r
library(dplyr)
trtmeans <- CTAcars %>%
  group_by(ftype) %>%
  summarise(means = mean(response)) %>%
  pull(means)
trtmeans
```

```
[1] 4.504388 4.361102 4.609048 4.720726 4.953424 6.098449
```

```r
fit <- aov(response ~ ftype, data = CTAcars)
ctrts <- c(-1/4, -1/4, 1/2, -1/4, -1/4, 1/2)

(est <- sum(ctrts * trtmeans))
```

```
[1] 0.7188387
```

```r
# Extract MSE from ANOVA table
MSE <- anova(fit) [2, 3]
# Sample size
ni <- 4
# Compute Standard Error (SE)
(se <- sqrt(MSE) * sqrt(sum(ctrts^2 / ni)))
```

```
[1] 0.1013458
```

```r
df_e <- anova(fit)[2, 1]
lower.ci <- est - qt(0.975, df_e) * se
upper.ci <- est + qt(0.975, df_e) * se
c(lower.ci, upper.ci)
```

```
[1] 0.5059191 0.9317583
```

```
(t_0 <- est/se)
```

```
[1] 7.09293
```

```
2*pt(-t_0, df_e)
```

```
[1] 1.301139e-06
```

```
SSw <- sum(ctrts*trtmeans)^2 / sum(ctrts^2/ni)
F0 <- SSw/MSE
pf(F0, 1, df_e, lower.tail = FALSE)
```

```
[1] 1.301139e-06
```

```
F0
```

```
[1] 50.30966
```

```
(t_0)^2
```

```
[1] 50.30966
```

The contrast estimate C = 0.7189 suggest that imported cars (Nissan and Mercedes) consume more oil per 100,000 miles than domestic cars (Ford, Chevrolet, Lincoln, and Cadillac). The 95% confidence interval (0.506, 0.932) does not include zaero, indicating that this difference is statistically significant. This provides evidence that, on average, imported cars exhibit higher oil consumption than domestic cars. Additionally, the p-value is effectively zero, leading us to reject the H_0. Therefore, we have strong statistical evidence that imported cars use significantly more oil compared to domestic cars.

(ii)

Cheap cars: Ford (1), Chevrolet (2), and Nissan (3)

Expensive cars: Lincoln (4), Cadillac (5), and Mercedes (6)

$C(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})$

```
# Contrast weights for Cheap vs Expensive Cars
ctrts2 <- c(1/3, 1/3, 1/3, -1/3, -1/3, -1/3)

# Compute contrast estimate
(est2 <- sum(ctrts2 * trtmeans))
```

```
[1] -0.7660205
```

```
# MSE from ANOVA table
MSE <- anova(fit)[2, 3]

# Compute Standard Error (SE)
(se2 <- sqrt(MSE) * sqrt(sum(ctrts2^2 / ni)))
```

```
[1] 0.09554973
```

```
# Compute CI
lower.ci_price <- est2 - qt(0.975, df_e) * se2
upper.ci_price <- est2 + qt(0.975, df_e) * se2
c(lower.ci_price, upper.ci_price)
```

```
[1] -0.9667630 -0.5652779
```

```
(t_0_price <- est2 / se2)
```

```
[1] -8.016982
```

```
# Compute p-value
2 * pt(-t_0_price, df_e)
```

```
[1] 2
```

The contrast estimate C = -0.766 indicates that inexpensive cars (Ford, Chevrolet, and Nissan) consume less oil per 100,000 miles than expensive cars (Lincoln, Cadillac, and Mercedes). The 95% confidence interval (-0.967, -0.565) does not include 0, we conclude that this difference is statistically significant. This confirms that, on average, cheap cars exhibit lower oil consumption compared to expensive cars. Additionally, the p-value is essentially 0, leading us to reject the null hypothesis. Thus, we have strong evidence that cheap cars use significantly less oil than expensive cars.

Exercise 2.

```
# Create weight loss data
response <- c(8, 9, 6, 7, 3, # Low Calorie
              2, 4, 3, 5, 1, # Low Fat
              3, 5, 4, 2, 3, # Low Carbs
              2, 2, -1, 0, 3) # Control (Placebo)

diet <- rep(c(1, 2, 3, 4), each = 5)

df <- data.frame(response, diet)
df$fdiet <- as.factor(df$diet)
fit.model <- lm(response ~ fdiet, data = df)
```

(a)

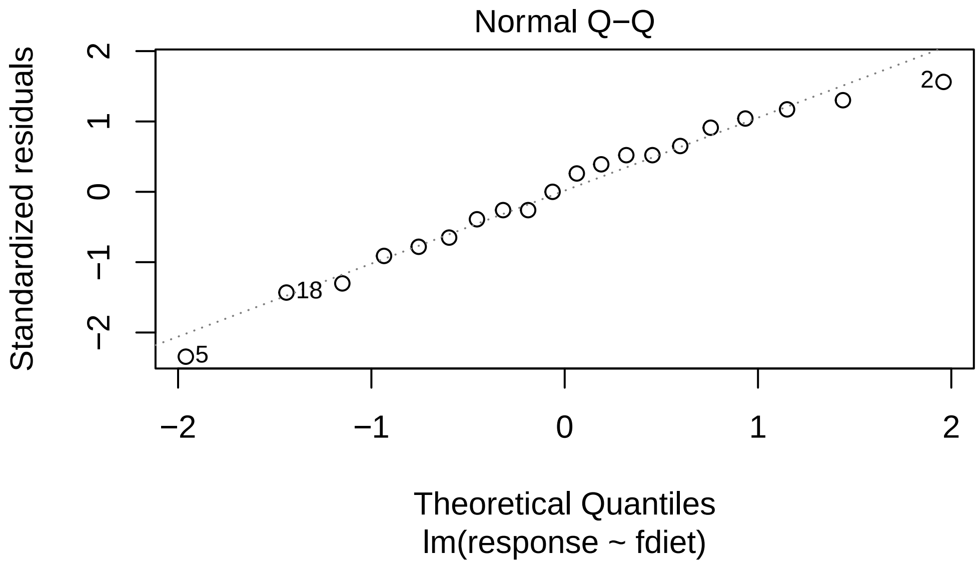
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

vs

H_A : At least one diet mean is different than the others.

- Independence: Since patients were randomly assigned to each diet and one individual's weight loss does not affect another's, we can assume independence both within and between groups.
- Normality: The residuals closely follow a linear trend along the Normal QQ plot, suggesting normality. This is further confirmed by the Shapiro-Wilk test, which indicates that the residuals are consistent with a normal distribution.
- Equal Constant Variance: The residuals vs fitted values scatterplot suggests that the assumption of equal variance holds across the four diet types, as the residuals display consistent spread within similar ranges.

```
# Normality
plot(fit.model, which = 2)
```



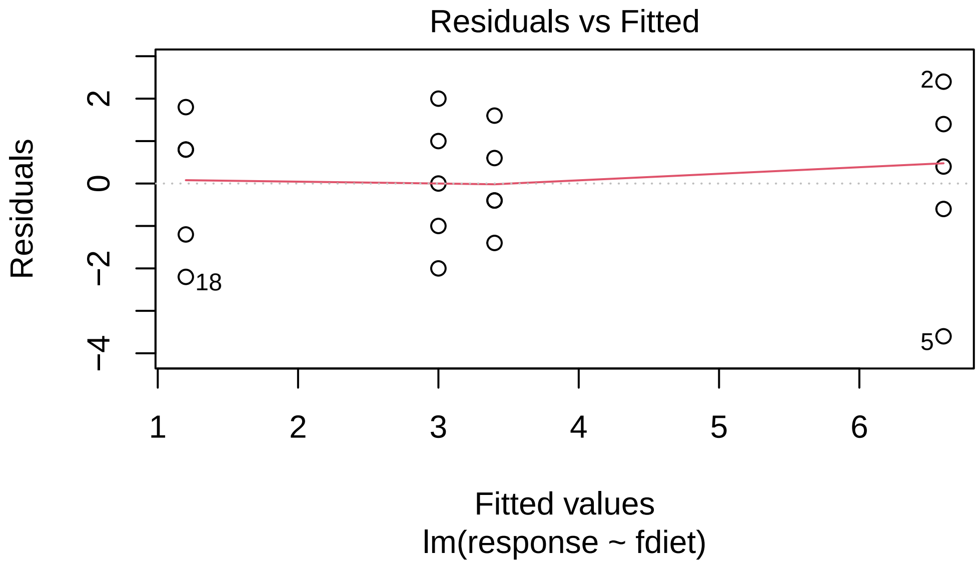
```
shapiro.test(resid(fit.model))
```

```

	Shapiro-Wilk normality test

data:  resid(fit.model)
W = 0.97079, p-value = 0.7714
```

```
# Constant variance
plot(fit.model, which = 1)
```



```
# Fit ANOVA table
anova(fit.model)
```

```
Analysis of Variance Table

Response: response
          Df  Sum Sq Mean Sq F value   Pr(>F)
fdiet      3  75.75   25.25  8.5593 0.001278 **
Residuals 16  47.20    2.95
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of 0.0013 is less than our significance level of $\alpha = 0.05$, leading us to reject H_0 which states that all weight loss diets yield the same mean weight loss. We conclude that at least one diet results in a mean weight loss that differs from the others.

(b)

Low Calorie diet

Control group (Placebo)

C(1,0,0,-1)

```r
library(emmeans)
```

```
Welcome to emmeans.
Caution: You lose important information if you filter this package's results.
See '? untidy'
```

```r
lsmDiet <- lsmeans(fit.model, ~ fdiet)
lsmDiet
```

```
 fdiet lsmean    SE df lower.CL upper.CL
 1        6.6 0.768 16    4.972     8.23
 2        3.0 0.768 16    1.372     4.63
 3        3.4 0.768 16    1.772     5.03
 4        1.2 0.768 16   -0.428     2.83

Confidence level used: 0.95
```

```r
ctrts_diet <- c(1, 0, 0, -1)


contrast(lsmDiet, method = list(ctrts = ctrts_diet), infer = c(T, T))
```

```
 contrast estimate   SE df lower.CL upper.CL t.ratio p.value
 ctrts         5.4 1.09 16      3.1      7.7   4.971  0.0001

Confidence level used: 0.95
```

Conducting a contrast analysis to compare Low Calorie diet with the Control group:

Since zero is not included in the confidence interval, we reject H_0, which states that there is no difference between the low-calorie diet and the control group. We have sufficient evidence to conclude that the weight loss effects between the low-calorie diet and the control group are significantly different.

(c)

Control

Average of the three treatments

C(1/3, 1/3, 1/3, -1)

```r
ctrts_lowfat <- c(1/3, 1/3, 1/3, -1)


contrast(lsmDiet, method = list(ctrts = ctrts_lowfat), infer = c(T, T))
```

```
 contrast estimate    SE df lower.CL upper.CL t.ratio p.value
 ctrts        3.13 0.887 16     1.25     5.01   3.533  0.0028

Confidence level used: 0.95
```

Since zero is not included in the confidence interval, we reject H_0 which states that there is no difference between the average weight loss from the low-calorie, low-fat, and low-carbohydrate diets and the control group. We have sufficient evidence to conclude that the weight loss effects of the diets differ significantly from the control group.