STATS_204_HW4

Qi Wang

Question 1: (9.1)

```
rounding <- read.table("D:/77/UCSC/study/204/HW/rounding.txt", header = TRUE)
str(rounding)
```

```
## 'data.frame':    66 obs. of  3 variables:
##  $ times : num  5.4 5.5 5.55 5.85 5.7 5.75 5.2 5.6 5.5 5.55 ...
##  $ method: chr  "RoundOut" "NarrowAngle" "WideAngle" "RoundOut" ...
##  $ block : int  1 1 1 2 2 2 3 3 3 4 ...
```

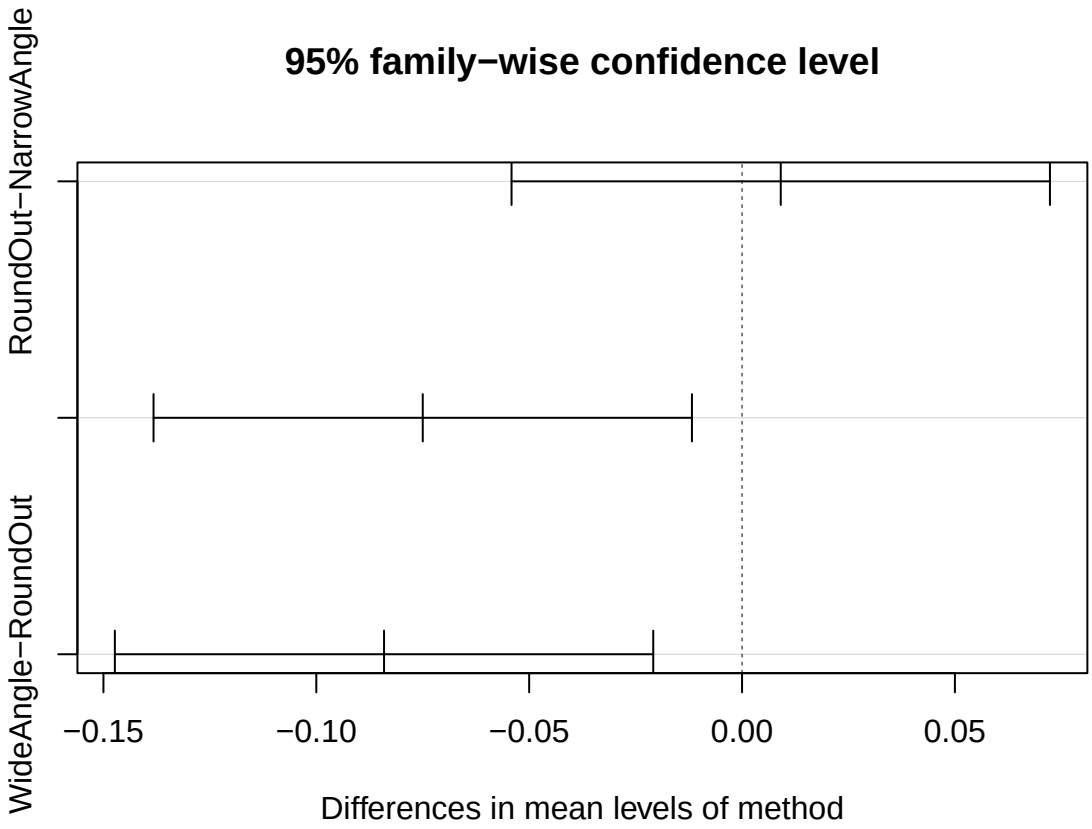
Here, we can see that the data has three variables and 66 observations. There are different methods and different blocks. Time is the response variable and it is continuous. So, for the random block design, we need to add block into the ANOVA.

```
rounding$block <- as.factor(rounding$block)
M1 <- aov(times ~ method + block, data = rounding)
summary(M1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## method       2  0.094 0.04686   6.288 0.00408 **
## block       21  4.219 0.20089  26.960 < 2e-16 ***
## Residuals   42  0.313 0.00745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, both the method and block are significant, which means that there is significant difference in means among different method groups, also the block is also significant therefore, different blocks will have significant differences in mean of times.

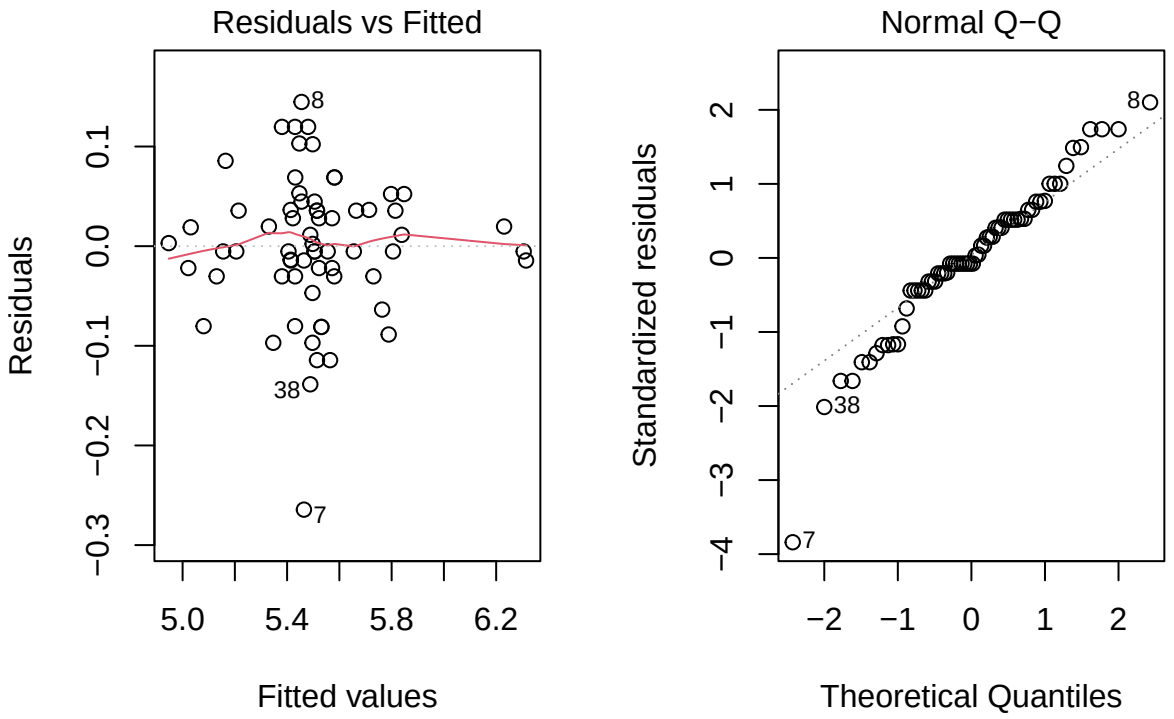
```
CI_1 <- TukeyHSD(M1, which = 1)
plot(CI_1)
```



Also, by Tukey HSD (I did not transform the name of the variable so it looks so long) we can see that wideanlge and roundout has significant difference, narrowangle and roundout also have significant difference, but wide angle and narrow angle does not have significant difference.

The residual plots are as follows:

```
par(mfrow = c(1,2))
plot(M1, which = 1:2)
```



Here, from the residual plot, there seems to be a decreasing trend of variance of residuals as the fitted values goes up. Therefore, we need to consider whether it has violated the assumption of constant variance. Also, in the qq plot, the tail on both sides deviated from the reference line, which make we suspicious about the normality assumption.

Question 2: (9.2)

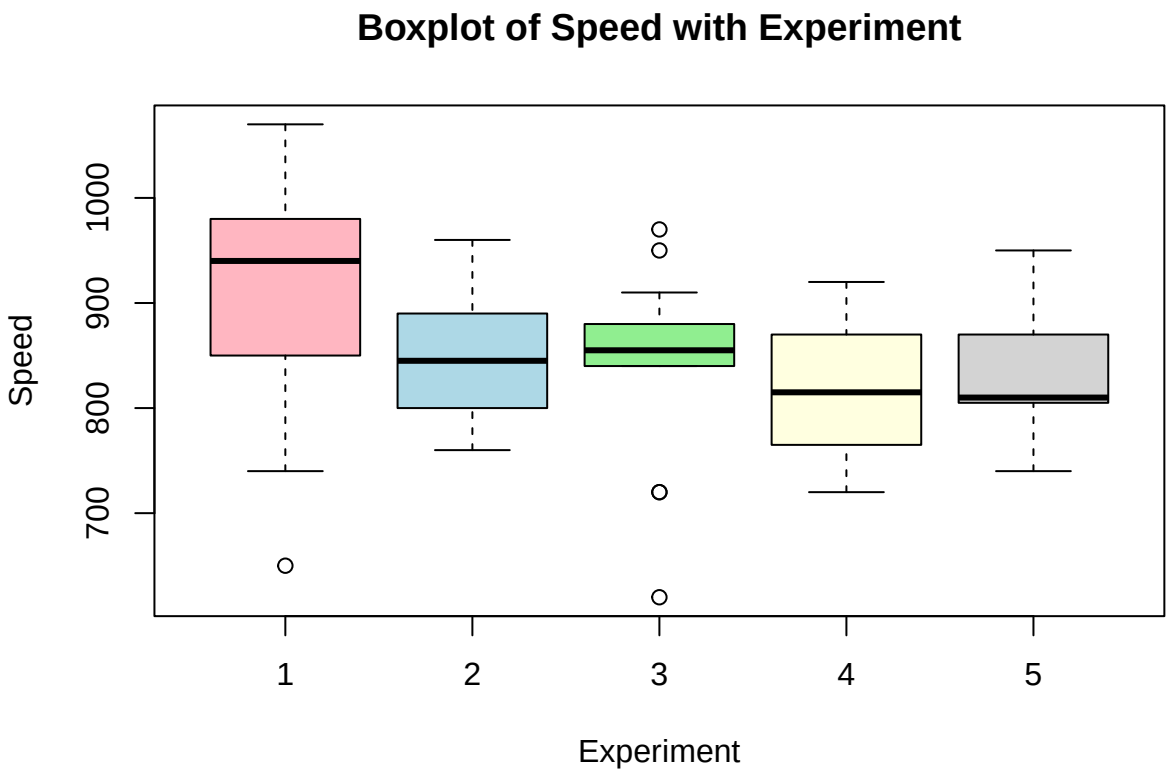
```
str(morley)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ Expt : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Run  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Speed: int  850 740 900 1070 930 850 950 980 980 880 ...
```

```
morley$Expt = factor(morley$Expt)
morley$Run = factor(morley$Run)
```

According to the output of str function in R, we have three variables and "Expt" indicates which group of experiment is this observation in and "Run" is the order of the output in each experiment.

```
boxplot(Speed ~ Expt, data = morley, col = c("lightpink", "lightblue", "lightgreen", "lightyellow", "lig
                                             ) , xlab = "Experiment", ylab = "Speed", main = "Boxplot of
```



If we regard this data set as a randomized block design, we can write a model as follows:

$$Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

Here, β_i means which group of experiment does this observation is from, and τ_j means which run of this observation is from. Runs here is the block variable and experiment is the treatment variable.

Our null hypothesis for β is:

$$\beta_i = 0 \quad for \;\; all \;\; i$$

The alternative is not all of β_i are 0. Our null hypothesis for τ is:

$$\tau_j = 0 \quad for \;\; all \;\; j$$

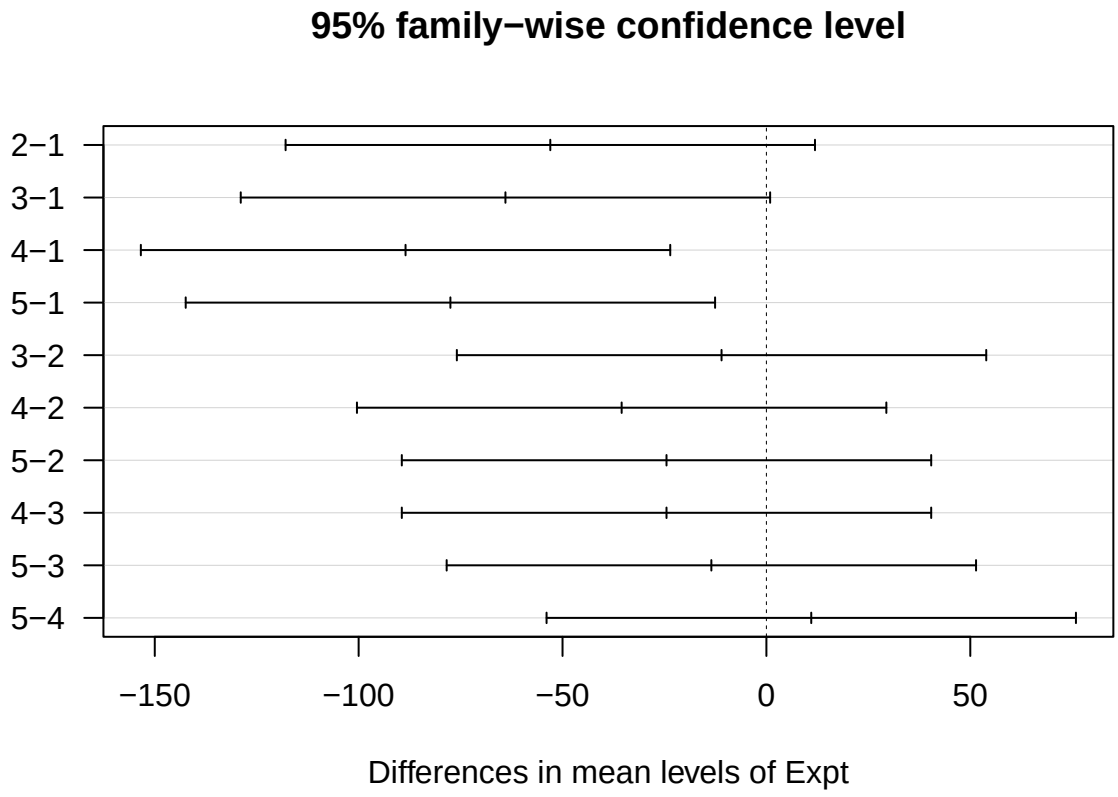
The alternative is not all of τ_j are 0.

```
M2 <- aov(Speed ~ Expt + Run, data = morley)
summary(M2)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Expt         4  94514   23629   4.378 0.00307 **
## Run         19 113344    5965   1.105 0.36321
## Residuals   76 410166    5397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

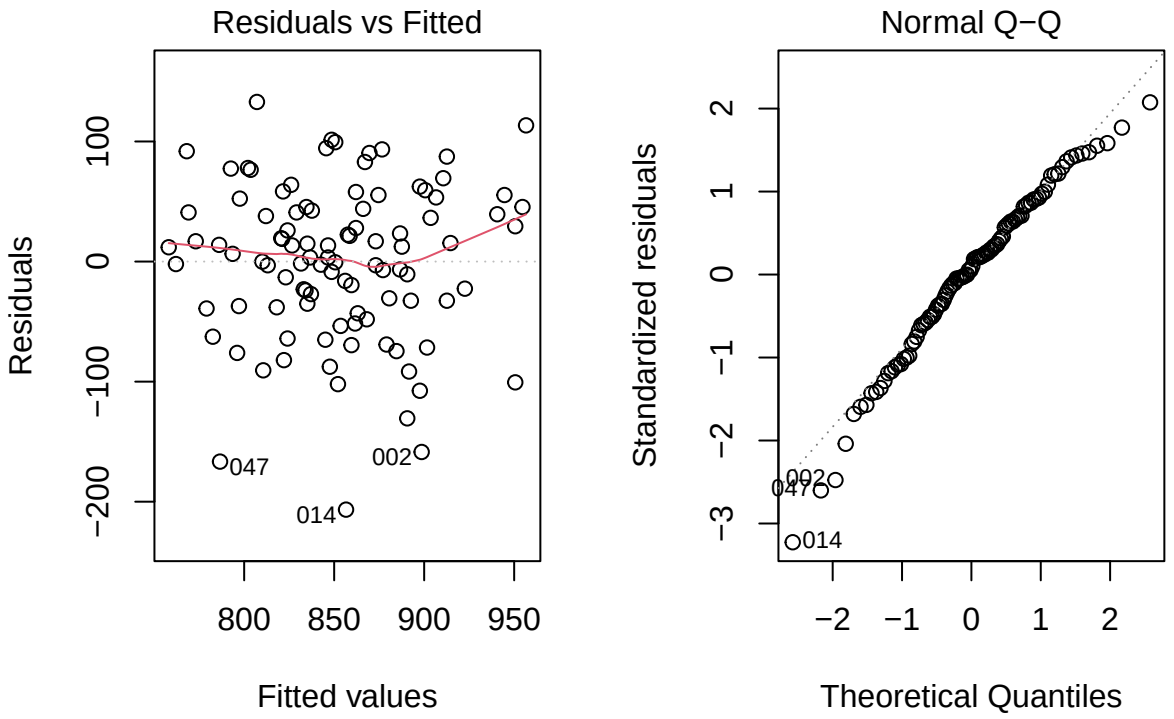
So since the "Expt" is significant, there are significant difference among mean speeds of different groups of experiments. However, there are no significant difference among mean speeds of different groups of run. This can be interpreted by the independence of each run in each experiment.

```
CI_2 <- TukeyHSD(M2, which = 1)
plot(CI_2, las = 1)
```



Here we can see that only 4-1 and 5-1 have significant difference, which means that the mean of the experiment 4 and experiment 1 and experiment 5 and experiment 1 are significantly different.

```
par(mfrow = c(1,2))
plot(M2, which =1:2)
```



There is a little quadratic form in the variances of residuals but not that obvious. In the qq plot, the tail on the left has some values that are a little far from the reference line. But overall, the data seems to perform okay.

Additional:

Question 1:

```
library(ISwR)
tb <- tb.dilute
tb$animal <- factor(tb$animal)
tb$logdose <- factor(tb$logdose)
```

The data has 18 observations with treatment logdose and have different animal categories. I am performing a 2-way anova. And first I need to transfer variables into factors. Then I will carry out a two-way ANOVA.

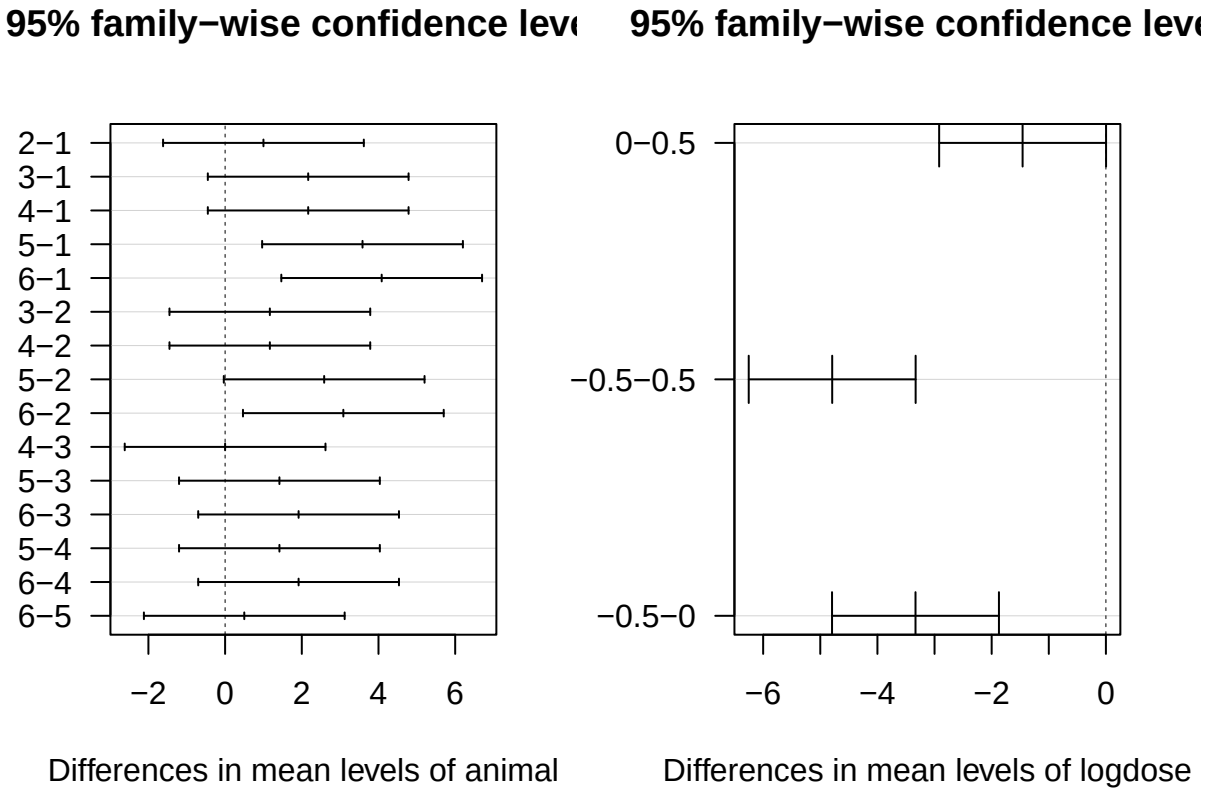
```
M3 <- aov(reaction ~ animal + logdose, data = tb)
anova(M3)
```

```
## Analysis of Variance Table
##
## Response: reaction
##           Df Sum Sq Mean Sq F value    Pr(>F)
## animal     5 35.208   7.042  8.2641  0.002527 **
## logdose    2 72.396  36.198 42.4817 1.295e-05 ***
## Residuals 10  8.521   0.852
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For here, we have only one observation in each group, so we cannot add interaction terms. Consider two of the Tukey HSD, we have:

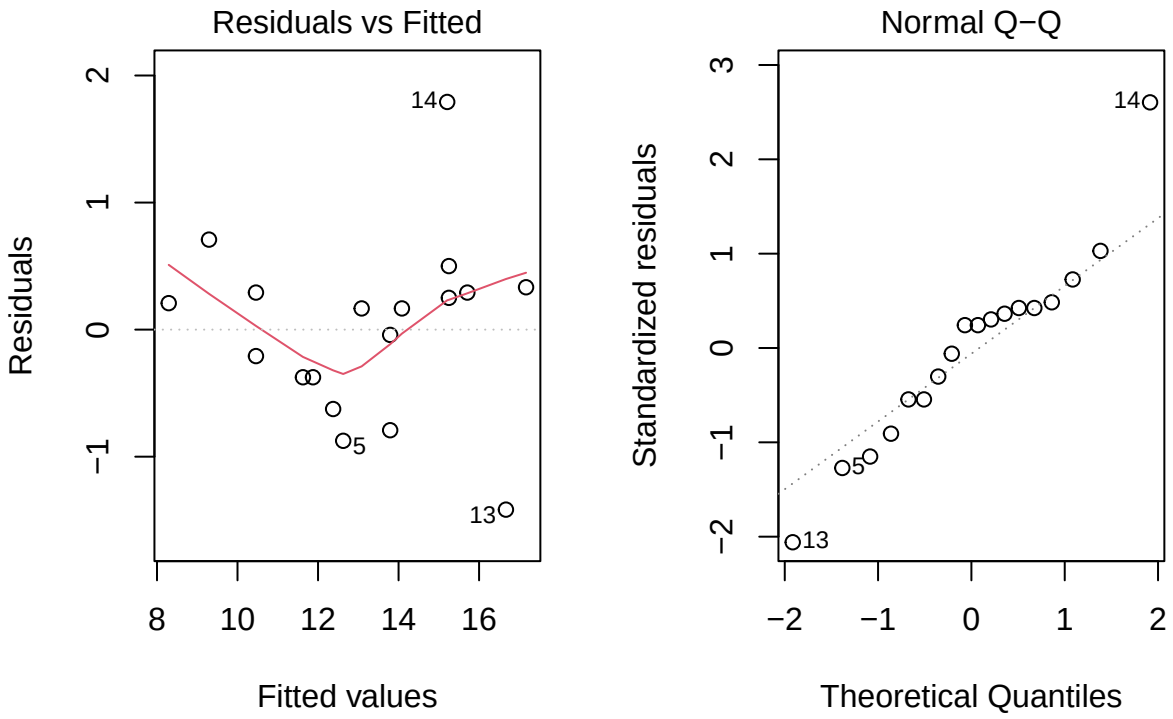
```
CI_3 <- TukeyHSD(M3, which = 1)
CI_4 <- TukeyHSD(M3, which = 2)
par(mfrow = c(1,2))
plot(CI_3, las = 1)
plot(CI_4, las = 1)
```



Here we can distinguish the animal groups that have significant difference in mean of reaction is group 5 compared with group 1, group 6 compared with group 1, group 6 compared with group 2. Also, the difference of mean is all significant for any two logdose groups.

Now, I will do the residual analysis.

```
par(mfrow = c(1,2))
plot(M3, which = 1:2)
```



The residuals performed very bad, it does not have constant variance since it performed like a quadratic trend, also it is not normally distributed since almost no points lie on the reference line.

2. Analysis of Covariance.

```
library(ISwR)
vitcap <- vitcap2
vitcap$group <- factor(vitcap$group)
M4 <- lm(vitcap$vital.capacity ~ vitcap$group * vitcap$age, data = vitcap)
anova(M4)
```

```
## Analysis of Variance Table
##
## Response: vitcap$vital.capacity
##                          Df  Sum Sq Mean Sq F value  Pr(>F)
## vitcap$group              2  2.7473  1.3737  3.8912 0.02450 *
## vitcap$age                1 14.8589 14.8589 42.0915 7.3e-09 ***
## vitcap$group:vitcap$age   2  2.4995  1.2497  3.5402 0.03376 *
## Residuals                78 27.5352  0.3530
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By the definition of analysis of covariance, the interaction term is significant, so the slope of different groups are different.

The model is:

$$Y_{ij} = (\mu + \delta_i) + (\beta + \gamma_i)X_{age} + \epsilon_{ij}$$

For the linear model:

```
summary(M4)
```

```
##
## Call:
## lm(formula = vitcap$vital.capacity ~ vitcap$group * vitcap$age,
##     data = vitcap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24497 -0.36929  0.01977  0.43681  1.13953
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 8.18344    0.99358   8.236 3.28e-12 ***
## vitcap$group2              -1.95341    1.10481  -1.768   0.0810 .
## vitcap$group3              -2.50315    1.04184  -2.403   0.0187 *
## vitcap$age                 -0.08511    0.01967  -4.327 4.44e-05 ***
## vitcap$group2:vitcap$age    0.03858    0.02327   1.658   0.1014
## vitcap$group3:vitcap$age    0.05450    0.02107   2.587   0.0116 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5942 on 78 degrees of freedom
## Multiple R-squared:  0.422,  Adjusted R-squared:  0.385
## F-statistic: 11.39 on 5 and 78 DF,  p-value: 2.871e-08
```

We have $\delta_1 = \gamma_1 = 0$ for group 1 in our linear regression model. For the ANCOVA model, we are testing that:

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$$

, the alternative is at least one of them is not zero. And the result is that we can reject the null hypothesis since the p-value is small enough. There are interaction effects, different groups have not all the same slope.

3. Weight Loss Experiment

```
L1 <- c(81, 88, 85, 84, 84)
L2 <- c(85, 80, 82, 80, 82)
L3 <- c(71, 77, 72, 80, 80)
L4 <- c(84, 84, 82, 81, 86)
L5 <- c(83, 88, 85, 86, 88)
L6 <- c(78, 75, 78, 79, 82)

method <- c(rep("Frying", 15), rep("Grilling", 15))
fat <- rep(c(rep(10,5),rep(15,5),rep(20,5)),2)
after <- c(L1, L2, L3, L4, L5, L6)
dif <- as.numeric(110 - after)
hamburger <- as.data.frame(cbind(dif, method, fat))
hamburger$dif <- as.numeric(hamburger$dif)
```

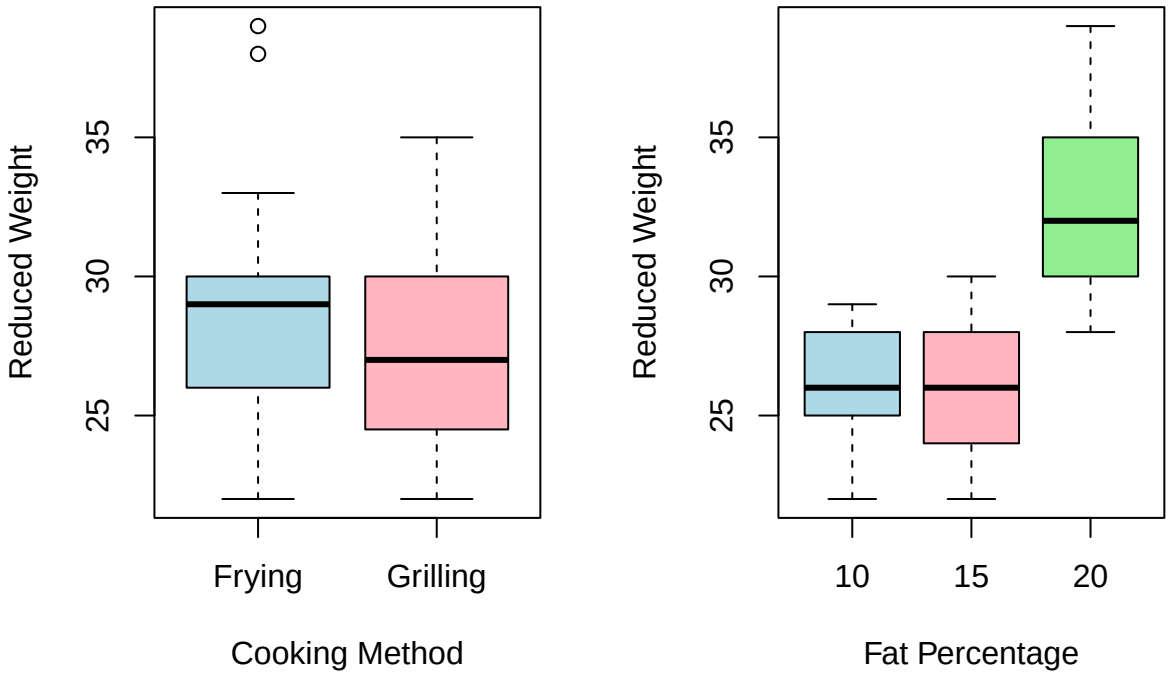
(a) Exploratory Data Analysis

```
summary(hamburger)
```

```
##       dif           method              fat
##  Min.   :22.00   Length:30          Length:30
##  1st Qu.:25.25   Class :character   Class :character
##  Median :28.00   Mode  :character   Mode  :character
##  Mean   :28.33
##  3rd Qu.:30.00
##  Max.   :39.00
```

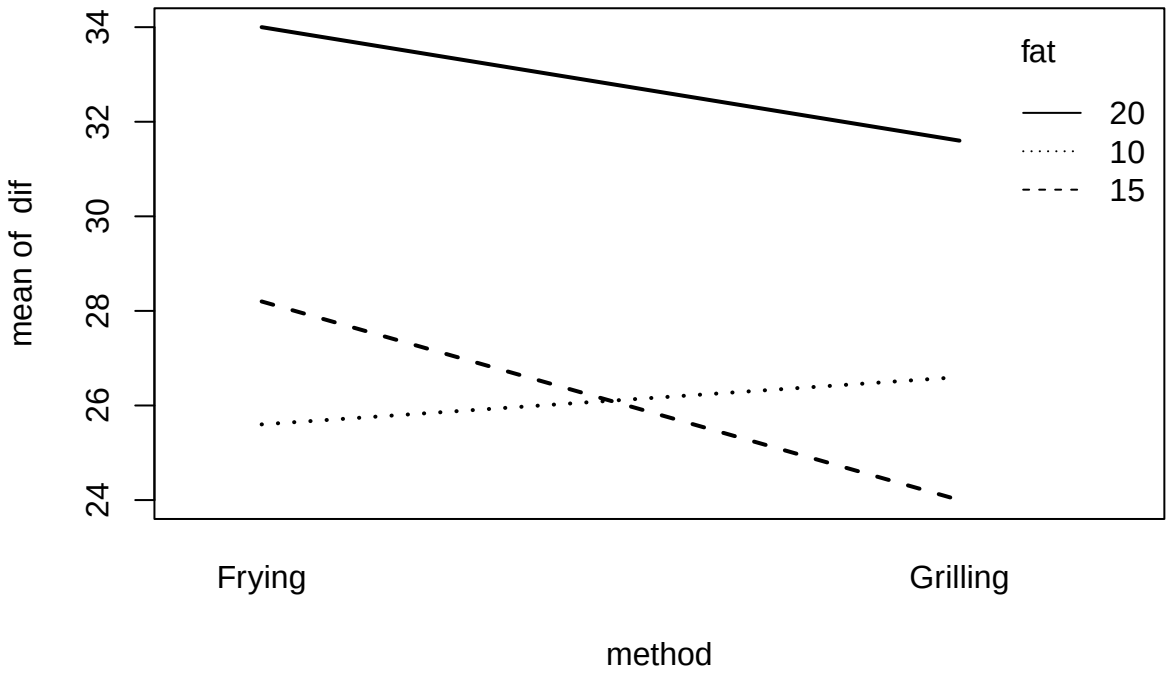
Here I regard fat percentage a categorical data since it only have three levels and seems not to be continuous. We can see the mean and median of the difference are both around 28. maximum is 39 and minimum is 22.

```
par(mfrow = c(1,2))
boxplot(dif ~ factor(method), col = c("lightblue", "lightpink"), xlab = "Cooking Method", ylab = "Reduc
boxplot(dif ~ factor(fat), col = c("lightblue", "lightpink","lightgreen"), xlab = "Fat Percentage", ylab
```



From the box plot we can see that the median of grilling is smaller than frying method, but for most part, they are almost the same. However, for fat percentage, it seems that 10 percent and 15 percent group does not have significant difference, but for 20 percent group, the reduced weight is significantly higher than the other two groups. Now let's consider the interaction:

```
with(data = hamburger, expr = {
  interaction.plot(method, fat, response =dif, lwd = 2)
})
```



It seems that there could be some interaction but we still need to use models to test.

```
M5 <- aov(dif ~ factor(method) * factor(fat))
anova(M5)
```

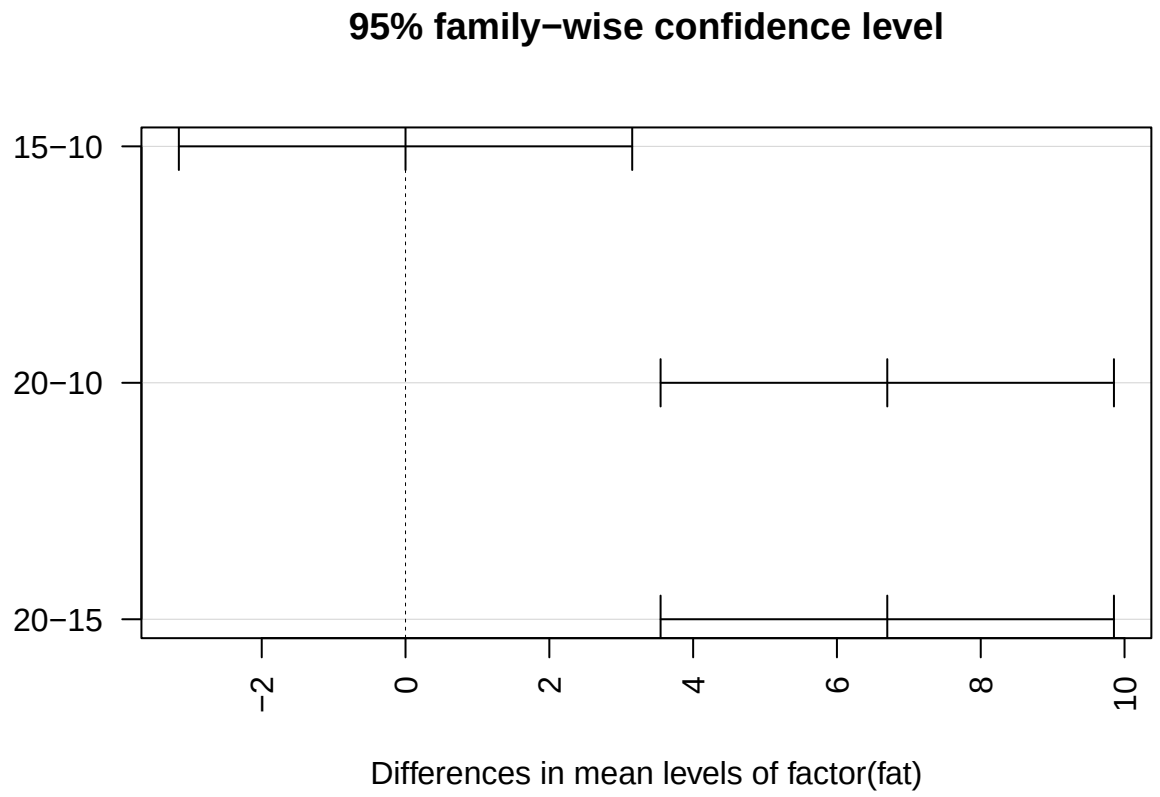
```
## Analysis of Variance Table
##
## Response: dif
##                            Df  Sum Sq Mean Sq F value    Pr(>F)
## factor(method)              1  26.133  26.133  3.5963   0.07001 .
## factor(fat)                 2 299.267 149.633 20.5917 6.207e-06 ***
## factor(method):factor(fat)  2  34.867  17.433  2.3991   0.11224
## Residuals                  24 174.400   7.267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here it seems that the mean of groups with different methods of cooking does not have significant difference. However, the mean of groups with different fat percentages has significant difference. Also the interaction seems not so significant. Then I prefer a model of no interaction terms.

```r
M5 <- aov(dif ~ factor(method) + factor(fat))
anova(M5)
```

```
## Analysis of Variance Table
##
## Response: dif
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## factor(method)  1  26.133  26.133  3.2469   0.08317 .
## factor(fat)     2 299.267 149.633 18.5910 9.704e-06 ***
## Residuals      26 209.267   8.049
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
CI_ham <- TukeyHSD(M5, which = 2, conf.level = 0.95)
plot(CI_ham, las = 2)
```



Here, there are significant difference in mean reduced weight between 10 and 20 percentage group, and 20 and 15 percentage group. But not significant difference in group 15 percentage and 10 percentage. And the mean of 20 percentage group is significantly higher than the others since the 95% CI does not include 0.

The 90% confidence interval for the difference in mean between difference methods of cooking is:

```r
TukeyHSD(M5, which = 1,conf.level = 0.9)
```

```
##   Tukey multiple comparisons of means
```

```
##     90% family-wise confidence level
##
## Fit: aov(formula = dif ~ factor(method) + factor(fat))
##
## $`factor(method)`
##                       diff       lwr         upr     p adj
## Grilling-Frying -1.866667 -3.633577 -0.09975664 0.0831663
```

So the difference of mean reduced weight between group grilling and group frying are significant if we set significance level to be 0.9. And the 90% CI is [-3.633577 -0.09975664].

4. Malaria Data Set

```
library(ISwR)
mala <- malaria
mala $ab <- log(mala$ab)
M_logistic <- glm(mal ~ age + ab, data = mala, family = binomial(link = "logit"))
summary(M_logistic)
```

```
##
## Call:
## glm(formula = mal ~ age + ab, family = binomial(link = "logit"),
##     data = mala)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8492  -0.7536  -0.4838   0.8809   2.5796
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.57234    0.95184   2.702 0.006883 **
## age         -0.06546    0.06772  -0.967 0.333703
## ab          -0.68235    0.19552  -3.490 0.000483 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.652  on 99  degrees of freedom
## Residual deviance:  98.017  on 97  degrees of freedom
## AIC: 104.02
##
## Number of Fisher Scoring iterations: 5
```

From the result above, we can see the age is not significant, but the log of antibody level is significant. And more antibody level means less probability to get malaria. Controlling all the other variables the same, the odds of people with one unit greater log of antibody level will be $e^{-0.68235}$ times compared with the odds of the people whose log of antibody level remain the same.

The deviance is 98.017 and not so far from the $N - p = 97$, which means that the model performed well.