Stats 204 | Data Analysis

Jordan Berninger

11/5/2019

Homework 4

Chapter 9 Problem 1

The data in "rounding.txt" gives the times required to round first base for 22 baseball players using three styles: rounding out, a narrow angle and a wide angle. The goal is to determine if the method of rounding first base has a significant effect on times to round first base.The data and the format of the data can be viewed using a text editor or a spreadsheet. With the data file in the current working directory, input the data using $rounding = read.table(''rounding.txt'', header = TRUE)$ Check using the str function that the data is in stacked format with three variables: time, method, player, where time is numeric and method and player are factors. Analyze the data assuming a randomized block design with time to round first base as the response, method of rounding first as the treatment, and player as the block variable. Plot residuals to check whether the assumptions for the distribution of random error appear to be valid.

We read in the data set, check the str() as instructed and note that the final column is called "block" instead of player, but that does not change how we proceed. We convert that field to a factor as instructed. We use str() again to confirm that the change was correctly made.

```
rounding = read.table("rounding.txt", header=TRUE)
str(rounding)
```

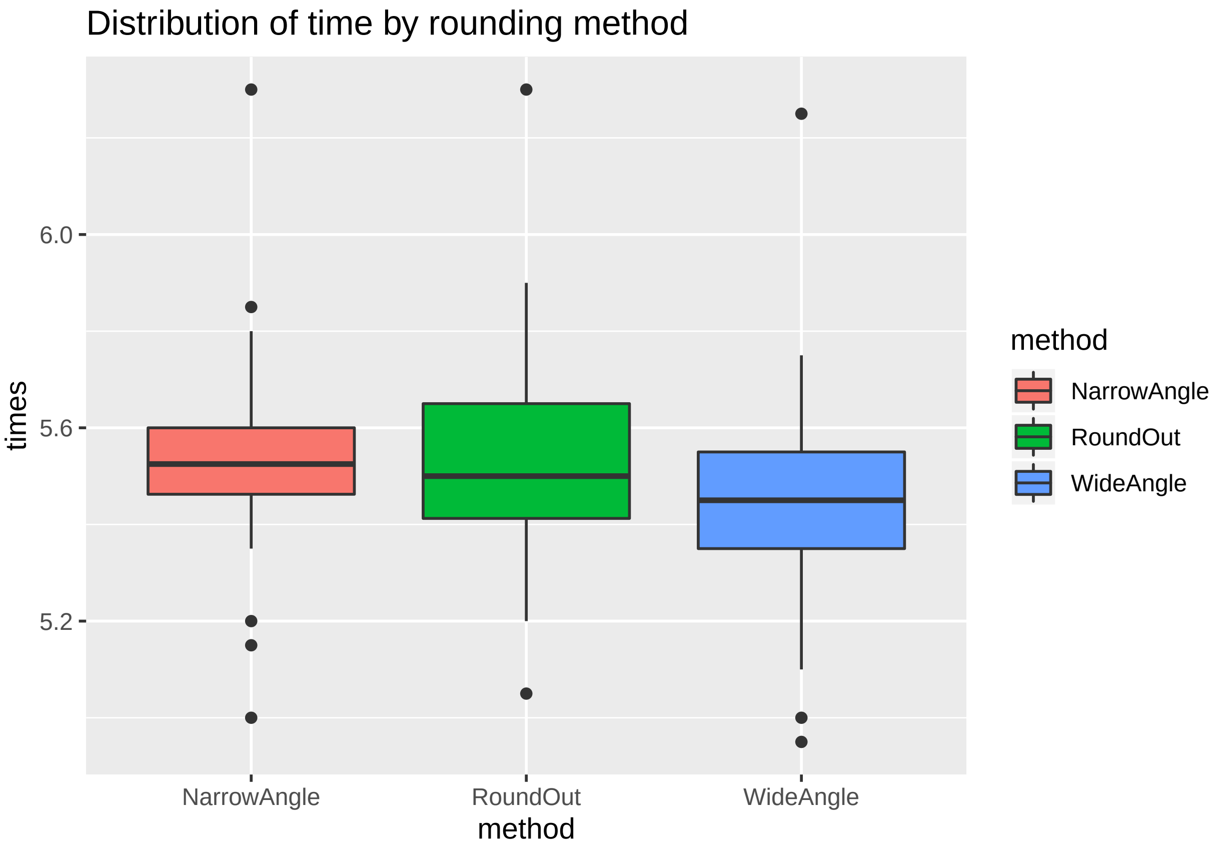
```
## 'data.frame':    66 obs. of  3 variables:
##  $ times : num  5.4 5.5 5.55 5.85 5.7 5.75 5.2 5.6 5.5 5.55 ...
##  $ method: Factor w/ 3 levels "NarrowAngle",..: 2 1 3 2 1 3 2 1 3 2 ...
##  $ block : int  1 1 1 2 2 2 3 3 3 4 ...
```

```
rounding <- rounding %>% mutate(block = factor(block, ordered = FALSE))
str(rounding)
```

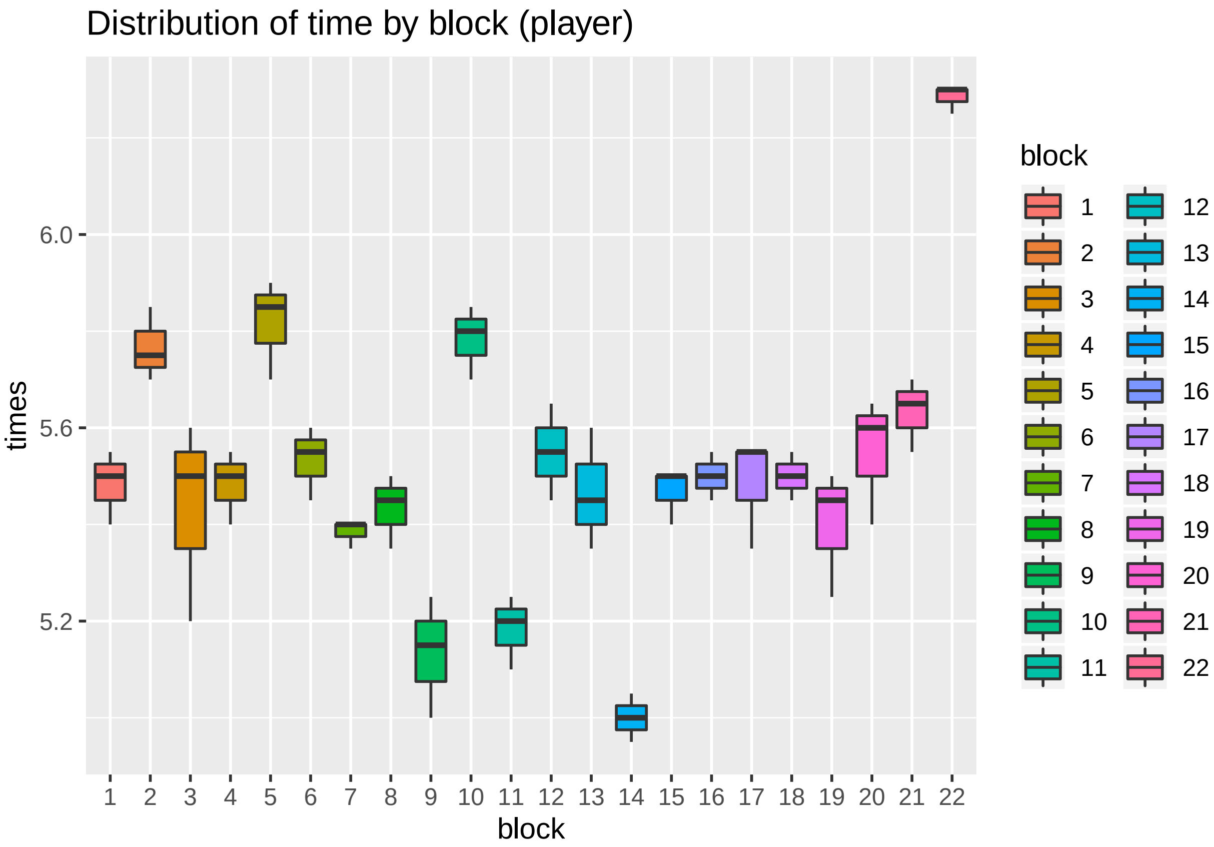
```
## 'data.frame':    66 obs. of  3 variables:
##  $ times : num  5.4 5.5 5.55 5.85 5.7 5.75 5.2 5.6 5.5 5.55 ...
##  $ method: Factor w/ 3 levels "NarrowAngle",..: 2 1 3 2 1 3 2 1 3 2 ...
##  $ block : Factor w/ 22 levels "1","2","3","4",..: 1 1 1 2 2 2 3 3 3 4 ...
```

Since we are interested in the effects of block and method on time, we should visualize how the distribution time across these different groups.

```
ggplot(data = rounding, aes(x = method, y = times, fill = method)) + geom_boxplot() +
  ggtitle("Distribution of time by rounding method")
```



```
ggplot(data = rounding, aes(x = block, y = times, fill = block)) + geom_boxplot() +
  ggtitle("Distribution of time by block (player)")
```



The boxplots hint that there will be significant differences between the mean times for blocks, however we need more rigorous methods to determine if there are significant differences between the mean times for the rounding methods.

Now we are ready to fit the randomized block model using the aov() function in R. The model we are considering is: $Time_{i,j} = \mu + block_i + method_j + \epsilon_{i,j}$ where $block_i$ is the block (player) effect for $i \in \{1, \ldots, 22\}$, $method_j$ is the method effect for $j \in \{1, 2, 3\}$ and indpeendent $\epsilon_{i,j}\ N(0, \sigma^2)$. We do not include an interaction effect in this model because there is only one data point for each combination of method and block, accordingly, our model would have more parameters than data points and we would achieve a perfect fit.

We look at the model summary below and note that both the method and block (player) variables have high F-statistics and low p-values which support the alternative hypotheses. There are 2 tests being evaluated in this model summary. The first has the null hypothesis that $block_i = 0 \forall i$ and the second is that $method_j = 0 \forall j$. Given the data, the F-test concludes that there are some non-zero block and method effects. In this case, we conclude that there are significant differences between the mean times for methods of running and significant differences between the mean times for the different runners. We look further to see which pairs have significant differences.

```
m1 <- aov(data = rounding, times ~ method + block)
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
```

The F-test showed us that there are significant differences in the method and player mean times, but it did not tell us which pairs have significant differences. The model.tables() function shows us the differences between the grand mean and the treatment and block means, respectively.

```
model.tables(m1, cterms="method", type = "mean")
```

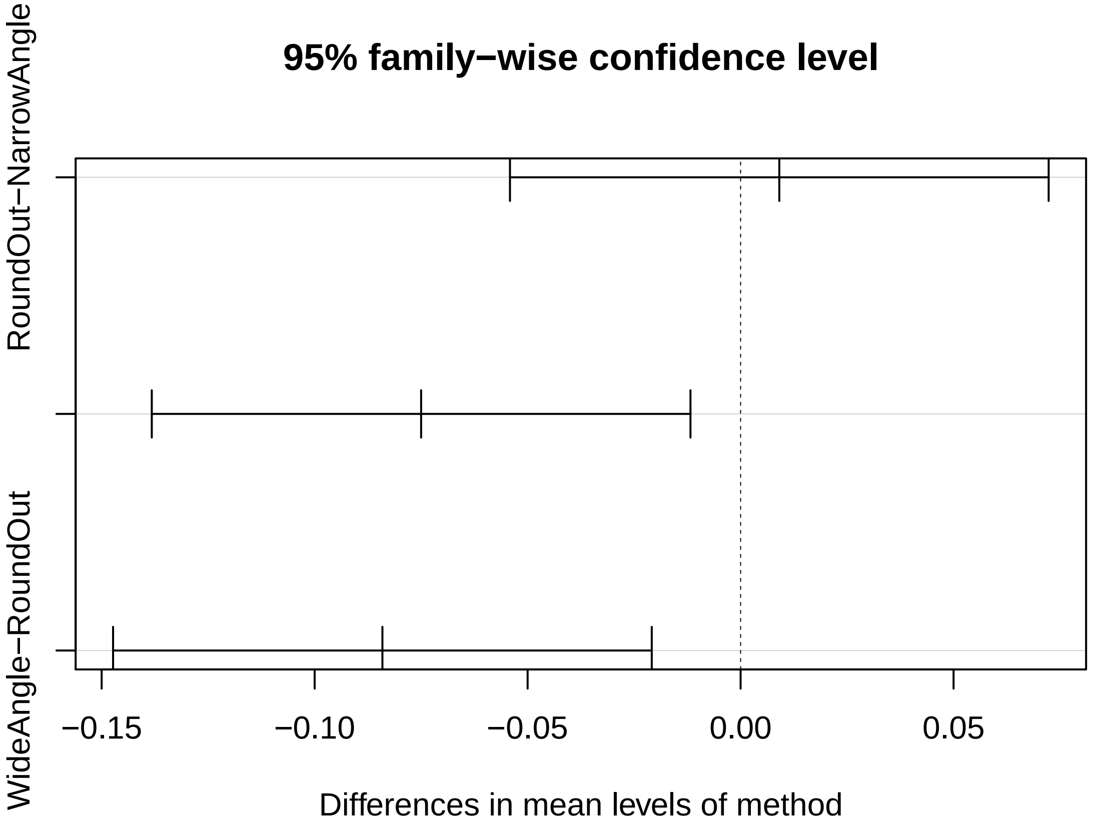
```
## Tables of means
## Grand mean
##
## 5.512121
##
##  method
## method
## NarrowAngle    RoundOut   WideAngle
##       5.534       5.543       5.459
```

```
model.tables(m1, cterms="block", type = "mean")
```

```
## Tables of means
## Grand mean
##
## 5.512121
##
##  block
## block
##     1     2     3     4     5     6     7     8     9    10    11    12
## 5.483 5.767 5.433 5.483 5.817 5.533 5.383 5.433 5.133 5.783 5.183 5.550
##    13    14    15    16    17    18    19    20    21    22
## 5.467 5.000 5.467 5.500 5.483 5.500 5.400 5.550 5.633 6.283
```

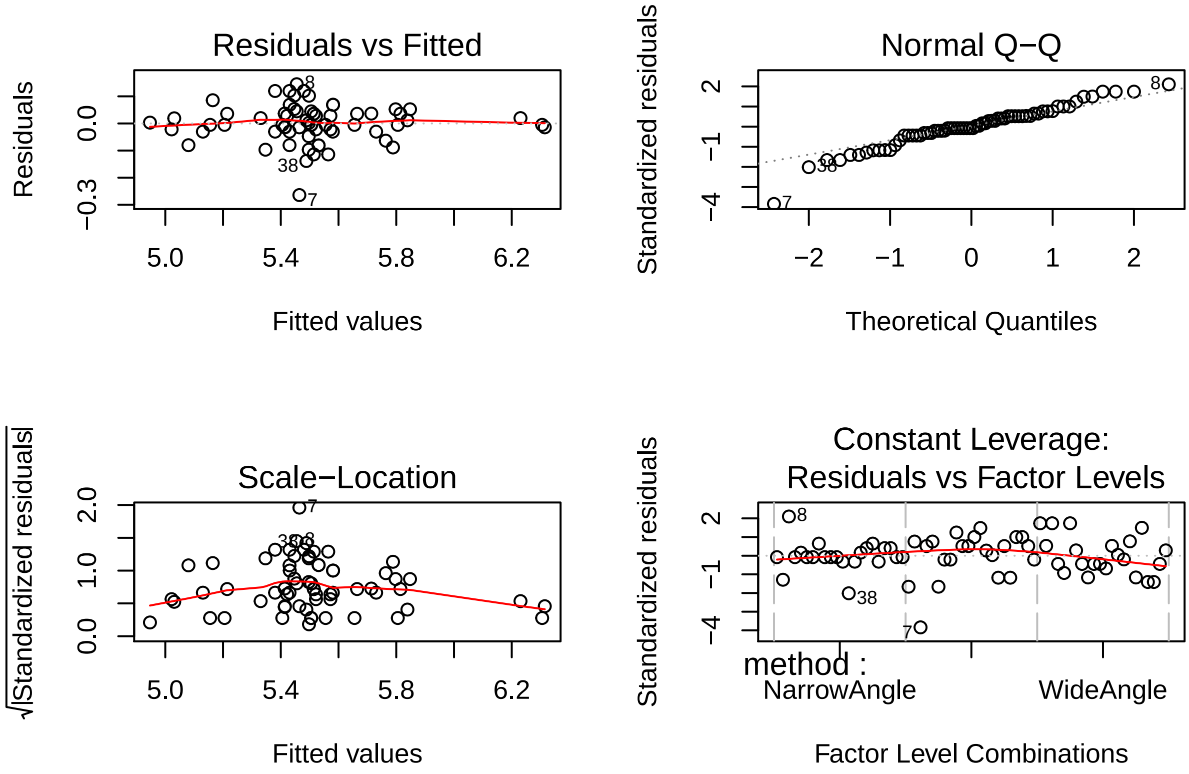
And the TukeyHSD() function in R returns the confidence interval for each pair of means. If the 95% confidence interval for the difference in means does not include zero, then we can conclude there is a significant difference between the corresponding pair of group means. We see WideAngle is significantly faster (lower time) than both narrow angle and roundout methods. Similarly, we see that some players are significantly faster than others, but we don't need to go through this with a fine-toothed comb - the list is too long to print since we are looking at all combinations of 2 players out of the 22.

```
TukeyHSD(m1, which = 1) %>% plot()
```



Now, we inspect the residual plots. We see in the top left plot that the mean of residuals is approximately zero and that there appears to be higher residual variance towards the middle section of fitted values. In the top right, we see that the standardized residuals roughly match the theoretical quantiles from the normal distribution, with some outliers on the tails - this indicates that the residuals roughly follow a normal distribution, which is a good sign for the model. In the bottom right plot, it looks like residuals have roughly the same mean and variance across the 3 treatment groups.

```r
par(mfrow = c(2,2))
plot(m1)
```



Overall, the F-tests indicate that there are significant differences between group means in for both methods and blocks (players). The specific treatments and blocks with significant differences were identified with the TukeyHSD() function, although we only did that for the treatments, for the sake of brevity. The residual plots indicate that the residuals are approximately normal, with mean zero and constant variance across treatment groups, which gives us confidence in our randomized block design model's conclusions.

Chapter 9 Problem 2

The morley data in R contains the classical data of Michaelson and Morley on the speed of light, recording five experiments of 20 consecutive runs each. The response is the speed of light measurement Speed. The experiment is Expt and the run is Run. See the documentation (?morley) and also http://lib.stat.cmu.edu/DASL/Stories/ SpeedofLight.html for more details about the experiments and the data set. Use the str function to check that there are 100 observations of the response Speed, Expt, and Run; all integer variables. Convert Expt and Run to factors using $morley\$Expt = factor(morley\$Expt)$, $morley\$Run = factor(morley\$Run)$ Display a boxplot of Speed by Expt. Speed of light is a constant, so we see there are some problems because the measurements of speed do not seem to be consistent across the five experiments. The data can be viewed as a randomized block experiment. What is the null hypothesis of interest? Analyze the data and residuals and summarize your conclusions.

We import the data, check the structure, and factorize the variables as instructed. We look at the boxplots of speed for each experiment group and note that the group means and variances have different characteristics. It is not clear if there are significant differences between experiement group and run means - more rigorous methods will be applied below.

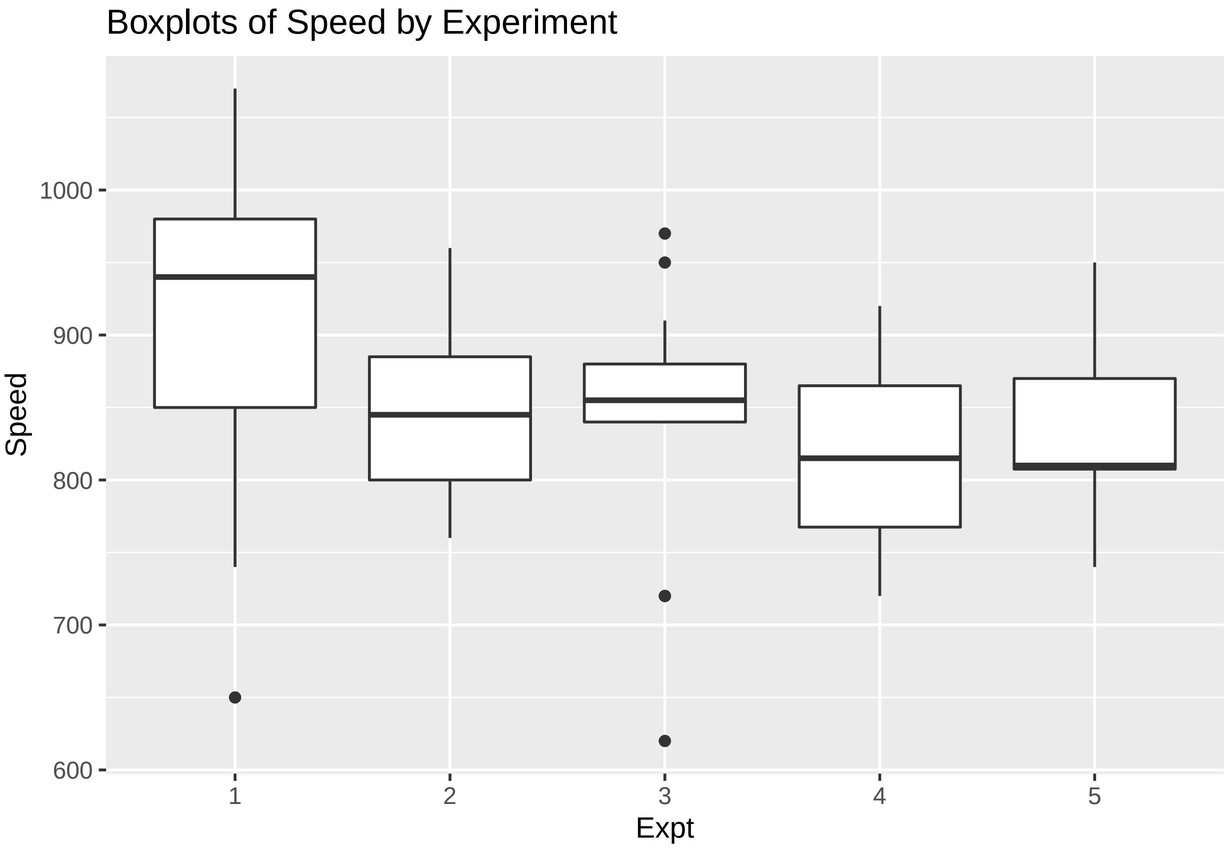
```
data(morley)
str(morley)
```

```
## 'data.frame':    100 obs. of  3 variables:
```

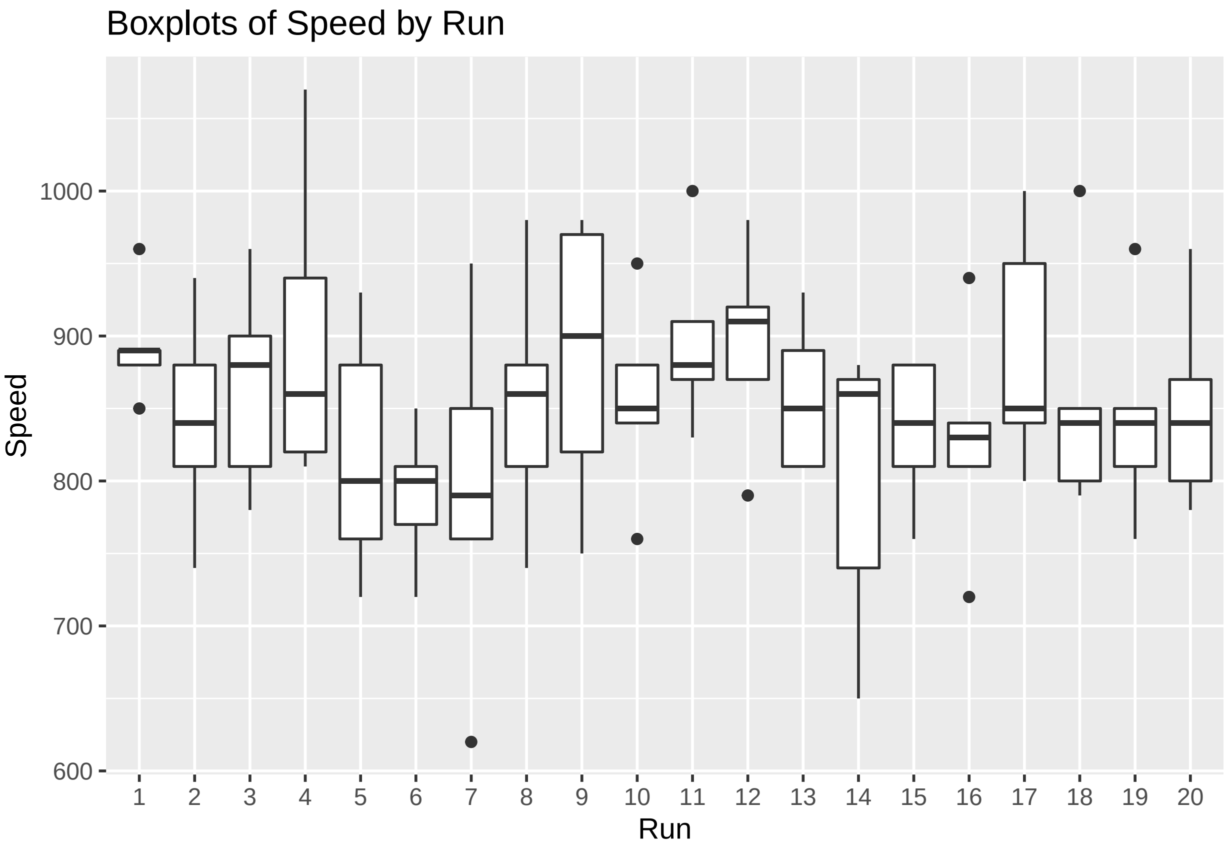
```
##  $ Expt : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Run  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Speed: int  850 740 900 1070 930 850 950 980 980 880 ...
```

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

ggplot(data = morley, aes(x = Expt, y = Speed)) + geom_boxplot() +
  ggtitle("Boxplots of Speed by Experiment")
```



```
ggplot(data = morley, aes(x = Run, y = Speed)) + geom_boxplot() +
  ggtitle("Boxplots of Speed by Run")
```



Now, according to physics, the speed of light through a vaccuum is a constant. Our data measures the speed of light for 5 experimental groups of 20 runs each. Accordingly, we can model Speed = μ + Expt effect + Run effect + error. We can think of the Expt effect parameters as modeling the differences between group means and the Run effect parameters as modeling the variance within experimental groups.

We code the randomized block deisgn model in R and return the model summary. The model we are employing is: $speed_{i,j} = \mu + expt_i + run_j + \epsilon_{i,j}$ where $expt_i$ is the experimental group effect for $i \in \{1, \ldots, 5\}$ and run_j is the run effect for $jin\{1, \ldots, 20\}$ and independent $\epsilon_{i,j}$ $N(0, \sigma^2)$.

There are 2 F-tests in the model summary with the null hypotheses : $0 = \beta_{Expt1} = \cdots = \beta_{Expt5}$ and $0 = \beta_{Run1} = \cdots = \beta_{Run20}$. In the F-Tests we are comparing the models with parameters against a model of the grand sample mean. The F-test numbers in the model summary indicate that Expt is a significant term, which means that there is at least one pair of experiment groups with a significant difference in group means. The model summary's low F-statistic for Run indicates that there are no significant differences between run means.

```
m2 <- aov(data = morley, Speed ~ Expt + Run)
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
```

Before we determine which Experimental groups have signicantly different group means we need to remove the Run effect from our model. What we are left with is no longer a randomized block design model, it is a one-way anova test.

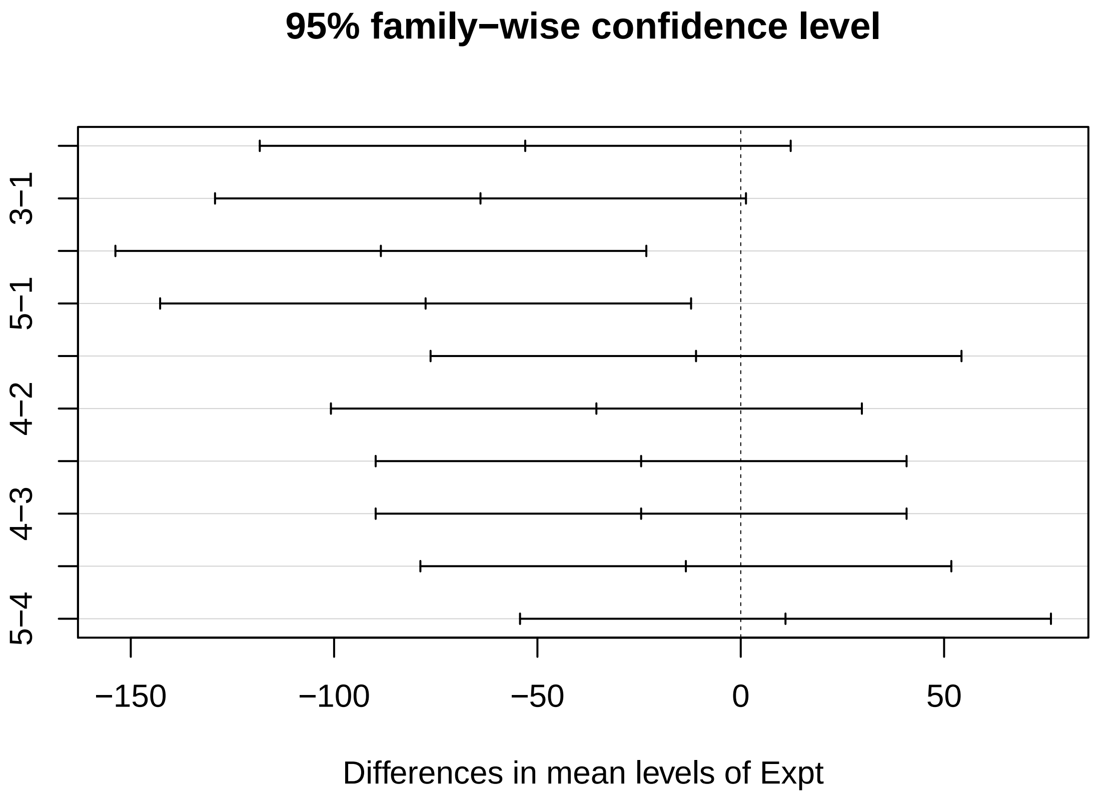
```r
m2 <- aov(data = morley, Speed ~ Expt)
```

The Tukey HSD gives us confidence intervals for the differences in experimental group means. At a 95% confidence level, we see that experimental group 1 records a significantly higher speed of light than groups 4 and 5.

```r
TukeyHSD(m2, which = 1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Speed ~ Expt, data = morley)
##
## $Expt
##      diff        lwr        upr     p adj
## 2-1 -53.0 -118.28006  12.280058 0.1679880
## 3-1 -64.0 -129.28006   1.280058 0.0574625
## 4-1 -88.5 -153.78006 -23.219942 0.0025733
## 5-1 -77.5 -142.78006 -12.219942 0.0115793
## 3-2 -11.0  -76.28006  54.280058 0.9899661
## 4-2 -35.5 -100.78006  29.780058 0.5571665
## 5-2 -24.5  -89.78006  40.780058 0.8343360
## 4-3 -24.5  -89.78006  40.780058 0.8343360
## 5-3 -13.5  -78.78006  51.780058 0.9784065
## 5-4  11.0  -54.28006  76.280058 0.9899661
```

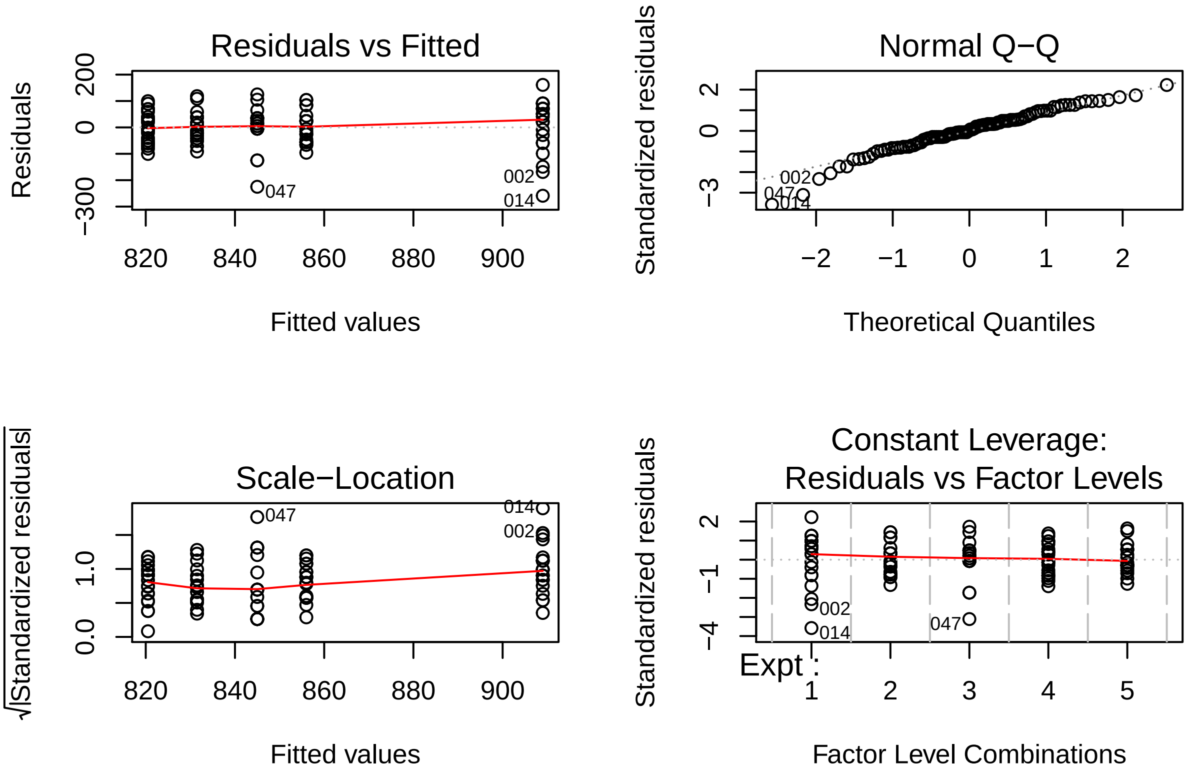
```r
TukeyHSD(m2, which = 1) %>% plot()
```



In terms of the original data, since we know the speed of light is a constant, we can interpret the experimental group effects as the estimated delay of each system. Since nothing can travel faster than the speed of light through a vaccuum, if we see that groups are recording speeds faster than c, we could incorporate that information into the model.

Inspecting the residual plot in the top left shows that the mean residual is approximately 0 with consistent variance across our range of fitted values. The QQ plot shows that the residuals match their theoretical quantiles fairly well, which means the residuals look like a sample from the normal distribution. The bottom right plot shows us that there is roughly equal residual variance across the experimental groups.

```r
par(mfrow = c(2,2))
plot(m2)
```



Collectively, the F-tests and residual plots show us that modeling speed with an intercept and experimental group factors leaves us with normally distributed residuals with consistent variance across groups. Since our residuals match our initial assumptions, there is no reason to second guess our model. This indicates there is a discernible group effect, negligible run effect, and random variation.

Additional Problem 1:

The data frames below are available in the ISwR package. Make sure you install the package and type library(ISwR) before using the data.

(a) Perform a two-way analysis of variance on the tb.dilute data. Explain your model, analysis and conclusions.

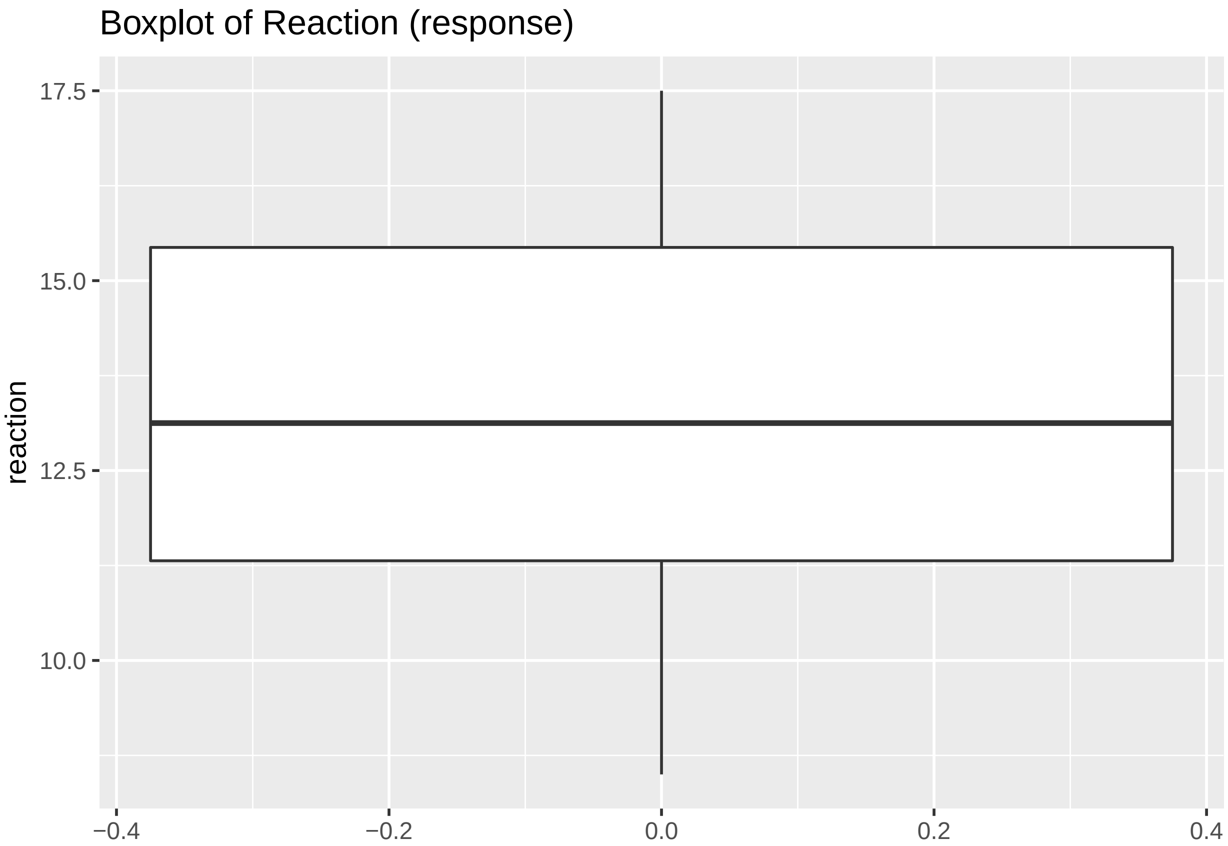
```r
library(ISwR)
data(tb.dilute)
```

```
tb.dilute <- tb.dilute %>%
  mutate(logdose = factor(logdose, ordered = TRUE, levels = c("-0.5", "0", "0.5")))
```

We have 3 columns in this dataset: reaction (numerical), animal (factor) and logdose (factor). I assume we want to model reaction with logdose and animal as the group effect and animal as the run effect. It makes sense to convert logdose into an ordered factor. We also want to check the distribution of our resposne variable. Our sample size is only 18, transformations on the response variable don't make sense.

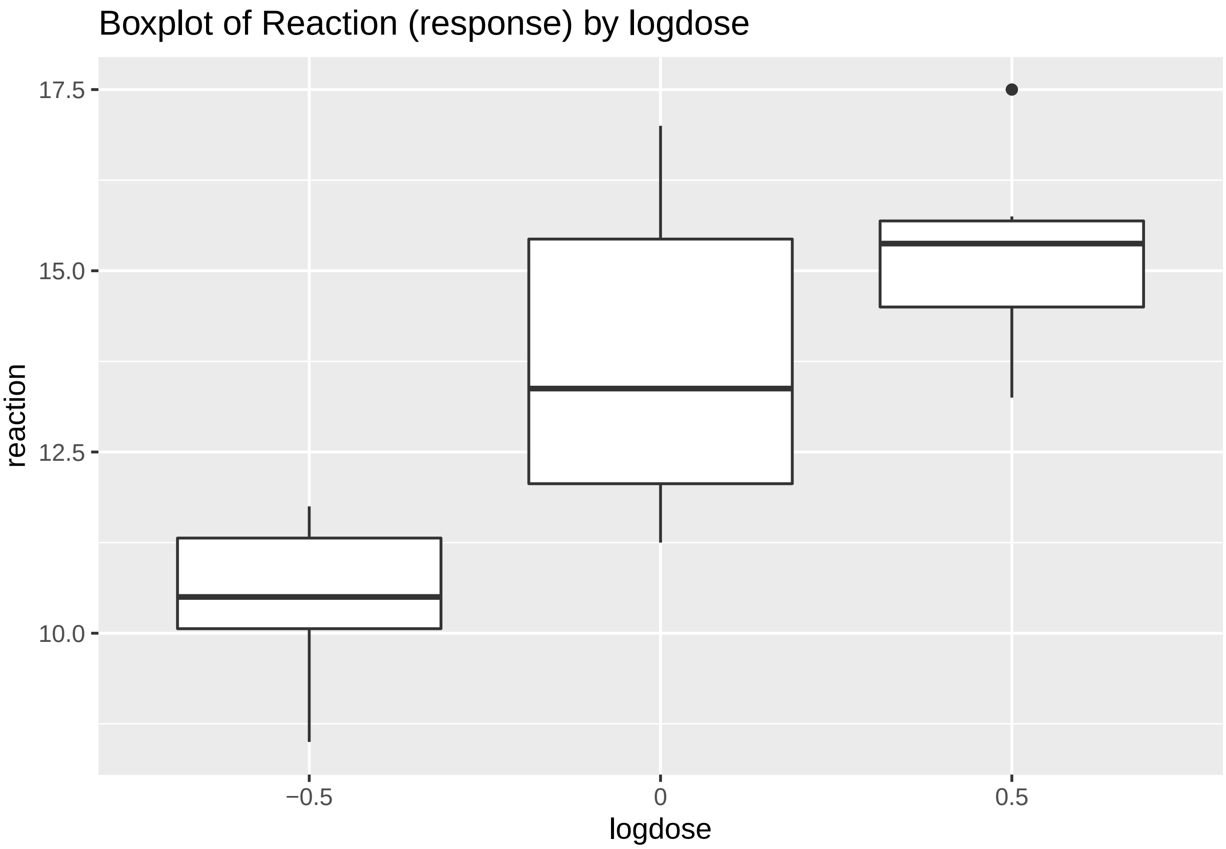
```
tb.dilute <- tb.dilute %>%
  mutate(logdose = factor(logdose, ordered = TRUE, levels = c("-0.5", "0", "0.5")))

ggplot(data = tb.dilute, aes(y = reaction)) + geom_boxplot() + ggtitle("Boxplot of Reaction (response)"
```



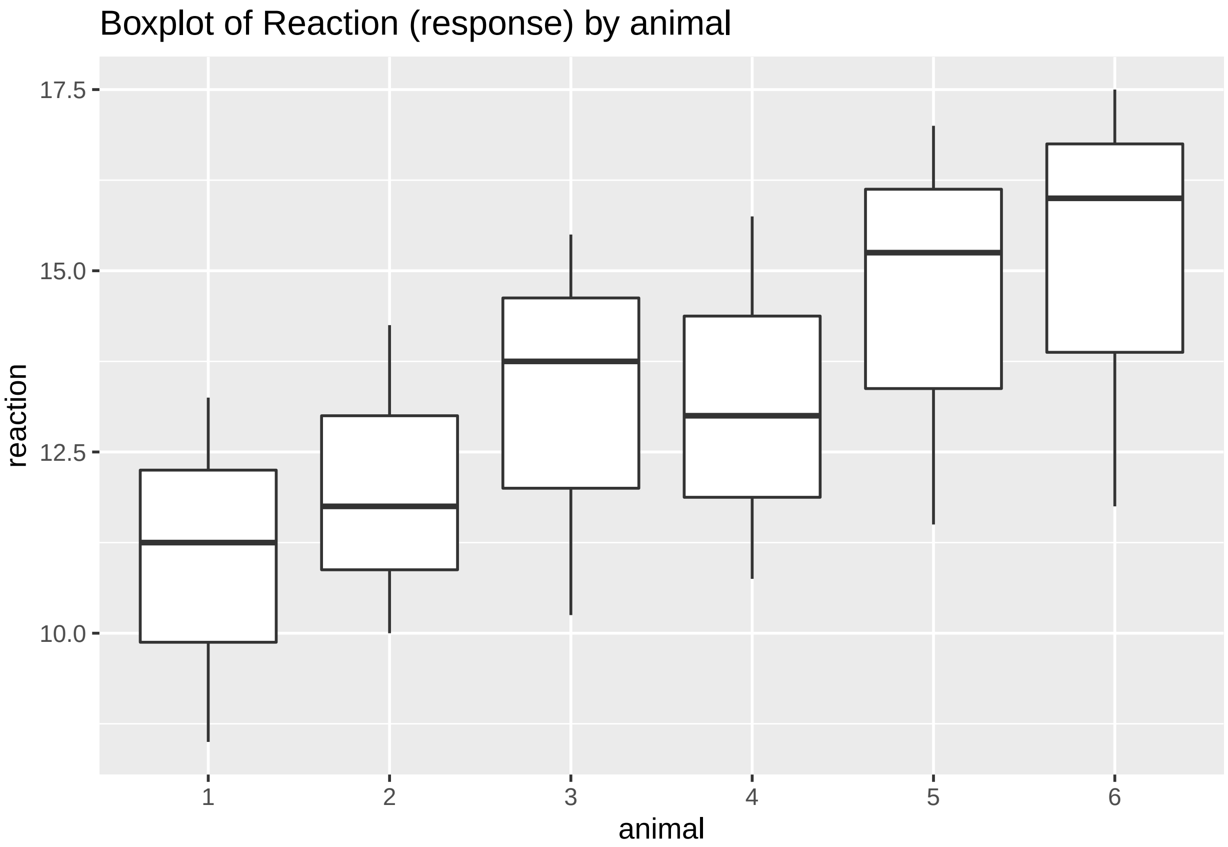
When we look at the boxplot of reaction for each logdose, we see indications that the groups have different reaction distributions.

```
ggplot(data = tb.dilute, aes(y = reaction, x = logdose)) + geom_boxplot() + ggtitle("Boxplot of Reactio
```



Looking at the boxplots of reaction for each animal, we see some variance across animals, but it is important to remember we only have 18 data points spread across 6 groups.

```
ggplot(data = tb.dilute, aes(y = reaction, x = animal)) + geom_boxplot() + ggtitle("Boxplot of Reaction
```



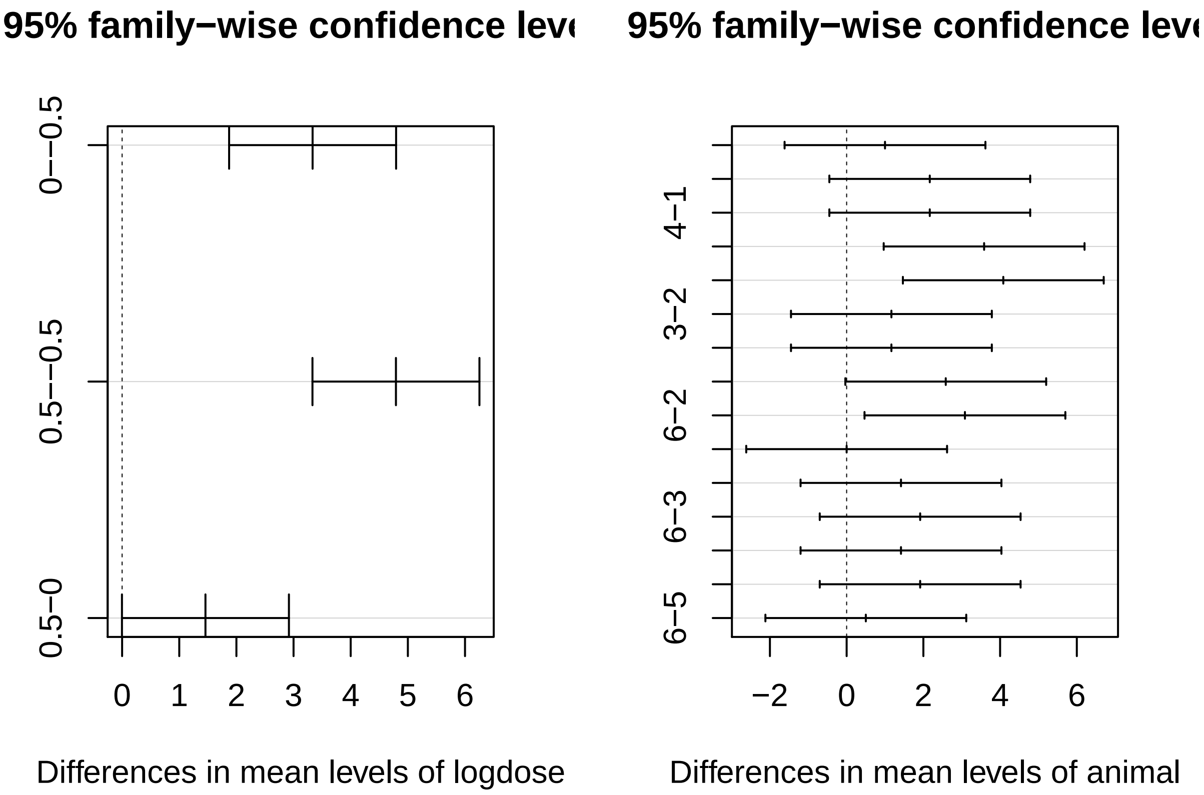
Now, we have inspected the response variable and its distributions across our two factor variables and see some visual evidence that there could be significant factor effects. We also noted there is no reason to transform our response variable. We plug the data into a two-way analysis of variance. We do not include an interaction effect, because there is only one occurance of each combination of logdose and animal, accordingly, if our model included parameters for the interaction effects we would have a perfect fit. We model we are fitting is $reaction_{i,j} = \mu + logdose_i + animal_j + \epsilon_{i,j}$, where, $logdose_i$ is the block (logdose) effect for $i \in \{1, 2, 3\}$, $animal_j$ is the method effect for $j \in \{1, \ldots, 6\}$ and indpendent $\epsilon_{i,j}$ $N(0, \sigma^2)$. The ANOVA table returned by the model has a very high F-statistic for logdose and a smaller, but substantial F-value for animal. The F-test here are evaluating the null hypotheses that $logdose_i = 0 \forall i$ and $animal_j = 0 \forall j$, respectively. We are comparing the models with parameters again the model of the grand sampke mean. The ANOVA table indicates that, at a 99% confidence level, we would reject both of these null hypotheses.

```r
m3 <- aov(data = tb.dilute, reaction ~ logdose + animal)
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: reaction
##           Df Sum Sq Mean Sq F value    Pr(>F)
## logdose    2 72.396  36.198 42.4817 1.295e-05 ***
## animal     5 35.208   7.042  8.2641  0.002527 **
## Residuals 10  8.521   0.852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

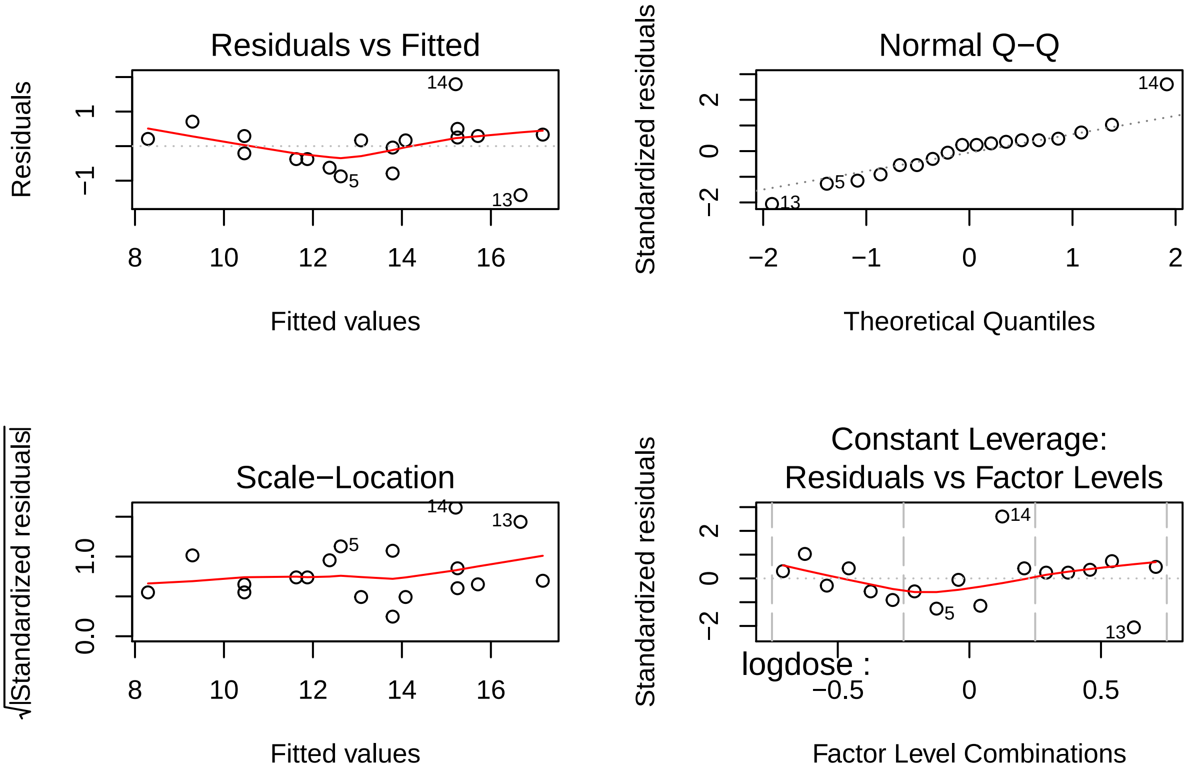
The F-test does not tell us which group means are significantly different. We can find this out from the Tukey HSD test, which tells us that the $logdose = -0.5$ is has a significantly lower group mean reaction than the $logdose = 0$ or $logdose = 0.5$ groups. Furthermore, at a 95% confidence level, we do not reject the null hypothesis that there is a significant different in mean reaction between the $logdose = 0$ and $logdose = 0.5$ groups. We also see that the model detects some significant differences between animal group means.

```r
par(mfrow = c(1,2))
plot(TukeyHSD(m3, which = 1))
plot(TukeyHSD(m3, which = 2))
```



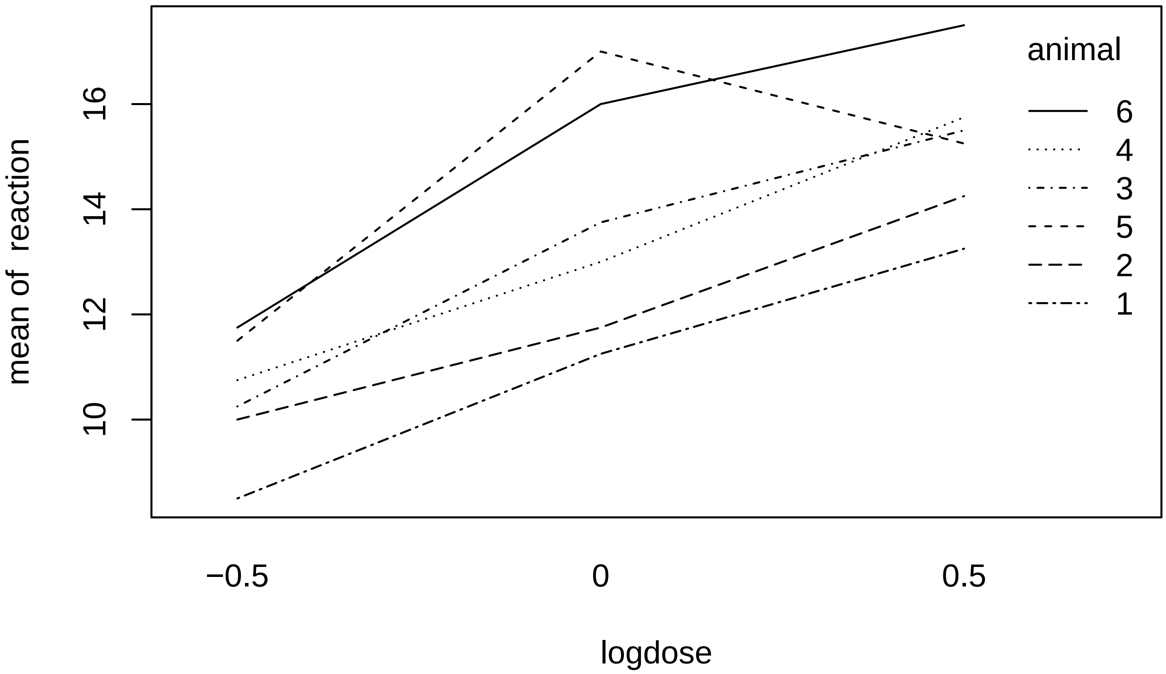
The residual plots from the model show that the residuals have mean = 0 and consistent variance across groups. The standardized residuals approximate the normal distribution, which is the sign of a good fit, even though there is some wiggle. The residual plots do not throw up an red flags for our model.

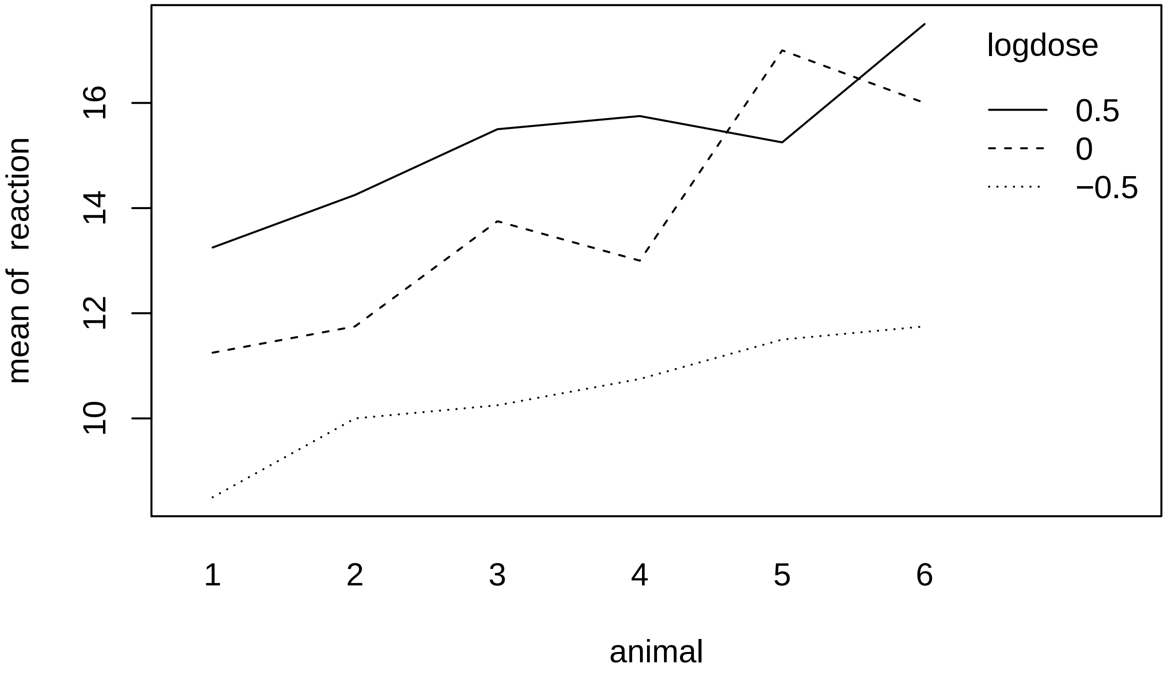
```r
par(mfrow = c(2,2))
plot(m3)
```



We can plot the interactions between logdose and animal factor variables. Recall that we could not parameterize the interaction effcet between these 2 variables because that would result in a model with perfect fit. The interaction plots tell us that $logdose = -0.5$ has the lowest mean reaction time across all animals.

```
with(data=tb.dilute, expr={
   interaction.plot(logdose, animal, response=reaction)
   interaction.plot(animal, logdose, response=reaction)
 })
```





In conclusion, our two-way analysis of variance model indicates that there are significnt group effects for both logdose and animal. Logdose accounts for a greater proportion of the variance in reaction time than animal and $logdose = -0.5$ has a significantly lower mean reaction time than either of the other logdose groups. At a 95% confidence level, we (barely) fail to reject the null hypothesis that $logdose = 0$ and $logdose = 0.5$ have a significant difference in mean reaction time. We also identify significant differences between several animal's mean reaction times. If we had a larger sample size, with multiple observations for various combinations of animal and logdose, we could 'afford' to parameterize the interaction effects.

(b) Analyze the vitcap2 dataset in the ISwR package using analysis of covariance. Explain your model, analysis and conclusions.

We load the data, check the structure, convert group into a factor variable and note that we are interested in modeling vital.capacity as our response variable. It does not make sense to convert age into a factor variable, so we are modeling vital.capacity with one factor and one quantitative variable.

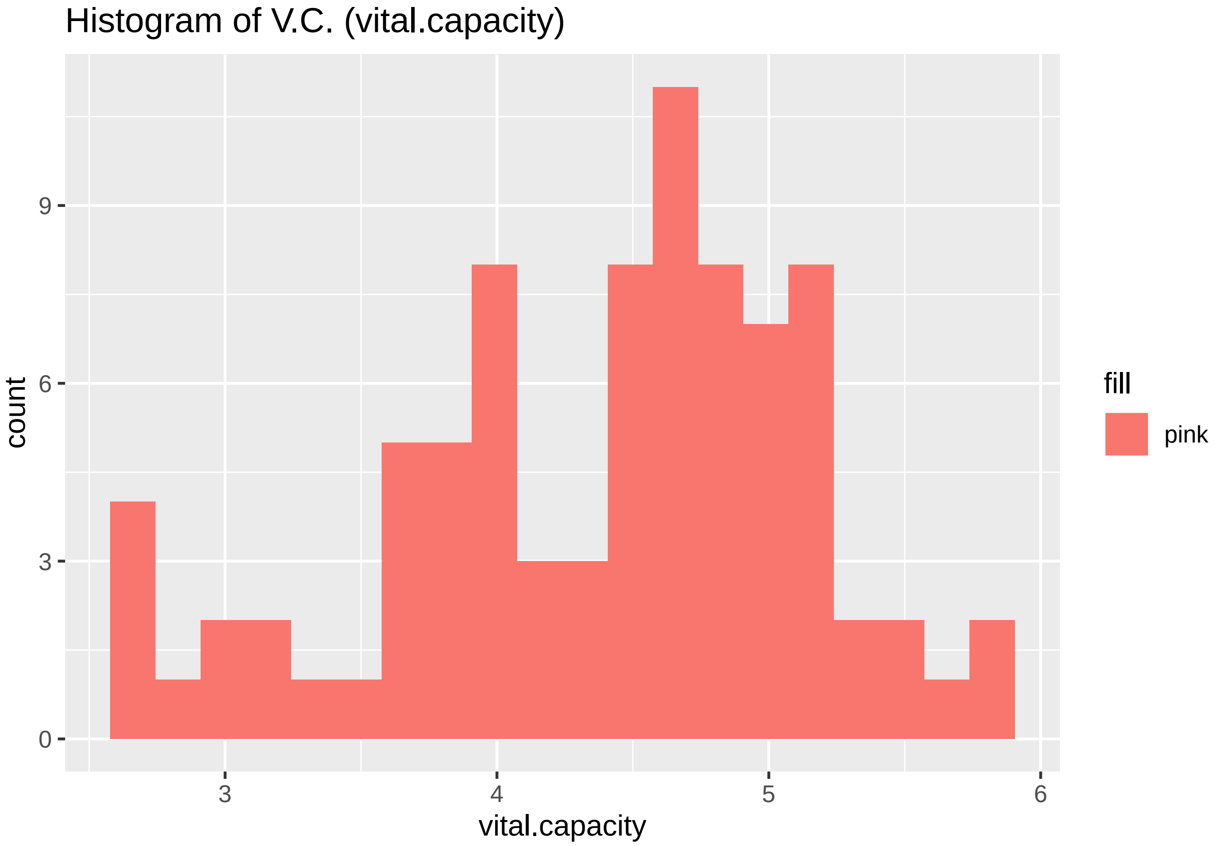
```
data(vitcap2)
str(vitcap2)
```

```
## 'data.frame':    84 obs. of  3 variables:
##  $ group         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age           : int  39 40 41 41 45 49 52 47 61 65 ...
##  $ vital.capacity: num  4.62 5.29 5.52 3.71 4.02 5.09 2.7 4.31 2.7 3.03 ...
```

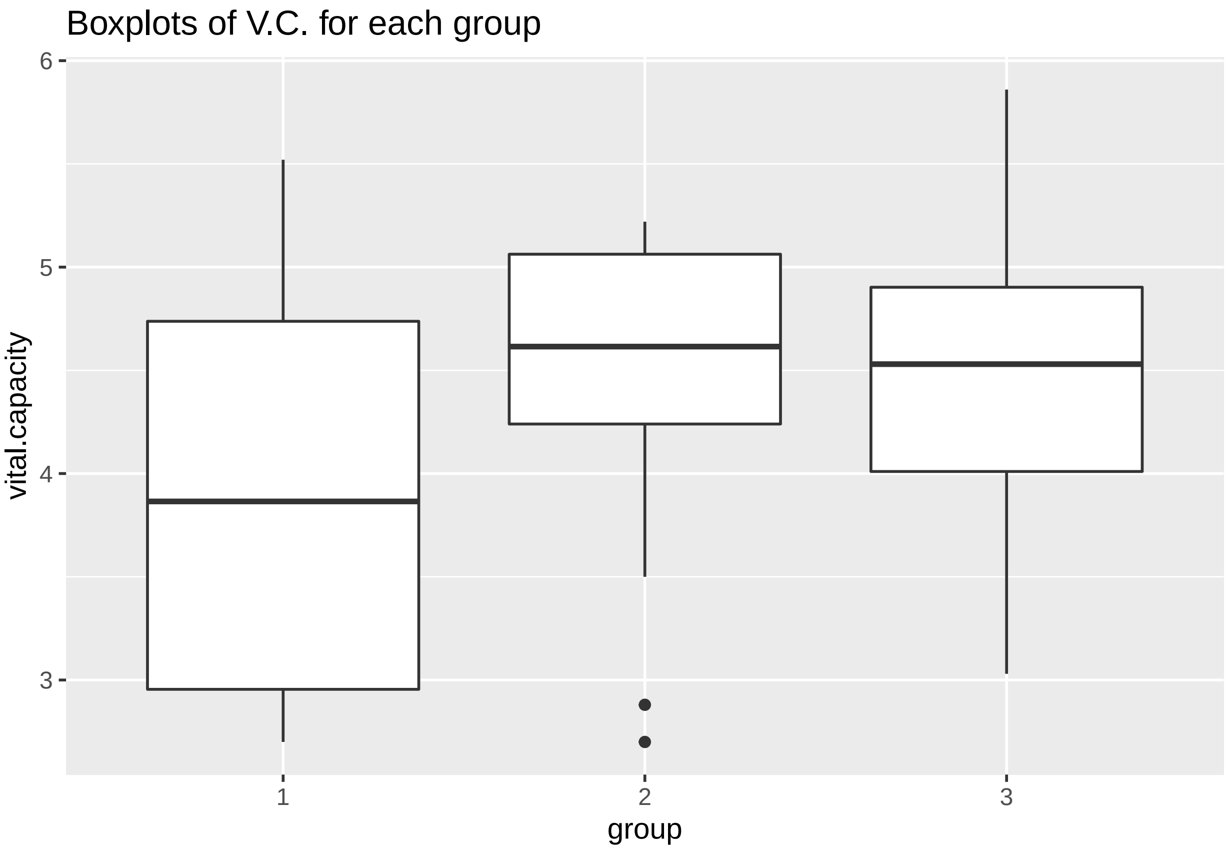
```
vitcap2 <- vitcap2 %>% mutate(group = factor(group))
```

Before we build a model, it is important to look at the distribution of our response variable, by itself and across our group variable. Looking at the following visualization, we note that vital.capacity appears to have a distribution that is symmetric around the median. There is no reason to transform the response variable. Looking at the boxplots for each group, we see that the group's have overlapping distribution, but it is not clear if there are significant differences between group means.

```
ggplot(data = vitcap2, aes(x = vital.capacity, fill = "pink")) + geom_histogram(bins = 20) +
  ggtitle("Histogram of V.C. (vital.capacity)")
```

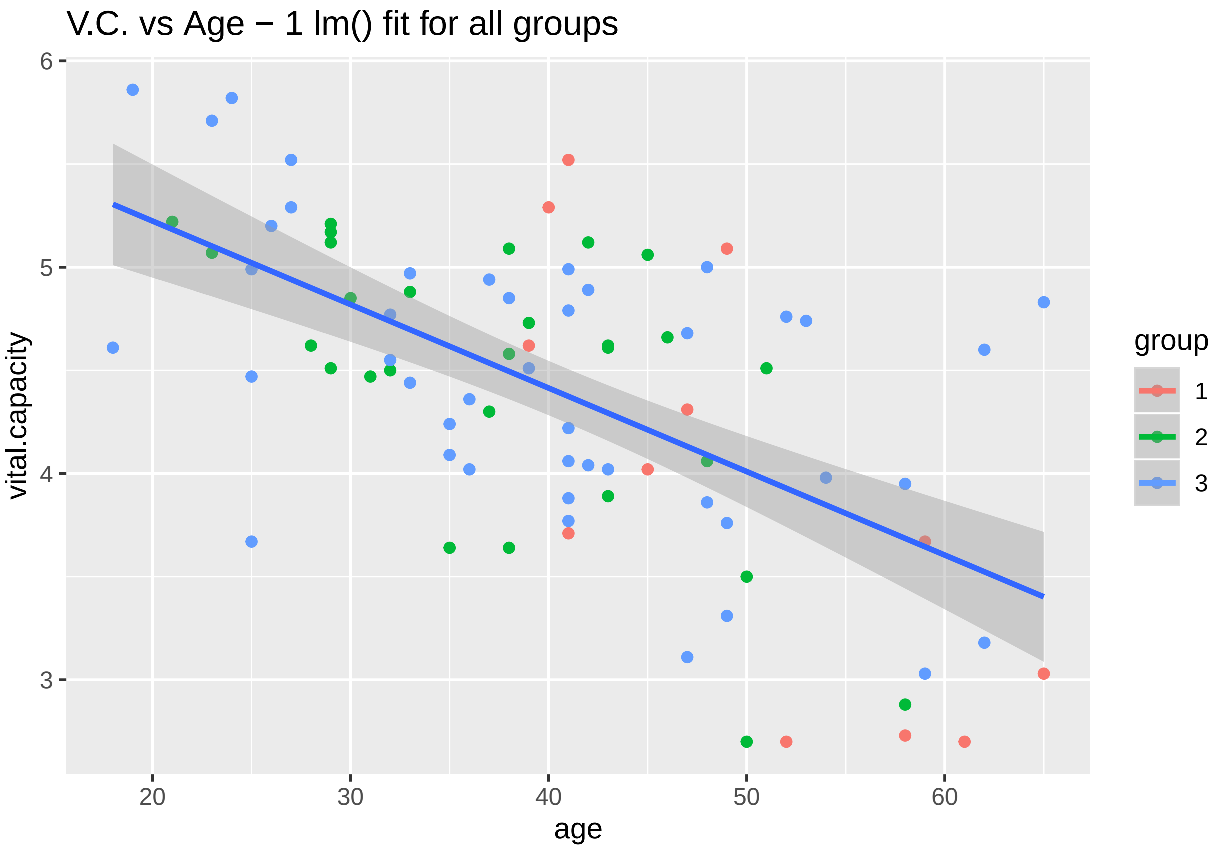


```
ggplot(data = vitcap2, aes(x = group, y = vital.capacity)) + geom_boxplot() +
  ggtitle("Boxplots of V.C. for each group")
```



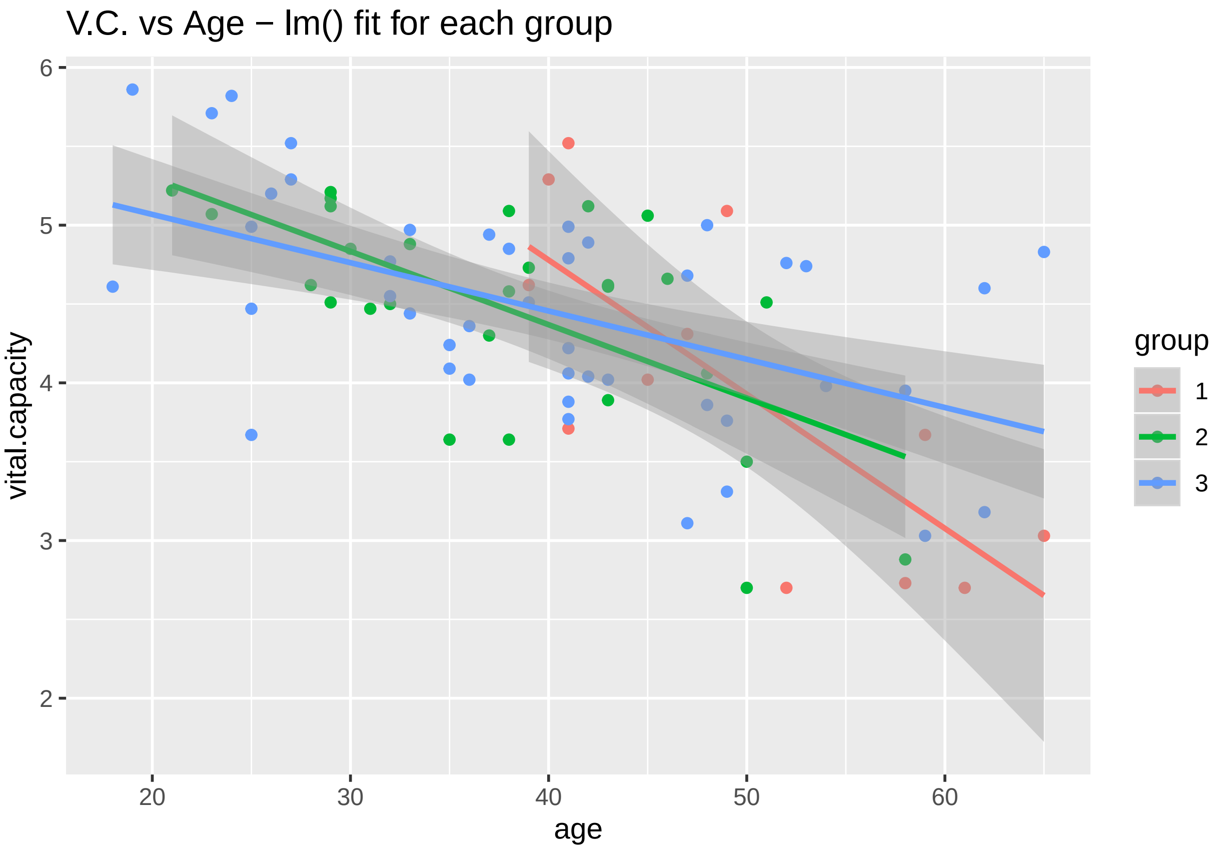
We are also interested in the effect of age. The following scatter plot shows indicates a negative correlation between age and vital capacity. The plot has has a one linear model for all the data points and it fits pretty well.

```
ggplot(data = vitcap2, aes(x = age, y = vital.capacity, col = group)) + geom_point() +
  geom_smooth(aes(group = 1), method = "lm") + ggtitle("V.C. vs Age - 1 lm() fit for all groups")
```



This next scatter plot has a different linear model for each group - we see that R fits slightly different slopes for each group. However, we cannot tell if the group slopes are significantly better than the global slope in the above plot. We will fit a model to help us determine if there should be a different parameter for age for the different groups.

```
ggplot(data = vitcap2, aes(x = age, y = vital.capacity, col = group)) + geom_point() +
  geom_smooth(method = "lm")  + ggtitle("V.C. vs Age - lm() fit for each group")
```



Now we are ready to code a model. As noted above, we are interested in the interaction effect between age and group here, so we will include that. We also note that we can afford to parameterize the interaction effect because we have repeated observations for many combinations of age and group.

```r
m4 <- aov(data = vitcap2, vital.capacity ~ age*group)
anova(m4)
```

```
## Analysis of Variance Table
##
## Response: vital.capacity
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## age        1 17.4446 17.4446 49.4159 6.918e-10 ***
## group      2  0.1617  0.0808  0.2290   0.79584
## age:group  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
```

The model summary gives us F-statistics that provide very strong evidence to reject the null hypothesis that the age parameter is equal to zero. It also gives weak evidence to reject the null hypothesis that group effects are all equal to zero. At a 95% confidence level, we also identify a significant interaction effect between age and group. We cannot include the interaction effect in the model if we do not include the original group term.

If we were using a 99% confidence level (alpha = 0.01), or if we had a 95% confidence level with corrections for multiple testing, then we would be left with the model $vc \sim \beta_0 + \beta_{age} * age$, so the same intercept and slope for all groups.

If we were using a 95% confidence interval and did not make corrections for multiple testing, then we would use the model include the group and group:age interaction efffects in the model, so each group would have its own intercept and age slope.

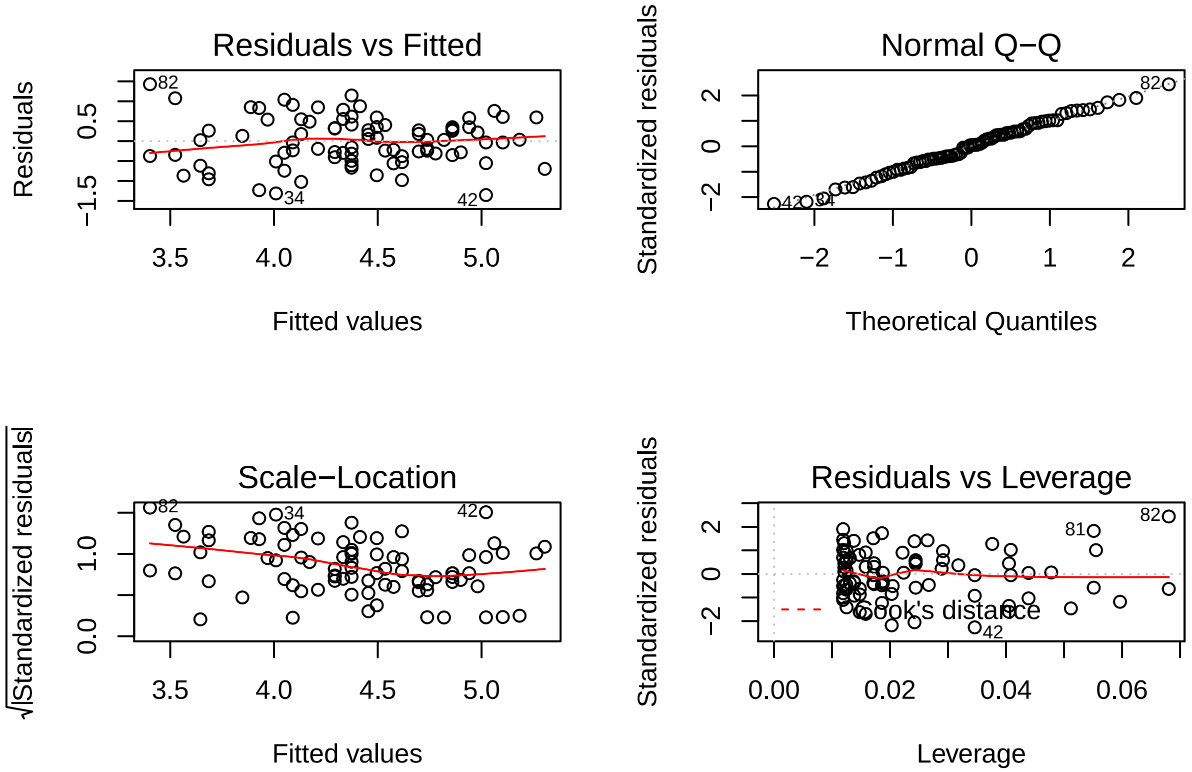
Using a 99% confidence level, the model we are left with has an models vital.capacity with an intercept term and a coefficent for age.

```
m5 <- lm(data = vitcap2, vital.capacity ~ age)
summary(m5)
```

```
##
## Call:
## lm(formula = vital.capacity ~ age, data = vitcap2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35136 -0.37332  0.02796  0.40735  1.42776
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.033316   0.247487  24.378  < 2e-16 ***
## age         -0.040478   0.005881  -6.883 1.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6068 on 82 degrees of freedom
## Multiple R-squared:  0.3662, Adjusted R-squared:  0.3584
## F-statistic: 47.37 on 1 and 82 DF,  p-value: 1.082e-09
```

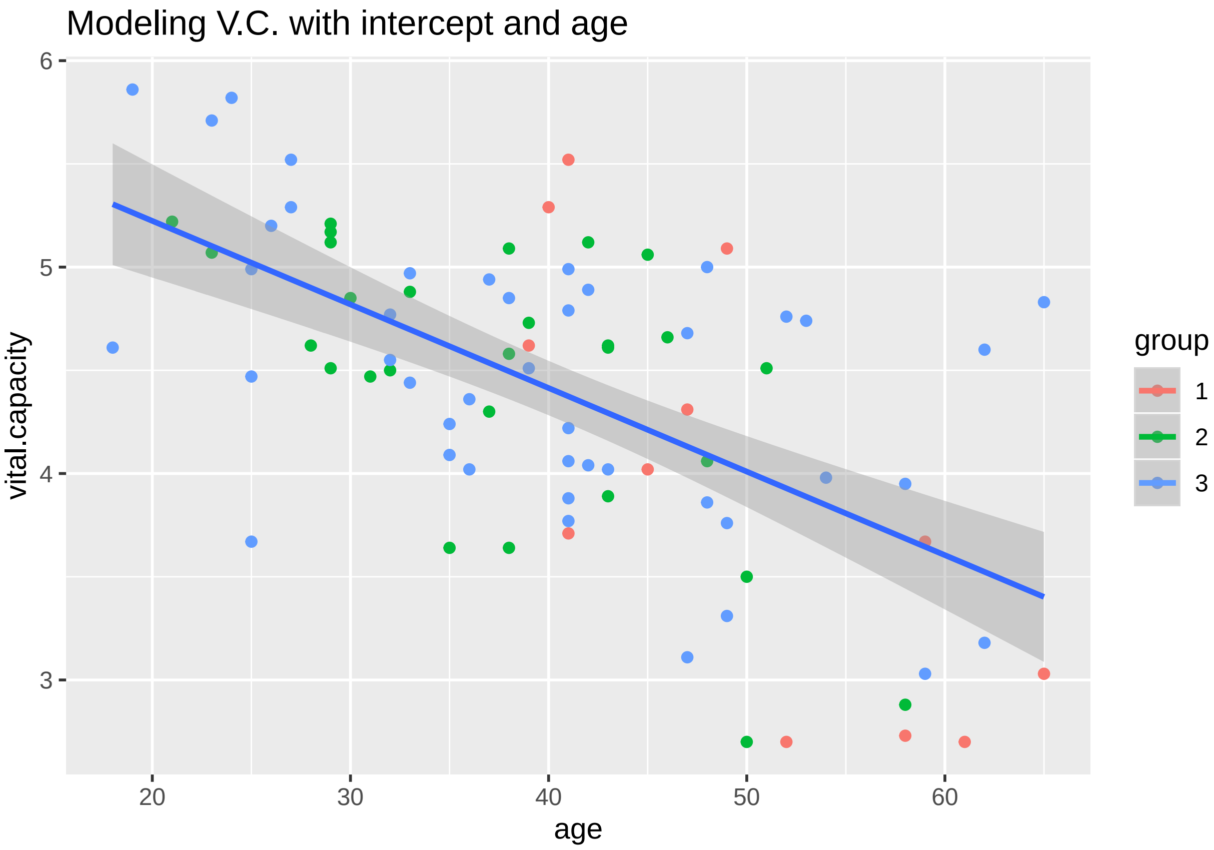
The residual plots from the model that follow show that the residuals have mean = 0 and consistent variance across groups. The standardized residuals approximate the normal distribution, which is the sign of a good fit. The residual plots do not throw up an red flags for our model.

```
par(mfrow = c(2,2))
plot(m5)
```



Finally, we bring back the visualization of the model we are happiest with: $vital.capacity = 6.033316 - 0.040478 * age$

```
ggplot(data = vitcap2, aes(x = age, y = vital.capacity, col = group)) + geom_point() +
  geom_smooth(aes(group = 1), method = "lm") + ggtitle("Modeling V.C. with intercept and age")
```



Additional Problem 2:

An experiment was run to investigate the amount of weight loss (in grams) by ground beef hamburgers after grilling or frying, and how much the weight loss is affected by the percentage of fat in the beef before cooking. The experiment involved 2 factors: cooking method (with two levels, frying and grilling) and fat content (with 3 levels: 10%, 15% and 20%). Hamburger patties weighing 110 g each were prepared from meat with the required fat content. There were 30 "cooking time slots which were randomly assigned to the treatments in such a way that each treatment was observed 5 times. The patty weights after cooking are shown below:

Method Frying, Fat Content 10%: 81, 88, 85, 84, 84

Method Frying, Fat Content 15%: 85, 80, 82, 80, 82

Method Frying, Fat Content 20%: 71, 77, 72, 80, 80

Method Grilling, Fat Content 10%: 84, 84, 82, 81, 86

Method Grilling, Fat Content 15%: 83, 88, 85, 86, 88

Method Grilling, Fat Content 20%: 78, 75, 78, 79, 82

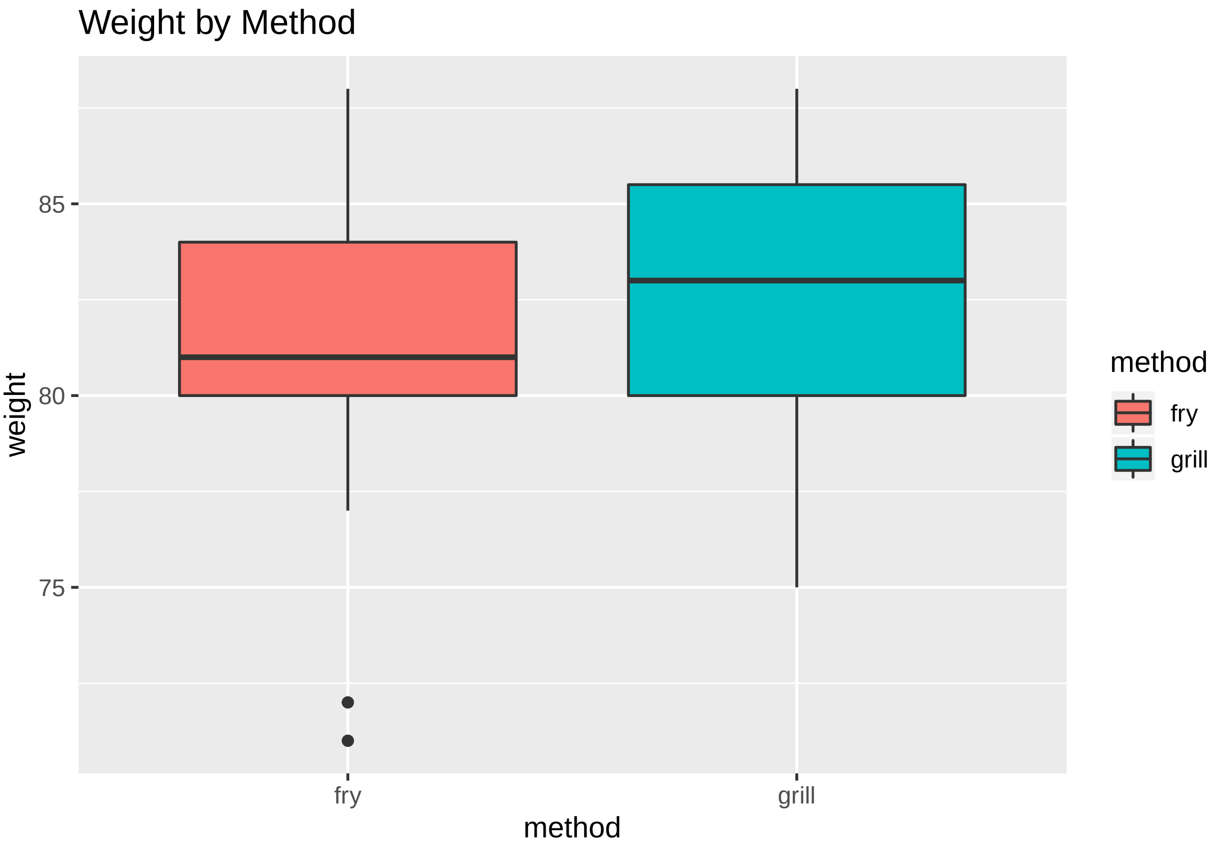
```
df <- data.frame(method = c(rep("fry", 15), rep("grill", 15)),
                 fat = c(rep(10, 5), rep(15, 5), rep(20, 5), rep(10, 5), rep(15, 5), rep(20, 5)),
                 weight = c(81, 88, 85, 84, 84,
                            85, 80, 82, 80, 82,
                            71, 77, 72, 80, 80,
                            84, 84, 82, 81, 86,
                            83, 88, 85, 86, 88,
                            78, 75, 78, 79, 82)) %>%
  mutate(fat = factor(fat))
str(df)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ method: Factor w/ 2 levels "fry","grill": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fat   : Factor w/ 3 levels "10","15","20": 1 1 1 1 1 2 2 2 2 2 ...
##  $ weight: num  81 88 85 84 84 85 80 82 80 82 ...
```

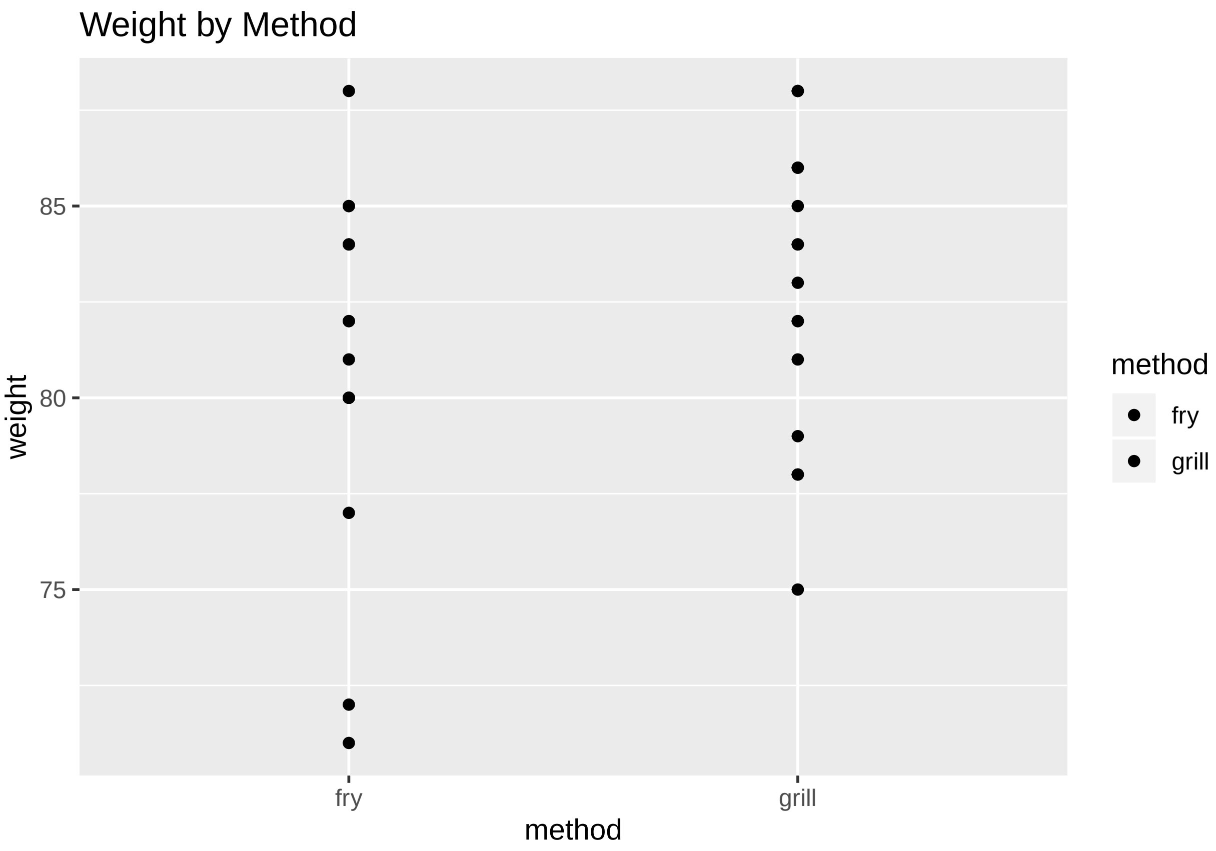
(i) Perform EDA

We begin by stating we want to model weight with fat (factor) and method (factor). Boxplots of our response variable (weight) broken down by each of the 2 factors is always good. In the boxplots we see similar distributions for the 2 cooking methods, and indications that the 20% fat group may have a smaller group mean than the other 2 levels of the fat factor variable.

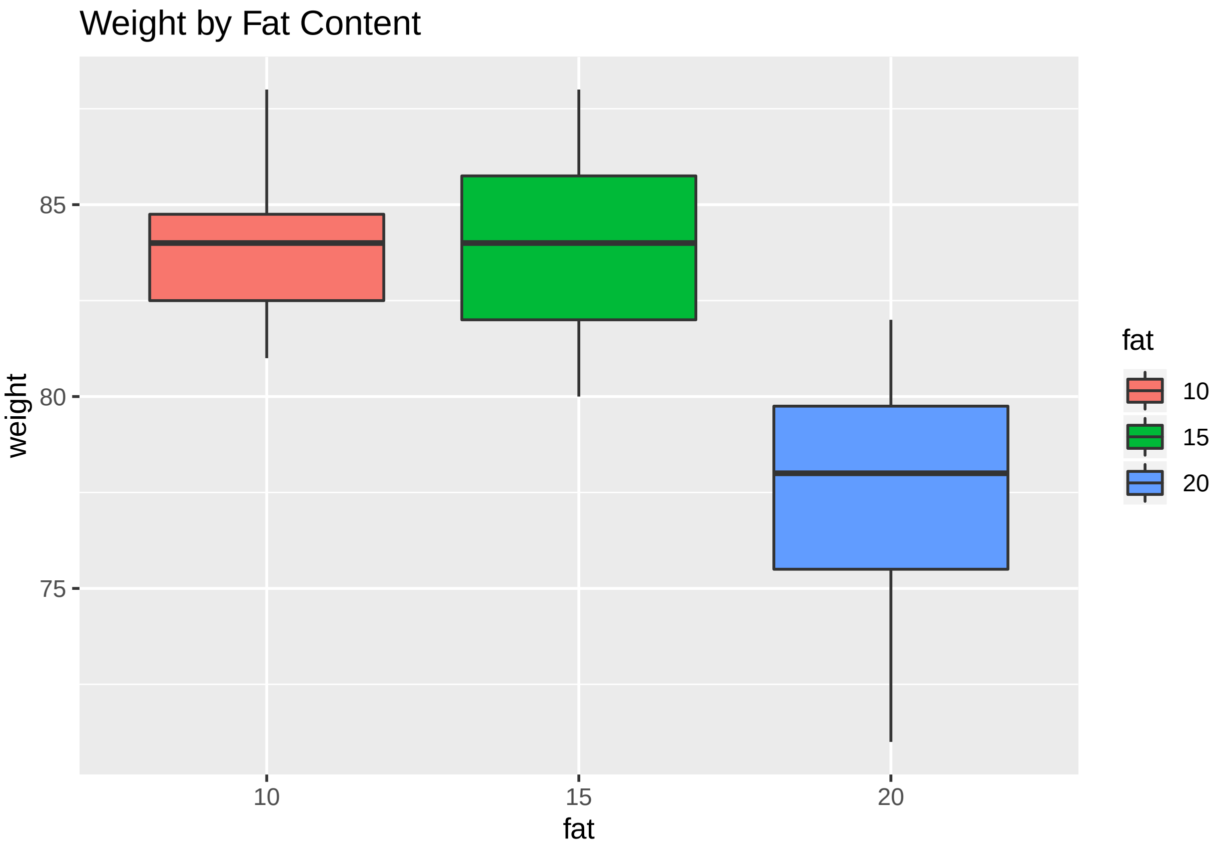
```
par(mfrow = c(2,1))
ggplot(data = df, aes(y = weight, x = method, fill = method)) + geom_boxplot() + ggtitle("Weight by Met
```



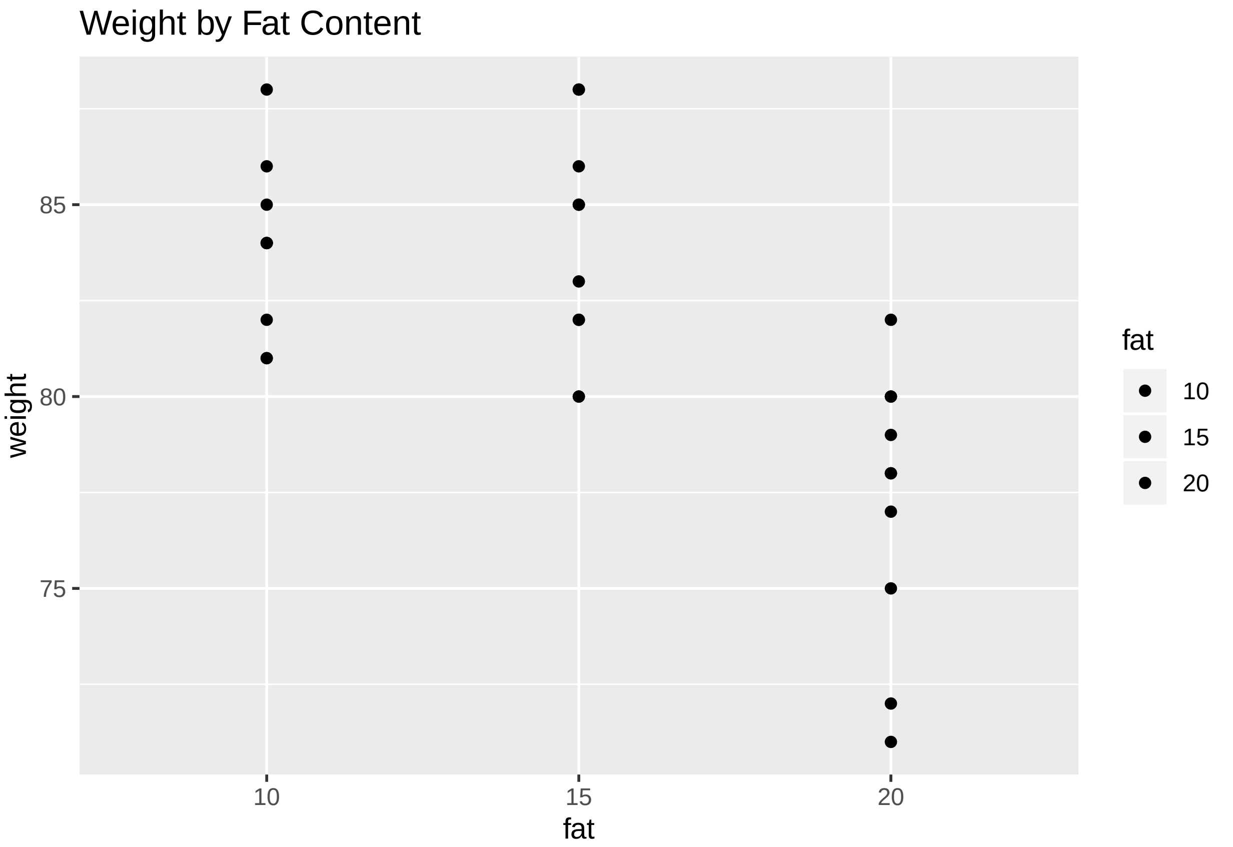
```
ggplot(data = df, aes(y = weight, x = method, fill = method)) + geom_point() + ggtitle("Weight by Metho
```



```
ggplot(data = df, aes(y = weight, x = fat, fill = fat)) + geom_boxplot() + ggtitle("Weight by Fat Conte
```



```
ggplot(data = df, aes(y = weight, x = fat, fill = fat)) + geom_point() + ggtitle("Weight by Fat Content
```



There is not much else we can visualize here, so we will proceed to the modeling.

(ii) Perform a two-way analysis of variance. Explain your model and your conclusions. In particular: (a) Compare the effects of the 3 levels of fat content pairwise, averaging over cooking methods and interpret your results; (b) Give a 95% confidence interval for the average difference in weight after cooking before frying and grilling 110g hamburgers.

We can start with the randomized block model, including the interaction term because we have repeated observations of the combinations of factor levels. Formally, we are evaluating the model: $weight_{i,j} = \mu + fat_i + method_j + interaction_{i,j} + \epsilon_{i,j}$ where fat_i is the effect of the fat content for $i \in \{1,2,3\}$ and $method_j$ is the method effect for $i \in \{1,2\}$ and indpendent $\epsilon_{i,j} \sim N(0, \sigma^2)$. The F-tests in the model are evualating the 3 null hypotheses that (1) $fat_i = 0 \forall i$, (2) $method_j = 0 \forall j$ and (3) $interaction_{i,j} = 0 \forall i, j$.

```
m6 <- aov(data = df, weight ~ method*fat)
summary(m6)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## method       1  26.13   26.13   3.596    0.070 .  
## fat          2 299.27  149.63  20.592 6.21e-06 ***
## method:fat   2  34.87   17.43   2.399    0.112    
## Residuals   24 174.40    7.27                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model summary indicates that the interaction effects are not significant, so we will re-run the model excluding the interaction effects.

```
m6 <- aov(data = df, weight ~ method + fat)
summary(m6)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## method       1  26.13   26.13   3.247  0.0832 .
## fat          2 299.27  149.63  18.591 9.7e-06 ***
## Residuals   26 209.27    8.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test from the second model indicates, that at a 95% confidence level, we should not reject the null hypothesis that $method_j = 0 \forall j$. We should, however, reject the null hypothesis that $fat_i = 0 \forall i$. Removing method, we evaluate a third model, $weight_{i,j} = \mu + fat_i + \epsilon_{i,j}$, which we note is a one-way anova model. The following F-test provides strong evidence to reject the null hypothesis that $fat_i = 0 \forall i$.

```
m6 <- aov(data = df, weight ~ fat)
summary(m6)
```

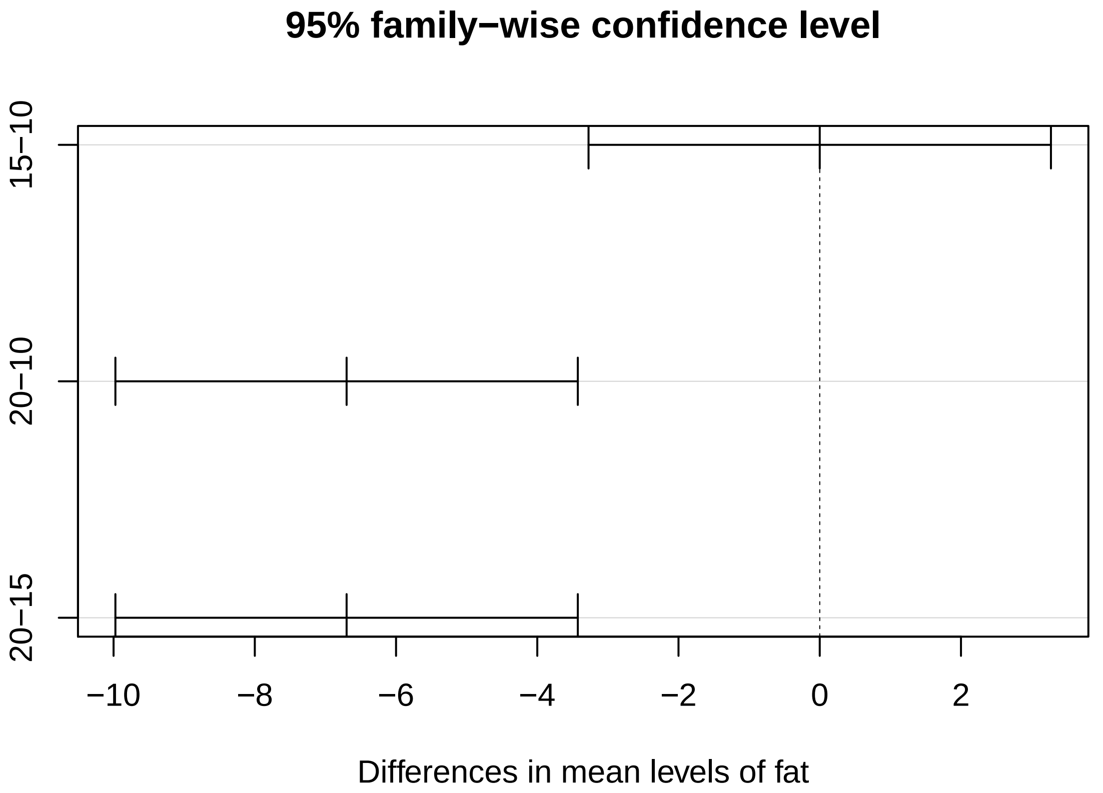
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## fat          2  299.3  149.63   17.16 1.55e-05 ***
## Residuals   27  235.4    8.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We run the TukeyHSD() test to see which groups of fat have significantly different group means of weight. We notice that there is no significant difference detected between 10% and 15%, and that 20% fat has a significantly lower mean weight than either 10% or 15%./

```
TukeyHSD(m6, which = 1)
```

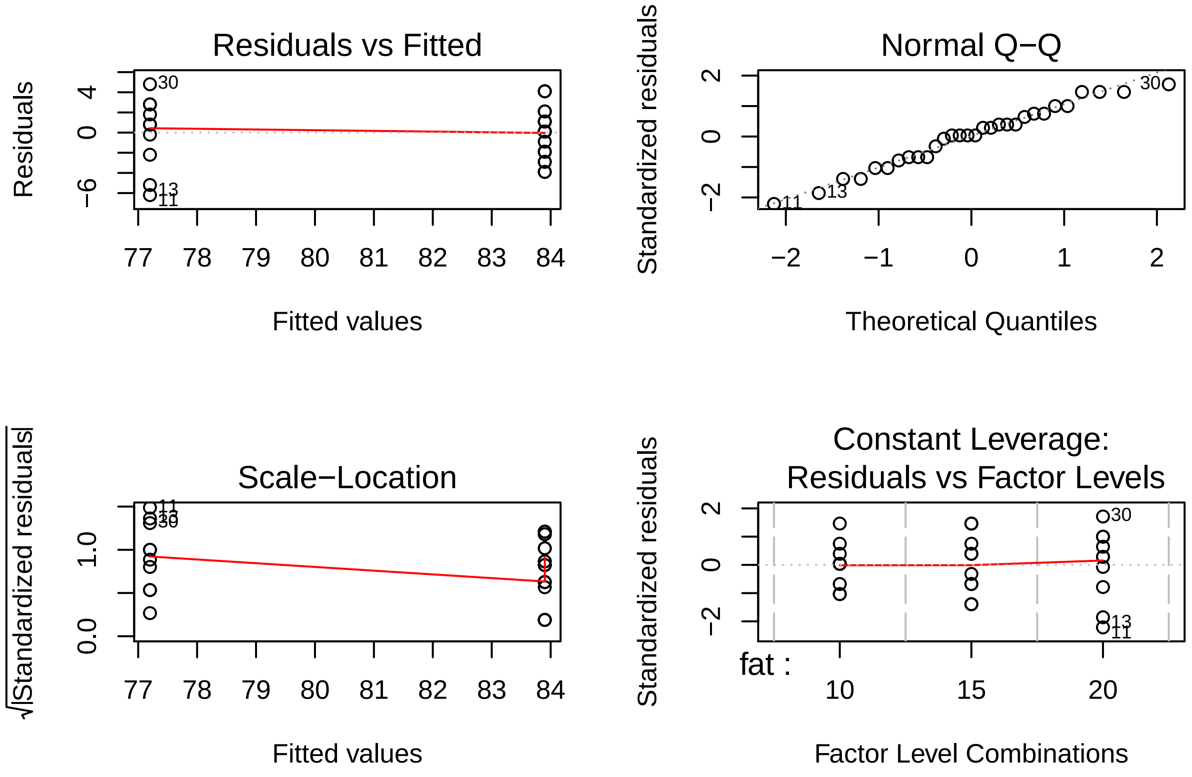
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ fat, data = df)
##
## $fat
##                diff       lwr       upr    p adj
## 15-10  1.421085e-14 -3.274056  3.274056 1.00e+00
## 20-10 -6.700000e+00 -9.974056 -3.425944 7.22e-05
## 20-15 -6.700000e+00 -9.974056 -3.425944 7.22e-05
```

```
plot(TukeyHSD(m6, which = 1))
```



The residual plots from the model show that residual mean and variance is constant across groups and there is very little deivation in the QQ plot. This indicates that the model fits well and we did not violate our initial assumptions about the $\epsilon_{i,j}$ distribution.

```r
par(mfrow = c(2,2))
plot(m6)
```



Finally, we want the 95% confidence interval for the average difference in weight after cooking before frying and grilling 110g hamburgers. Although we determined that there were no significant difference in the mean weights for the different cooking methods

Although we determined that the cooking method was not significant, we can compose a linear model with fat content and method, and then return the confidence interval as desired.

```
confint(lm(data = df, weight ~ method + fat))
```

```
##                   2.5 %    97.5 %
## (Intercept) 80.8372707 85.096063
## methodgrill -0.2627293  3.996063
## fat15       -2.6079668  2.607967
## fat20       -9.3079668 -4.092033
```

This shows us that the 95% confidence interval for the difference in cooking method is $-0.2627293, 3.996063$.