STAT 516 hw 5

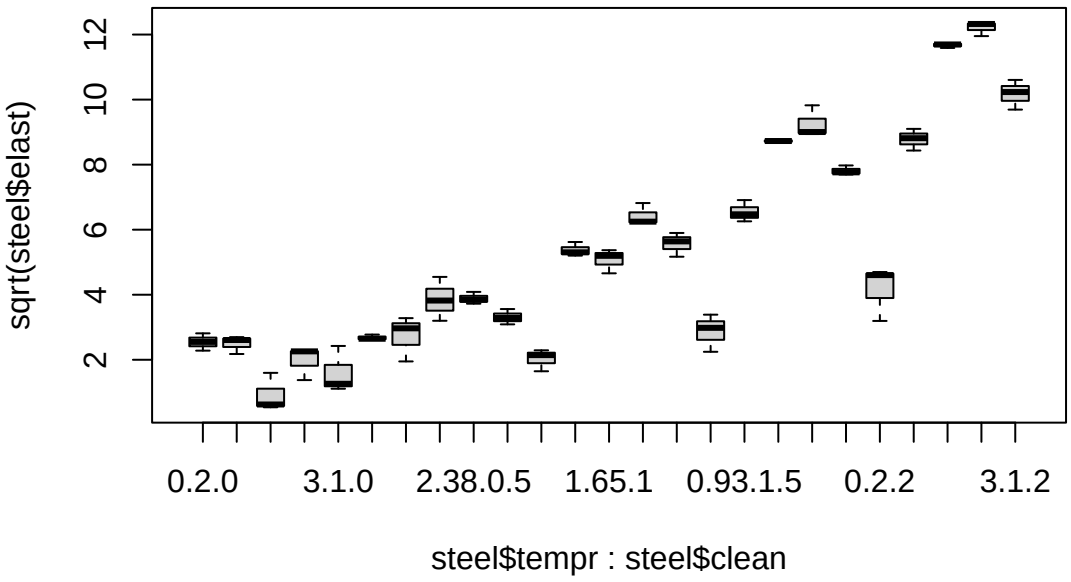
Solutions

Chp 9 Ex 7

```
file <- "Data Tables 4th edition/Chapter 9/datatab_9_27.prn"
steel <- read.csv(file, sep = " ",colClasses = c("factor","factor","numeric"))
```

Some exploration of the data shows that a square-root transformation of the response makes the response values more normally distributed around the treatment means. We therefore use the square root of the elasticity values throughout.

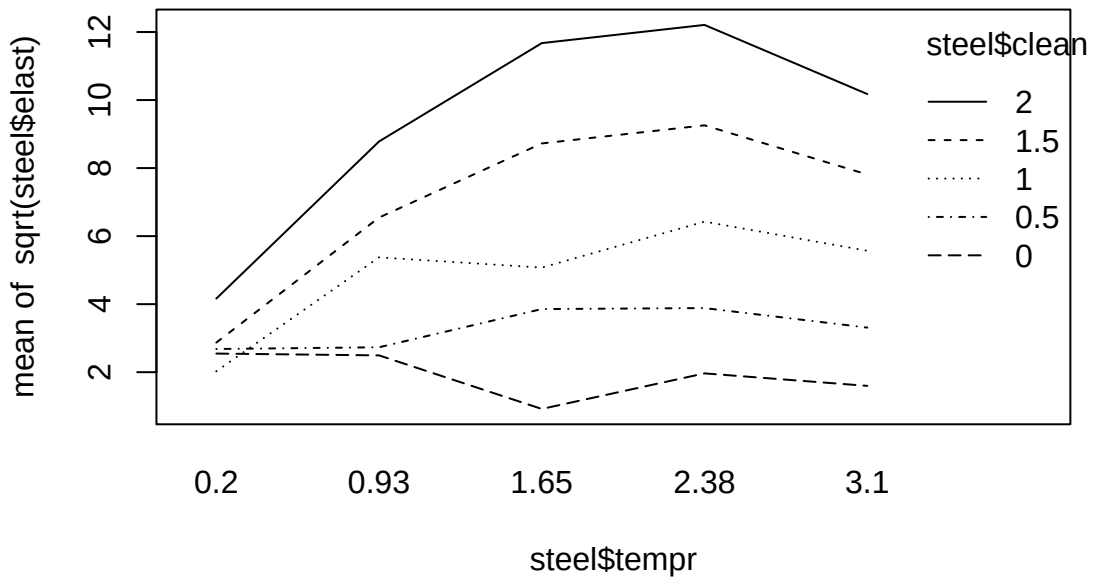
```
boxplot(sqrt(steel$elast) ~ steel$tempr + steel$clean)
```



There are only three replicates at each treatment combination, so each boxplot is constructed with only three values, so viewing boxplots is not the most sensible way to visualize this data.

Because of the small number of replicates it may be better to look at an interaction plot right away (though the interaction plot itself does not tell us whether an interaction is significant).

```
interaction.plot(steel$tempr,steel$clean,sqrt(steel$elast))
```



The interaction plot shows some evidence of interaction in that the slopes of the lines corresponding to the different levels of the cleaning agent factor are not all the same across the levels of the temperature factor.

Now we produce the ANOVA table.

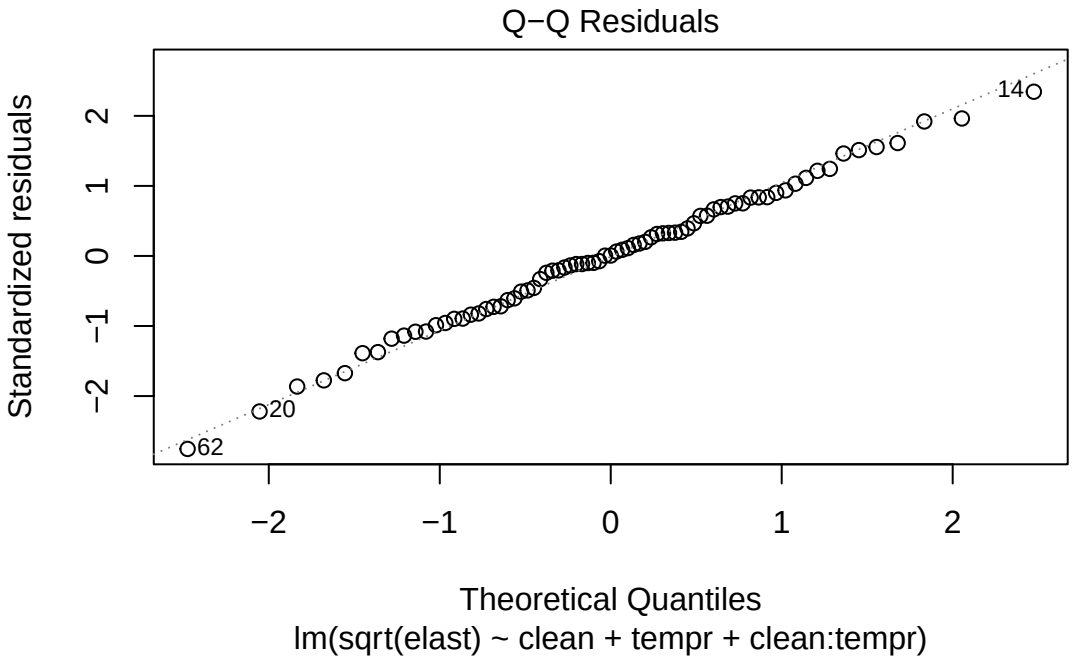
```
lm_out <- lm(sqrt(elast) ~ clean + tempr + clean:tempr, data = steel)
anova(lm_out)
```

```
Analysis of Variance Table

Response: sqrt(elast)
            Df Sum Sq Mean Sq F value    Pr(>F)
clean        4 533.62 133.405 715.540 < 2.2e-16 ***
tempr        4 131.83  32.958 176.774 < 2.2e-16 ***
clean:tempr 16 113.84   7.115  38.162 < 2.2e-16 ***
Residuals   50   9.32   0.186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

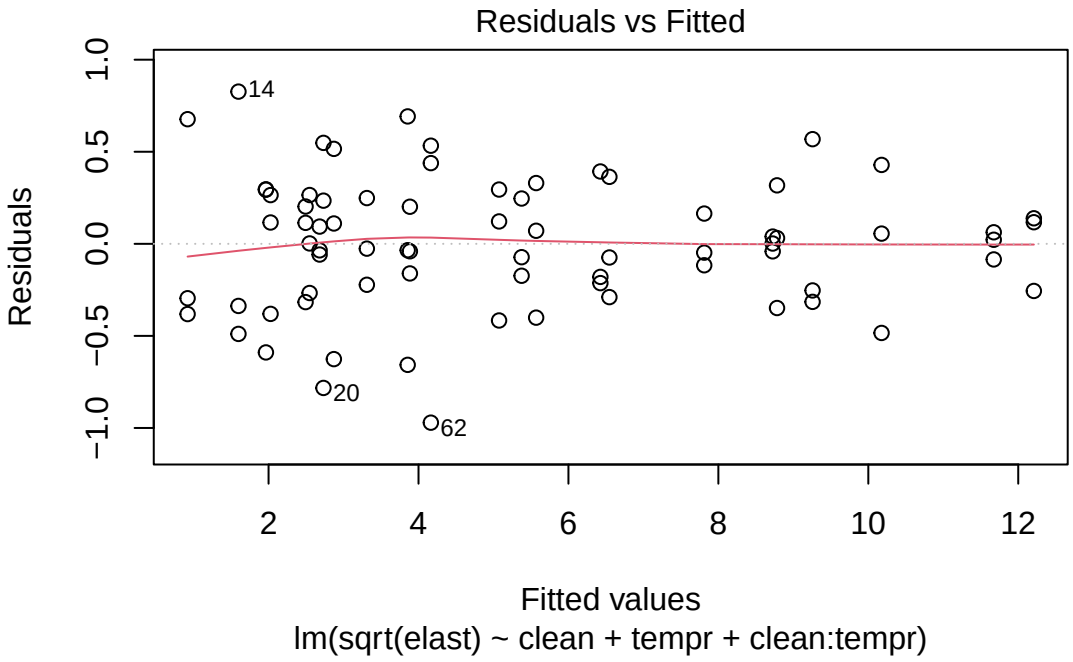
The ANOVA table shows us that the interaction is statistically significant; though before we trust this result, we should look at some diagnostic plots.

```
plot(lm_out,which = 2)
```



The Normal Q-Q plot suggests that the transformed response values are approximately Normally distributed around the treatment means (the residuals were heavy-tailed when the response values were not transformed).

```
plot(lm_out,which = 1)
```



The residuals versus fitted values plot shows approximately equal variances across the fitted values, although the variance may be somewhat higher at the lower fitted values.

The p value for the interaction is very small, so we should not try to tell stories about main effects.

Let's fix the cleaning agent concentration at the highest level and make pairwise comparisons of the means across the temperature factor.

```
a <- 4
b <- 5
n <- 3
y11. <- mean(sqrt(steel$elast[steel$clean == 2 & steel$tempr == 0.2]))
y12. <- mean(sqrt(steel$elast[steel$clean == 2 & steel$tempr == 0.93]))
y13. <- mean(sqrt(steel$elast[steel$clean == 2 & steel$tempr == 1.65]))
y14. <- mean(sqrt(steel$elast[steel$clean == 2 & steel$tempr == 2.38]))
y15. <- mean(sqrt(steel$elast[steel$clean == 2 & steel$tempr == 3.1]))
levsA <- levels(steel$tempr)

MSE <- anova(lm_out)$`Mean Sq`[4]
me <- qtukey(0.95,b,a*b*(n-1)) * sqrt(MSE) * sqrt(1/n)
to_compare <- c(y11.,y12.,y13.,y14.,y15.)
CIs <- matrix(NA,choose(b,2),2)
```

```
comp <- numeric(choose(b,2))
k <- 1
for(i in 1:(b-1))
  for(j in (i+1):b){

    dij <- to_compare[i] - to_compare[j]
    CIs[k,] <- c(dij - me, dij + me)
    comp[k] <- paste(levsA[i],"-",levsA[j])
    k <- k + 1
  }
colnames(CIs) <- c("lower","upper")
rownames(CIs) <- comp
round(CIs,3)
```

```
             lower  upper
0.2 - 0.93  -5.624 -3.611
0.2 - 1.65  -8.516 -6.502
0.2 - 2.38  -9.050 -7.036
0.2 - 3.1   -7.019 -5.005
0.93 - 1.65 -3.898 -1.884
0.93 - 2.38 -4.433 -2.419
0.93 - 3.1  -2.401 -0.387
1.65 - 2.38 -1.542  0.472
1.65 - 3.1   0.490  2.504
2.38 - 3.1   1.025  3.039
```

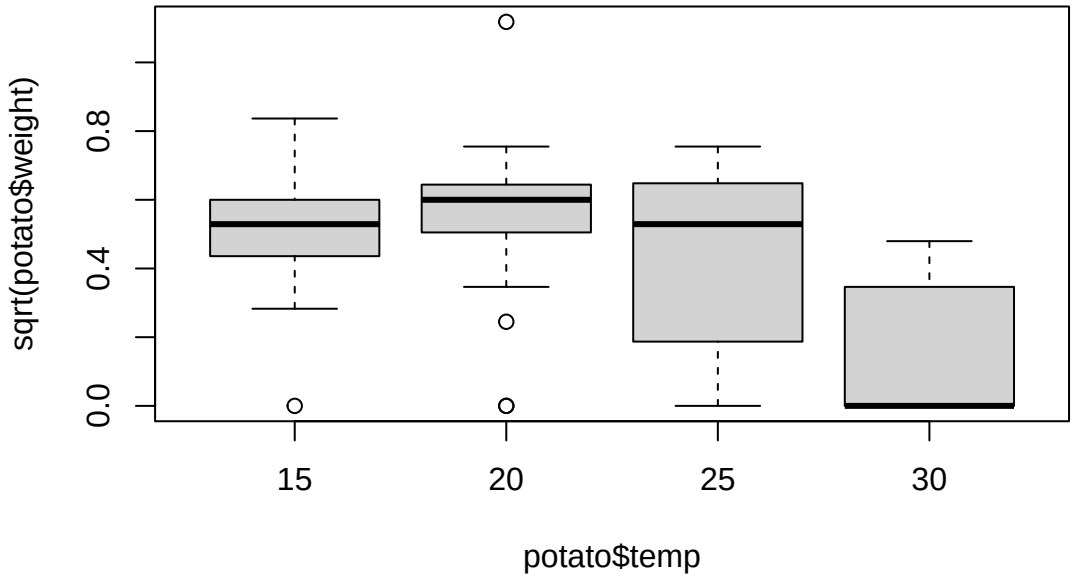
Temperature levels 1.65 and 2.38 are achieve statistically significantly higher elasticity on average than the other temperatures, but the mean elasticities at these two temperature levels may be equal.

Chp 9 Ex 10

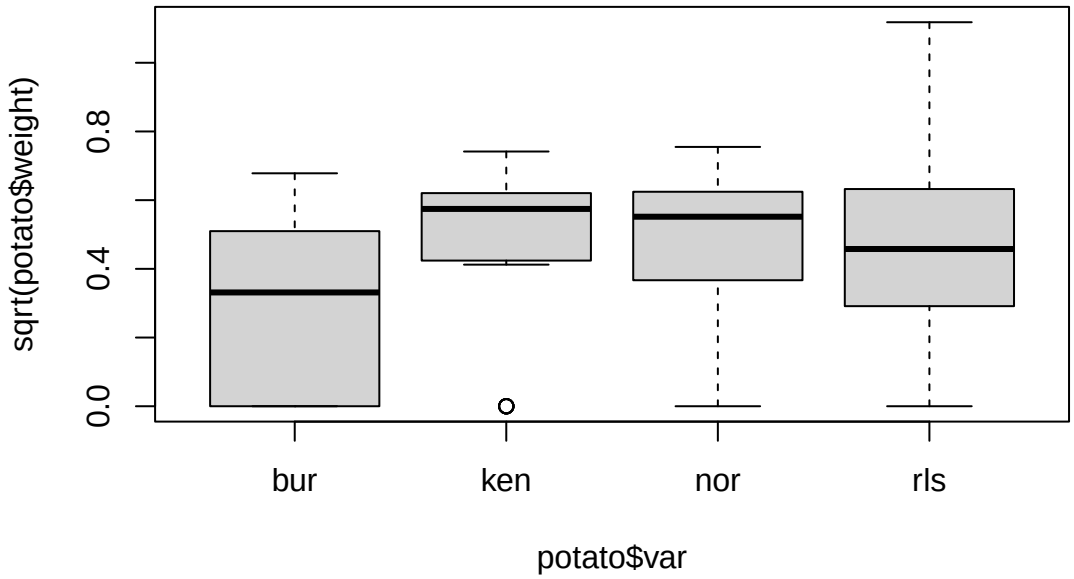
```
file <- "Data Tables 4th edition/Chapter 9/datatab_9_30.prn"
potato <- read.csv(file, sep = " ", colClasses = c("factor","factor","numeric"))
```

After some preliminary analysis, it seems beneficial to (like in the previous question) take the square root of the response values—otherwise there is some fanning in the residuals towards greater variance at greater fitted values.

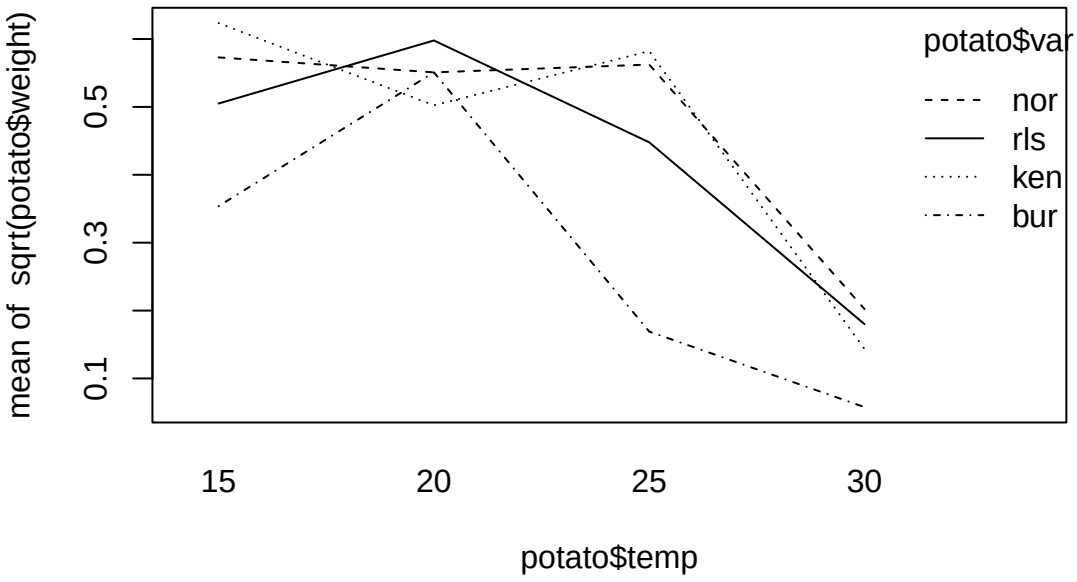
```
boxplot(sqrt(potato$weight) ~ potato$temp)
```




```
boxplot(sqrt(potato$weight) ~ potato$var)
```




```
interaction.plot(potato$temp,potato$var,sqrt(potato$weight))
```

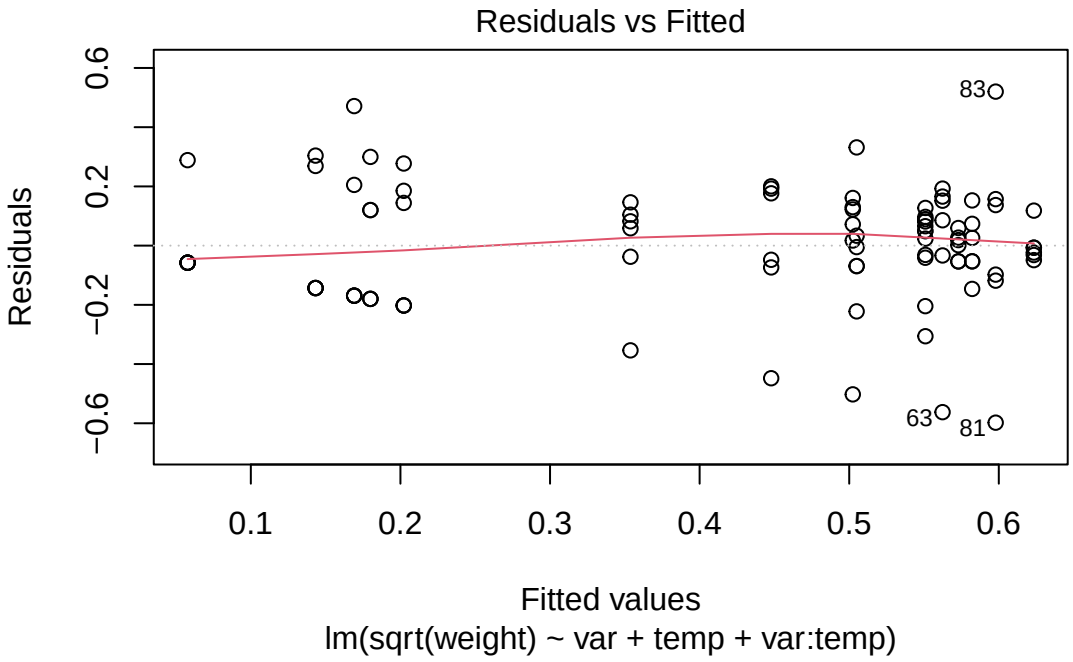



```
lm_out <- lm(sqrt(weight) ~ var + temp + var:temp, data = potato)
anova(lm_out)
```

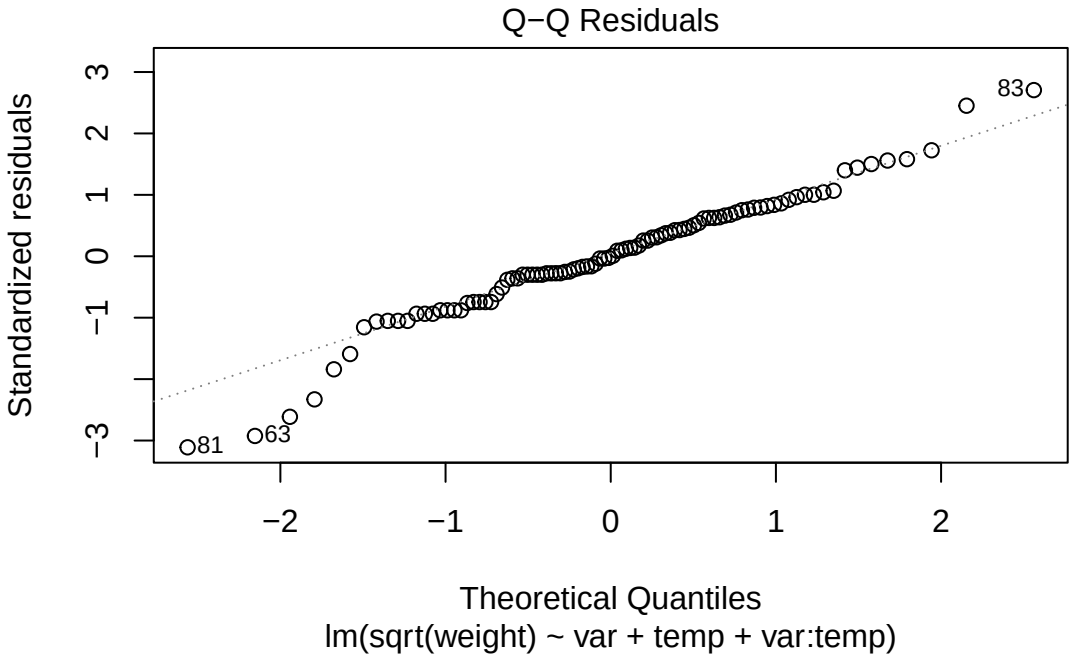
```
Analysis of Variance Table

Response: sqrt(weight)
          Df Sum Sq Mean Sq F value    Pr(>F)
var        3 0.5590 0.18633  4.2023   0.00818 **
temp       3 2.4294 0.80979 18.2632 4.045e-09 ***
var:temp   9 0.4402 0.04891  1.1030   0.37037
Residuals 80 3.5472 0.04434
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(lm_out,which = 1)
```




```
plot(lm_out,which = 2)
```



There is some suggestion that the responses may not be normally distributed around the

treatment means, but let's proceed with the analysis as though we were not concerned about this.

The p value for the interaction is quite large, so we can ignore interaction effects and go ahead and give interpretations to main effects; that is, we can make meaningful comparisons of marginal means.

To illustrate a comparison of marginal means, suppose we use Dunnett's procedure to compare all temperatures to the hottest temperature, treating the hottest temperature as the baseline.

The following code carries this out:

```
a <- 4
b <- 4
n <- 6
y.1. <- mean(potato$weight[potato$temp == 15])
y.2. <- mean(potato$weight[potato$temp == 20])
y.3. <- mean(potato$weight[potato$temp == 25])
y.4. <- mean(potato$weight[potato$temp == 30])

# error df is a*b*(n-1) = 4*4*5 = 80, b = 4 means including baseline
# the highest error df in dunnett's table is 60. Use this.
MSE <- anova(lm_out)$`Mean Sq`[4]
me <- 2.41 * sqrt(MSE)* sqrt(2/(n*a))
CIs <- rbind(c(y.1. - y.4. - me,y.1. - y.4. + me),
             c(y.2. - y.4. - me,y.2. - y.4. + me),
             c(y.3. - y.4. - me,y.3. - y.4. + me))
colnames(CIs) <- c("lower","upper")
rownames(CIs) <- c("15 - 30","20 - 30","25 - 30")
round(CIs,3)
```

```
        lower upper
15 - 30 0.085 0.378
20 - 30 0.149 0.442
25 - 30 0.065 0.358
```

According to Dunnett's method, the three temperatures 15, 20, and 25 achieve higher mean weights than the temperature 30, averaging over the four varieties.

Chp 9 Ex 12

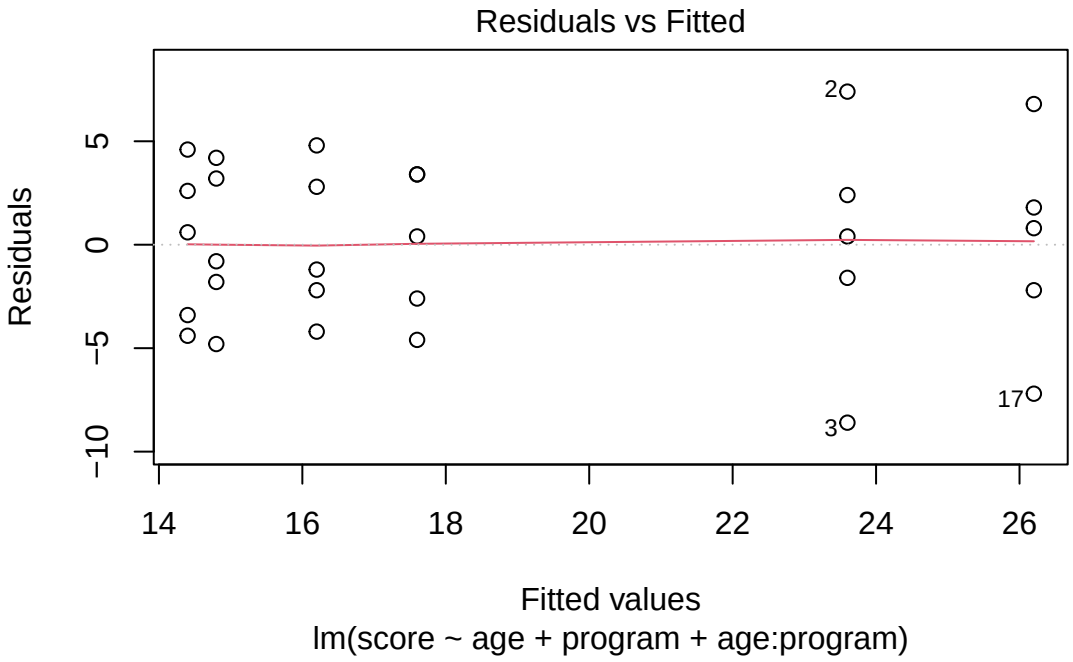
```r
file <- "Data Tables 4th edition/Chapter 9/datatab_9_32.prn"
risk <- read.csv(file, sep = " ", header = FALSE,
                 colClasses = c("factor","factor","numeric"))
colnames(risk) <- c("age","program","score")
```

```r
lm_out <- lm(score ~ age + program + age:program,data = risk)
anova(lm_out)
```

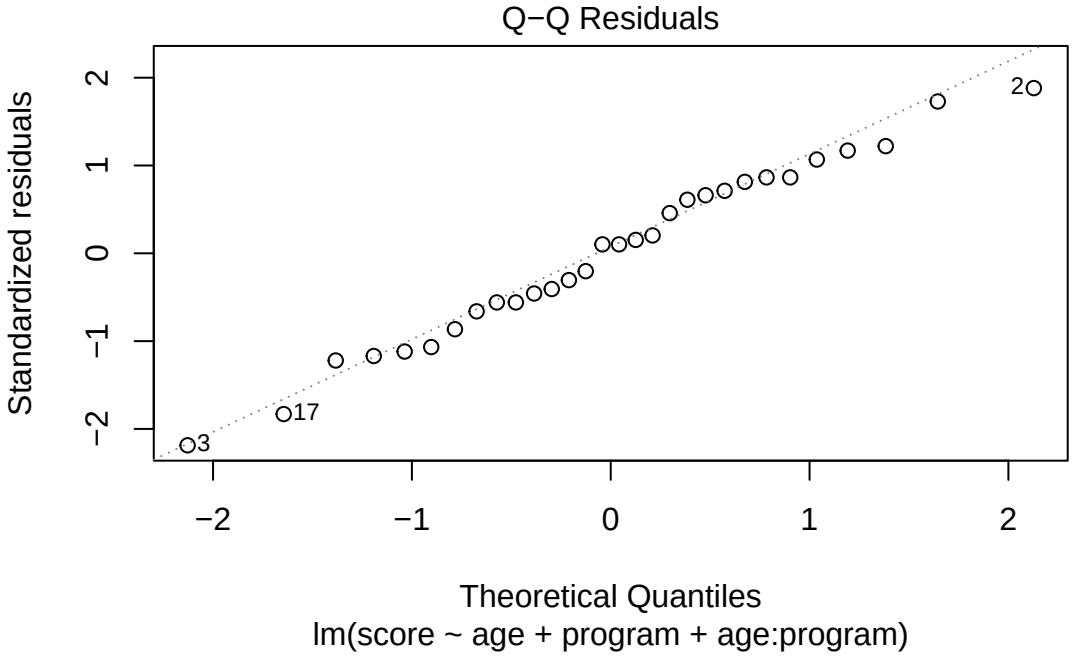
```
Analysis of Variance Table

Response: score
            Df Sum Sq Mean Sq F value    Pr(>F)
age          1   3.33   3.333  0.1724    0.6817
program      2 562.20 281.100 14.5397 7.302e-05 ***
age:program  2  41.27  20.633  1.0672    0.3597
Residuals   24 464.00  19.333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(lm_out,which = 1)
```



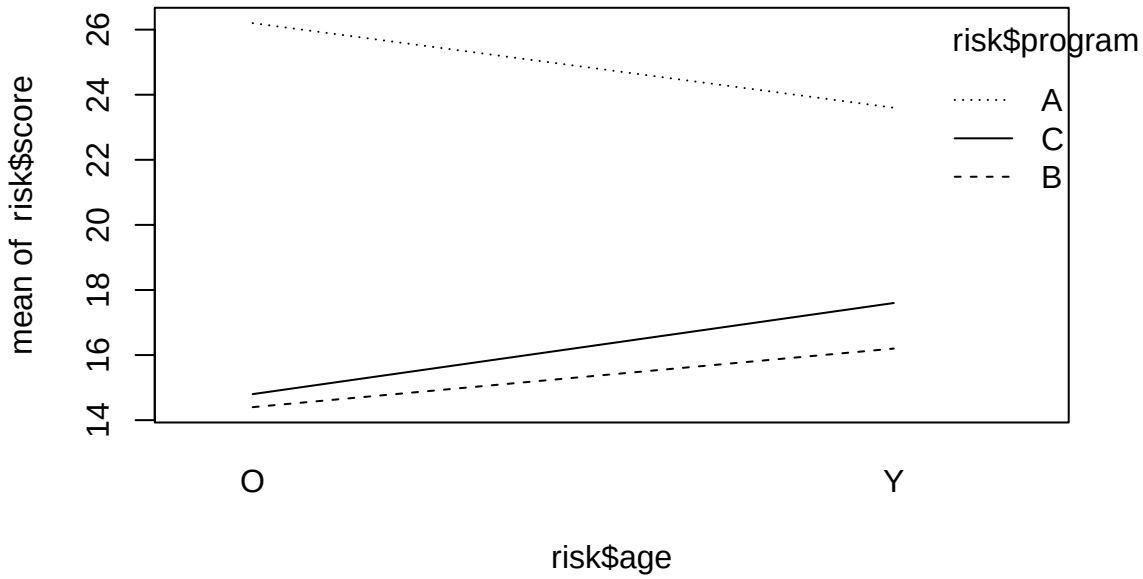

```
plot(lm_out,which = 2)
```



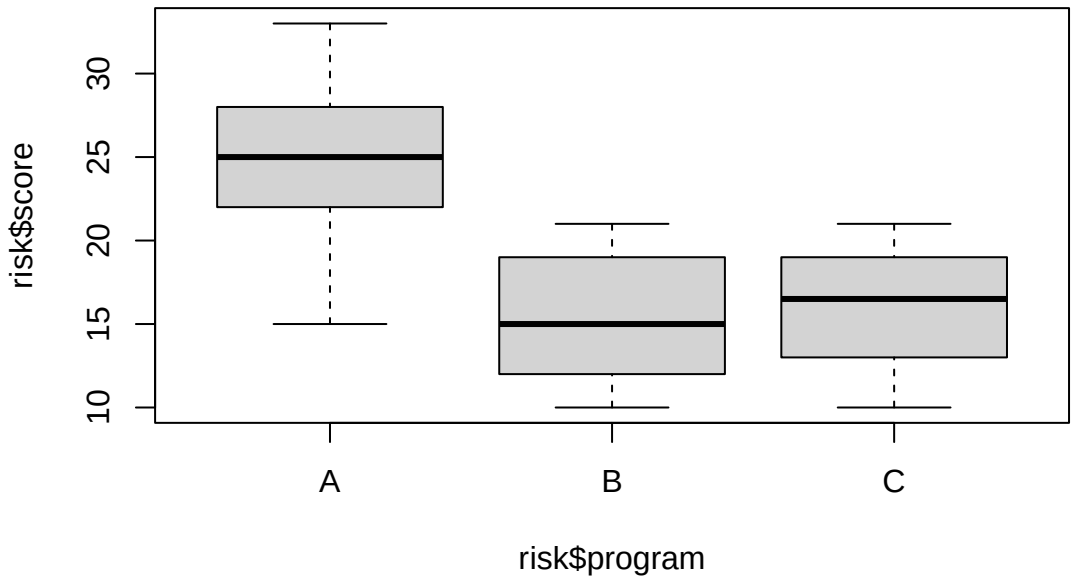
The diagnostic plots look good. So we will go ahead and interpret the results in the ANOVA

table: The interaction appears not to be significant—even though the interaction plot suggests there may be an interaction.

```
interaction.plot(risk$age,risk$program,risk$score)
```




```
boxplot(risk$score~risk$program)
```



We can go ahead and compare marginal means as in the previous question.

As the age factor appears to be insignificant, we will use Tukey's method to compare the marginal means of the programs.

```
a <- 2
b <- 3
n <- 5
y.1. <- mean(risk$score[risk$program == "A"])
y.2. <- mean(risk$score[risk$program == "B"])
y.3. <- mean(risk$score[risk$program == "C"])
MSE <- sum(lm_out$residuals^2) / ( a*b*(n-1))
me <- qtukey(0.95,b,a*b*(n-1)) * sqrt(MSE) * sqrt(1/(a*n))
CIs <- rbind(c(y.1. - y.2. - me,y.1. - y.2. + me),
             c(y.1. - y.3. - me,y.1. - y.3. + me),
             c(y.2. - y.3. - me,y.2. - y.3. + me))
colnames(CIs) <- c("lower","upper")
rownames(CIs) <- c("A - B","A - C","B - C")
round(CIs,3)
```

```
       lower  upper
A - B  4.689 14.511
A - C  3.789 13.611
B - C -5.811  4.011
```

We see from the pairwise comparisons that there seems to be no difference between programs B and C, but program A is better than both C and B. So program A emerges as a clear winner.