Stat 5309 Midterm Project

Tom Wilson

Mar 21, 2019

1

The effective life of insulating luids at an accelerated load of 35 kV is being studied. Test fluid_data have been obtained for four types of fluids. The results were as follows:

a

Either read fluid_data into R or create the fluid_dataframe.

```
fluidtypes <- c("1","2","3","4")
fluid_data <- data.frame(fluid = rep(fluidtypes,each=6)
                    ,lifetime=c(17.6,18.9,16.3,17.4,20.1,21.6,
                                16.9,15.3,18.6,17.1,19.5,20.3,
                                21.4,23.6,19.4,18.5,20.5,22.3,
                                19.3,21.1,16.9,17.5,18.3,19.8
                                ))
fluid_data %>% kable()
```

fluid	lifetime
1	17.6
1	18.9
1	16.3
1	17.4
1	20.1
1	21.6
2	16.9
2	15.3
2	18.6
2	17.1
2	19.5
2	20.3
3	21.4
3	23.6
3	19.4
3	18.5
3	20.5
3	22.3
4	19.3
4	21.1
4	16.9
4	17.5
4	18.3
4	19.8

b

Build a linear model, using aov. Is there a significant difference among treatment means? which fluid gives the longer life?

```
insulation_life_model <- aov(formula = lifetime ~ fluid,data=fluid_data)
summary(insulation_life_model)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fluid        3  30.17   10.05   3.047 0.0525 .
## Residuals   20  65.99    3.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Given that there is no difference in lifetime between fluid types, we would expect a result at least this extreme 5.25% of the time. At a confidence level of 95% we fail to reject the null hypothesis and conclude that any difference observed is due to chance.

c

Construct a 95% Confidence Interval for themean life of fluid 2.

```
anova_of_insulation_life <- anova(insulation_life_model)
MSError=anova_of_insulation_life$`Mean Sq`[2]
LSD.test(insulation_life_model, "fluid", MSerror = MSError,console = TRUE)
```

```
##
## Study: insulation_life_model ~ "fluid"
##
## LSD t Test for lifetime
##
## Mean Square Error:  3.299667
##
## fluid,  means and individual ( 95 %) CI
##
##   lifetime      std r      LCL      UCL  Min  Max
## 1 18.65000 1.952178 6 17.10309 20.19691 16.3 21.6
## 2 17.95000 1.854454 6 16.40309 19.49691 15.3 20.3
## 3 20.95000 1.879096 6 19.40309 22.49691 18.5 23.6
## 4 18.81667 1.554885 6 17.26975 20.36358 16.9 21.1
##
## Alpha: 0.05 ; DF Error: 20
## Critical Value of t: 2.085963
##
## least Significant Difference: 2.187666
##
## Treatments with the same letter are not significantly different.
##
##   lifetime groups
## 3 20.95000      a
## 4 18.81667     ab
## 1 18.65000      b
## 2 17.95000      b
```

16.40 to 19.50 is a 95% confidence interval for the mean of fluid type 2.

Construct a 99% Confidence Interval for the difference between the lives of Fluids 2 and 3.

```
TukeyHSD(insulation_life_model, conf.level=0.99)
```

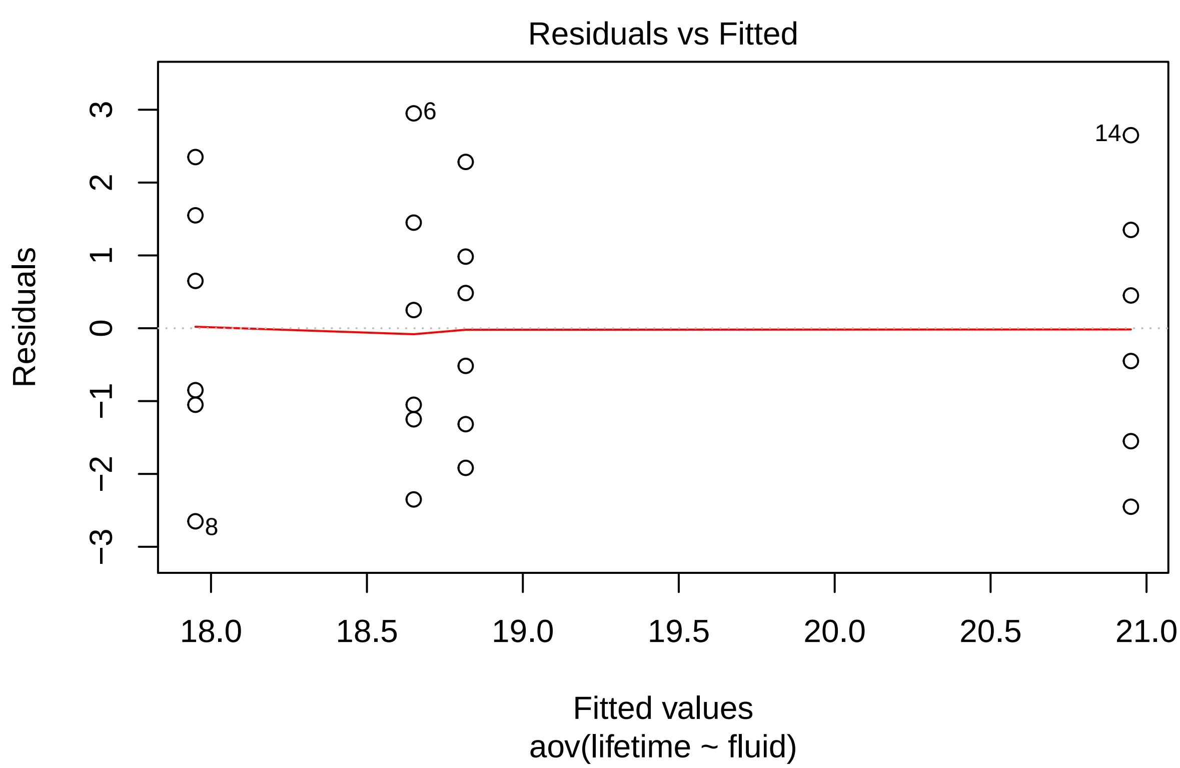
```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = lifetime ~ fluid, data = fluid_data)
##
## $fluid
##           diff        lwr      upr     p adj
## 2-1 -0.7000000 -4.4212724 3.021272 0.9080815
## 3-1  2.3000000 -1.4212724 6.021272 0.1593262
## 4-1  0.1666667 -3.5546057 3.887939 0.9985213
## 3-2  3.0000000 -0.7212724 6.721272 0.0440578
## 4-2  0.8666667 -2.8546057 4.587939 0.8413288
## 4-3 -2.1333333 -5.8546057 1.587939 0.2090635
```

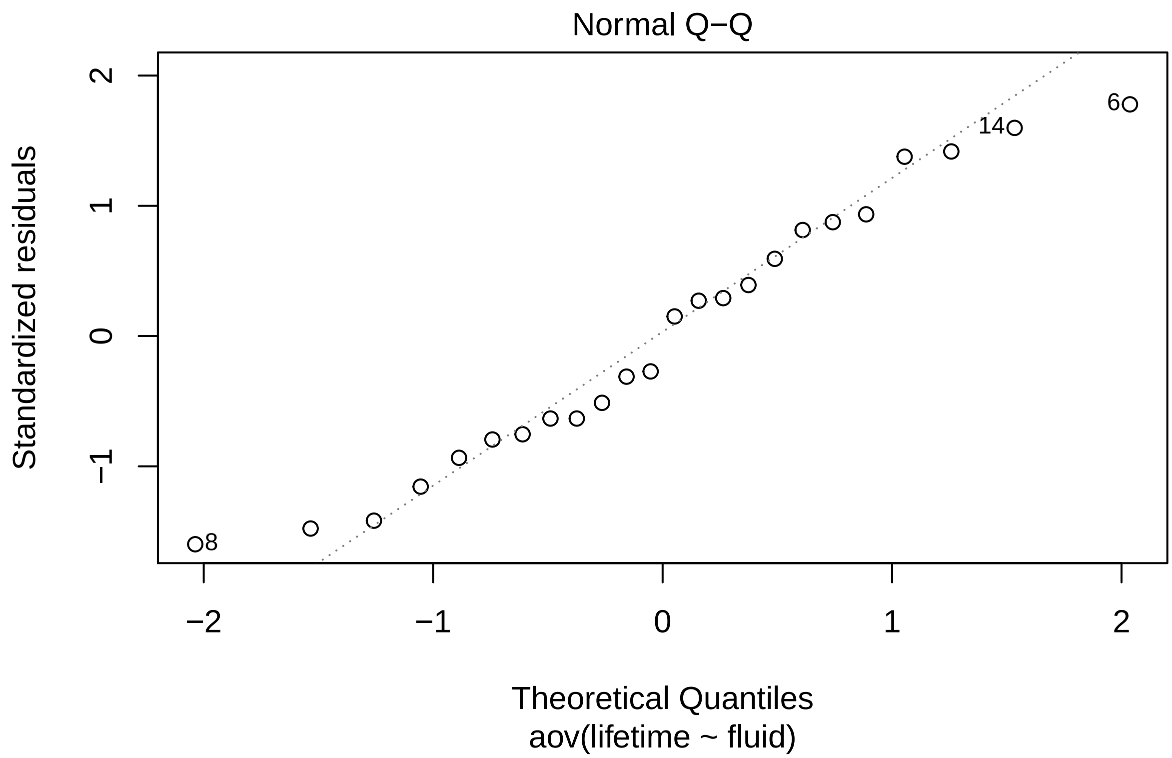
The difference between type 4 and type 1 is between -0.7213 and 6.7213 with 99% confidence.

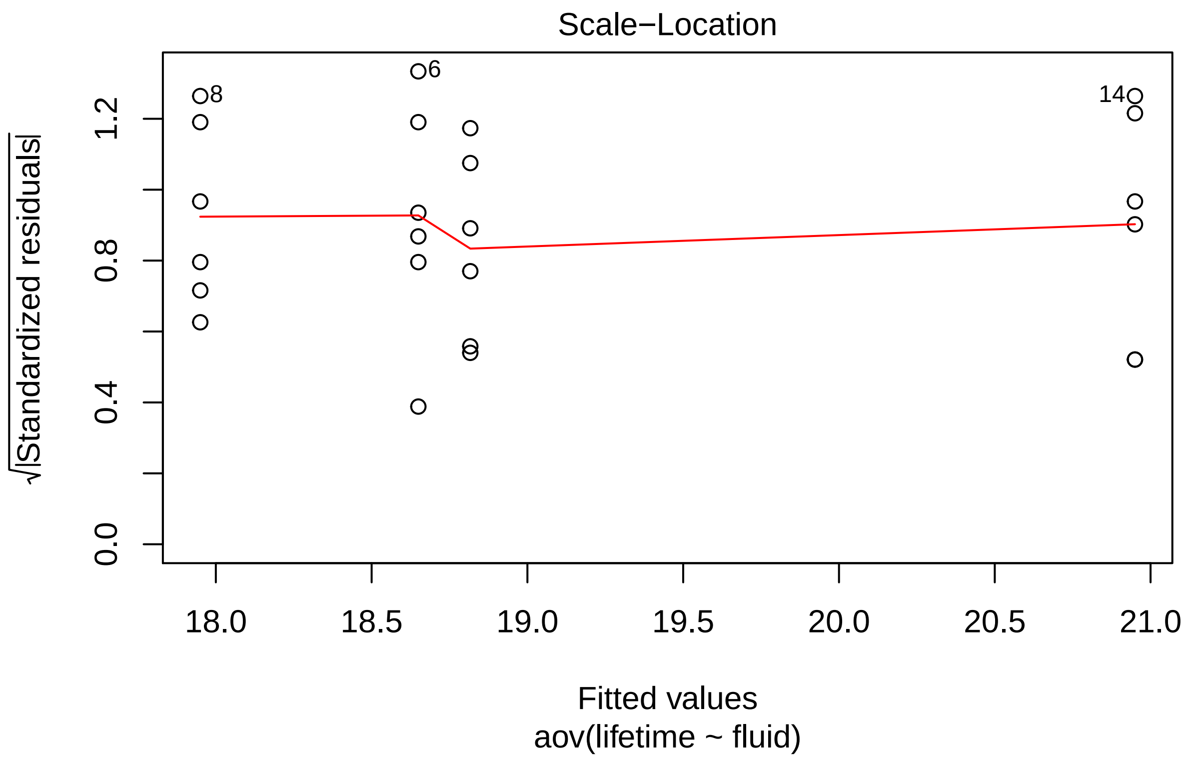
d

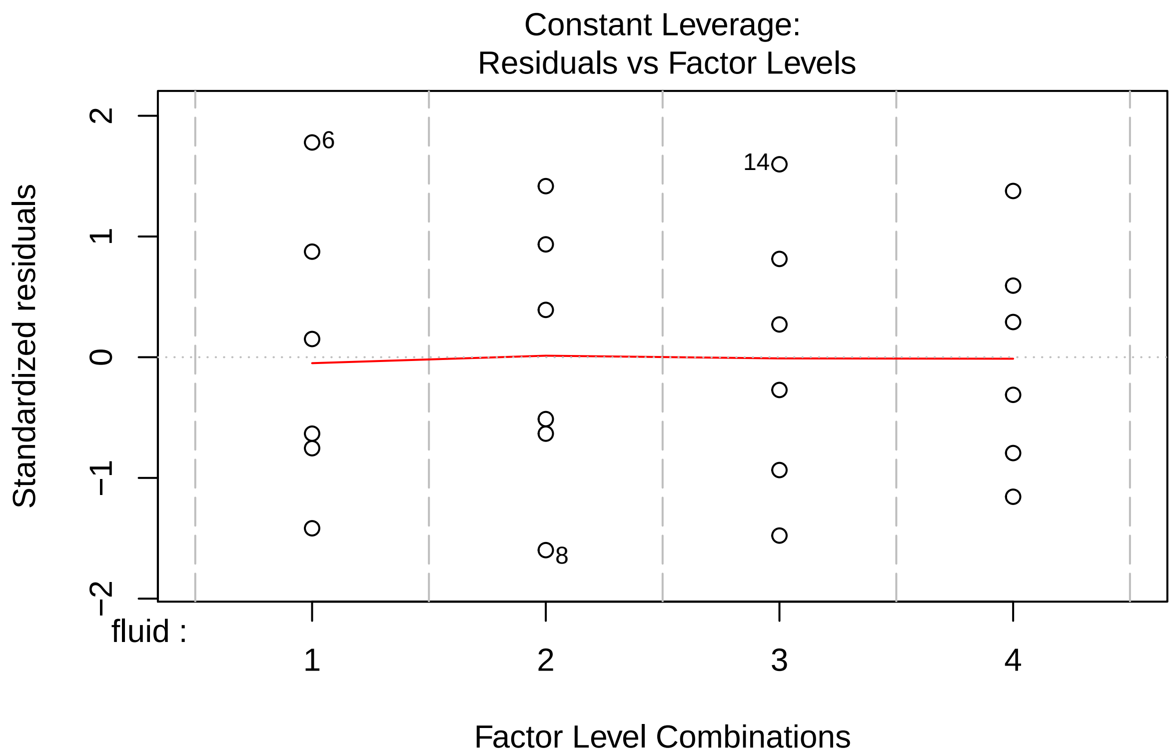
Perform a complete 3-part residuals check.

```
plot(insulation_life_model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(lifetime ~ fluid)
−2
−1
0
1
2
8
14
6


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
18.0
18.5
19.0
19.5
20.0
20.5
21.0
8
6
14
Fitted values
aov(lifetime ~ fluid)



based on visual analysis, the residuals are close to normally distributed with a mean of zero and a constant variance.

e

Calculate the number of replicates for a power of 0.99

```
fluid_means <- fluid_data %>% group_by(fluid) %>% summarise(trt_mean = mean(lifetime))
power.anova.test(groups = 4,
                 between.var = var(fluid_means$trt_mean),
                 within.var = anova_of_insulation_life$`Mean Sq`[2],
                 power=0.99
                 )
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 4
##               n = 16.45871
##     between.var = 1.675833
##      within.var = 3.299667
##       sig.level = 0.05
##           power = 0.99
##
## NOTE: n is number in each group
```

17 replicates are needed to achieve a power of 0.99 .

2

a

Either read data into R or create the dataframe.

```
oils <- c("1","2","3")
trucks <- c("1","2","3","4","5")
data <- expand.grid(truck=trucks,oil=oils)
data <- cbind(data,fuel_consumption = c(0.5,
                                        0.634,
                                        0.487,
                                        0.329,
                                        0.512,
                                        0.535,
                                        0.675,
                                        0.52,
                                        0.435,
                                        0.54,
                                        0.513,
                                        0.595,
                                        0.488,
                                        0.4,
                                        0.51))
data
```

```
##    truck oil fuel_consumption
## 1      1   1            0.500
## 2      2   1            0.634
## 3      3   1            0.487
## 4      4   1            0.329
## 5      5   1            0.512
## 6      1   2            0.535
## 7      2   2            0.675
## 8      3   2            0.520
## 9      4   2            0.435
## 10     5   2            0.540
## 11     1   3            0.513
## 12     2   3            0.595
## 13     3   3            0.488
## 14     4   3            0.400
## 15     5   3            0.510
```

b

Build a linear model. Is there any significant difference of means about the oil types? Which oil type gies the lowest fuel consumption?

c

Is the blocking approach effective?

d

Do a complete residual assumption check.

3

Suppose that in in Problem 4-15, the engineer suspects that the workplaces used by the four operators may represent an additional source of variation. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

a

Set up a dataframe with 2 blocking factors (order and operator) and treatment (A,B,C,D)

```
orders <- c("1st","2nd","3rd","4th")
operators <- c("op1","op2","op3","op4")
data <- expand.grid(operator=operators,order_of_assembly=orders)
data <- cbind(data,workplace=c("C","B","D","A",
                               "B","C","A","D",
                               "A","D","B","C",
                               "D","A","C","B"
                               ),
              observation=c(11,10,14,8,
                            8,12,10,12,
                            9,11,7,15,
                            9,8,18,6
                            )
              )
data %>% kable()
```

operator	order_of_assembly	workplace	observation
op1	1st	C	11
op2	1st	B	10
op3	1st	D	14
op4	1st	A	8
op1	2nd	B	8
op2	2nd	C	12
op3	2nd	A	10
op4	2nd	D	12
op1	3rd	A	9
op2	3rd	D	11
op3	3rd	B	7
op4	3rd	C	15
op1	4th	D	9
op2	4th	A	8
op3	4th	C	18
op4	4th	B	6

b

Use Latin Square to analyze the treatment means.

c

Which level combination brings the lowest time?

4

The factors that influence the breaking strength of a syntheti fiber are being studied. Four production machines and three operators are chosen and a factorial experiment is run using fiber from the same production batch. The results follow.

a

Either read data into R or create the dataframe.

```
machines <- c("1","2","3","4")
operators <- c("o1","o2","o3")
data <- expand.grid(machine=rep(machines,2),operator=operators)
data <- cbind(data,strength = c(109,110,108,110,
                                110,115,109,108,
                                110,110,111,114,
                                112,111,109,112,
                                116,112,114,120,
                                114,115,119,117
                                )
              )
data %>% kable()
```

machine	operator	strength
1	o1	109
2	o1	110
3	o1	108
4	o1	110
1	o1	110
2	o1	115
3	o1	109
4	o1	108
1	o2	110
2	o2	110
3	o2	111
4	o2	114
1	o2	112
2	o2	111
3	o2	109
4	o2	112
1	o3	116
2	o3	112
3	o3	114

machine	operator	strength
4	o3	120
1	o3	114
2	o3	115
3	o3	119
4	o3	117

b

Build a linear model. Any interaction between operator and machine?

c

Build a reduced model.

d

Do a complete 3-part residual assumption check.

5

An experiment is conducted to study the influence of operating temperature and three types of face-plate glass in the light output of an oscilloscope tube. The following data are collected.

a

Either read data into R or create the dataframe.

```
temperatures <- c(100,125,150)
glasses <- c("t1","t2","t3")
data <- expand.grid(temperature=rep(temperatures,3),
                    glass = glasses)
data <- cbind(data,output=c(580,1090,1392,
                            568,1087,1380,
                            570,1085,1386,

                            550,1070,1328,
                            530,1035,1312,
                            579,1000,1299,

                            546,1045,867,
                            575,1053,904,
                            599,1066,889))
data %>% kable()
```

temperature	glass	output
100	t1	580
125	t1	1090

temperature	glass	output
150	t1	1392
100	t1	568
125	t1	1087
150	t1	1380
100	t1	570
125	t1	1085
150	t1	1386
100	t2	550
125	t2	1070
150	t2	1328
100	t2	530
125	t2	1035
150	t2	1312
100	t2	579
125	t2	1000
150	t2	1299
100	t3	546
125	t3	1045
150	t3	867
100	t3	575
125	t3	1053
150	t3	904
100	t3	599
125	t3	1066
150	t3	889

b

Build a linear model. Any interaction between glass type and temperature?

c

Build a reduced model.

d

Do a complete 3-part residual assumption check.

6

Sludge is the dried product remaining from processed sewage. It can be used as fertilizer on crops. However, it contains heavy metals. They hypothesized the concentration of certain heavy metals in sludge differ among the metropolitan areas from which the sludge is obtained. The sludge was aded to the sand at 3 different rates: 0.5,1.0,1.5 metric tons/acre. The zinc levels were recorded.

a

Set up a dataframe named metals. Use factos city (A,B,C), rate (0.5,1.0,1.5), and zinc for the observations.

```
cities <- c("A","B","C")
rates <- c(0.5,1.0,1.5)
data <- expand.grid(rate=rates,city=cities)
data <- cbind(data,zinc=c(26.4,25.2,26.0, 30.1,47.7,73.8, 19.4,23.2,18.9,
                          23.5,39.2,44.6, 31.0,39.1,71.1, 19.3,21.3,19.8,
                          25.4,25.5,35.5, 30.8,55.3,68.4, 18.7,23.2,19.6,
                          22.9,31.9,38.6, 32.8,50.7,77.1, 19.0,19.9,21.9
                          )
              )
data %>% kable()
```

rate	city	zinc
0.5	A	26.4
1.0	A	25.2
1.5	A	26.0
0.5	B	30.1
1.0	B	47.7
1.5	B	73.8
0.5	C	19.4
1.0	C	23.2
1.5	C	18.9
0.5	A	23.5
1.0	A	39.2
1.5	A	44.6
0.5	B	31.0
1.0	B	39.1
1.5	B	71.1
0.5	C	19.3
1.0	C	21.3
1.5	C	19.8
0.5	A	25.4
1.0	A	25.5
1.5	A	35.5
0.5	B	30.8
1.0	B	55.3
1.5	B	68.4
0.5	C	18.7
1.0	C	23.2
1.5	C	19.6
0.5	A	22.9
1.0	A	31.9
1.5	A	38.6
0.5	B	32.8
1.0	B	50.7
1.5	B	77.1
0.5	C	19.0
1.0	C	19.9
1.5	C	21.9

b

Build an aov model, using zinc as the response. Which factors are significant? Interaction is significant? Perform an interaction plot.

c

List all the factor means and effects. using tapply() or model.table().

d

calculate the interaction sum squares from scratch.

1.

The yield of a chemical process is being studied. The two most important variables are thought ot be th pressure and the temperature. Three levels of each factor are selected and a factorial experiment with two replicates is performed. The yield data follows.

a

Set up the dataframe.

```
temperatures <- c("150","160","170")
pressures <- c("200","215","230")
process <- expand.grid(pressure = rep(pressures,2),
                       temperature = temperatures)

process <- cbind(process,yield=c(90.4,90.7,90.2,
                                  90.2,90.6,90.4,
                                  90.1,90.5,89.9,
                                  90.3,90.6,90.1,
                                  90.5,90.8,90.4,
                                  90.7,90.9,90.1))
process %>% kable()
```

pressure	temperature	yield
200	150	90.4
215	150	90.7
230	150	90.2
200	150	90.2
215	150	90.6
230	150	90.4
200	160	90.1
215	160	90.5
230	160	89.9
200	160	90.3
215	160	90.6
230	160	90.1
200	170	90.5

pressure	temperature	yield
215	170	90.8
230	170	90.4
200	170	90.7
215	170	90.9
230	170	90.1

b

Build a linear model using aov(). Are the pressure means significant? Are the temp means significant? Is the interaction significant?

```
yield_model <- aov(yield ~ temperature * pressure, data=process)
summary(yield_model)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## temperature           2 0.3011  0.1506   8.469 0.008539 **
## pressure              2 0.7678  0.3839  21.594 0.000367 ***
## temperature:pressure  4 0.0689  0.0172   0.969 0.470006
## Residuals             9 0.1600  0.0178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction between temperature and pressure is not significant. The main effects of temperature and pressure are significant.

```
yield_model <- aov(yield ~ temperature + pressure, data=process)
summary(yield_model)
```

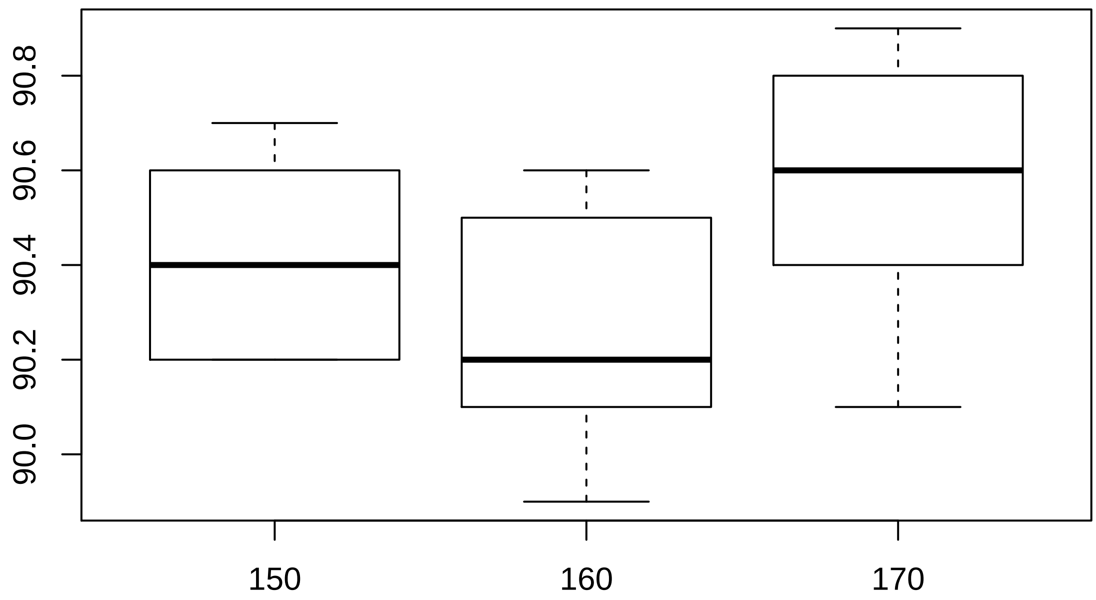
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## temperature  2 0.3011  0.1506   8.551  0.00426 **
## pressure     2 0.7678  0.3839  21.803 7.03e-05 ***
## Residuals   13 0.2289  0.0176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c

Create a boxplot of

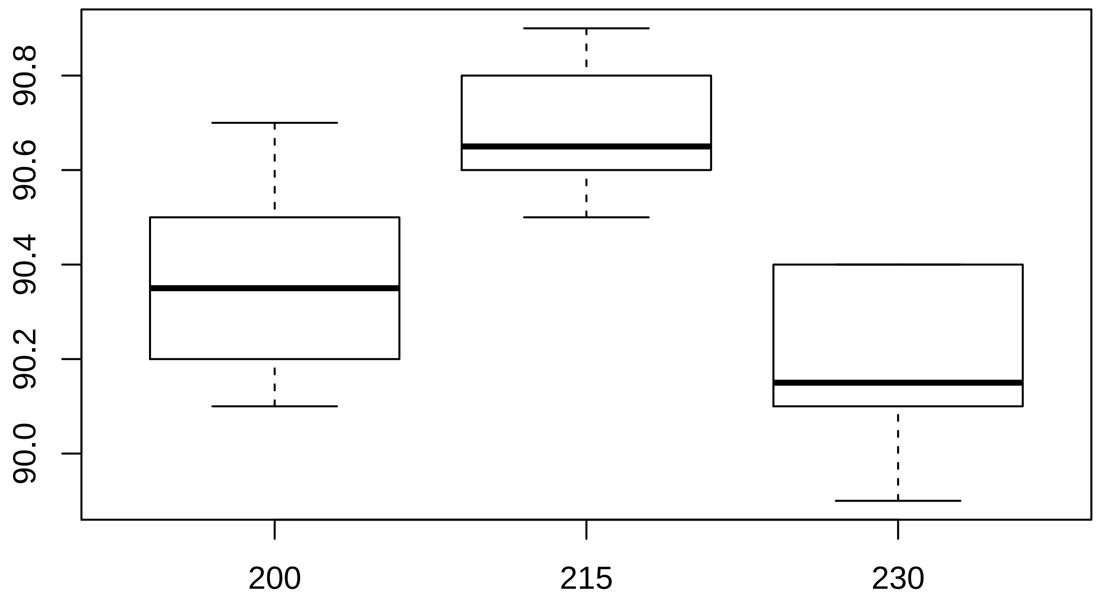
yield vs temp

```
boxplot(yield~temperature, data=process)
```



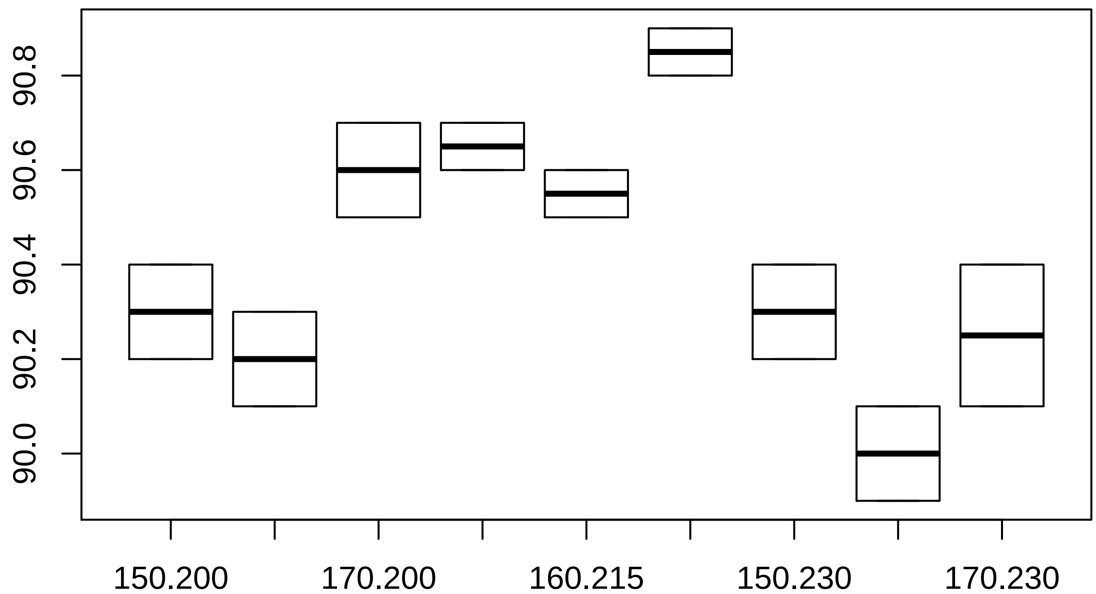
yield vs pressure

```
boxplot(yield~pressure, data=process)
```



yield vs temp and pressure

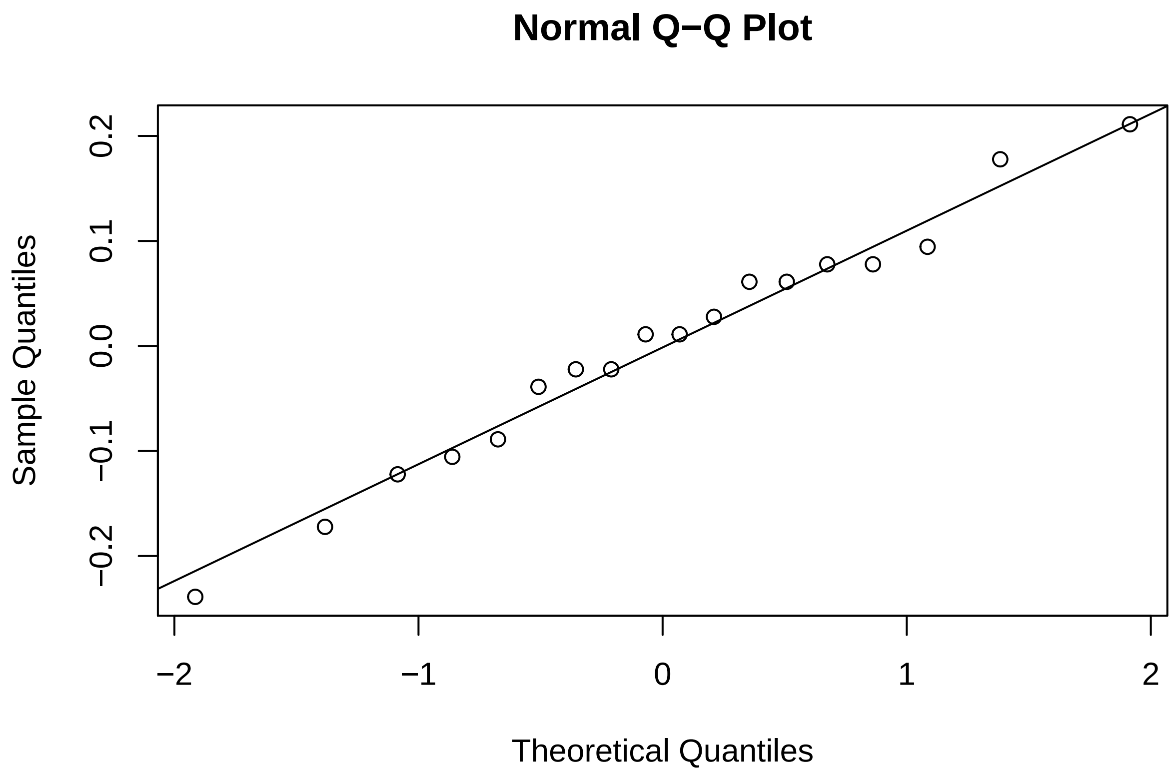
```
boxplot(yield~temperature*pressure, data=process)
```



d

Perform a residuals assumption check

```
qqnorm(yield_model$residuals)
qqline(yield_model$residuals)
```



The residuals are consistent with a normal distribution centered on zero with constant variance.

2

Johnson an dLeone describe an experiment to investigate warping of copper plates. The two factors studied were the temperature and the copper content of the plates. The response variable was a meaure of the amont of warping. The data were as follows.

a

Set up the dataframe

```
temperatures <- c("50","75","100","125")
copper_contents <- c("40","60","80","100")
copperplate <- expand.grid(copper_content=rep(copper_contents,2),
                           temperature=temperatures
                           )
copperplate <- cbind(copperplate,warping = c(17,16,24,28,
                                              20,21,22,27,
                                              12,18,17,12,
                                              9,13,12,31,
                                              16,18,25,30,
                                              12,21,23,23,
                                              21,23,23,29,
```

```
                                                    17,21,22,31))
copperplate %>% kable()
```

copper_content	temperature	warping
40	50	17
60	50	16
80	50	24
100	50	28
40	50	20
60	50	21
80	50	22
100	50	27
40	75	12
60	75	18
80	75	17
100	75	12
40	75	9
60	75	13
80	75	12
100	75	31
40	100	16
60	100	18
80	100	25
100	100	30
40	100	12
60	100	21
80	100	23
100	100	23
40	125	21
60	125	23
80	125	23
100	125	29
40	125	17
60	125	21
80	125	22
100	125	31

b

build a response model surface (RSM) with warpage as response, use rsm().

```
temp_nums <- copperplate$temperature %>% as.numeric()
cu_nums <- copperplate$copper_content %>% as.numeric()
response <- copperplate$warping
copper_model<- rsm(response ~ SO(temp_nums, cu_nums), data=copperplate)
summary(copper_model)
```

```
##
## Call:
## rsm(formula = response ~ SO(temp_nums, cu_nums), data = copperplate)
##
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)        23.68750    6.63925  3.5678 0.001427 **
## temp_nums         -10.33750    3.90821 -2.6451 0.013670 *
## cu_nums             0.57500    3.90821  0.1471 0.884167
## temp_nums:cu_nums   0.16000    0.57274  0.2794 0.782181
## temp_nums^2         2.18750    0.71593  3.0555 0.005142 **
## cu_nums^2           0.50000    0.71593  0.6984 0.491131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.6165, Adjusted R-squared:  0.5427
## F-statistic: 8.358 on 5 and 26 DF,  p-value: 8.125e-05
##
## Analysis of Variance Table
##
## Response: response
##                         Df Sum Sq Mean Sq F value    Pr(>F)
## FO(temp_nums, cu_nums)   2 523.03 261.513 15.9442 3.027e-05
## TWI(temp_nums, cu_nums)  1   1.28   1.280  0.0780    0.7822
## PQ(temp_nums, cu_nums)   2 161.12  80.562  4.9118    0.0155
## Residuals               26 426.45  16.402
## Lack of fit             10 145.44  14.544  0.8282    0.6093
## Pure error              16 281.00  17.563
##
## Stationary point of response surface:
##  temp_nums    cu_nums
##  2.3979170 -0.9586667
##
## Eigenanalysis:
## eigen() decomposition
## $values
## [1] 2.1912841 0.4962159
##
## $vectors
##                  [,1]        [,2]
## temp_nums -0.99888317  0.04724851
## cu_nums   -0.04724851 -0.99888317
```