Stat 5309 Midterm Project

Tom Wilson

Mar 21, 2019

1

The effective life of insulating fluids 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 data into R or create the 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.

```
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
```

Is there a significant difference among treatment means?

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.

which fluid gives the longer life?

```
tapply(fluid_data$lifetime, fluid_data$fluid, mean)
```

```
##        1        2        3        4
## 18.65000 17.95000 20.95000 18.81667
```

Although it is not significant, fluid 3 is associated with the longest life.

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
```

16.4 to 19.5 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)
```

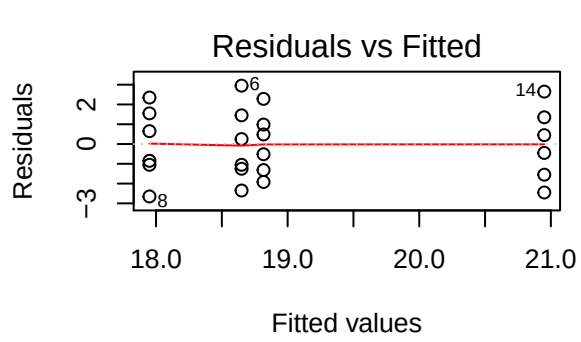
```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = lifetime ~ fluid, data = fluid_data)
##
## $fluid
##           diff        lwr      upr     p adj
## 2-1 -0.7000000 -4.4212724 3.021272 0.9080815
## 3-1  2.3000000 -1.4212724 6.021272 0.1593262
## 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
```

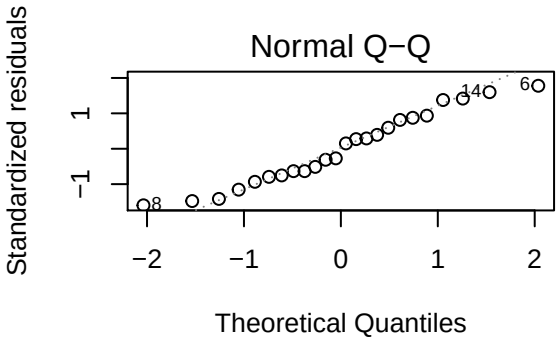
The difference between type 4 and type 1 is between -0.7213 and 6.7213 with 99% confidence.

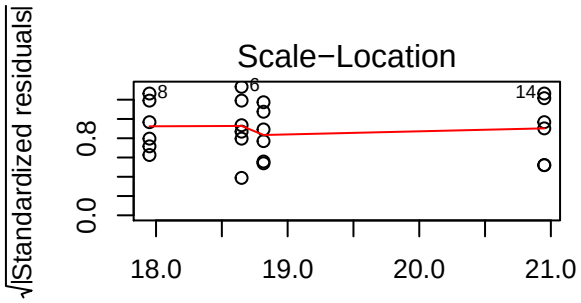
d

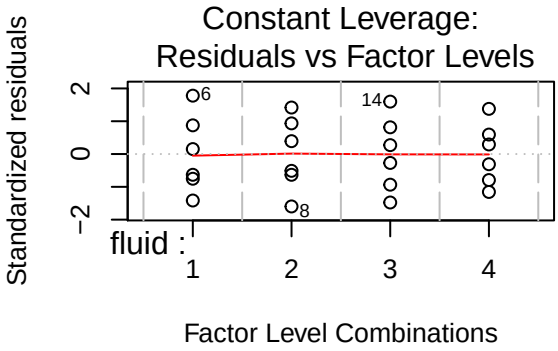
Perform a complete 3-part residuals check.

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









Independence

Based on a plot of standardized residual vs factor level (in this case, fluid type) the residuals are independent of fluid type.

Normality

Based on a quantile-quantile plot te residuals are very close to normally distrbuted with a mean of zero.

Homoscedasticity

based on a plot of residual vs fitted values, variation in residual is constant.

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")
fuel_data <- expand.grid(truck=trucks,oil=oils)
fuel_data <- cbind(fuel_data,fuel_consumption = c(0.500,0.634,0.487,
                                                  0.329,0.512,0.535,
                                                  0.675,0.520,0.435,
                                                  0.540,0.513,0.595,
                                                  0.488,0.400,0.510
                                                  )
                   )
fuel_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.

```
fuel_consumption_model <- aov(formula = fuel_data$fuel_consumption ~ oil+truck,data=fuel_data)
summary(fuel_consumption_model)
```

```
##             Df  Sum Sq  Mean Sq F value   Pr(>F)
## oil          2 0.00671 0.003353   6.353   0.0223 *
## truck        4 0.09210 0.023025  43.626 1.78e-05 ***
## Residuals    8 0.00422 0.000528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Is there any significant difference of means about the oil types?

Both truck and oil have significantly different means.

Which oil type gies the lowest fuel consumption?

```
tapply(fuel_data$fuel_consumption,fuel_data$oil,mean)
```

```
##      1      2      3
## 0.4924 0.5410 0.5012
```

Oil type 1 is associated with the lowest fuel consumption.

c

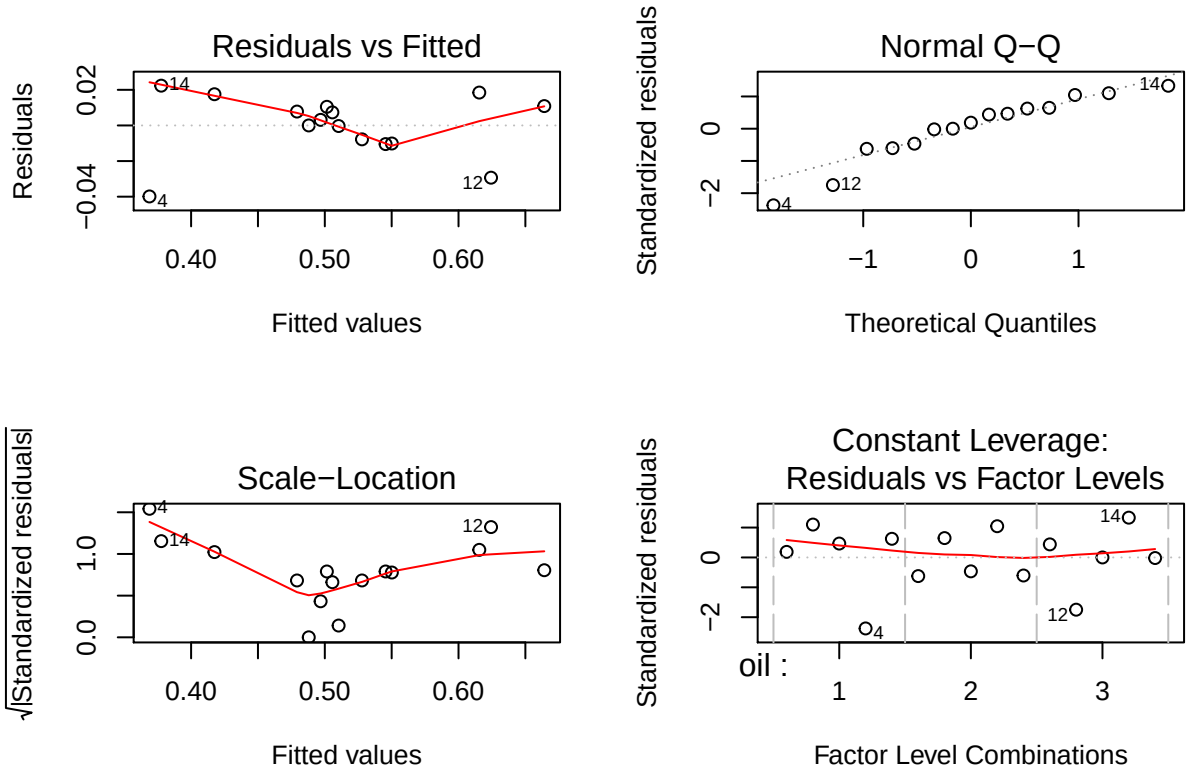
Is the blocking approach effective?

The blocking variable of truck type has a significant effect; therefore, blocking is effective. Without blocking, the effect of oil on fuel consumption could not be decoupled from trucktype.

d

Do a complete residual assumption check.

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



Independence

Based on a plot of standardized residual vs factor level (in this case, oil type) the residuals are nearly independent of oil type.

Normality

Based on a quantile-quantile plot te residuals are very close to normally distrbuted with a mean of zero.

Homoscedasticity

based on a plot of residual vs fitted values, variation in residual is not constant. There may be a more complicated relationship between oil, truck, and fuel consumption.

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_3 <- c("1st","2nd","3rd","4th")
operators_3 <- c("op1","op2","op3","op4")
workplaces <- c("A","B","C","D")
workplace_data <- expand.grid(operator=operators_3,
                              order_of_assembly=orders_3
                              )
workplace_data <- cbind(workplace_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
                                      )
                        )
workplace_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.

```
workplace_model <- aov(formula = observation ~ order_of_assembly+operator+workplace ,data=workplace_data)
summary(workplace_model)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## order_of_assembly  3    0.5    0.17   0.029 0.9928
## operator           3   19.0    6.33   1.086 0.4240
## workplace          3   95.5   31.83   5.457 0.0377 *
```

```
## Residuals          6   35.0    5.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At a confidence level of 95%, there are differences between workplaces.

```
TukeyHSD( workplace_model, "workplace")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = observation ~ order_of_assembly + operator + workplace, data = workplace_data)
##
## $workplace
##      diff        lwr       upr     p adj
## B-A -1.00 -6.9119977  4.911998 0.9328804
## C-A  5.25 -0.6619977 11.161998 0.0787664
## D-A  2.75 -3.1619977  8.661998 0.4393220
## C-B  6.25  0.3380023 12.161998 0.0398761
## D-B  3.75 -2.1619977  9.661998 0.2264008
## D-C -2.50 -8.4119977  3.411998 0.5098871
```

Pairwise, C is different from both B and A. Other pairs of workplaces are not significantly different from each other.

c

Which level combination brings the lowest time?

```
workplace_model_tables <- model.tables( workplace_model, type = "means" )
workplace_model_tables$tables$workplace
```

```
## workplace
##     A     B     C     D
##  8.75  7.75 14.00 11.50
```

Workplace B is associated with the lowest time.

4

The factors that influence the breaking strength of a synthetic 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_4 <- c("o1","o2","o3")
fiber_data <- expand.grid(machine=rep(machines,2),operator=operators_4)
fiber_data <- cbind(fiber_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
                                        )
                     )
fiber_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
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?

```
fiber_model <- aov(formula = strength ~ operator*machine ,data=fiber_data)
summary(fiber_model)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## operator          2 160.33   80.17  21.143 0.000117 ***
## machine           3  12.46    4.15   1.095 0.388753
## operator:machine  6  44.67    7.44   1.963 0.150681
## Residuals        12  45.50    3.79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is no significant interaction between operator and machine. Only operator is significant.

c

Build a reduced model.

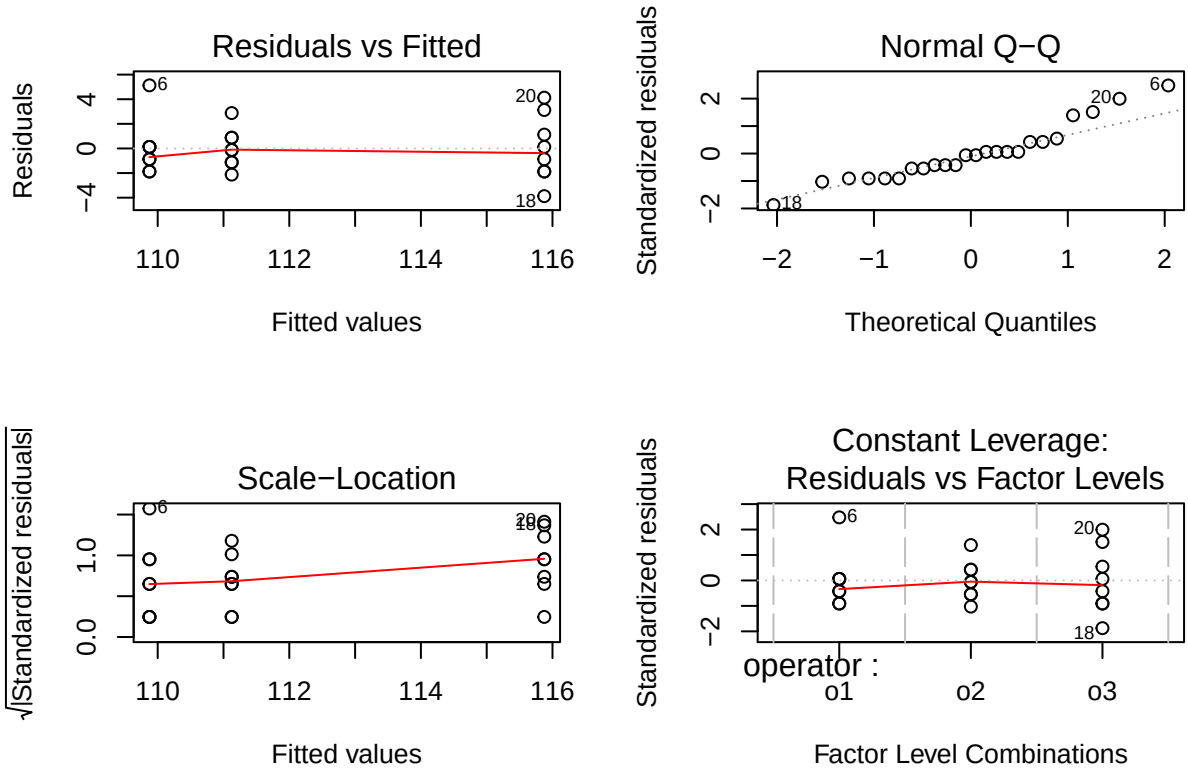
```
fiber_model_reduced <- aov(formula = strength ~ operator ,data=fiber_data)
summary(fiber_model_reduced)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## operator     2  160.3   80.17    16.4 5.12e-05 ***
## Residuals   21  102.6    4.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d

Do a complete 3-part residual assumption check.

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



Independence

Based on a plot of standardized residual vs factor level (in this case, fluid type) the residuals are independent of fluid type.

Normality

Based on a quantile-quantile plot te residuals are very close to normally distrbuted with a mean of zero.

Homoscedasticity

based on a plot of residual vs fitted values, variation in residual is constant.

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")
glass_data <- expand.grid(temperature=rep(temperatures,3),
                          glass = glasses)
glass_data <- cbind(glass_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
                                        )
                    )
glass_data %>% kable()
```

temperature	glass	output
100	t1	580
125	t1	1090
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

temperature	glass	output
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?

```
light_output_model <- aov(formula = output~temperature*glass,data=glass_data)
summary(light_output_model)
```

```
##                   Df  Sum Sq Mean Sq F value   Pr(>F)
## temperature        1 1779756 1779756 142.907 7.81e-11 ***
## glass              2  150865   75432   6.057  0.00838 **
## temperature:glass  2  226178  113089   9.081  0.00144 **
## Residuals         21  261532   12454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant interaction between temperature and glass type.

c

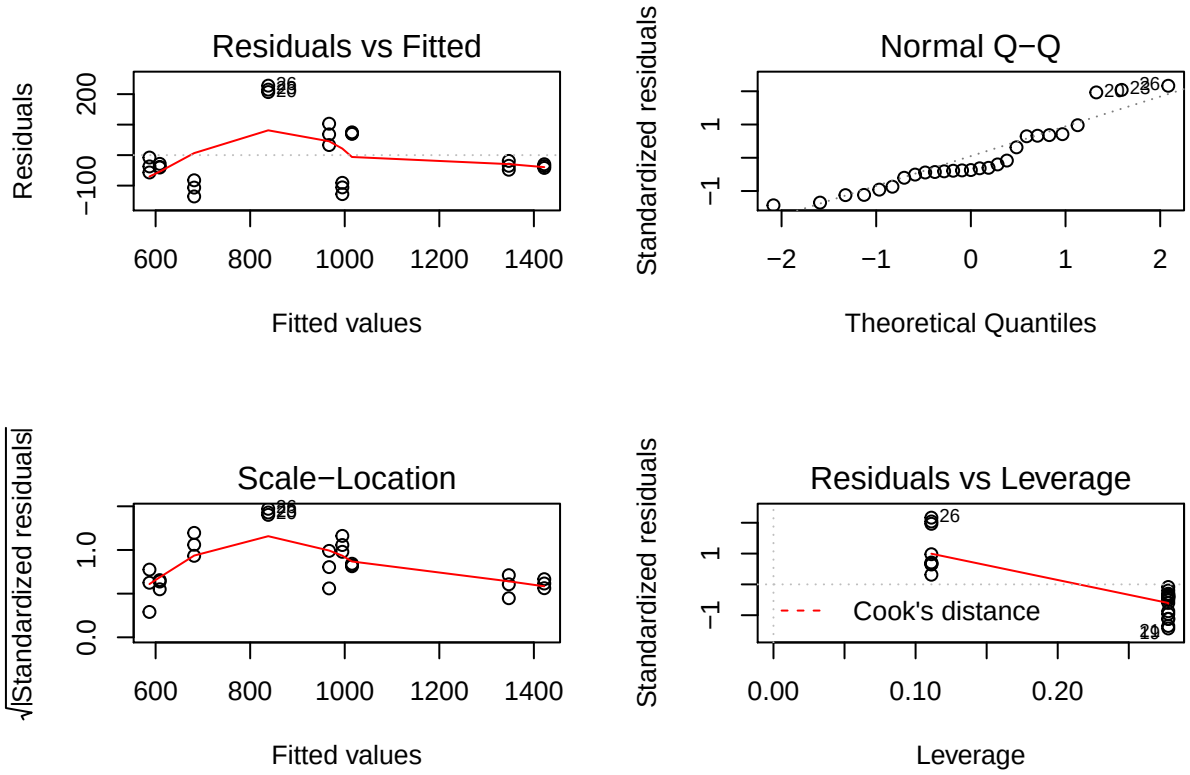
Build a reduced model.

Since the interaction is significant, the full interactive model is appropriate.

d

Do a complete 3-part residual assumption check.

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



Independence

Based on a plot of standardized residual vs factor level (in this case, glass type and temperature) the residuals not independent of factor level.

Normality

Based on a quantile-quantile plot te residuals are very close to normally distrbuted with a mean of zero.

Homoscedasticity

based on a plot of residual vs fitted values, variation in residual mostly constant except for a cluster of high residuals at $\hat{y} = 800$.

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")
zinc_data <- expand.grid(rate=rates,city=cities)
zinc_data <- cbind(zinc_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
                                    )
                   )
zinc_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.

```
zinc_model <- aov(formula = zinc ~ rate*city ,data=zinc_data)
summary(zinc_model)
```

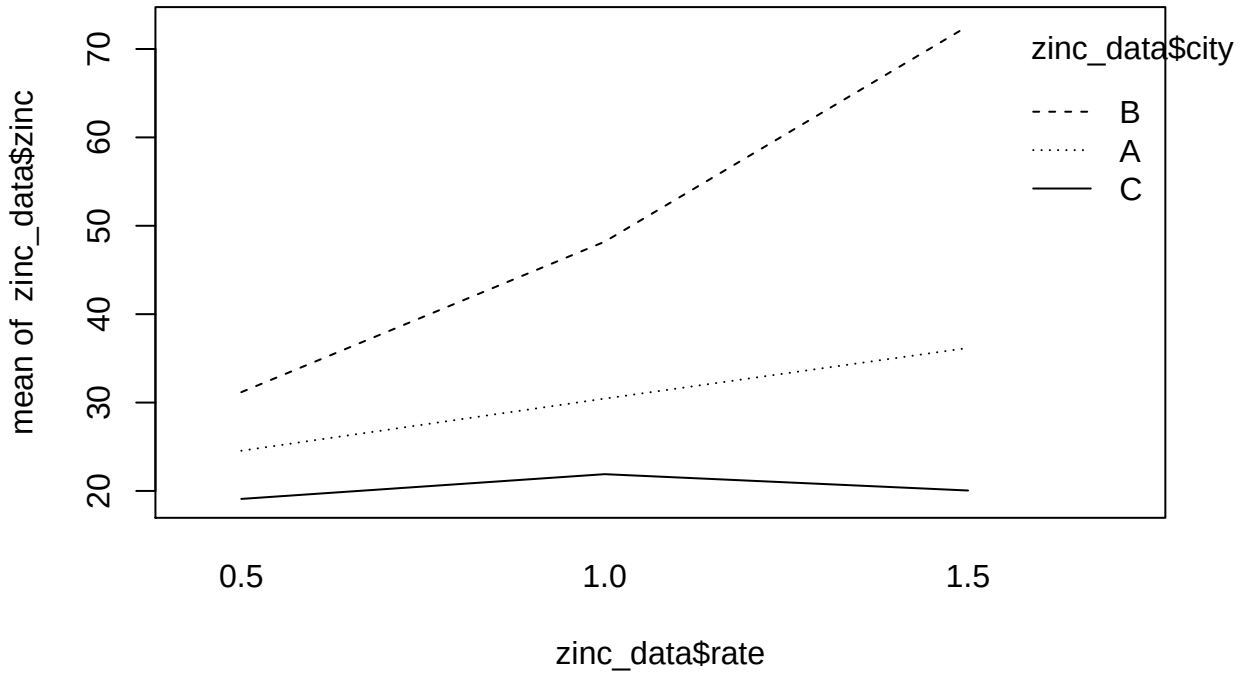
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## rate         2   1945   972.7   50.72 7.18e-10 ***
## city         2   5721  2860.3  149.13 2.56e-15 ***
## rate:city    4   1809   452.3   23.58 1.78e-08 ***
## Residuals   27    518    19.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which factors are significant? Interaction is significant?

Rate, City, and the interaction between rate and city are all significant.

Perform an interaction plot.

```
interaction.plot(x.factor = zinc_data$rate,
                 trace.factor = zinc_data$city,
                 response = zinc_data$zinc)
```



Different cities have different response in zinc concentration to increased rate.

c

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

```
zinc_model_tables <- model.tables( zinc_model, type = "means" )
zinc_model_tables$tables$"rate:city"
```

```
##      city
## rate  A      B      C
##   0.5 24.550 31.175 19.100
##   1.0 30.450 48.200 21.900
##   1.5 36.175 72.600 20.050
```

d

calculate the interaction sum squares from scratch.

```
grand_mean <- mean(zinc_data$zinc)
rate_means <- zinc_data %>% group_by(rate) %>% summarise(rate_mean = mean(zinc))
city_means <- zinc_data %>% group_by(city) %>% summarise(city_mean = mean(zinc))
cell_means <- zinc_data %>% group_by(city,rate) %>% summarise(cell_mean = mean(zinc))

zinc_data_means <- zinc_data %>%
  cbind(grand_mean = grand_mean) %>%
  merge(rate_means,by="rate") %>%
  merge(city_means,by="city") %>%
  merge(cell_means,by=c("rate","city"))

total_df <- nrow(zinc_data) - 1
rate_df <- nrow(rate_means) - 1
city_df <- nrow(city_means) - 1
interaction_df <- rate_df*city_df

total_deviations <- zinc_data_means$zinc - zinc_data_means$grand_mean
total_sum_of_squares <- sum(total_deviations^2)
total_mean_square <- total_sum_of_squares / total_df

rate_deviations <- zinc_data_means$zinc - zinc_data_means$rate_mean
rate_sum_of_squares <- sum(rate_deviations^2)
rate_mean_square <- rate_sum_of_squares / rate_df

city_deviations <- zinc_data_means$zinc - zinc_data_means$city_mean
city_sum_of_squares <- sum(city_deviations^2)
city_mean_square <- city_sum_of_squares / city_df

interaction_deviations <- zinc_data_means$cell_mean - zinc_data_means$city_mean - zinc_data_means$rate_
interaction_sum_of_squares <- sum(interaction_deviations^2)
interaction_mean_square <- interaction_sum_of_squares/interaction_df
interaction_sum_of_squares
```

```
## [1] 1809.398
```

the interaction sum of squares is 1809.398 which is consistent with the summary output in part b.