Homework 4

Igor Kuivjogi Fernandes

2023-02-16

1 The ANOVA from a randomized complete block experiment output is shown below. In this experiment, 30 experimental units were evaluated.

Source	SS	DF	MS	F	P
Treatment	1010.56	4	-	29.84	-
Block	-	-	64.765	-	-
Error	169.33	20	-		
Total	1503.71	-			

a) Fill in the blanks.

```
pf(q = 29.84, df1 = 4, df2 = 20, lower.tail = F)  # for treatment
```

```
## [1] 3.544848e-08
```

```
pf(q = 7.64956, df1 = 5, df2 = 20, lower.tail = F)  # for block
```

```
## [1] 0.0003688504
```

Source	SS	DF	MS	F	P
Treatment	1010.56	4	252.64	29.84	3.544848e-08
Block	323.82	5	64.765	7.64956	0.0003688504
Error	169.33	20	8.4665		
Total	1503.71	29			

b) How many blocks were used in this experiment?
6 blocks

c) What conclusions can you draw?
The treatment effect is significant when at a significance level of $\alpha = 0.05$.
The blocking effect in this case is useful to reduce the error sum of squares.

2 An experiment with 12 hybrids of Brachiaria spp was carried out in a randomized block design with three replications. The variable measured was the leaf protein content (P %).

```r
df <- data.frame(
  hybrid = 1:12,
  b1 = c(6.8, 5.8, 6.8, 5.6, 6.9, 3.9, 6, 4.5, 6.1, 5.3, 5.9, 5.2),
  b2 = c(8.9, 6.4, 8.9, 6.2, 6.1, 4.9, 5.5, 5, 5.3, 6.5, 9, 6.4),
  b3 = c(10, 9, 11, 6.9, 7, 5.2, 7.9, 6.1, 8.5, 9.7, 11.2, 7.6)
)
df_long <- reshape(df, direction = 'long', idvar = 'hybrid', varying = c('b1', 'b2', 'b3'),
                   timevar = 'block', v.names = 'protein')
rownames(df_long) <- 1:nrow(df_long)
df_long$hybrid <- as.factor(df_long$hybrid)
df_long$block <- as.factor(df_long$block)
tibble::glimpse(df_long)
```

```
## Rows: 36
## Columns: 3
## $ hybrid  <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5, 6, 7, 8,~
## $ block   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ protein <dbl> 6.8, 5.8, 6.8, 5.6, 6.9, 3.9, 6.0, 4.5, 6.1, 5.3, 5.9, 5.2, 8.~
```

a) Formulate the statistical hypotheses H0 and H1 related to the hybrids.
In this example, the hybrid is a treatment, so we can build a hypothesis on this treatment:

H_0 : the mean leaf protein content is equal across all the hybrids
H_1 : at least one mean differs

b) Check the basic assumptions at 5% probability for the purpose of performing the ANAVA (normality of errors: Q-Q Plot; additivity of effects: Tukey test; homoscedasticity: Anscombe and Tukey test (1963)). Interpret the results. Perform the analysis of variance (ANAVA).

First, let's see whether using a blocking effect reduces error variance:

```r
fit <- aov(protein ~ hybrid, data = df_long)
summary(fit)
```

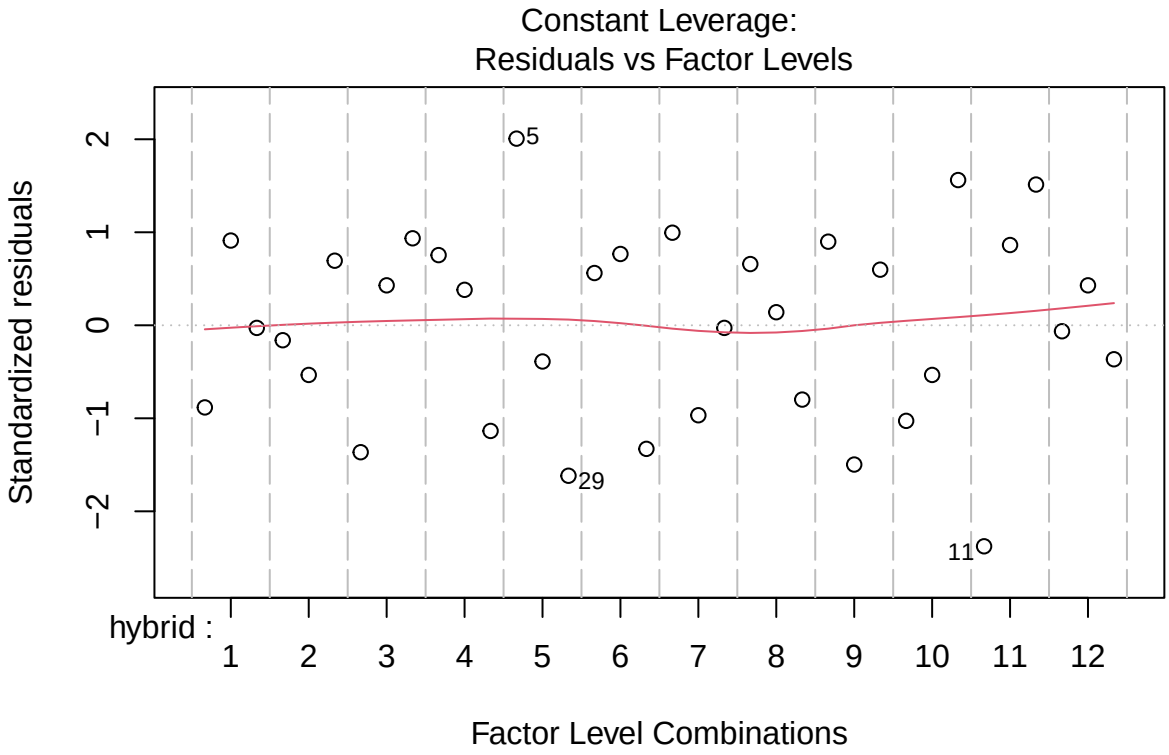
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## hybrid      11  57.00   5.182   2.085 0.0642 .
## Residuals   24  59.65   2.486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fit_block <- aov(protein ~ hybrid + block, data = df_long)
summary(fit_block)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## hybrid      11  57.00   5.182   6.612 9.00e-05 ***
## block        2  42.41  21.205  27.056 1.18e-06 ***
## Residuals   22  17.24   0.784
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding a blocking effect is useful to reduce error variance. In fact, without a blocking effect the treatment effect would not be significant at a significance level of $\alpha = 0.05$ because p-value $= 0.0642 > \alpha$. Let's check the assumptions of ANOVA now:

```
plot(fit_block, which = 5)
```



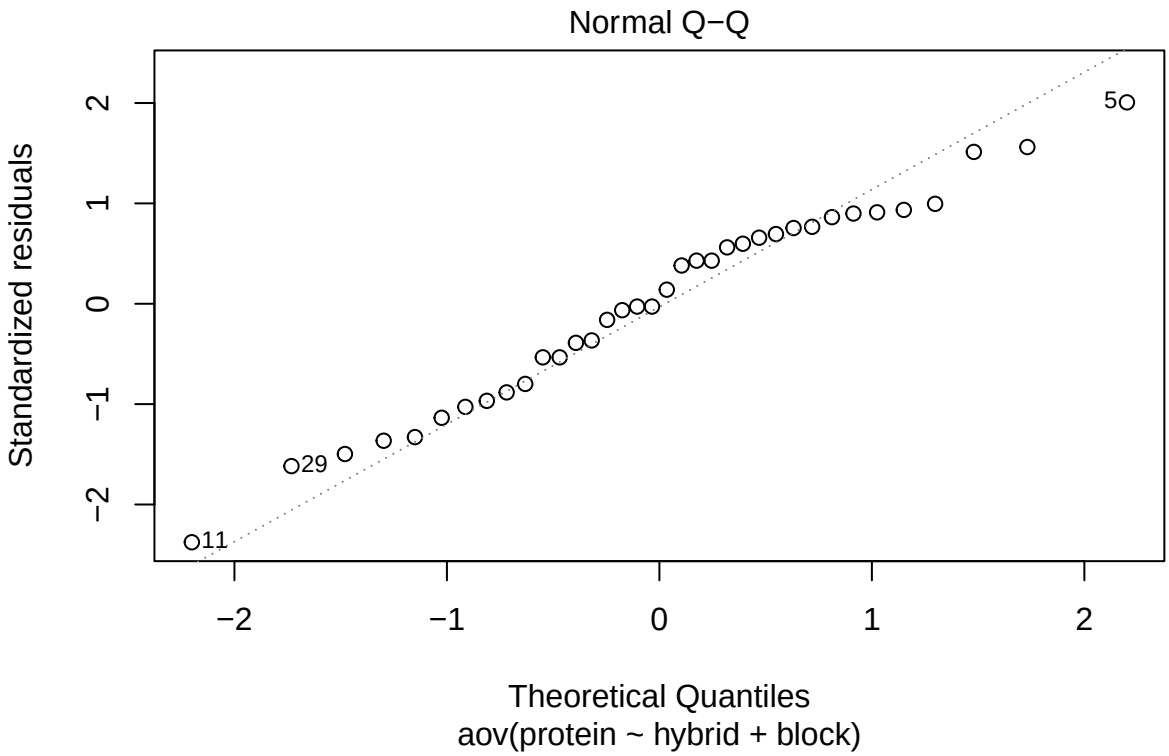
For the Levene's test, the null hypothesis is that the variances are equal across different levels.

```
car::leveneTest(protein ~ hybrid, data = df_long)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 11  0.5719 0.8324
##       24
```

From the plot, we can see that the variance are homogeneous across the different levels. From the Levene's test, we don't reject the null hypothesis that the variances are equal using a significance level of $\alpha = 0.05$.

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



For the Shapiro Wilk test, the null hypothesis in this case is that the residuals come from a normal distribution.

```
shapiro.test(fit_block$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit_block$residuals
## W = 0.97646, p-value = 0.6254
```

The Q-Q plot shows that the central points are around the line but there are some points in the tails more far away from the line. From the test, we conclude that the residuals are normally distributed using a significance level of $\alpha = 0.05$, i.e. we don't reject the null hypothesis that the residuals are normally distributed.

Let's now check the additivity of effects using the tukey test for additivite effects:

```
daewr::Tukey1df(data.frame(df_long$protein, df_long$hybrid, df_long$block))
```

```
## Registered S3 method overwritten by 'DoE.base':
##   method           from
##   factorize.factor conf.design

## Source           df     SS        MS        F     Pr>F
## A                11   57.0022    5.182
```

```
## B                2   42.4106    21.2053
## Error           22   17.2428    3.1351
## NonAdditivity    1   6.1902    6.1902    11.76    0.0025
## Residual        21   11.0526    0.5263
```

We reject the null hypothesis that the effects are additive, i.e. we can see that there's interaction between the treatment and the block using a significance level of $\alpha = 0.05$.

c) Which hybrid performed best?

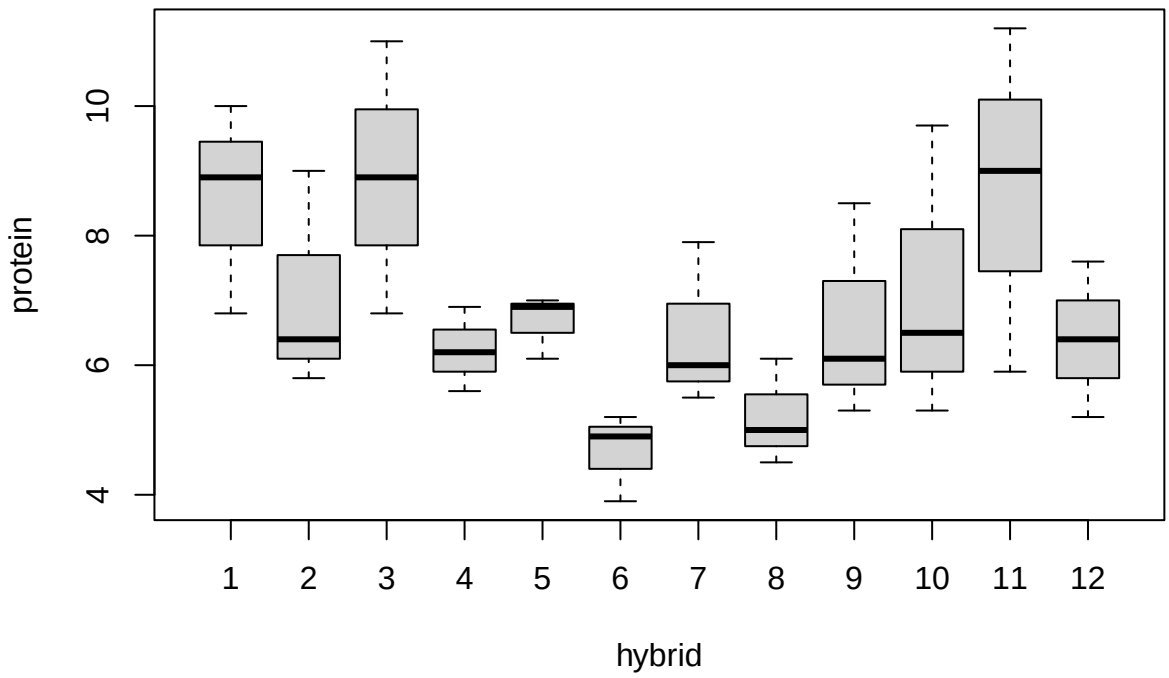
```
lsmeans::lsmeans(fit, ~hybrid)
```

```
##  hybrid lsmean   SE df lower.CL upper.CL
##  1        8.57 0.91 24     6.69    10.45
##  2        7.07 0.91 24     5.19     8.95
##  3        8.90 0.91 24     7.02    10.78
##  4        6.23 0.91 24     4.35     8.11
##  5        6.67 0.91 24     4.79     8.55
##  6        4.67 0.91 24     2.79     6.55
##  7        6.47 0.91 24     4.59     8.35
##  8        5.20 0.91 24     3.32     7.08
##  9        6.63 0.91 24     4.75     8.51
##  10       7.17 0.91 24     5.29     9.05
##  11       8.70 0.91 24     6.82    10.58
##  12       6.40 0.91 24     4.52     8.28
##
## Confidence level used: 0.95
```

The hybrid with the highest protein mean (i.e. performed the best) was the 3rd one.

d) Create a graph that shows the performance of different hybrids.

```
boxplot(protein ~ hybrid, data = df_long)
```



3 [Use data set: HW4_Q3.csv] The investigators (K. Blenk, M. Chen, G. Evans, J. Chen Ibinson, J. Lamack, and E. Scott, 2000) planned an experiment to investigate how rapid-rise yeast and regular yeast differ in terms of their rate of rising. They were also interested in finding out whether temperature had a significant effect on the rising rate. For each observation, 0.3 gm of yeast and 0.45 gm of sugar were mixed and added to a test tube, together with 6 ml of water. The test tube was placed into a water bath of a specified temperature. The level (height) of the mixture in the test tube was recorded immediately and then again after 15 minutes. Each response is the percentage gain in the height of the mixture in the test tube after 15 minutes. There were three treatment factors:

- Factor C: Initial temperature of water mixed with the yeast and flour (3 levels: $100 \circ$F, $115 \circ$F, $130 \circ$F)
- Factor D: Type of yeast (2 levels: Rapid rise, Regular)
- Factor E: Temperature of water bath (2 levels: $70 \circ$F, $85 \circ$F)

a) Explain in at most two sentences why the treatment combinations should be randomly ordered in each block before measurements.
The idea of using a block is to control the error. If you know a priori that there's some factor affecting your experiment but this effect is not intended to be analysed you can use it as a block to control the variance. The block controls the variance by grouping more homogeneous experimental units inside the same block, so inside each block you have to randomize the samples to account for independence of experimental units. If you don't randomize the experimental units within each block, you could favor some levels to be in a specific (e.g. spatially) portion of a block.

b) Obtain the analysis of variance table and explain what conclusions you can draw from it.

```
df <- read.csv('HW4_Q3.csv')
df_long <- reshape(
  df,
  direction = 'long',
  idvar = c('water', 'yeast', 'bath'),
  varying = c('block1', 'block1.1', 'block3'),
  timevar = 'block',
  v.names = 'rising'
)
rownames(df_long) <- 1:nrow(df_long)
df_long$water <- as.factor(df_long$water)
df_long$yeast <- as.factor(df_long$yeast)
df_long$bath <- as.factor(df_long$bath)
df_long$block <- as.factor(df_long$block)
tibble::glimpse(df_long)
```

```
## Rows: 36
## Columns: 5
## $ water  <fct> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 1, 1, 1, 1, 2, 2, 2, 2, 3, ~
## $ yeast  <fct> 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, ~
## $ bath   <fct> 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, ~
## $ block  <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ rising <dbl> 8.2, 30.0, 12.6, 64.7, 18.1, 63.5, 4.2, 96.8, 44.4, 58.2, 19.8,~
```

```
fit <- aov(rising ~ water * yeast * bath + block, data = df_long)
summary(fit)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## water             2    553     277   1.704 0.205019
## yeast             1    351     351   2.163 0.155545
## bath              1  12447   12447  76.711 1.27e-08 ***
## block             2   3494    1747  10.766 0.000549 ***
## water:yeast       2     17       8   0.052 0.949153
## water:bath        2     99      50   0.305 0.739961
## yeast:bath        1   1472    1472   9.072 0.006414 **
## water:yeast:bath  2    113      57   0.348 0.709600
## Residuals        22   3570     162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

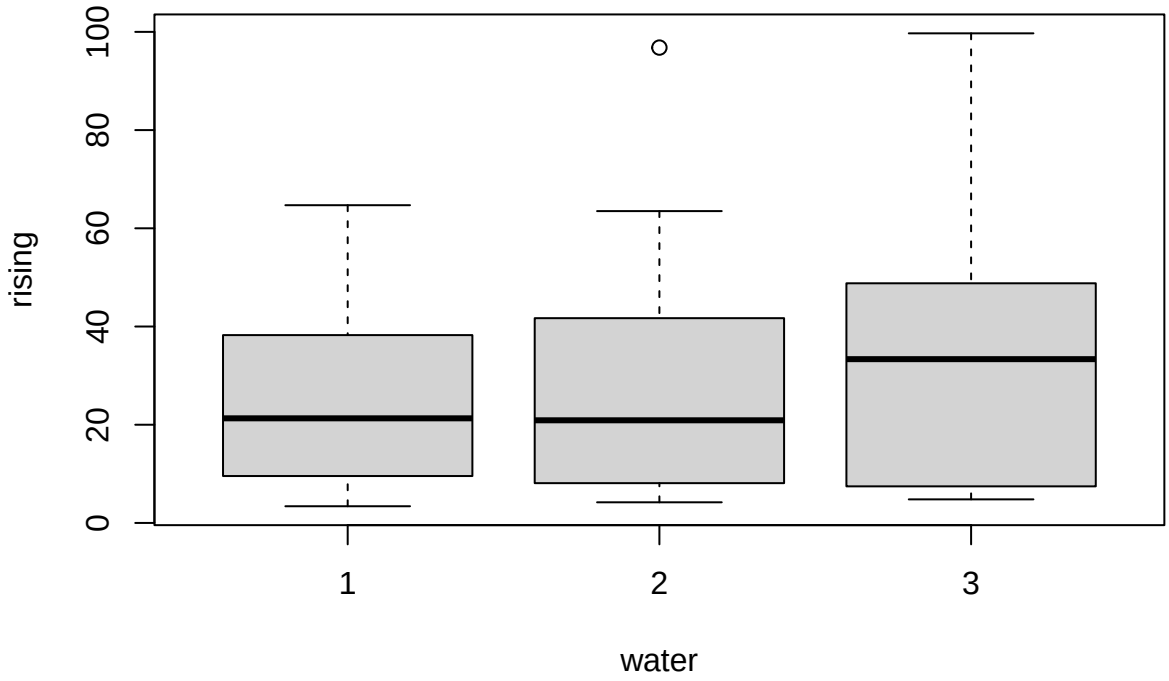
The "bath" effect is significant, but all his related interactions "water:bath" and "water:yeast:bath" are not.

The "water" effect is not significant, neither all his related interactions "water:yeast", "water:yeast", "water:bath", and "water:yeast:bath", so the "water" factor is not being significant at all.

Although the "yeast" effect is not significant, the interaction effect "yeast:bath" is significant so I would keep it in the model.

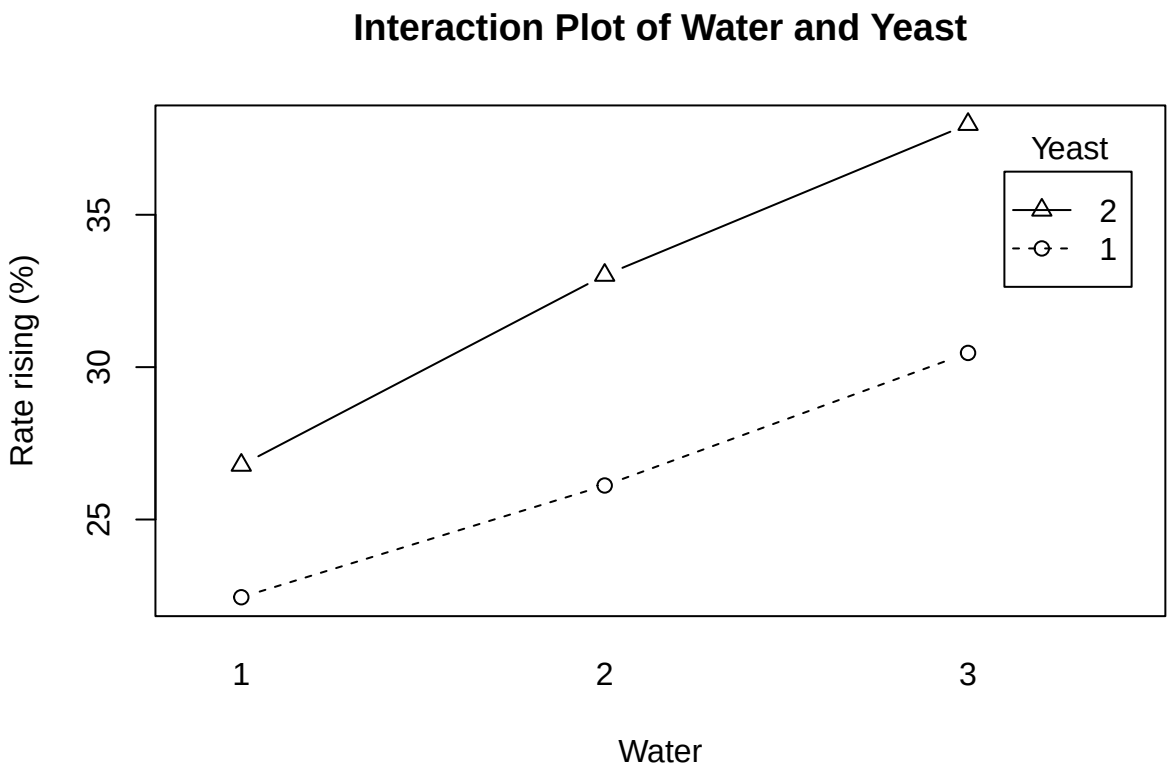
c) Create a figure that illustrates the effect of the factor "water".

```
boxplot(rising ~ water, data = df_long)
```



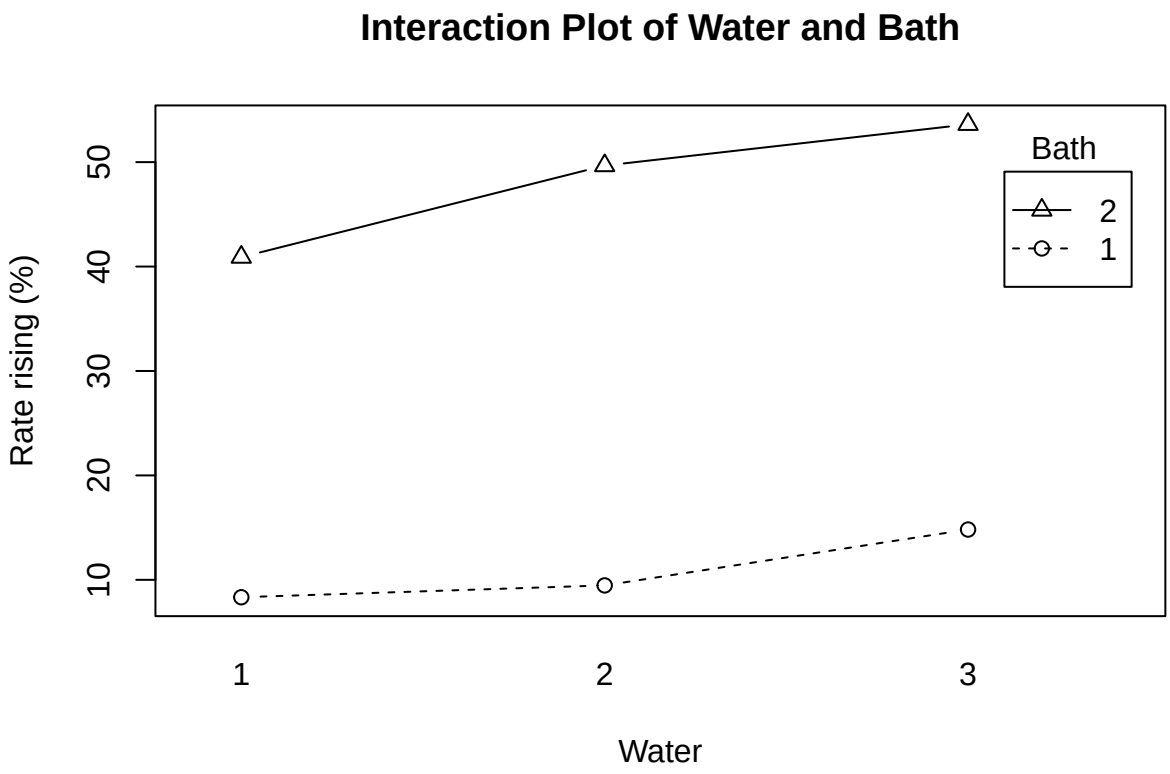
The water seems not to be affecting the rate of rising, as we just saw in the ANOVA results, although the 3rd water level has a slightly larger rate rising median.

```
with(df_long, {interaction.plot(water, yeast, rising, type = 'b',
                            pch = c(1, 2), leg.bty = 'o',
                            main = 'Interaction Plot of Water and Yeast',
                            xlab = 'Water', ylab = 'Rate rising (%)',
                            trace.label = 'Yeast')})
```



The interaction plot between "water" and "yeast" agrees with the ANOVA results: there are not interaction between them because the lines are quite parallel.

```
with(df_long, {interaction.plot(water, bath, rising, type = 'b',
                            pch = c(1, 2), leg.bty = 'o',
                            main = 'Interaction Plot of Water and Bath',
                            xlab = 'Water', ylab = 'Rate rising (%)',
                            trace.label = 'Bath')})
```



The same occurs for "water" and "bath" terms, i.e., there's no interaction effect.

4 The data set presented in HW4_Q4.csv comes from an experiment that aimed to evaluate the effect of three doses of herbicide and four different fertilizers on yield. It is a factorial CRD. Run the proper analysis and make interpretations.

```
df <- read.csv('HW4_Q4.csv')
df$yield <- abs(df$yield)  # fix one negative response point
df$herbicide_dose <- as.factor(df$herbicide_dose)
df$fertilizer <- as.factor(df$fertilizer)
tibble::glimpse(df)
```

```
## Rows: 36
## Columns: 5
## $ plot           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
## $ herbicide_dose <fct> 0, 100, 200, 0, 100, 200, 0, 100, 200, 0, 100, 200, 0, ~
## $ fertilizer     <fct> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 1, 1, 1, 2, 2, 2, 3~
## $ rep            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2~
## $ yield          <dbl> 13.140612, 10.967572, 6.724322, 69.950929, 44.421650, 2~
```

```
with(df, table(herbicide_dose, fertilizer))
```

```
##               fertilizer
```

```
## herbicide_dose 1 2 3 4
##            0   3 3 3 3
##            100 3 3 3 3
##            200 3 3 3 3
```

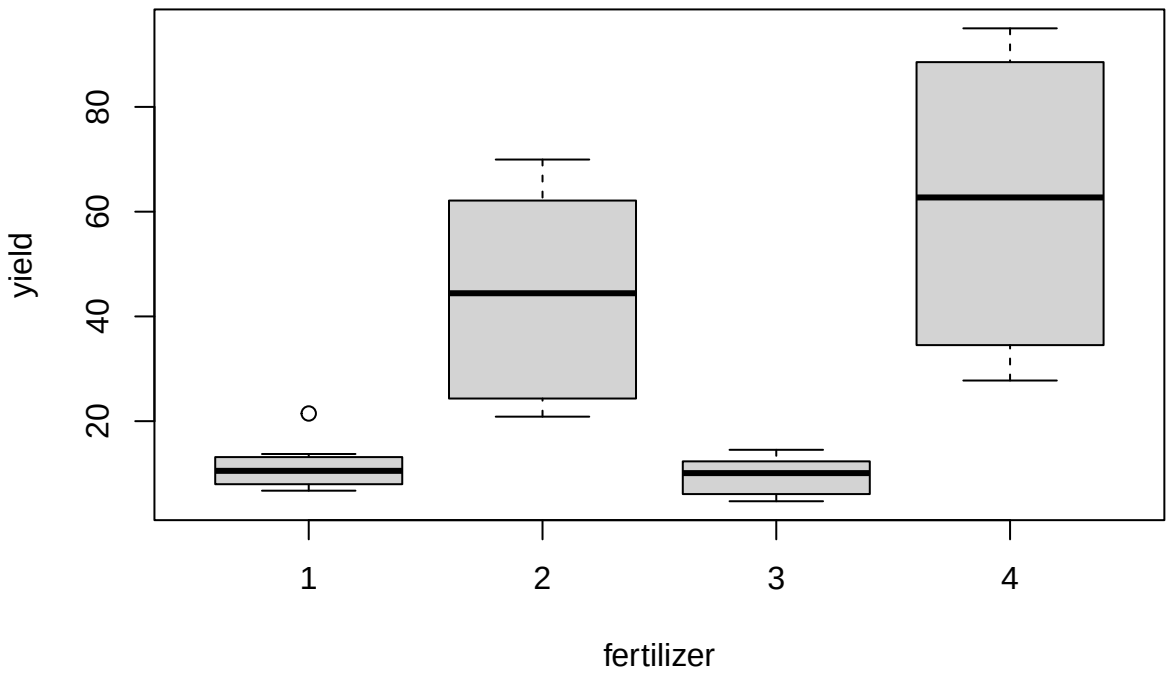
We have the same number of replications for each combination level.

Let's see how each factor seems to be affecting the yield:

```
boxplot(yield ~ herbicide_dose, data = df)
```

Seems that as you increase the herbicide dose, the yield decreases.

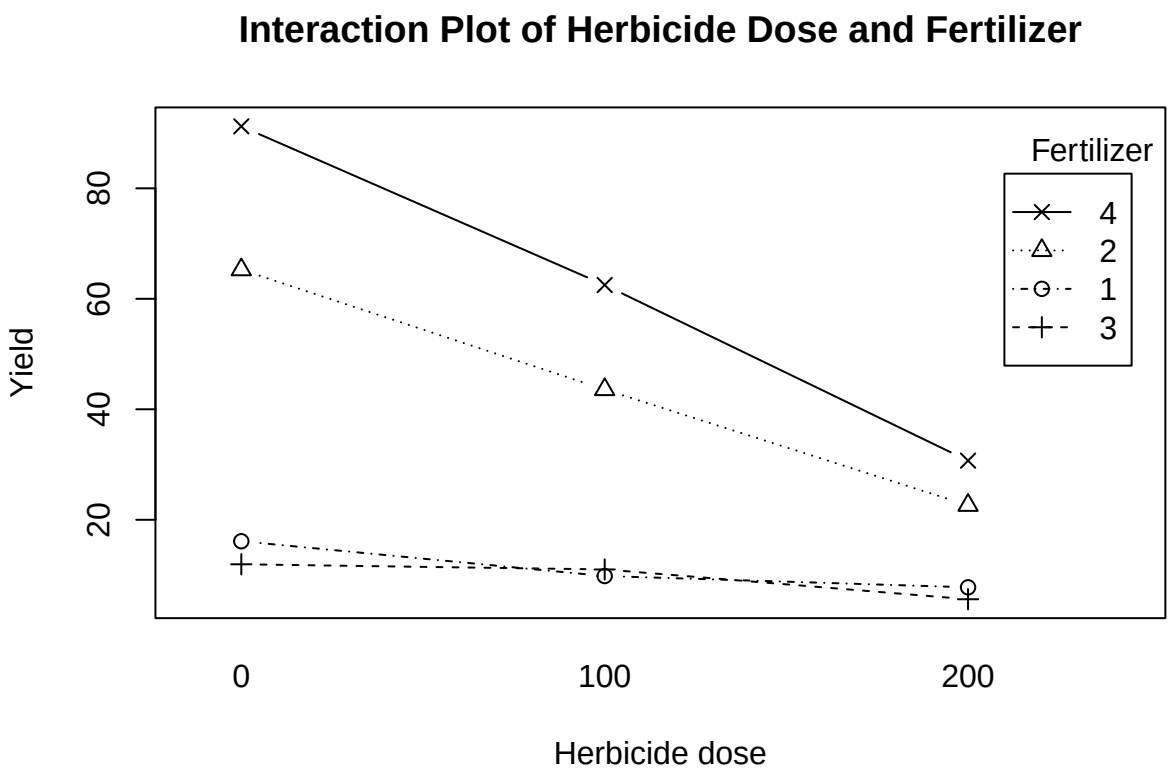
```
boxplot(yield ~ fertilizer, data = df)
```



The type of fertilizer seems to be important to the mean yield as well.

What about the interactions?

```
with(df, {interaction.plot(herbicide_dose, fertilizer, yield, type = 'b',
                           pch = c(1, 2, 3, 4), leg.bty = 'o',
                           main = 'Interaction Plot of Herbicide Dose and Fertilizer',
                           xlab = 'Herbicide dose', ylab = 'Yield',
                           trace.label = 'Fertilizer')})
```



There seems to be an interaction effect as well, because as long you increase the herbicide dose the yield decreases in different rates depending on the fertilizer type.

Let's check the significance of effects with an ANOVA:

```
fit <- aov(yield ~ herbicide_dose * fertilizer, data = df)
summary(fit)
```

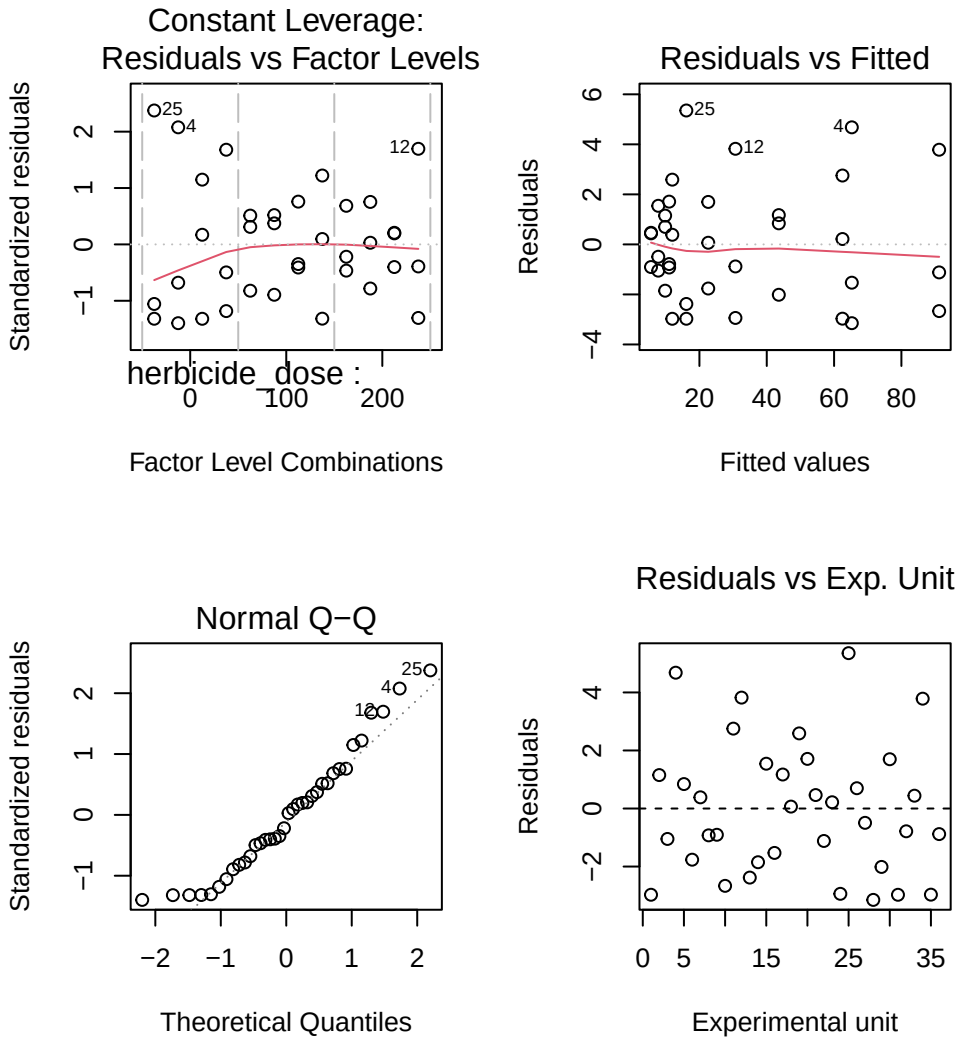
```
##                            Df Sum Sq Mean Sq F value   Pr(>F)
## herbicide_dose              2   5207    2604  341.60  < 2e-16 ***
## fertilizer                  3  17494    5831  765.09  < 2e-16 ***
## herbicide_dose:fertilizer  6   3203     534   70.03 5.08e-14 ***
## Residuals                  24    183       8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results confirms all the interpretations for the aforementioned plots: both treatments and the interaction are significant using a significance level of $\alpha = 0.05$ because all the $p - values < \alpha$.

What about the residuals?

```
par(mfrow = c(2, 2))
plot(fit, which = 5)
plot(fit, which = 1)
plot(fit, which = 2)
plot(residuals(fit) ~ plot, main = 'Residuals vs Exp. Unit',
```

```
    font.main = 1, data = df, xlab = 'Experimental unit', ylab = 'Residuals')
abline(h = 0, lty = 2)
```



The residuals seems to have homogeneous variance across different herbicide doses.
The Q-Q plot shows some lines getting away from the Q-Q line.
The Residuals vs Experiment Unit plot shows a very good horizontal pattern, i.e. which shows an independence of experimental units.
In general, I would say the model adequacy is good.

5 A researcher wants to conduct an experiment to evaluate the effect of irrigation and cover crops on rice yield. Three levels of irrigation were selected (1, 2, and 3 irrigations/per week). Four cover crops were selected (A, B, C, D). They would like to use three replicates. There is a gradient of fertility in the field making it slightly heterogeneous. With the information above, design the proper experiment. Made up some data to analyze this experiment and present the design and the analysis below.