Homework 5 - AGST 5014

Igor Kuivjogi Fernandes and Ashmita Upadhyay

2023-03-13

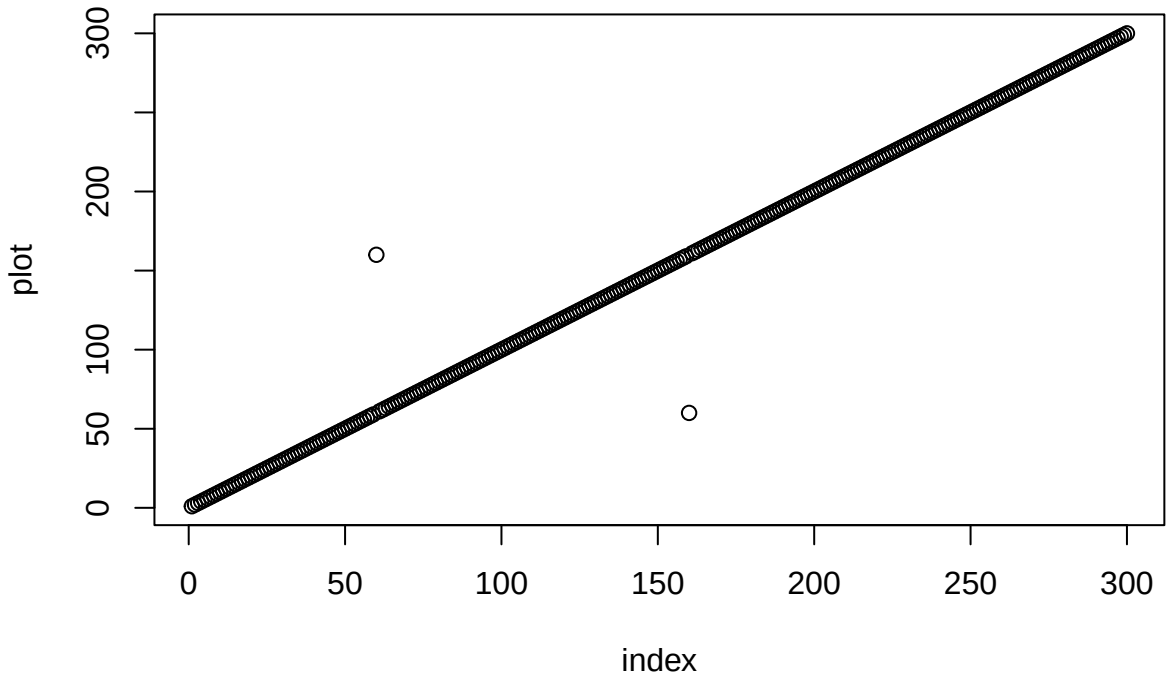
1. The following data set has 2 mistakes. Find the issue and analyze the experiment. Include all the necessary analysis and interpretation

```
q1 <- read.csv("HW5_Q1_fixed.csv")
q1$fertilizer_application <- NULL
q1$rep <- NULL
factor_cols <- c("row", "column", "cultivar", "fer", "block")
q1[factor_cols] <- lapply(q1[factor_cols], as.factor)
str(q1)
```

```
## 'data.frame':    300 obs. of  7 variables:
##  $ plot    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ row     : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ column  : Factor w/ 25 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ cultivar: Factor w/ 20 levels "1","2","3","4",..: 1 1 1 2 2 2 3 3 3 4 ...
##  $ y       : num  82.4 67.2 255.1 186.6 139.1 ...
##  $ block   : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ fer     : Factor w/ 3 levels "April","June",..: 1 2 3 1 2 3 1 2 3 1 ...
```

We have 12 rows and 25 columns.

```
plot(q1$plot, xlab = 'index', ylab = 'plot')
```

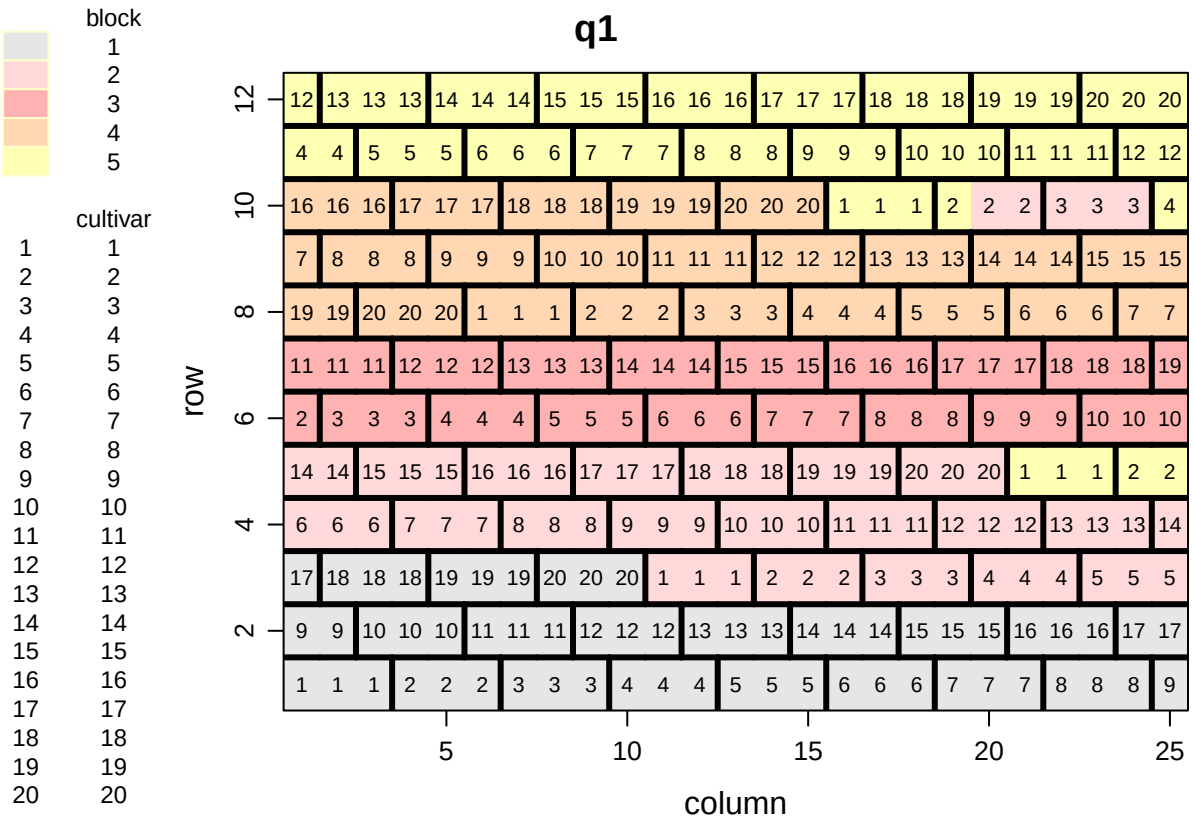



```
q1[rownames(q1) %in% c(60, 160), ]
```

```
##     plot row column cultivar         y block   fer
## 60   160   3     10       20 104.71737     1   May
## 160   60   7     10       14  50.57193     3 April
```

The plot 160 should be labelled as 60, and the plot 60 should be labelled as 160.

```
desplot::desplot(
  q1,
  block ~ column * row,
  cex = 0.7,
  text = cultivar,
  out1 = cultivar,
  ticks = T
)
```

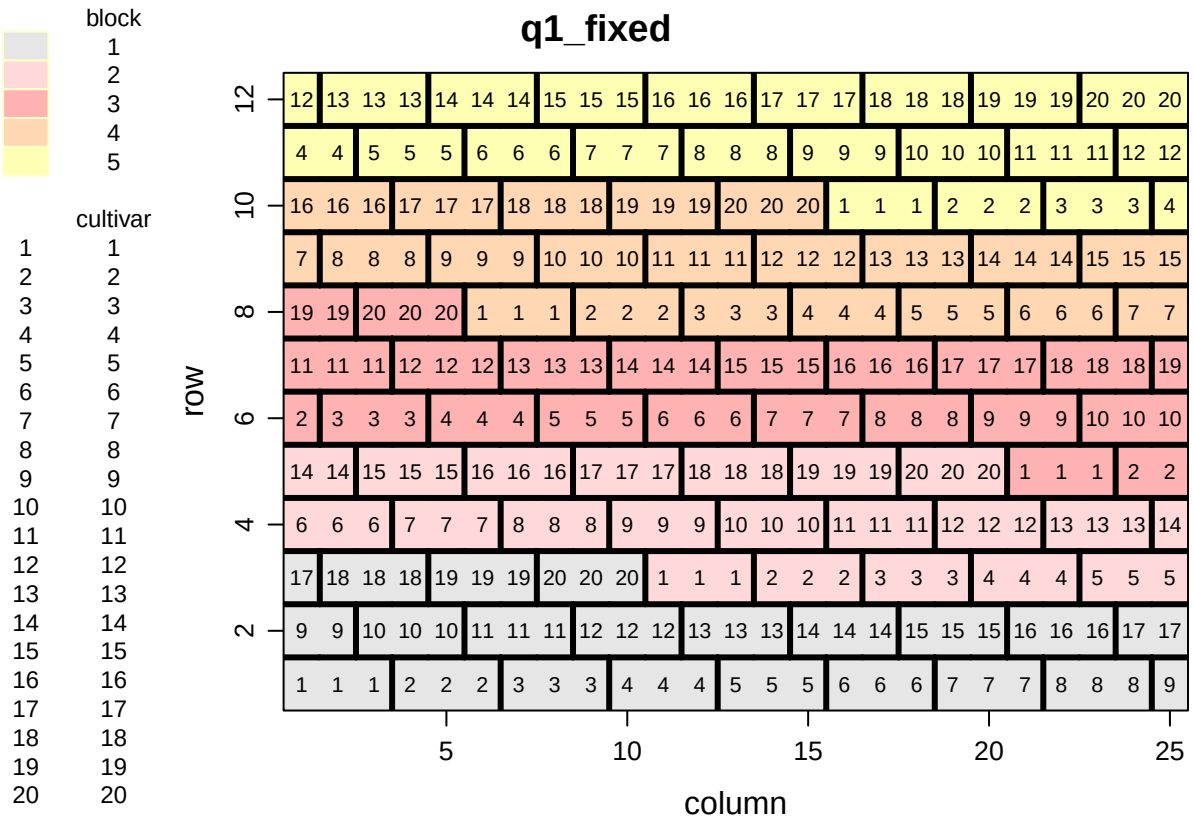


The row 5 with columns 21 to 25 should be labelled as block 3 rather than block 5.
The row 8 with columns 1 to 5 should be labelled as block 3 rather than block 4.
The row 10 with columns 20 to 24 should be labelled as block 5 rather than block 2.
Fixing this, now each cultivar appears only once within each block.

Fixing errors:

```
q1_fixed <- q1

# fixing plots
q1_fixed[rownames(q1_fixed) == 60, 'plot'] <- 60
q1_fixed[rownames(q1_fixed) == 160, 'plot'] <- 160

# fixing blocks
q1_fixed[(q1_fixed$row == 5) & (q1_fixed$column %in% 21:25), 'block'] <- 3
q1_fixed[(q1_fixed$row == 8) & (q1_fixed$column %in% 1:5), 'block'] <- 3
q1_fixed[(q1_fixed$row == 10) & (q1_fixed$column %in% 20:24), 'block'] <- 5
```

```
desplot::desplot(
  q1_fixed,
  block ~ column * row,
  cex = 0.7,
  text = cultivar,
  out1 = cultivar,
  ticks = T
)
```



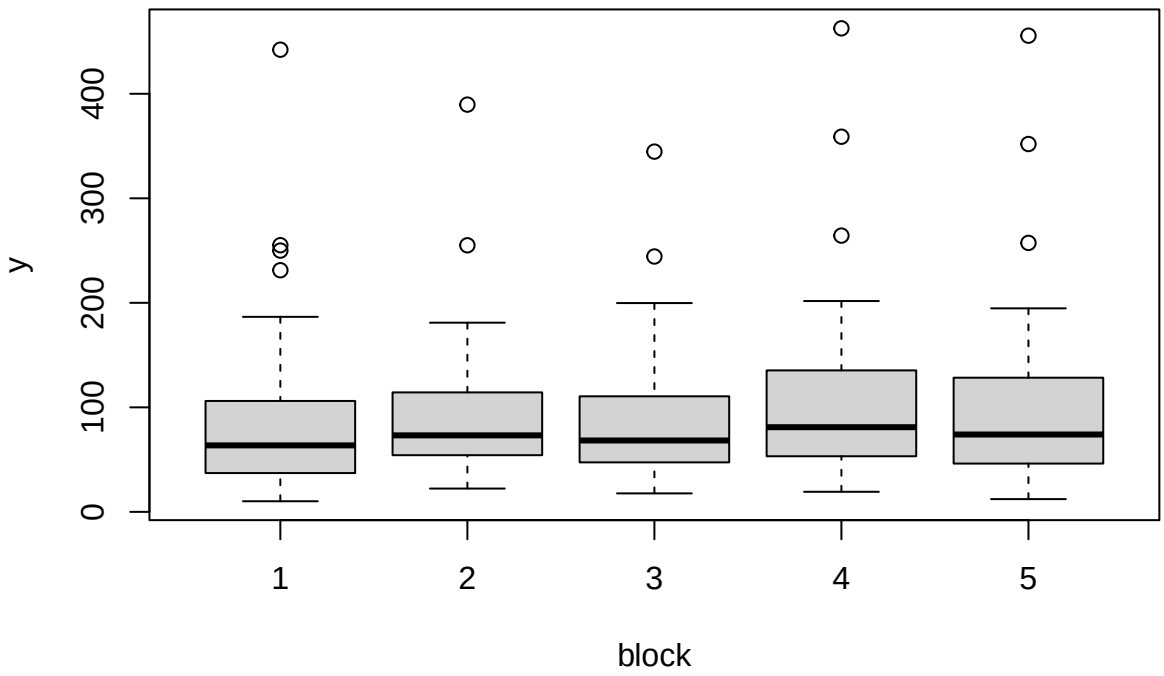

```
table(q1_fixed$block, q1_fixed$cultivar, dnn = c('block', 'cultivar'))
```

```
##       cultivar
## block 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
##     1 3 3 3 3 3 3 3 3 3  3  3  3  3  3  3  3  3  3  3  3
##     2 3 3 3 3 3 3 3 3 3  3  3  3  3  3  3  3  3  3  3  3
##     3 3 3 3 3 3 3 3 3 3  3  3  3  3  3  3  3  3  3  3  3
##     4 3 3 3 3 3 3 3 3 3  3  3  3  3  3  3  3  3  3  3  3
##     5 3 3 3 3 3 3 3 3 3  3  3  3  3  3  3  3  3  3  3  3
```

Now we have the 3 eu's for each block/cultivar combination.

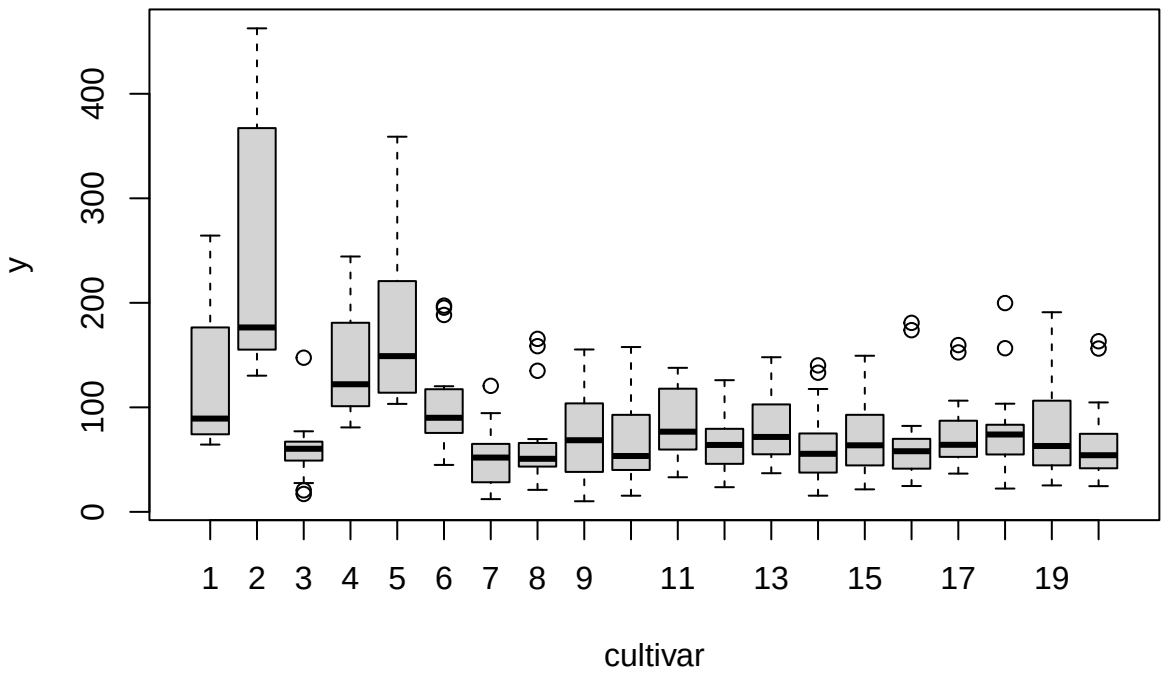
Let's do the analysis now.

```
boxplot(y ~ block, data = q1_fixed)
```



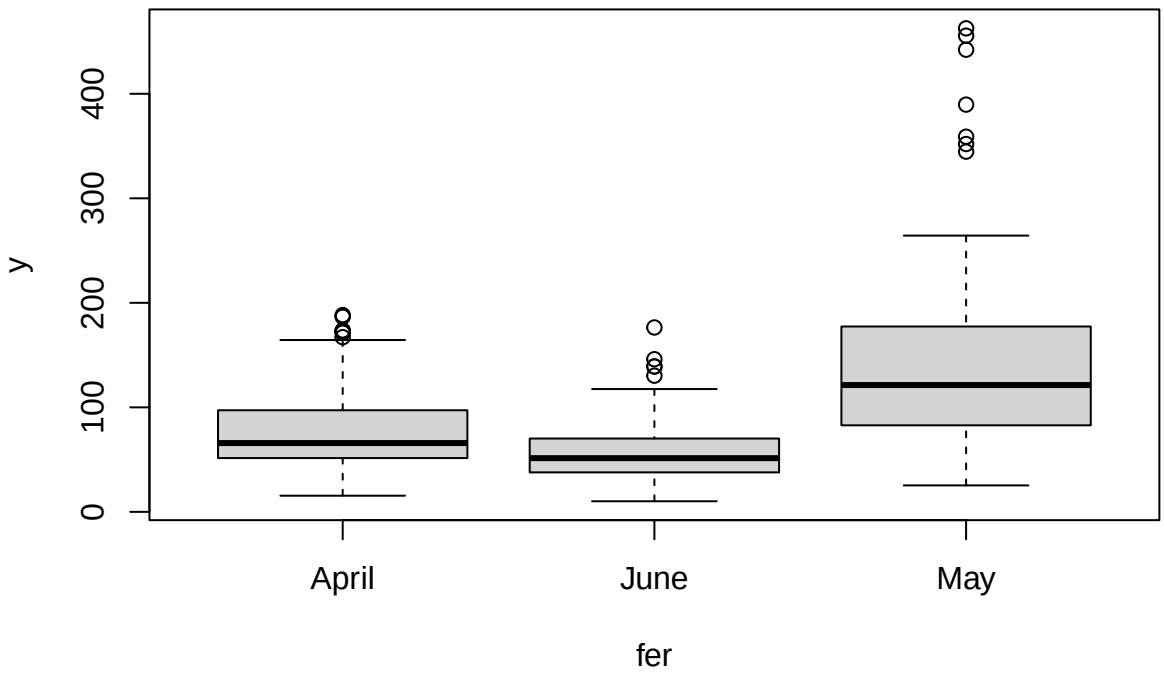
The yield has very similar distribution between blocks.

```
boxplot(y ~ cultivar, data = q1_fixed)
```



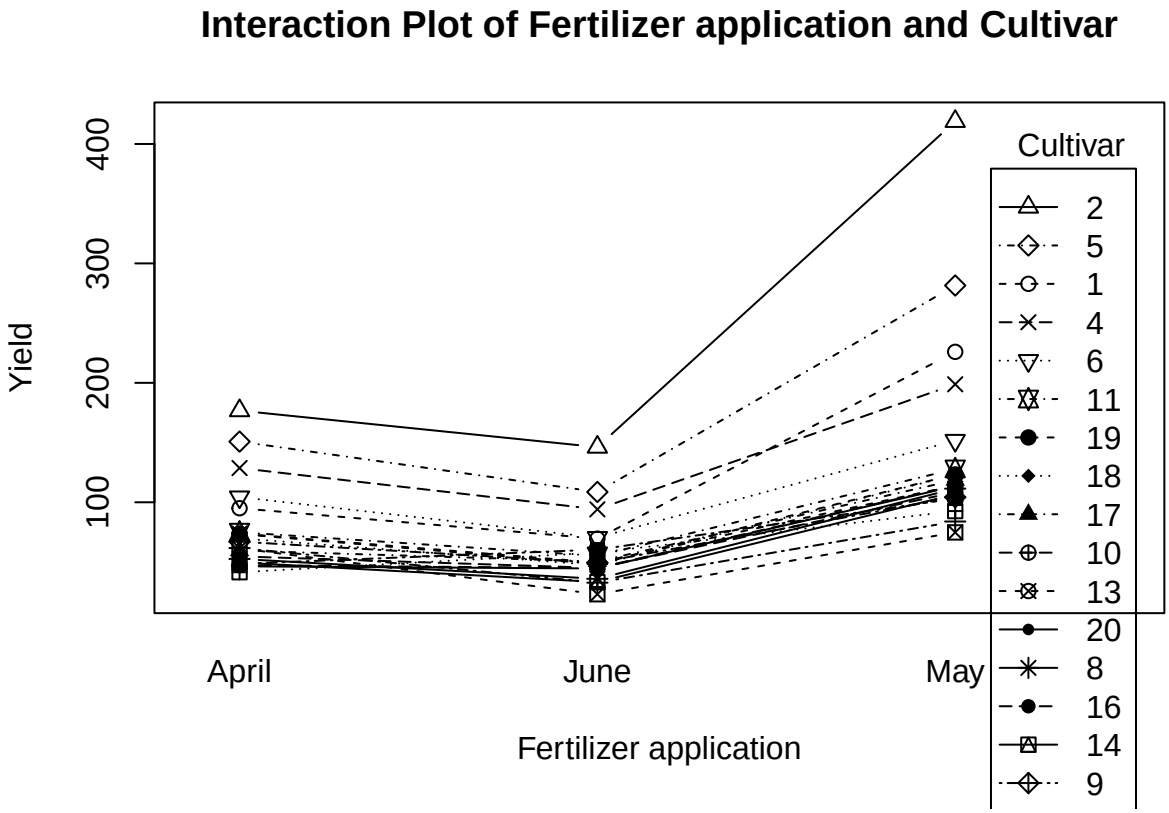
Some cultivar (e.g. 1, 2, and 5) had a larger yield than others.

```
boxplot(y ~ fer, data = q1_fixed)
```



The yield in May was larger than the other months.

```
with(q1_fixed, {interaction.plot(fer, cultivar, y, type = 'b',
                            pch = 1:20, leg.bty = 'o',
                            main = 'Interaction Plot of Fertilizer application and Cultivar',
                            xlab = 'Fertilizer application', ylab = 'Yield',
                            trace.label = 'Cultivar')})
```



When comparing June and May, for example, the Cultivar 2 had a larger yield difference when comparing to other cultivars, so seems there's an interaction between cultivar and fertilizer application.

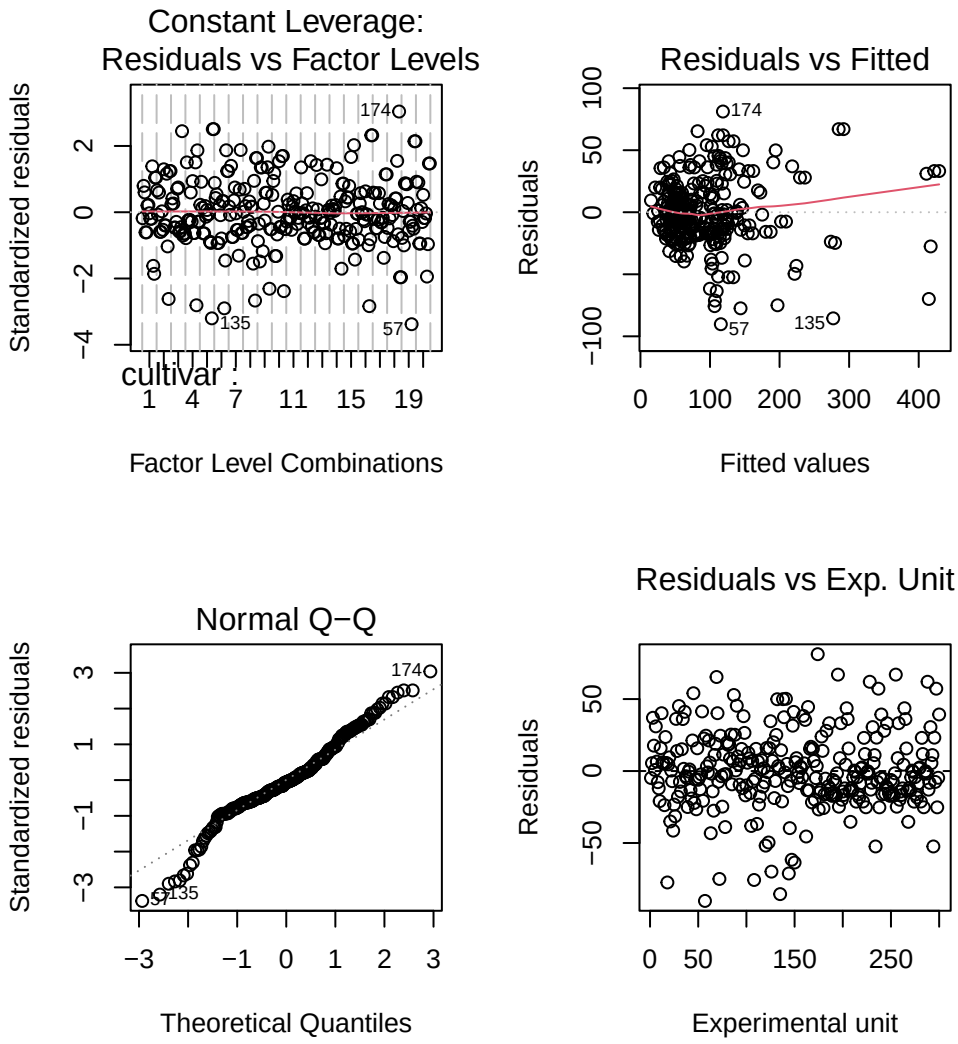
Let's fit an ANOVA now:

```
fit <- aov(y ~ cultivar * fer + block, data = q1_fixed)
summary(fit)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## cultivar      19 658986   34683  38.254  < 2e-16 ***
## fer            2 414700  207350 228.698  < 2e-16 ***
## block          4  12242    3060   3.376   0.0104 *
## cultivar:fer  38 183379    4826   5.323 2.54e-16 ***
## Residuals    236 213971     907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using an significance level of $\alpha = 0.05$, the cultivar, fertilizer application, and the interaction are significant, because all the p-values are lower than α.

```
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 = q1_fixed, xlab = 'Experimental unit', ylab = 'Residuals')
abline(h = 0, lty = 2)
```




```
shapiro.test(residuals(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.97318, p-value = 2.142e-05
```

The homogeneous of variance between different cultivar levels seems to be present.
The Residuals vs Fitted plot shows we have many fitted values between 0 and 100, and from 100 the residual's variance is larger. We have a horizontal pattern across the experimental units.
The normality of residuals was not met, maybe because the yield distribution is very skewed but we don't know whether the big values (from May fertilizer application) of yield are outliers or not. Maybe we could try to transform the data or use a GLM, for example.

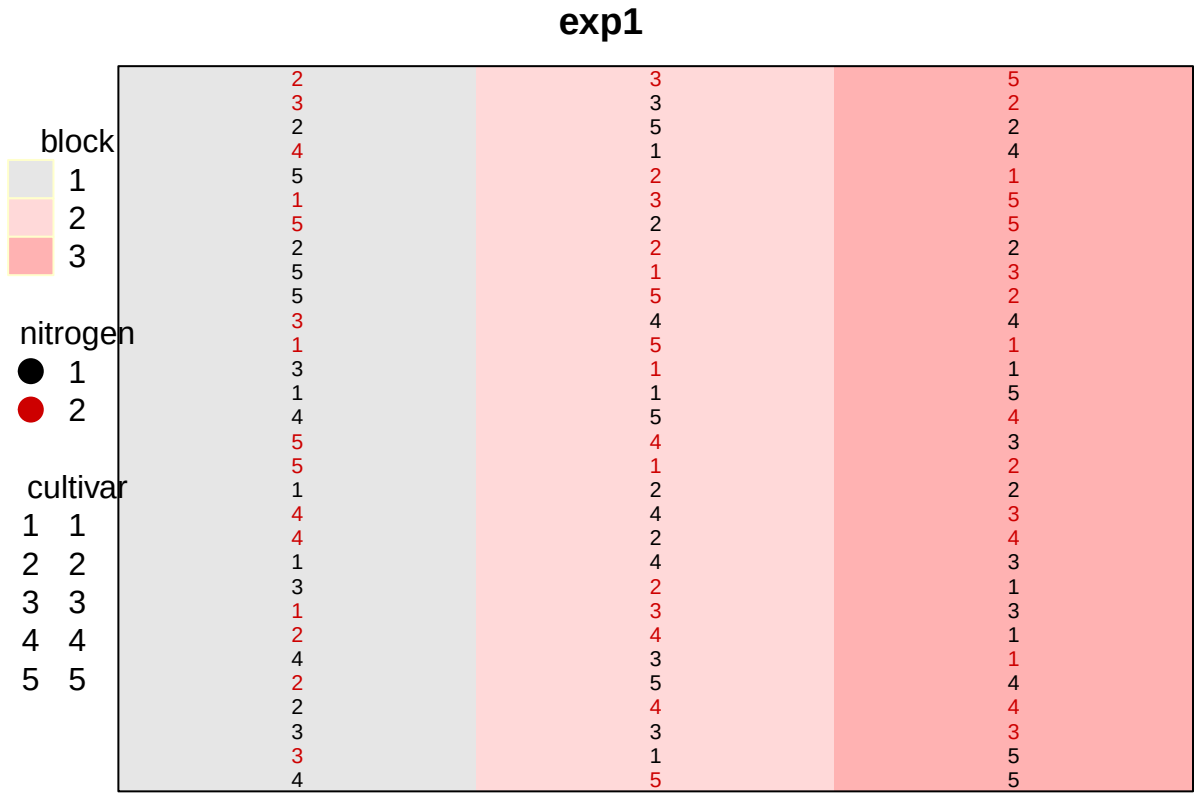
2. Design an experiment to answer the following questions (Attach the CSV file and include below a figure with the layout of your experiment):

a) Does increasing the dose of nitrogen affect yield? Does it depend on the frequency of irrigation? or on the cultivar? In the allocated field, I have some areas that are more fertile than others. I have resources for a total of 90 EU.
Let's say we have two doses of nitrogen 1 and 2, frequency of irrigation 0 and 10, 5 different cultivars, and a block to account for 3 subgroups (different fertile area gradients).

```
set.seed(2023)
grid <- expand.grid(nitrogen = 1:2, freq_irrigation = c(0, 10, 20), cultivar = 1:5)

# randomize block 1
b1 <- grid
b1$block <- 1
b1 <- b1[sample(1:nrow(b1)), ]
b1$row <- 1:nrow(b1)
b1$col <- 1
b1$yield <- rnorm(nrow(b1), 10, 1)  # low yield

# randomize block 2
b2 <- grid
b2$block <- 2
b2 <- b2[sample(1:nrow(b2)), ]
b2$row <- 1:nrow(b2)
b2$col <- 2
b2$yield <- rnorm(nrow(b2), 14, 1.2)  # medium yield

# randomize block 3
b3 <- grid
b3$block <- 3
b3 <- b3[sample(1:nrow(b3)), ]
b3$row <- 1:nrow(b3)
b3$col <- 3
b3$yield <- rnorm(nrow(b2), 18, 1.1)  # high yield

# bind all blocks
exp1 <- rbind(b1, b2, b3)
rownames(exp1) <- 1:nrow(exp1)
factors <- c('nitrogen', 'freq_irrigation', 'cultivar', 'block', 'row', 'col')
exp1[factors] <- lapply(exp1[factors], as.factor)

exp1[exp1$nitrogen == 2, 'yield'] <- (
  exp1[exp1$nitrogen == 2, 'yield'] + rnorm(sum(exp1$nitrogen == 2), 1, 1)
)  # nitrogen dose 2 favors yield

exp1[exp1$freq_irrigation == 10, 'yield'] <- (
  exp1[exp1$freq_irrigation == 10, 'yield'] + rnorm(sum(exp1$freq_irrigation == 10), 1, 1)
)  # frequency of irrigation 10 favors the yield

exp1[exp1$cultivar == 4, 'yield'] <- (
  exp1[exp1$cultivar == 4, 'yield'] + rnorm(sum(exp1$cultivar == 4), 2, 1)
)  # cultivar 4 favors the yield
```

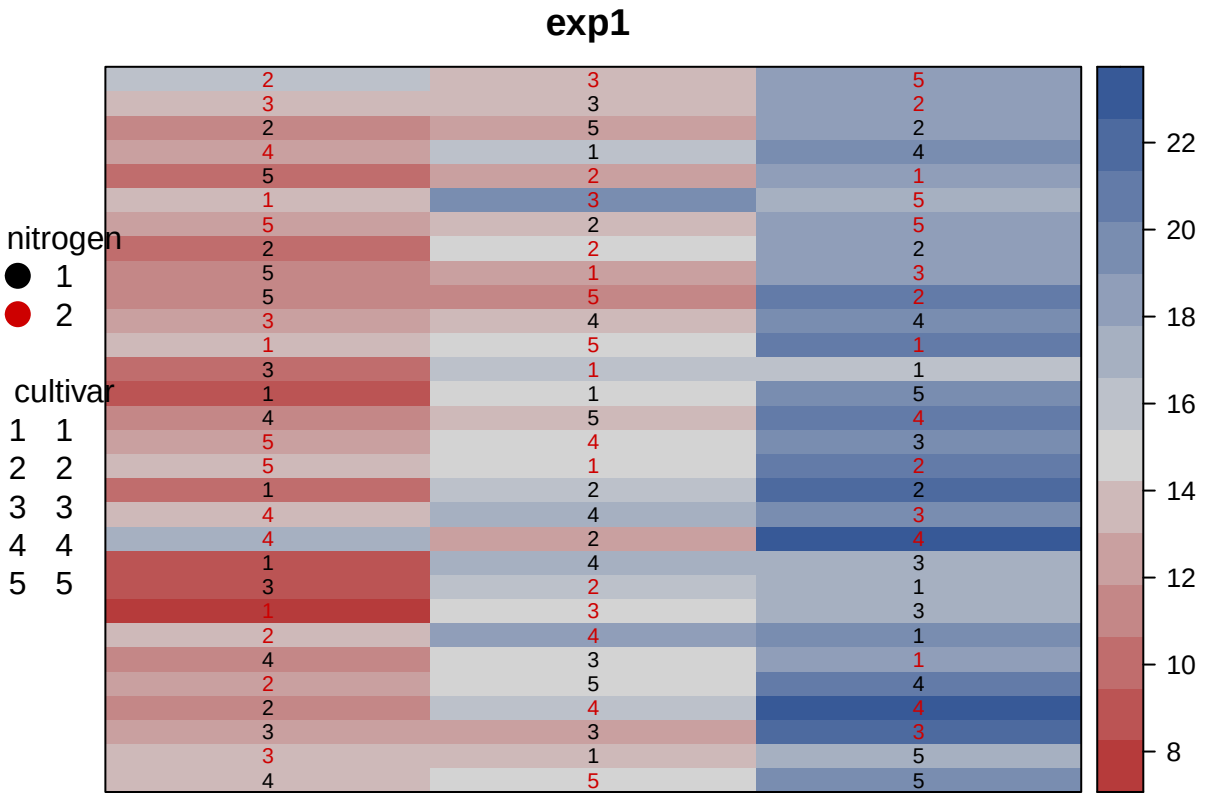
The layout:

```
desplot::desplot(
  exp1,
  block ~ col * row,
  text = cultivar,
  col = nitrogen,
  cex = 0.7
)
```



Each block represent a different fertile area that affects the yield. This effect could be seem using the color as the response:

```
desplot::desplot(
  exp1,
  yield ~ col * row,
  text = cultivar,
  col = nitrogen,
  cex = 0.7
)
```




```
fit <- aov(yield ~ nitrogen * freq_irrigation * cultivar + block, data = exp1)
summary(fit)
```

```
##                                    Df Sum Sq Mean Sq F value   Pr(>F)
## nitrogen                            1   44.4    44.4  23.240 1.07e-05 ***
## freq_irrigation                     2   30.5    15.3   7.992  0.00086 ***
## cultivar                            4   68.3    17.1   8.939 1.08e-05 ***
## block                               2  825.6   412.8 216.143  < 2e-16 ***
## nitrogen:freq_irrigation            2    4.5     2.2   1.174  0.31637
## nitrogen:cultivar                   4    9.6     2.4   1.257  0.29740
## freq_irrigation:cultivar            8   34.1     4.3   2.231  0.03782 *
## nitrogen:freq_irrigation:cultivar   8   20.2     2.5   1.322  0.25096
## Residuals                          58  110.8     1.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

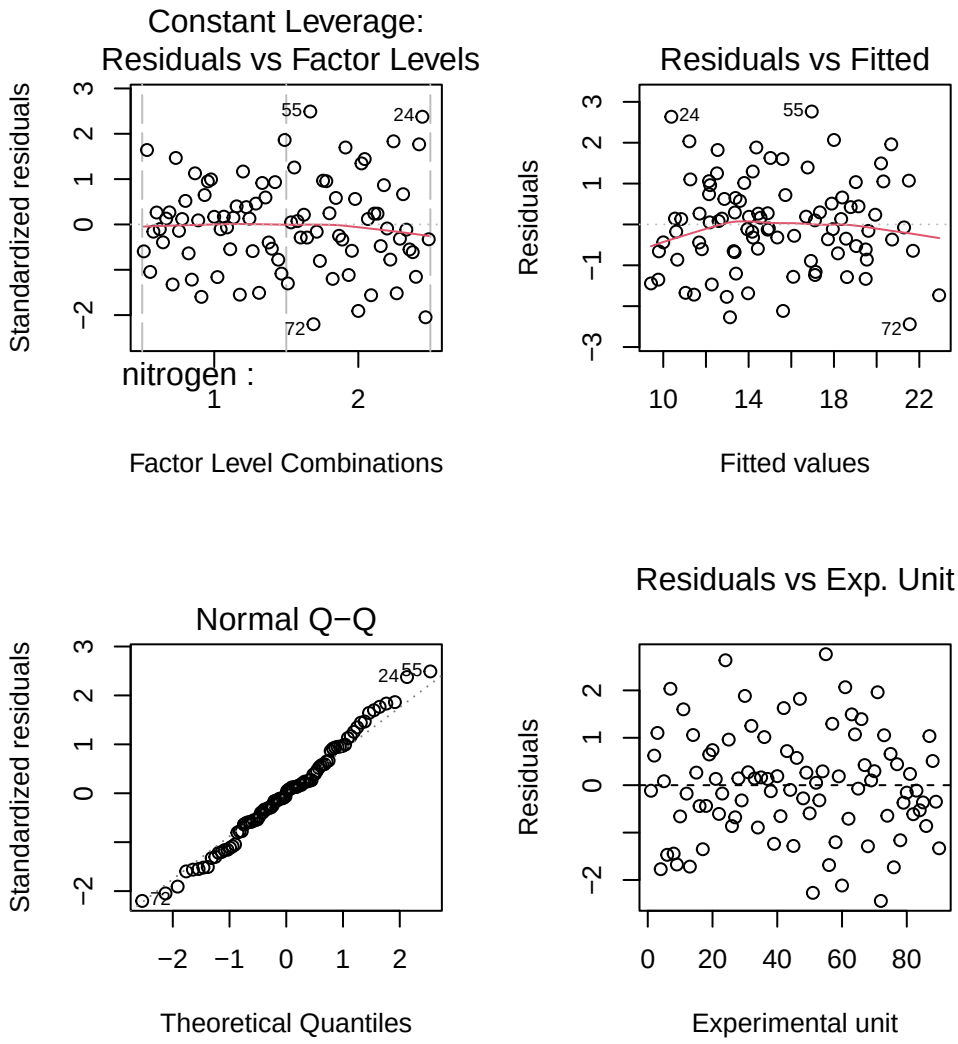
Using a significance level of $\alpha = 0.05$, the dose of nitrogen affects the yield, the frequency of irrigation also affects the yield, so as the cultivar. The interaction 'freq_irrigation:cultivar' is also significant, but the others interactions are not.
The block effect was important to control the errors because it has a large mean sum of squares compared to the residuals.
In fact, if we do not account for a blocking effect, all the terms would not be significant:

```
summary(aov(yield ~ nitrogen * freq_irrigation * cultivar, data = exp1))
```

```
##                                    Df Sum Sq Mean Sq F value Pr(>F)
## nitrogen                            1   44.4   44.38   2.844 0.0969 .
## freq_irrigation                     2   30.5   15.26   0.978 0.3819
## cultivar                            4   68.3   17.07   1.094 0.3678
## nitrogen:freq_irrigation            2    4.5    2.24   0.144 0.8665
## nitrogen:cultivar                   4    9.6    2.40   0.154 0.9606
## freq_irrigation:cultivar            8   34.1    4.26   0.273 0.9723
## nitrogen:freq_irrigation:cultivar   8   20.2    2.52   0.162 0.9950
## Residuals                          60  936.3   15.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's look to the residuals:

```
par(mfrow = c(2, 2))
plot(fit, which = 5)
plot(fit, which = 1)
plot(fit, which = 2)
plot(residuals(fit) ~ rownames(exp1), main = 'Residuals vs Exp. Unit',
     font.main = 1, data = exp1, xlab = 'Experimental unit', ylab = 'Residuals')
abline(h = 0, lty = 2)
```



We have homogeneous of variance between nitrogen doses, the residuals look normal, and there's a horizontal pattern across the experimental units.

b) Is there a variation in bacteria/fungus growth with different agar types (semi-solid, gel-like state)? I have 30 plates (experimental unit) available.

If we have 30 experimental units and two levels, so the rep is 2.

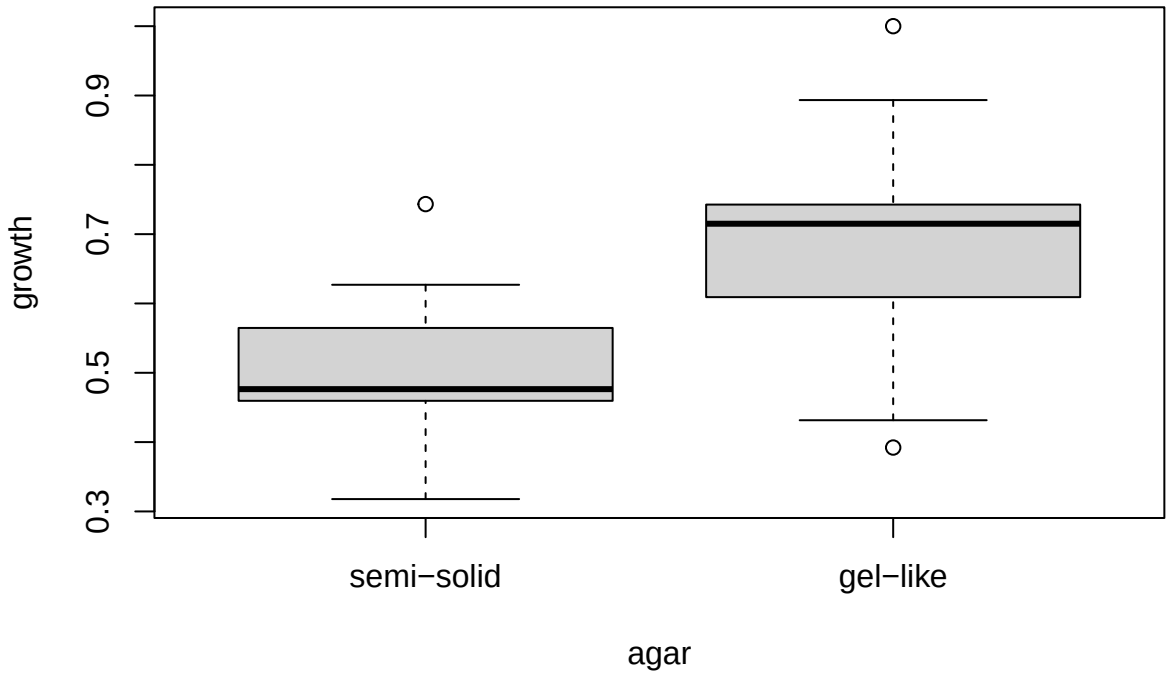
```
set.seed(2023)
f <- c(rep('semi-solid', 15), rep('gel-like', 15))
sampled <- sample(f)
eu <- 1:length(sampled)
exp2 <- data.frame(eu = eu, agar = sampled, growth = rnorm(30, 5, 1))
exp2$agar <- factor(exp2$agar, levels = c('semi-solid', 'gel-like'))  # change level order

# gel-like agar favors yield
exp2[exp2$agar == 'gel-like', 'growth'] <- (
```

```
  exp2[exp2$agar == 'gel-like', 'growth'] + rnorm(15, 1, 1)
)

# normalize growth to the interval [0, 1]
exp2$growth <- exp2$growth / max(exp2$growth)
str(exp2)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ eu    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ agar  : Factor w/ 2 levels "semi-solid","gel-like": 2 2 1 1 1 2 1 1 2 1 ...
##  $ growth: num  0.729 0.739 0.463 0.474 0.627 ...
```

Let's do the analysis.

```
boxplot(growth ~ agar, data = exp2)
```



The bacteria growth for gel-like agar is larger than for semi-solid.
Fitting an one-way ANOVA:

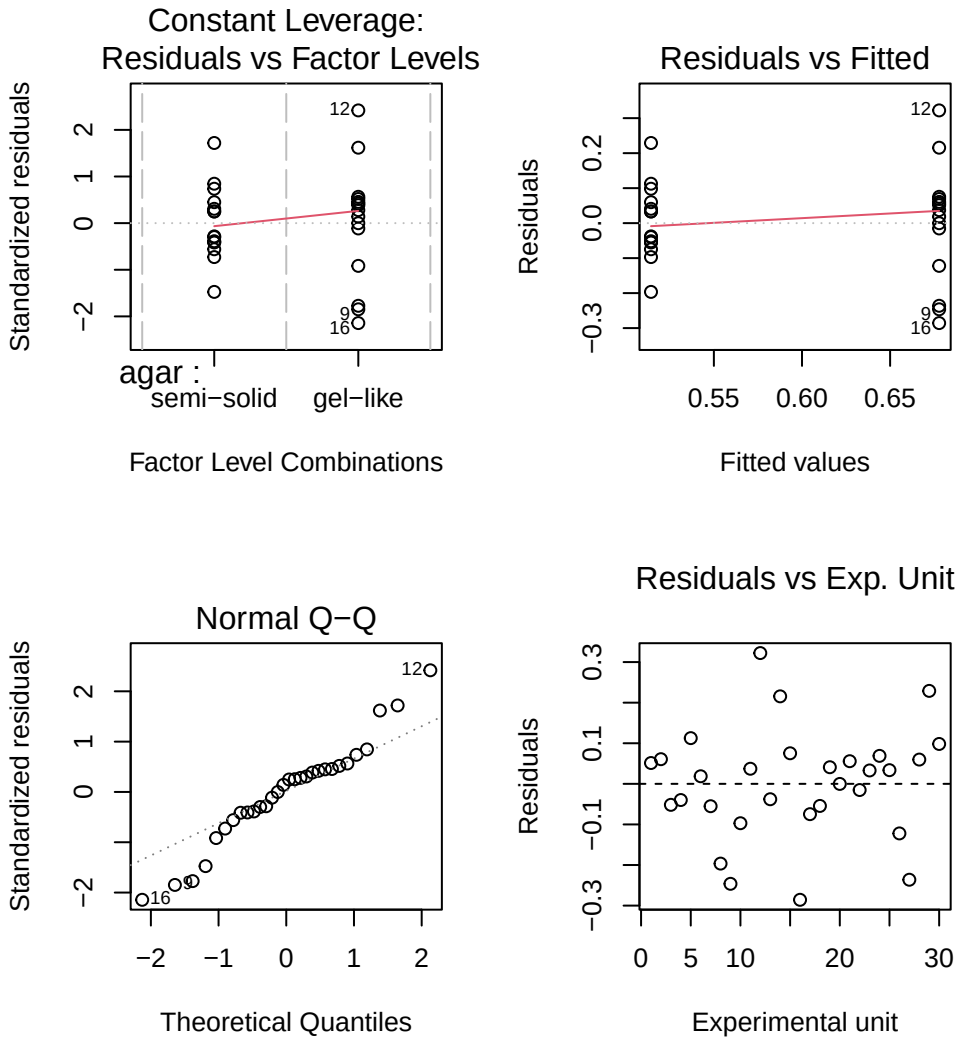
```
fit <- aov(growth ~ agar, data = exp2)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## agar         1 0.2005 0.20051   10.54 0.00303 **
```

```
## Residuals   28 0.5327 0.01902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The agar type affects the bacteria growth, using an significance level of $\alpha = 0.05$, because the p-value $< \alpha$. Residuals:

```
par(mfrow = c(2, 2))
plot(fit, which = 5)
plot(fit, which = 1)
plot(fit, which = 2)
plot(residuals(fit) ~ rownames(exp2), main = 'Residuals vs Exp. Unit',
     font.main = 1, data = exp2, xlab = 'Experimental unit', ylab = 'Residuals')
abline(h = 0, lty = 2)
```



The variances are homogeneous between different agar types, the residuals seems to be normally distributed (although there are some deviance in the tails), and there is a horizontal pattern across the experimental units.

```
shapiro.test(residuals(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.96158, p-value = 0.3397
```

The Shapiro-Wilk test confirms the residuals are normally distributed, because the p-value $> \alpha$, using a significance level of $\alpha = 0.05$, which means we do not reject the null hypothesis that the residuals are normally distributed.

3.The data set `mcconway.turnip` from the package `agridat` presents us with an RCBD experiment of turnips with 16 treatments allocated at random to each of four blocks The 16 treatments were combinations of two varieties, two planting dates, and four densities.

a) Run anova as usual. Are the requirements met?

```
q3 <- agridat::mcconway.turnip
q3[c('gen', 'date', 'density')] <- lapply(q3[c('gen', 'date', 'density')], as.factor)
str(q3)
```

```
## 'data.frame':    64 obs. of  5 variables:
##  $ gen    : Factor w/ 2 levels "Barkant","Marco": 1 1 1 1 1 1 1 1 1 1 ...
##  $ date   : Factor w/ 2 levels "21Aug1990","28Aug1990": 1 1 1 1 1 1 1 1 1 1 ...
##  $ density: Factor w/ 4 levels "1","2","4","8": 1 1 1 1 2 2 2 2 3 3 ...
##  $ block  : Factor w/ 4 levels "B1","B2","B3",..: 1 2 3 4 1 2 3 4 1 2 ...
##  $ yield  : num  2.7 1.4 1.2 3.8 7.3 3.8 3 1.2 6.5 4.6 ...
```

```
# checking frequency tables
with(q3, table(gen, date))
```

```
##          date
## gen       21Aug1990 28Aug1990
##   Barkant        16        16
##   Marco          16        16
```

```
with(q3, table(gen, density))
```

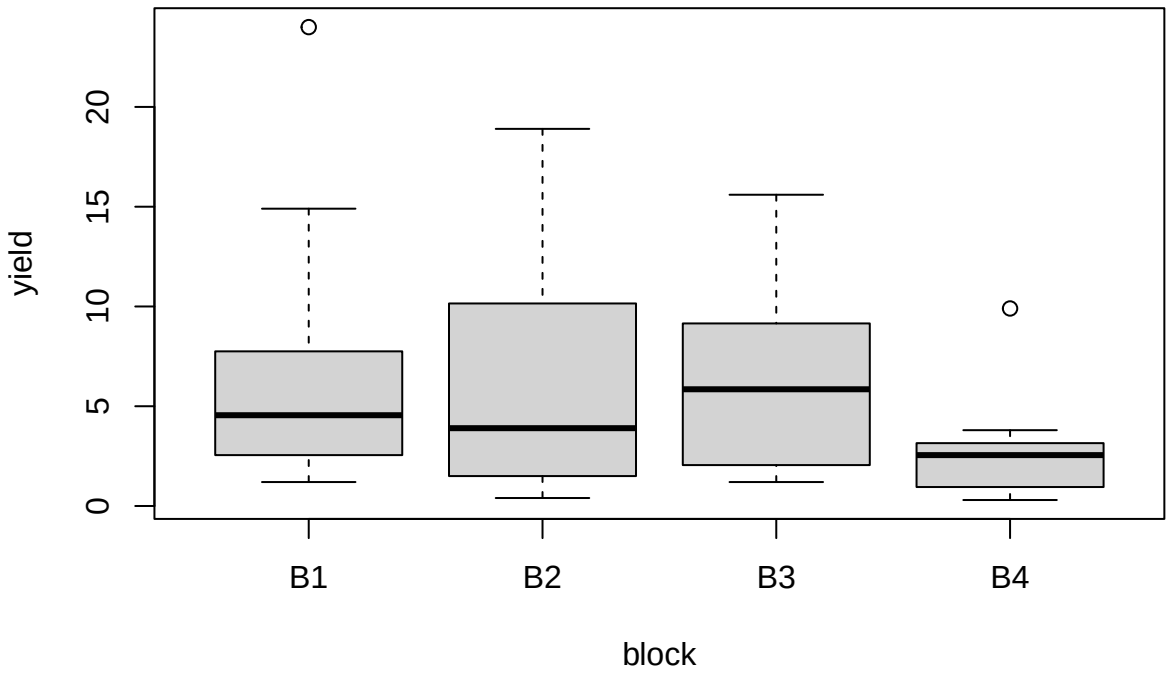
```
##          density
## gen       1 2 4 8
##   Barkant 8 8 8 8
##   Marco   8 8 8 8
```

```
with(q3, table(date, density))
```

```
##            density
## date        1 2 4 8
##   21Aug1990 8 8 8 8
##   28Aug1990 8 8 8 8
```

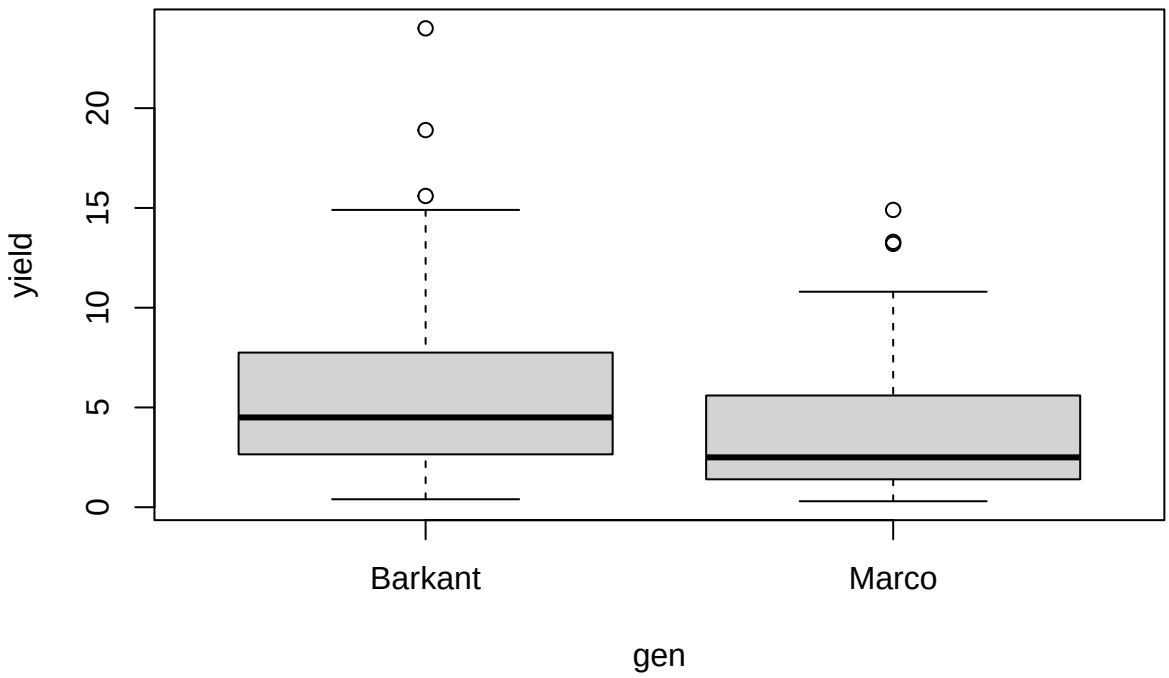
The data is balanced.

```
boxplot(yield ~ block, data = q3)
```



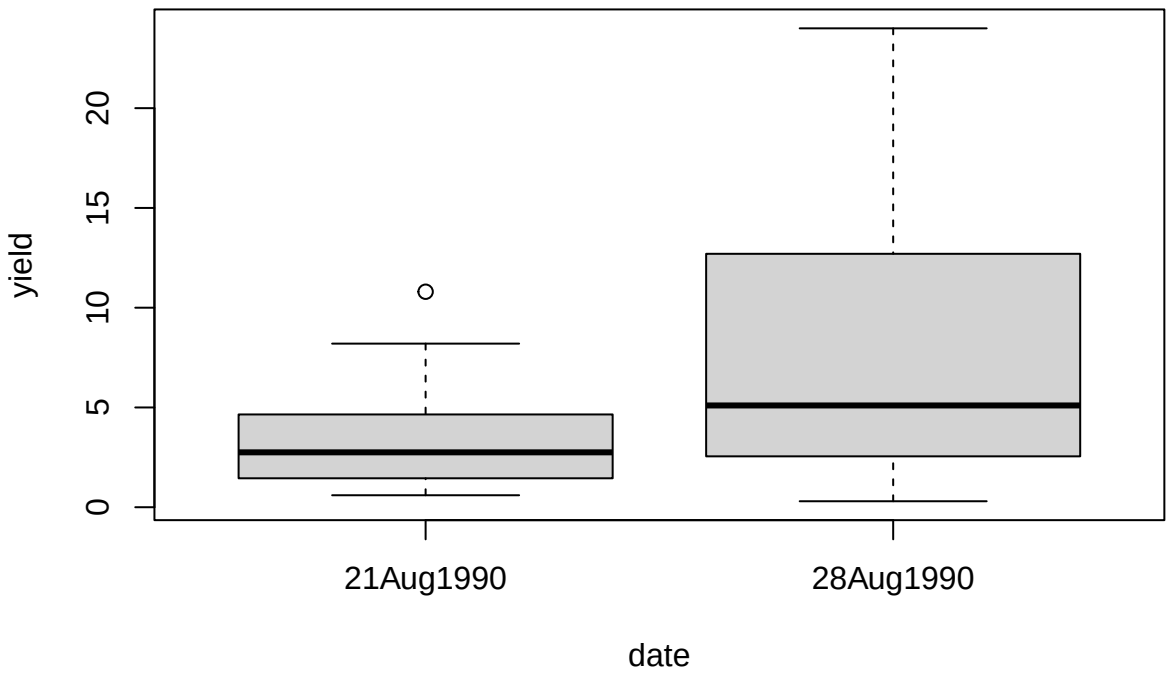
For the block B4, we have very low yield compared to the other ones.

```
boxplot(yield ~ gen, data = q3)
```



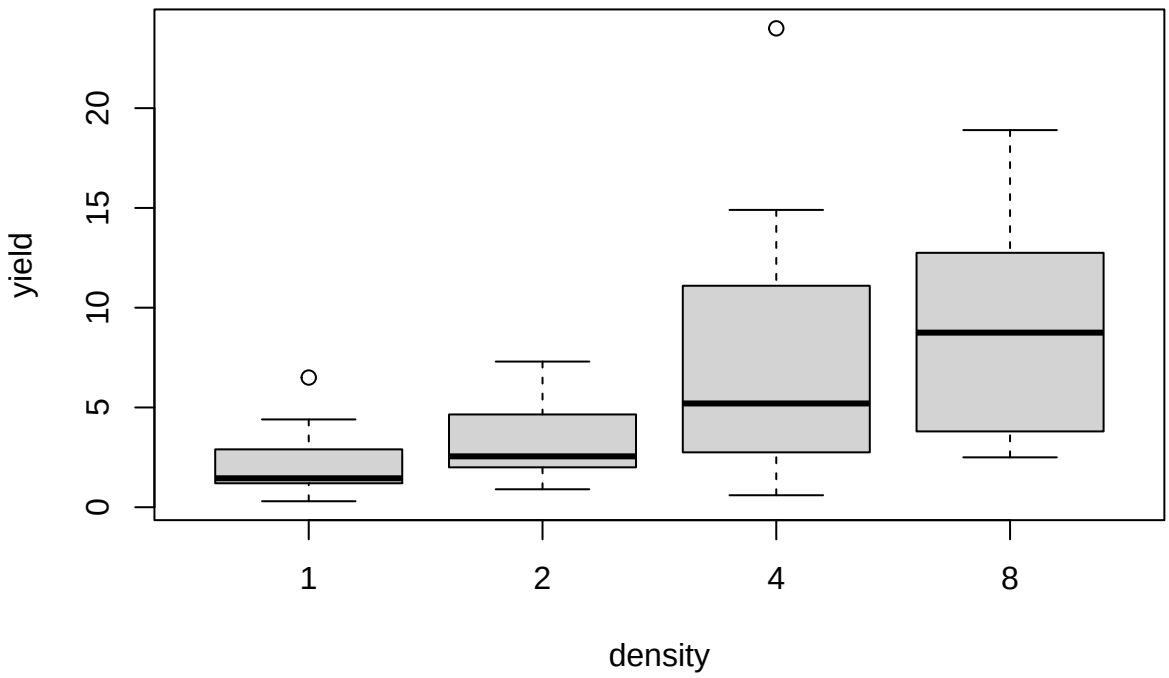
The yield is slightly higher for Barkant.

```
boxplot(yield ~ date, data = q3)
```



The yield was way larger for the 28 Aug 1990.

```
boxplot(yield ~ density, data = q3)
```



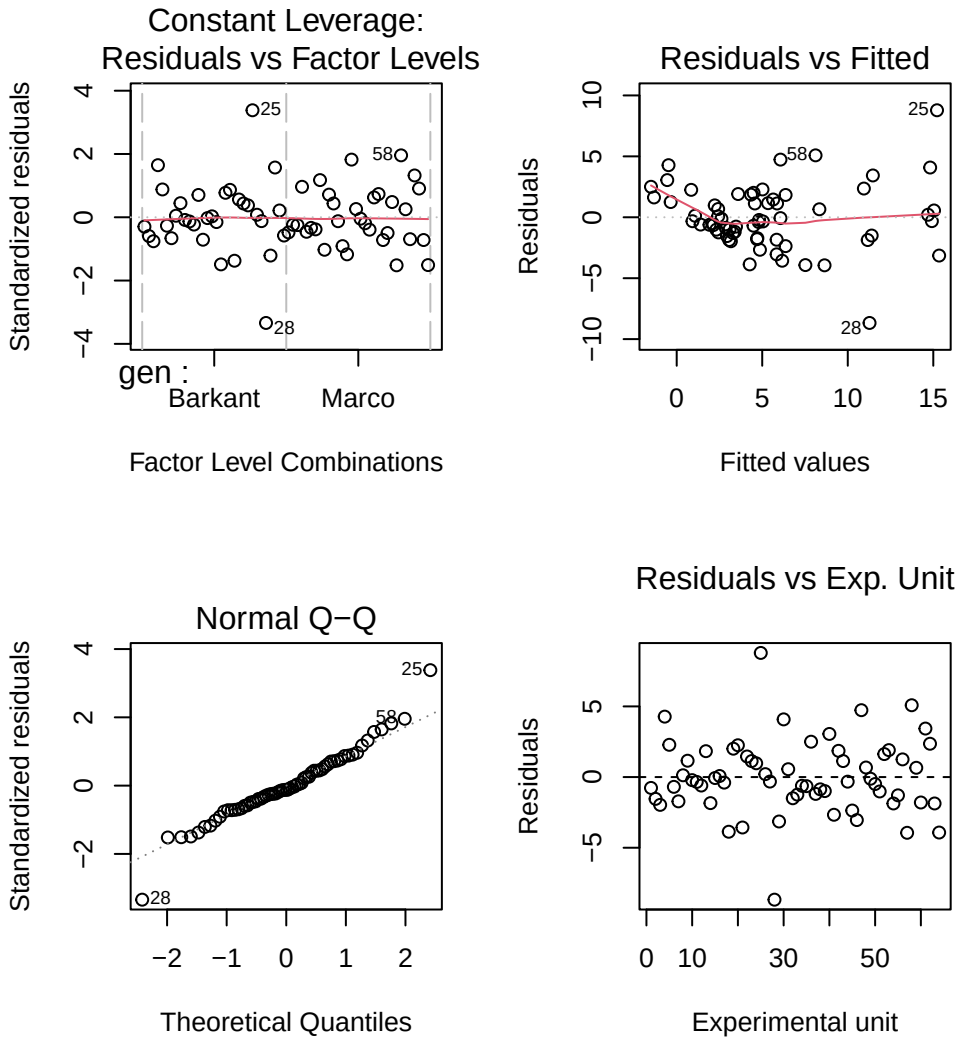
Seems the yield increases as the density increases.

Let's fit an ANOVA:

```
fit <- aov(yield ~ gen * date * density + block, data = q3)
summary(fit)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## gen               1   84.0   83.95   8.753  0.00491 **
## date              1  233.7  233.71  24.367 1.14e-05 ***
## density           3  470.4  156.79  16.347 2.51e-07 ***
## block             3  163.7   54.58   5.690  0.00216 **
## gen:date          1   36.5   36.45   3.800  0.05749 .
## gen:density       3    8.6    2.88   0.301  0.82485
## date:density      3  154.8   51.60   5.380  0.00299 **
## gen:date:density  3   18.0    6.00   0.626  0.60224
## Residuals        45  431.6    9.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(2, 2))
plot(fit, which = 5)
plot(fit, which = 1)
plot(fit, which = 2)
plot(residuals(fit) ~ rownames(q3), main = 'Residuals vs Exp. Unit',
     font.main = 1, data = q3, xlab = 'Experimental unit', ylab = 'Residuals')
abline(h = 0, lty = 2)
```



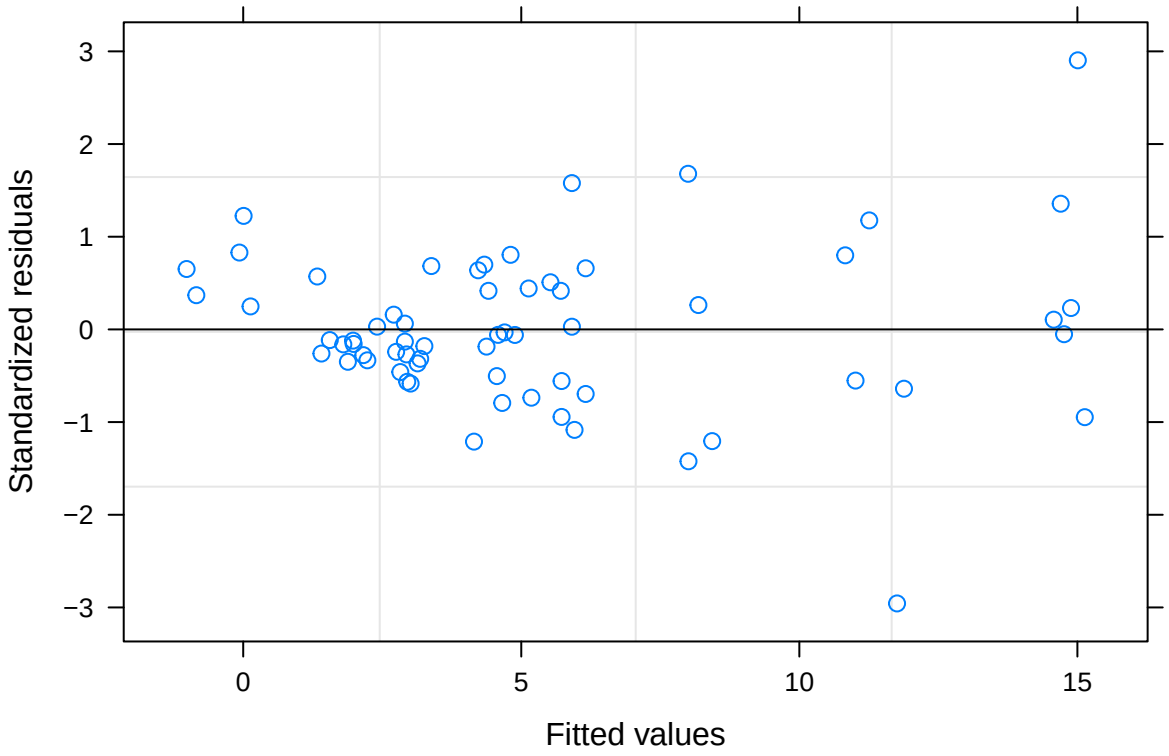

```
shapiro.test(residuals(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.96283, p-value = 0.05116
```

Some residuals points has deviations from the QQ line, but from the Shapiro-Wilk test we do not reject the null hypothesis that the residuals are normally distributed using a significance level of $\alpha = 0.05$.
The residuals seems to have homogeneous variance between the varieties, and the residuals show horizontal pattern across the experimental units.
One problem is that seems the residual's variance increases as the fitted values increase (funnel shape pattern), so the ANOVA assumptions were not met, i.e., there's a non linear relationship between fitted values and residuals.

b) Run a mixed model considering block as a random term.

```
library(nlme)

mod1.nlme <- lme(
  fixed = yield ~ gen * date * density,
  random = ~1 | block,  # block as random
  weights = NULL,  # homoscedastic errors
  data = q3
)
plot(mod1.nlme)
```



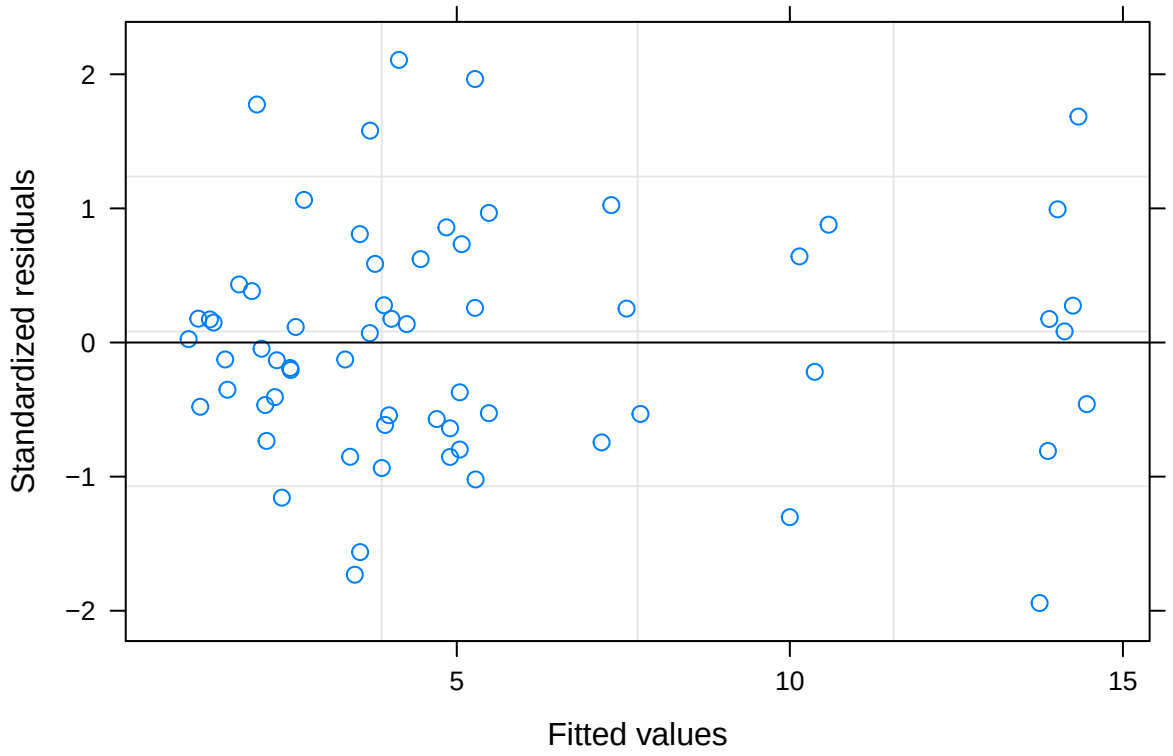
We still have a funnel shape for the Residuals VS Fitted values plot.

c) If there is any issue with the data that result in not meeting the ANOVA assumptions, use a mixed model to solve it.

We saw that yield distribution is very different between `date` and `density` levels, so we could use a heterogeneous variances for both factors.

```
# heterogeneous variance for date and density
# when "form" includes a grouping factor with M > 1 levels, the variance function allows
# M different variances, one for each level of the factor
mod2.nlme <- update(
  mod1.nlme,
```

```
  weights = varComb(varIdent(form =  ~1 | date), varIdent(form = ~1 | density))
)
plot(mod2.nlme)
```



Now the Residuals vs Fitted values is way better, showing a horizontal pattern.

```
shapiro.test(
  residuals(mod2.nlme, type = 'pearson')  # standardized residuals
)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod2.nlme, type = "pearson")
## W = 0.98326, p-value = 0.5365
```