Homework 7 - AGST 5014

Igor Kuivjogi Fernandes and Ashmita Upadhyay

2023-04-12

1. A soil scientist conducted an experiment to evaluate the effects of soil compaction and soil moisture on the activity of soil microbes. Reduced levels of microbe activity will occur in poorly aerated soils. The aeration levels can be restricted in highly saturated or compacted soils. Treated soil samples were placed in airtight containers and incubated under conditions conducive to microbial activity. The microbe activity in each soil sample was measured as the percent increase in CO2 produced above atmospheric levels. The treatment design was a 3x3 factorial with three levels of soil compaction (bulk density = mg soil/m^3) and three levels of soil moisture (kg water/kg soil). There were two replicate soil container units prepared for each treatment. The CO2 evolution/kg soil/day was recorded on three successive days. The data for each soil container unit are shown below:

```
library(tidyverse)
library(nlme)
```

```
q1 <- read.csv('HW7_Q1.csv')
q1 <- transform(
  q1, Density = factor(Density), Moisture = factor(Moisture), Unit = factor(Unit)
)
str(q1)
```

```
## 'data.frame':    18 obs. of  6 variables:
##  $ Density : Factor w/ 3 levels "1.1","1.4","1.6": 1 1 1 1 1 1 2 2 2 2 ...
##  $ Moisture: Factor w/ 3 levels "0.1","0.2","0.24": 1 1 2 2 3 3 1 1 2 2 ...
##  $ Unit    : Factor w/ 18 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ day1    : num  2.7 2.9 5.2 3.6 4 4.1 2.6 2.2 4.3 3.9 ...
##  $ day2    : num  0.34 1.57 5.04 3.92 3.47 3.47 1.12 0.78 3.36 2.91 ...
##  $ day3    : num  0.11 1.25 3.7 2.69 3.47 2.46 0.9 0.34 3.02 2.35 ...
```

a) Describe this experiment (treatment and experimental design) and write the statistical model.

This is a Repeated Measures design, because the data for each soil container was collected on three successive days.
We have two treatments:
- soil compaction (density) with three levels (1.1, 1.4, 1.6)
- soil moisture with three levels (0.1, 0.2, 0.24)

Statistical model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \alpha\tau_{ik} + \beta\tau_{jk} + \alpha\beta_{ij} + \varepsilon_{ijk},$$

where y_{ijk} is the response, mu is the grand mean, α_i is the density treatment, with $i = 1, 2, 3$, β_j is the moisture treatment, with \$j = 1, 2, 3, \$$\tau_k$ is the day (time) of measurement, with $k = 1, 2, 3$, $\alpha\tau_{ik}$ is the density interacting with time, $\beta\tau_{jk}$ is the moisture interacting with time, $\alpha\beta_{ij}$ in the interaction between density and moisture, and ε_{ijk} is the error term.

b) Conduct the proper statistical analysis. Are the assumptions for this type of analysis met?

Firstly, we can try a multivariate model and check the assumption of sphericity.

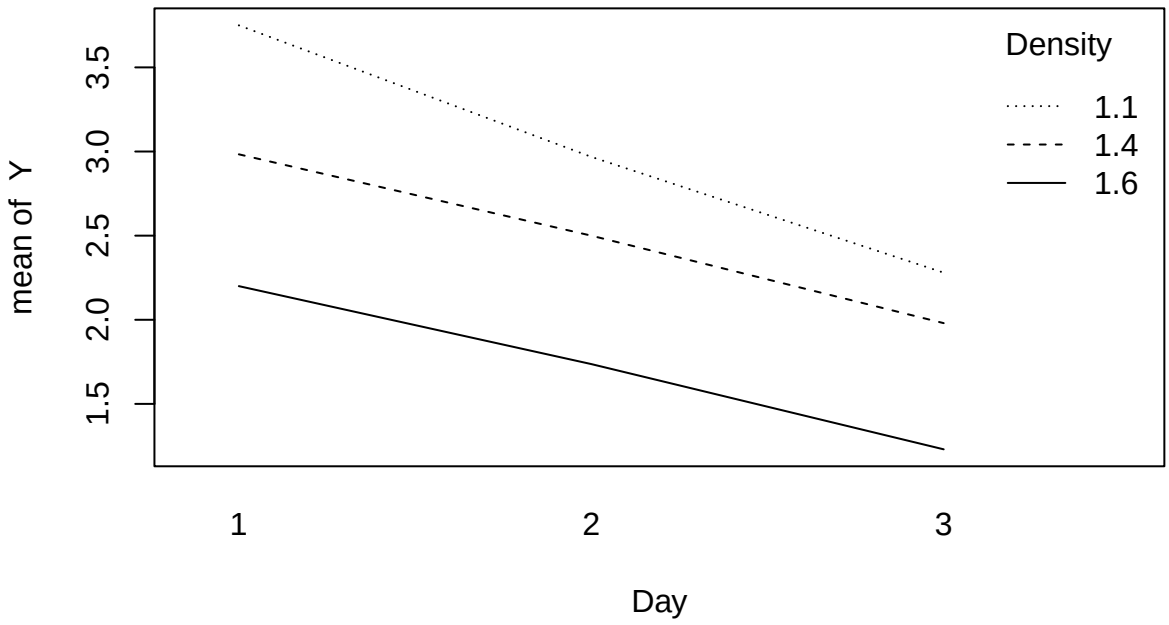
```
mod_multi <- lm(cbind(day1, day2, day3) ~ Density * Moisture, data = q1)
mod <- car::Anova(mod_multi, idata = data.frame(day = factor(1:3)), idesign = ~day)
summary(mod, multivariate = F)$sphericity.tests
```

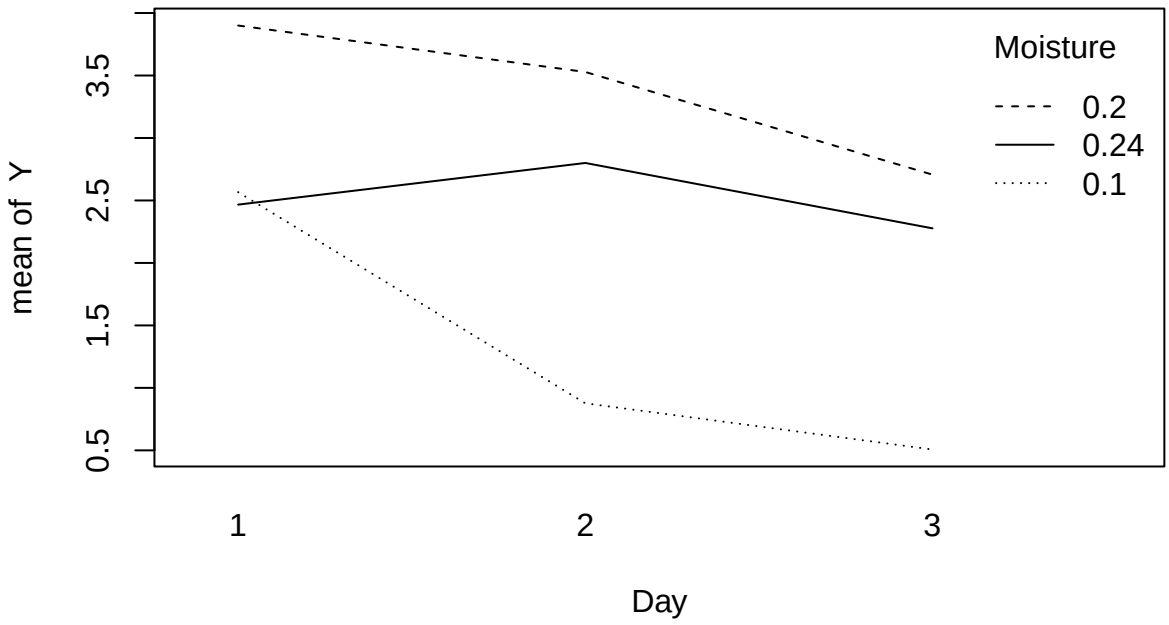
```
##                       Test statistic  p-value
## day                             0.3787 0.020568
## Density:day                     0.3787 0.020568
## Moisture:day                    0.3787 0.020568
## Density:Moisture:day            0.3787 0.020568
```

For Mauchly Tests for Sphericity, we reject the null hypothesis that the covariance matrix of the repeated measures obeys the Huynh-Feldt condition. In this case, we cannot use the split-plot framework and it's better to stick with a model that is able to account for variance-covariance structures.

Let's do some plots for checking the trend of response.

```
# wide to long format
q1_long <- q1 %>%
  pivot_longer(-c(Density, Moisture, Unit), names_to = 'Day', values_to = 'Y') %>%
  mutate(Day = factor(stringr::str_sub(Day, 4)))

par(mfrow = c(2, 1))
with(q1_long, interaction.plot(Day, Density, Y))
with(q1_long, interaction.plot(Day, Moisture, Y))
```





Seems the trend in response for density is quite the same for every density level, whereas for moisture the trends are different.

Let's fit linear models with different variance-covariance structures.

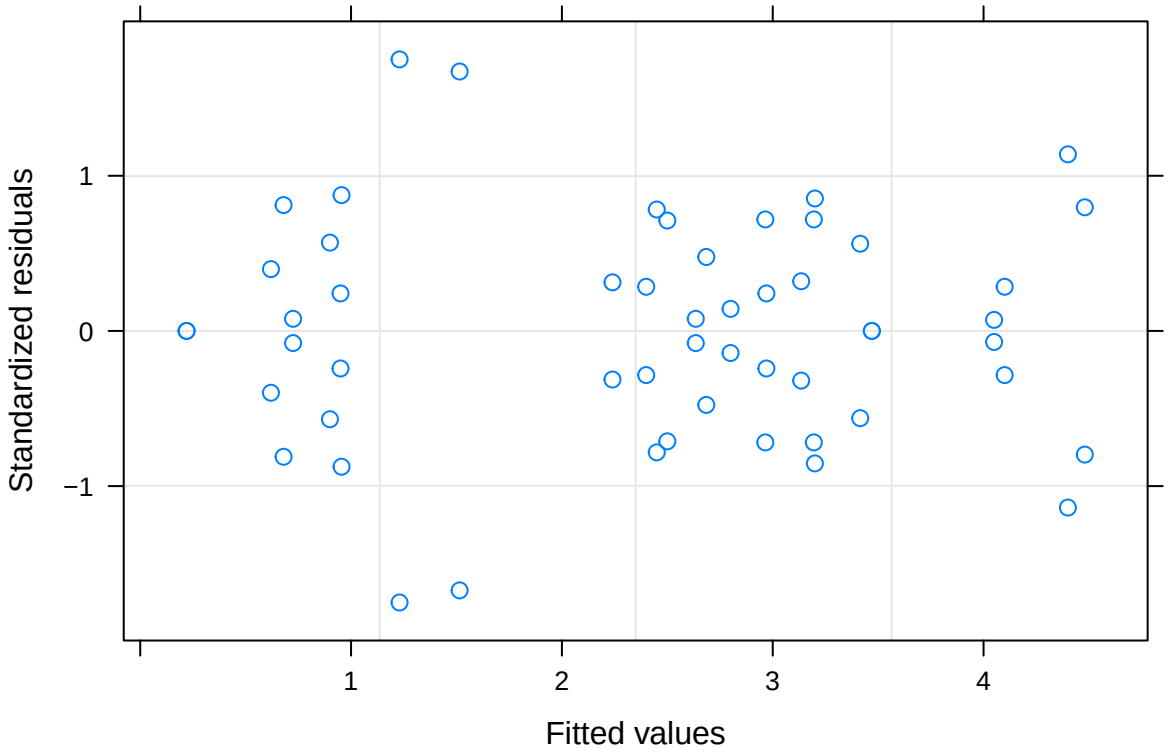
```
# compound symmetry
mod_cs <- gls(
  Y ~ Density * Moisture * Day,
  corr = corCompSymm(form = ~ 1 | Unit),
  data = q1_long,
)

# diagonal
mod_diag <- gls(
  Y ~ Density * Moisture * Day,
  weights = varIdent(form = ~ 1 | Day),
  data = q1_long,
)

# 1st order auto regressive
mod_ar1 <- gls(
  Y ~ Density * Moisture * Day,
  corr = corAR1(form = ~ 1 | Unit),
  data = q1_long,
)

# unstructured
mod_us <- gls(
  Y ~ Density * Moisture * Day,
  corr = corSymm(form = ~ 1 | Unit),
  weights = varIdent(form = ~ 1 | Day),
  data = q1_long,
)

# comparing models
anova(mod_cs, mod_diag, mod_ar1, mod_us)
```

```
##          Model df      AIC      BIC    logLik   Test   L.Ratio p-value
## mod_cs       1 29 122.9437 160.5230 -32.47186
## mod_diag     2 30 136.2583 175.1334 -38.12913 1 vs 2 11.314534  0.0008
## mod_ar1      3 29 118.8532 156.4325 -30.42662 2 vs 3 15.405018  0.0001
## mod_us       4 33 120.9361 163.6987 -27.46807 3 vs 4  5.917098  0.2054
```

The AIC and BIC tell us that the model 3 (1st order auto regressive) is the best model. The LR tests also shows us that the model 3 is better than model 2 and 4. Hence, we chose model 3.

Let's check some assumptions:

```
plot(mod_ar1)
```



Seems we have homogeneous variance across fitted values.

```
shapiro.test(residuals(mod_ar1, type = 'pearson'))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod_ar1, type = "pearson")
## W = 0.9853, p-value = 0.7455
```

We also don't reject the hypothesis that the residuals are normally distributed, using a significance level of $\alpha = 0.05$.

```
anova(mod_ar1)
```

```
## Denom. DF: 27
##                       numDF   F-value p-value
## (Intercept)               1 268.32804  <.0001
## Density                   2   6.54128  0.0048
## Moisture                  2  13.41938  0.0001
## Day                       2  28.29815  <.0001
## Density:Moisture          4   1.41012  0.2573
## Density:Day               4   0.59290  0.6707
## Moisture:Day              4  13.79650  <.0001
## Density:Moisture:Day      8   2.22168  0.0578
```

Both `Density` and `Moisture` treatments are significant, as well the interaction `Moisture:Day`. The interactions `Density:Moisture`, `Density:Day` and `Density:Moisture:Day` are not significant, all using a significance level of $\alpha = 0.05$.

2. An agronomist conducted a yield trial with five alfalfa cultivars in a randomized complete block design with three replications. Each plot was harvested four times in each of two years. The plot yield (lb/plot) from two harvests from each plot in each of two years are shown in the table below. Compare the yield of the cultivars (notice that you need adjusted means for that).

```
q2 <- read.csv('HW7_Q2.csv')
q2 <- transform(q2, Cultivar = factor(Cultivar), Block = factor(Block))
str(q2)
```

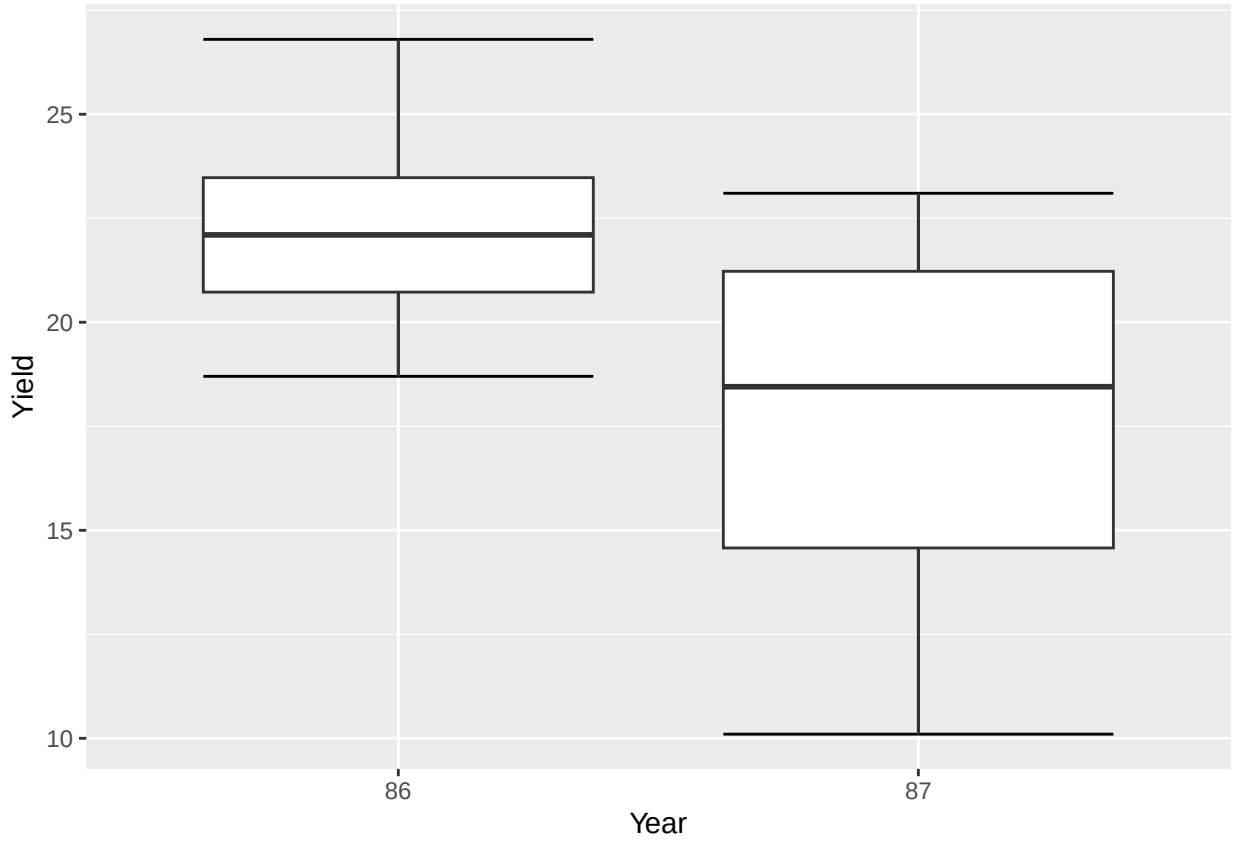
```
## 'data.frame':    15 obs. of  6 variables:
##  $ Cultivar: Factor w/ 5 levels "1","2","3","4",..: 1 1 1 2 2 2 3 3 3 4 ...
##  $ Block   : Factor w/ 3 levels "1","2","3": 1 2 3 1 2 3 1 2 3 1 ...
##  $ Apr.86  : num  20.4 21.5 21.1 19.1 20.8 20.5 19.3 19.8 20.5 23.2 ...
##  $ May.86  : num  23.2 23.7 23.4 22.4 22.1 23.5 22.1 25.4 24.8 25.6 ...
##  $ Apr.87  : num  14.8 18.8 14.3 14.5 10.1 12 14.5 16.9 16.7 14.9 ...
##  $ May.87  : num  22.9 22.6 22.1 19.2 22 21.5 19.5 23.1 20.1 19.5 ...
```

```
# pivot to long with 2 measurements
q2_long <- q2 %>%
  pivot_longer(Apr.86:May.87, names_to = 'harvesttime', values_to = 'Yield') %>%
  separate_wider_delim(harvesttime, '.', names = c('Month', 'Year')) %>%
  arrange(Year, Cultivar, Block) %>%
  mutate(Unit = factor(rep(1:30, each = 2)),
         Cultivar = factor(Cultivar),
         Block = factor(Block),
         Month = factor(Month),
         Year = factor(Year))

str(q2_long)
```

```
## tibble [60 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Cultivar: Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ Block   : Factor w/ 3 levels "1","2","3": 1 1 2 2 3 3 1 1 2 2 ...
##  $ Month   : Factor w/ 2 levels "Apr","May": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Year    : Factor w/ 2 levels "86","87": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Yield   : num [1:60] 20.4 23.2 21.5 23.7 21.1 23.4 19.1 22.4 20.8 22.1 ...
##  $ Unit    : Factor w/ 30 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
```

We split the year as a factor and used only two measurements (April and May).
As we have only two measurements the compound symmetry condition is met, but we still need to account for the possible heterogeneity of variances in the year.

```
ggplot(q2_long, aes(x = Year, y = Yield)) +
  stat_boxplot(geom = 'errorbar') +
  geom_boxplot()
```



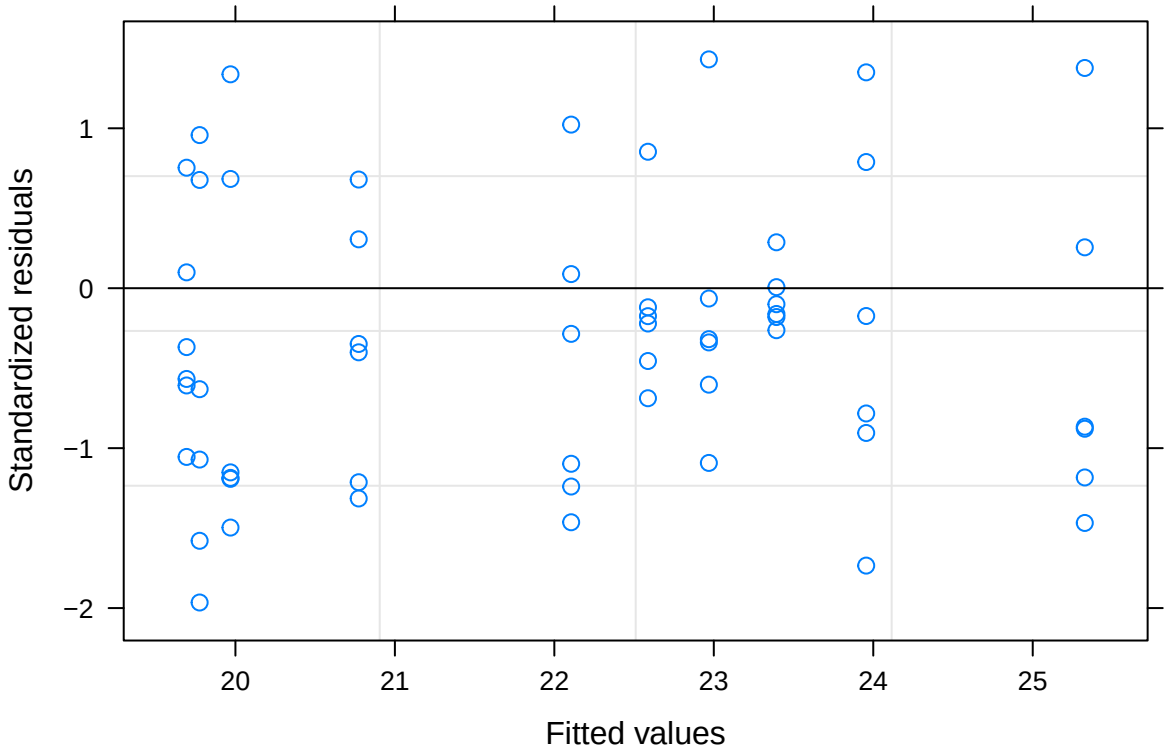
In general, the yield in 86 was higher than in 87.

Let's do an inter-block analysis, where block is random.
We also put a diagonal matrix for `Year` to account for heterogeneous variances.

```
mm <- lme(
  fixed = Yield ~ Cultivar * Month,
  random = ~ 1 | Block,
  weights = varIdent(form = ~ 1 | Year),
  data = q2_long
)
```

Let's check some assumptions:

```
plot(mm)
```



We have homogeneous variance across fitted values.

```
shapiro.test(residuals(mm, type = 'pearson'))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mm, type = "pearson")
## W = 0.96741, p-value = 0.1089
```

We don't reject the null hypothesis that the residuals are normally distributed, using a significance level of $\alpha = 0.05$.

```
anova(mm)
```

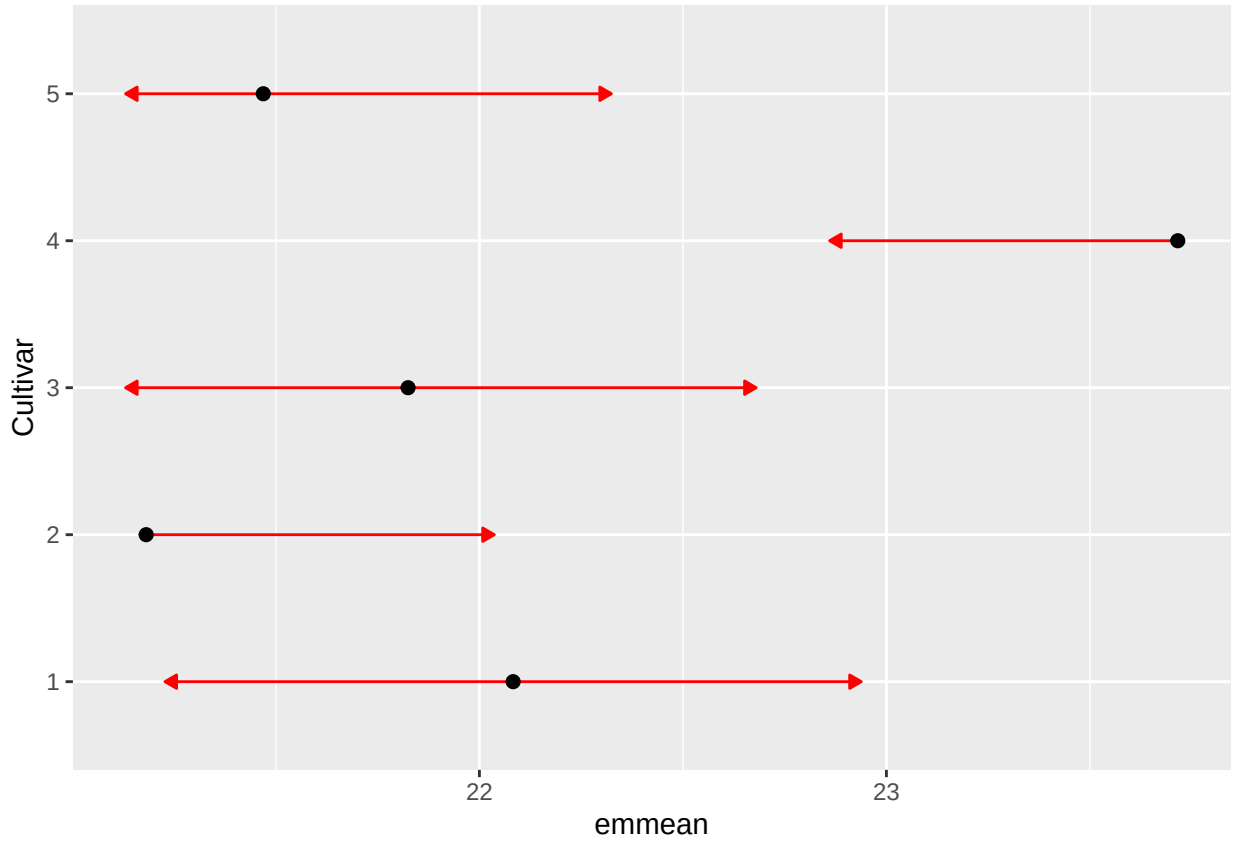
```
##                numDF denDF   F-value p-value
## (Intercept)        1    48 13347.640  <.0001
## Cultivar           4    48     5.377  0.0012
## Month              1    48    69.496  <.0001
## Cultivar:Month     4    48     0.567  0.6875
```

The `Cultivar` is significant, using a significance level of $\alpha = 0.05$. Which one performs the best?

```
emm <- emmeans::emmeans(mm, pairwise ~ Cultivar)
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```
plot(emm, comparisons = T, CI = F)
```



The `Cultivar` 4 performs the best, although it's slightly overlapping with `Cultivar` 1 (i.e. the difference between 4 and 1 is not significant).

```
as_tibble(emm$contrasts)[3, ]
```

```
## # A tibble: 1 x 6
##   contrast              estimate    SE    df t.ratio p.value
##   <chr>                    <dbl> <dbl> <dbl>   <dbl>   <dbl>
## 1 Cultivar1 - Cultivar4    -1.63 0.604    48   -2.71  0.0679
```

Actually, the difference would be significant using an $\alpha = 0.10$ because p-value for the contrast `Cultivar1 - Cultivar4` is 0.0679.

3. One experiment was set up to assess possible phytotoxicity effects relating to an excessive persistence of herbicide residues in soil. The three crops were sown 40 days after a herbicide treatment (a check was included as the second herbicide treatment).

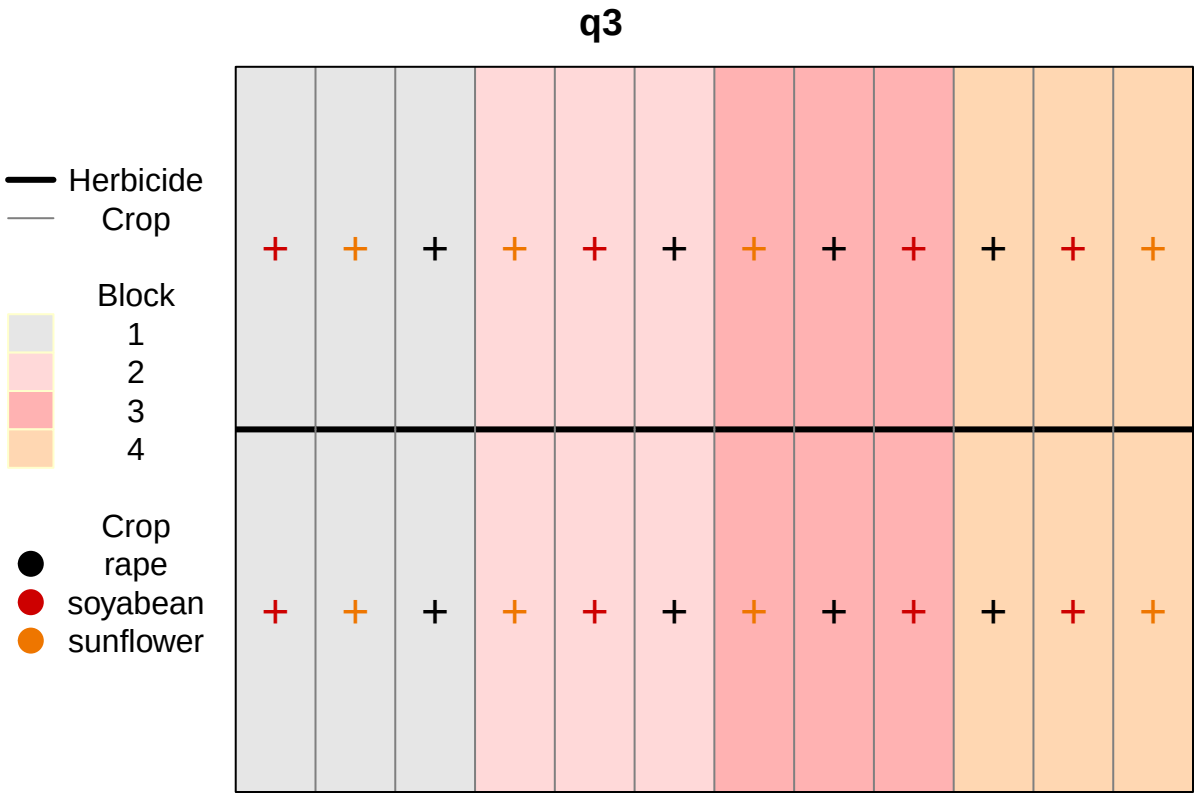
```r
q3 <- read.csv('HW7_Q3.csv')
q3 <- transform(q3,
  Herbicide = as.factor(Herbicide),
  Crop = as.factor(Crop),
  row = as.factor(row),
  col = as.factor(col),
  Block = as.factor(Block)
)
str(q3)
```

```
## 'data.frame':    24 obs. of  6 variables:
##  $ Herbicide  : Factor w/ 2 levels "Check","rimsulfuron": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Crop       : Factor w/ 3 levels "rape","soyabean",..: 2 2 3 3 1 1 2 2 3 3 ...
##  $ Block      : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 2 2 2 2 ...
##  $ CropBiomass: num  199 226 201 290 157 ...
##  $ col        : Factor w/ 2 levels "1","2": 1 2 1 2 1 2 1 2 1 2 ...
##  $ row        : Factor w/ 12 levels "1","2","3","4",..: 1 1 2 2 3 3 5 5 4 4 ...
```

a) Describe this experiment and write the statistical model.

Let's check the field plot:

```r
desplot::desplot(
  q3,
  Block ~ row + col,
  col = Crop,
  cex = 1.5,
  out1 = Herbicide,
  out2 = Crop,
  out2.gpar = list(col = 'gray50', lwd = 1, lty = 1),
)
```



This is a strip plot design where two levels of herbicide were applied horizontally and after that the three crops were randomized vertically within blocks. Both treatments (herbicide and crop) are the whole-plots, whereas the interaction between herbicide and crop is the sub-plot, which means we have three error terms.

Statistical model:

$$y_{ijk} = \mu + b_i + \alpha_j + \beta_k + b\alpha_{ij} + b\beta_{ik} + \alpha\beta_{jk} + \varepsilon_{ijk},$$

where y_{ijk} is the crop biomass, μ is the grand mean, b_i is the random block, with $i = 1, 2, 3, 4$, α_j is the herbicide treatment (main-plot), which $j = 1, 2$, β_k is crop treatment (main-plot), with $k = 1, 2, 3$, $\alpha\beta_{jk}$ is the interaction between herbicide and crop (sub-plot), the block interactions with each treatment are the error terms, and ε is the sub-plot error term.

b) Run the proper analyze on experiment and proceed with interpretations.

```
mod_strip <- aov(CropBiomass ~ Herbicide * Crop + Error(Block/(Herbicide * Crop)), data = q3)
summary(mod_strip)
```

```
##
## Error: Block
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals  3  70739   23580
##
## Error: Block:Herbicide
##           Df Sum Sq Mean Sq F value Pr(>F)
## Herbicide  1    346     346   0.052  0.834
```

```
## Residuals  3  20000    6667
##
## Error: Block:Crop
##           Df Sum Sq Mean Sq F value Pr(>F)
## Crop       2  85978   42989   7.693 0.0221 *
## Residuals  6  33527    5588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Block:Herbicide:Crop
##                Df Sum Sq Mean Sq F value Pr(>F)
## Herbicide:Crop  2  12549    6274     2.2  0.192
## Residuals       6  17109    2851
```

Using a significance level of $\alpha = 0.05$, we can see that the `Crop` is significant, but `Herbicide` and the interaction between `Herbicide` and `Crop` are not.

We can also compare the cultivars' performances:

```
emmeans::emmeans(mod_strip, pairwise ~ Crop)$contrasts
```

```
## Note: re-fitting model with sum-to-zero contrasts
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```
##  contrast             estimate   SE df t.ratio p.value
##  rape - soyabean         -33.6 37.4  6  -0.898  0.6610
##  rape - sunflower       -140.4 37.4  6  -3.756  0.0221
##  soyabean - sunflower   -106.8 37.4  6  -2.858  0.0649
##
## Results are averaged over the levels of: Herbicide
## P value adjustment: tukey method for comparing a family of 3 estimates
```

The sunflower was better than rape, but sunflower is not different from soybean, using a significance level of $\alpha = 0.05$.