Homework 7 - AGST 5014

Igor Kuivjogi Fernandes and Ashmita Upadhyay

2023-04-08

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:

```
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:

$$y = \mu + \alpha + \beta + \tau + \alpha\tau + \beta\tau + \alpha\beta + \alpha\beta\tau + \varepsilon,$$

where y is the response, α is the density treatment, β is the moisture treatment, τ is the day (time) of measurement, $\alpha\tau$ is the density interacting with time, $\beta\tau$ is the moisture interacting with time, $\alpha\beta$ in the interaction between density and moisture, $\alpha\beta\tau$ is the interaction between density, moisture and time, and ε 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.

```r
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
```

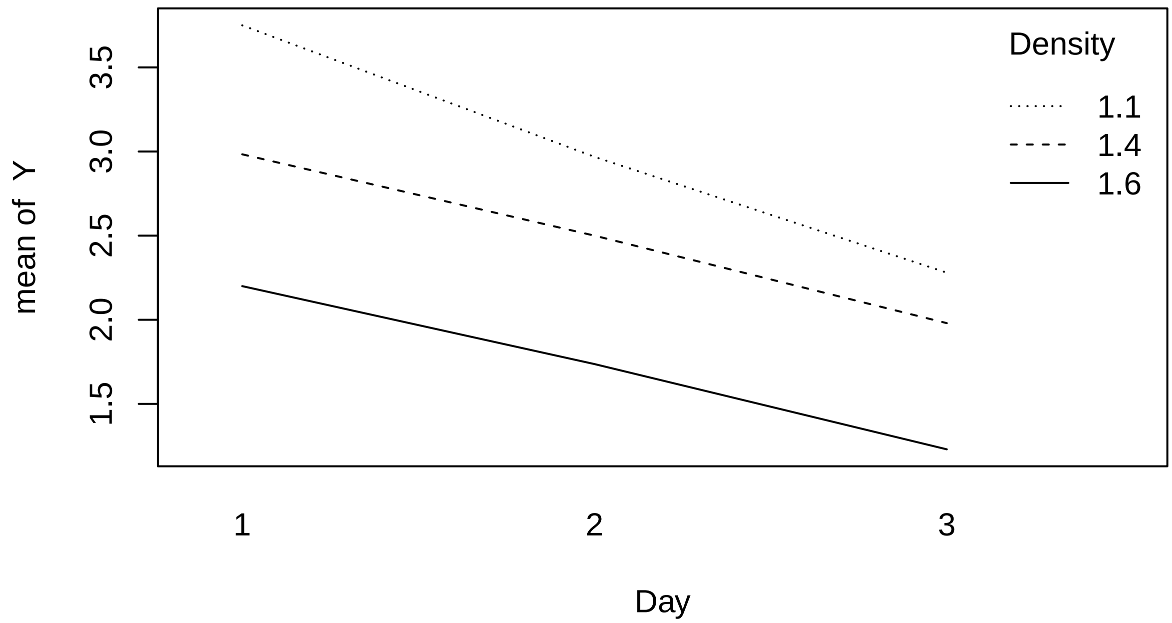
```
##                       Test statistic  p-value
## day                           0.3787 0.020568
## Density:day                   0.3787 0.020568
## Moisture:day                  0.3787 0.020568
## Density:Moisture:day          0.3787 0.020568
```

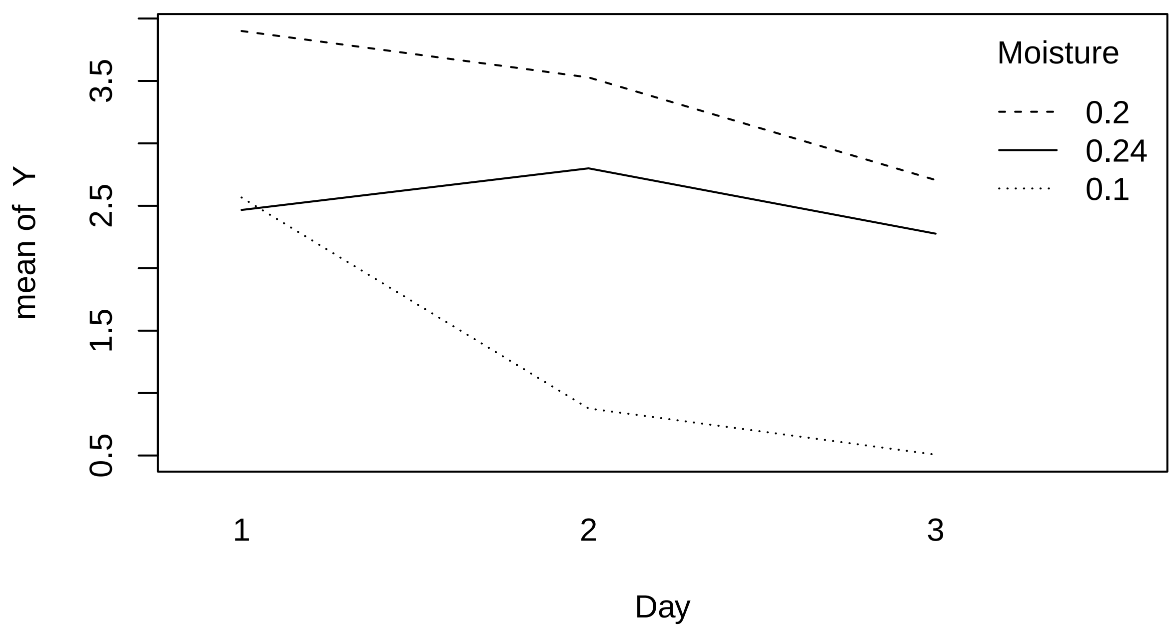
For Mauchly Tests for Sphericity, we reject the null hypothesis that the covariance matrix of the repeated measures obeys the Huynh-Feldt condition. In this case, we cannot use the split-plot framework and it's better to stick with a model that is able to account for variance-covariance structures.

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

```r
# wide to long format
q1_long <- q1 |>
  tidyr::pivot_longer(-c(Density, Moisture, Unit), names_to = 'Day', values_to = 'Y') |>
  dplyr::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.

```
library(nlme)

# identity
mod_id <- gls(
  Y ~ Density * Moisture * Day,
  data = q1_long,
)

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

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

```
##         Model df      AIC      BIC    logLik   Test   L.Ratio p-value
## mod_id      1 28 132.5619 168.8453 -38.28096
## mod_cs      2 29 122.9437 160.5230 -32.47186 1 vs 2 11.618193  0.0007
## mod_ar1     3 29 118.8532 156.4325 -30.42662
## mod_us      4 31 117.3137 157.4846 -27.65685 3 vs 4  5.539533  0.0627
```

The AIC and log likelihood tell us the unstructured model (model 4) is the best, whereas from the AIC the 1st auto regressive model (model 3) is the best. The p-value for testing model 3 vs 4 shows us that we do not reject the hypothesis (using a significance level of $\alpha = 0.05$) that the model 4 (unstructured) is better than the model 3 (1st order auto regressive), so we stick with the 3th model.

```
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).

```r
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 ...
```

```r
# pivot to long with 4 measurements
q2_long <- q2 |>
  tidyr::pivot_longer(-c(Cultivar, Block), names_to = 'Date', values_to = 'Y')

# pivot to long with 2 measurements
# q2_long <- q2 %>%
#   tidyr::pivot_longer(-c(Cultivar, Block), names_to = 'Date', values_to = 'Y') %>%
#   dplyr::mutate(Year = factor(ifelse(Date %in% c('Apr.86', 'May.86'), 86, 87))) %>%
#   dplyr::mutate(Month = stringr::str_sub(Date, end = 3)) %>%
#   dplyr::select(-Date) %>%
#   tidyr::pivot_wider(id_cols = c(Cultivar, Block, Year), names_from = 'Month', values_from = 'Y') %>%
#   dplyr::arrange(Year, Cultivar, Block)
```

We could try include the year as a factor and assume there are only 2 measurements in time, but we can also analyse it with 4 measurements. We stick with the latter.
The 4 measurements are not evenly spaced.
Let's fit some linear models with variance-covariance structures.

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).

a) Describe this experiment and write the statistical model.

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

```r
q3 <- read.csv("HW7_Q3.csv")
```