Homework II

J Lacasa

2026-02-15

Read the exercises below, answer them, and knit the .Rmd file into an HTML of PDF file. Rename it, including your last name (e.g., "Smith_hw2.Rmd") and submit it on CANVAS by Tue, February 24th.

The exercises below aim to (i) review the most important aspects/benefits of multilevel models, (ii) review the analysis of classic designed experiments, and (iii) review the analysis of any data with a clear/"obvious" data architecture.

1 Multi-level models

Consider a data set collected to study the water quality for different agricultural practices (e.g., conservation practices, intensive practices, etc.) across an area covering different watersheds. In each (jth) watershed, water samples were taken to measure nitrate concentrations for a field under the ith agricultural practice. Note that multiple fields (with different agricultural practices) may fit in the same watershed. Then, we have multiple observations y_{ijk} for the ith agricultural practice in the jth watershed and kth field. The main objective of the study is to quantify the water quality for the different agricultural practices. All watersheds are equally important and their size can be assumed similar, but some watersheds have more observations than others, and they also have different proportions of the presence of the agricultural practices.

1.1 Data architecture

Draw a schematic representation of how the architecture in the data looks like and embed a picture of that representation in this document.

1.2 Recovery of inter-group information

Consider the model

$$y_{ijk} = \mu + P_i + w_j + \varepsilon_{ijk}, \varepsilon_{ijk} \sim N(0, \sigma^2),$$

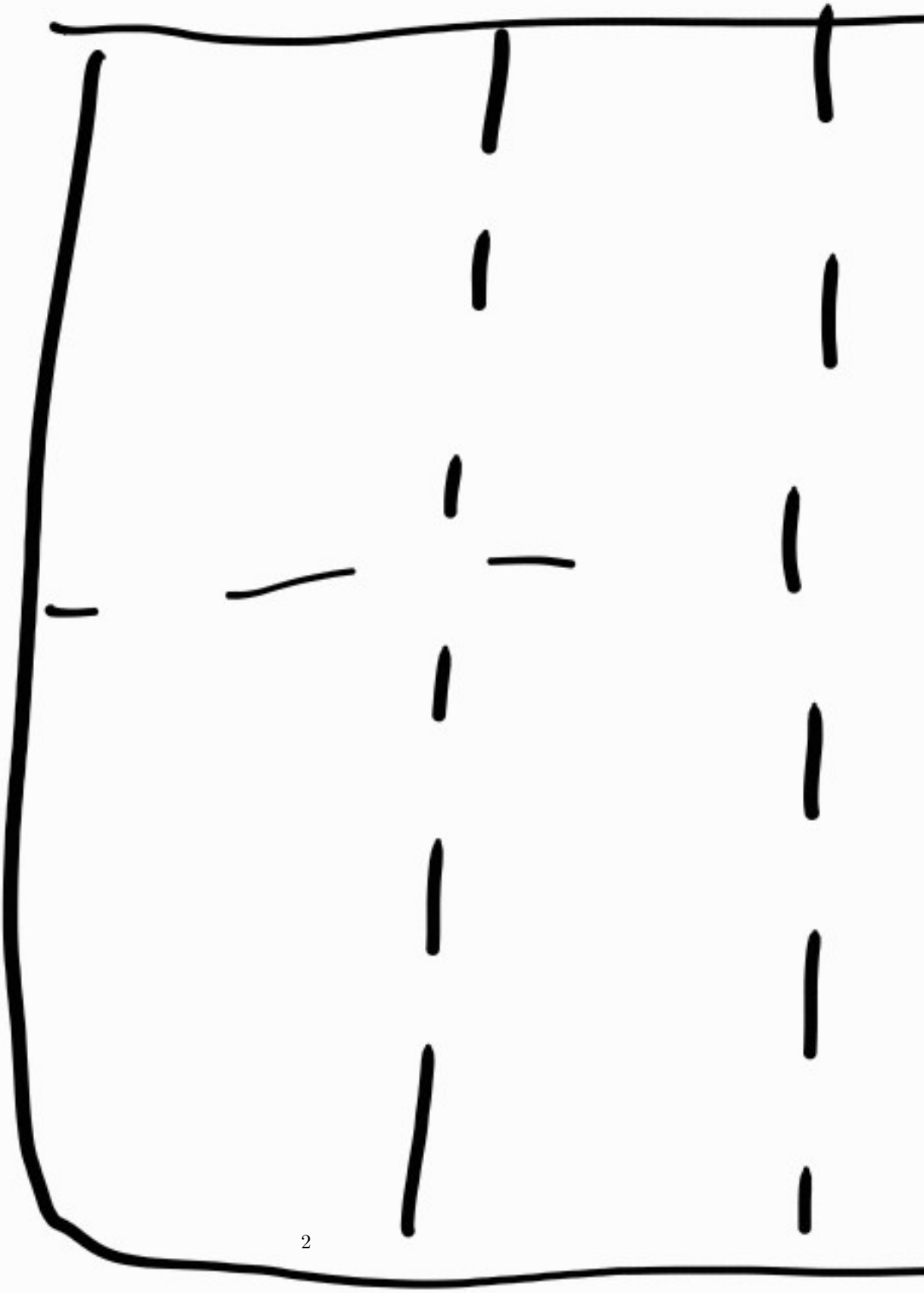
where:

- y_{ijk} is the observation of the ith agricultural practice in the jth watershed, and kth field,
- P_i is the effect of the ith agricultural practice,
- w_j is the effect of the jth watershed, and
- ε_{ijk} is the residual.

Something we learned in this course is that the assumption behind w_j can affect our results and our inference. In this case, how does modeling w_j as a fixed effect versus as a random effect affect inference and results?

Answer:

Modeling w_j as a fixed effect constrains the inference to that individual



2 Analysis of an incomplete block design

The data below were generated by a designed experiment with a one-way treatment structure, where genotype is the treatment factor and has 13 levels, in an incomplete block design.

```
url <- "https://raw.githubusercontent.com/stat799/spring2026/refs/heads/main/data/bibd.csv"
dat <- read.csv(url)
head(dat)
```

```
##   block gen    yield
## 1   B01 G03 26.42798
## 2   B01 G06 25.78419
## 3   B01 G09 30.27279
## 4   B01 G11 21.32960
## 5   B02 G03 19.22134
## 6   B02 G04 28.44798
```

2.1 Model fitting

Write a statistical model to describe the data generating process and fit said model using statistical software.

Justify the different components of the model (e.g., fixed and/or random effects).

Make sure (and show) that the model is reliable, and show why it is reliable.

Answer: The model can be written out as

$$y_{ijk}|\boldsymbol{u} \sim N(\mu_{ijk}, \sigma^2)$$

3 Analysis of a split-plot designed experiment.

The data below were generated for a study aiming to evaluate the individual plant performance (in grams of grain per plant) for different fertilizer sources (8 different sources) and moments of application (4 different sources). Scientists ran a designed experiment with a 8×4 factorial treatment structure, in a split plot in an RCBD. Fertilizer source was the whole-plot treatment factor, and the application moment was the split-plot treatment factor. Note that two plants were observed per plot (indicated as "sample").

```
url <- "https://raw.githubusercontent.com/stat799/spring2026/refs/heads/main/data/splitplot_subs.csv"
dat <- read.csv(url) |>
  mutate(block = as.factor(block),
         fertilizer_source = as.factor(fertilizer_source),
         application_moment = as.factor(application_moment))
str(dat)
```

```
## 'data.frame':    192 obs. of  5 variables:
##  $ block             : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ grams_plant       : num  109 108.8 97.3 108.1 102.5 ...
##  $ fertilizer_source : Factor w/ 8 levels "Fert1","Fert2",..: 1 2 3 4 5 6 7 8 1 2 ...
##  $ application_moment: Factor w/ 4 levels "M1","M2","M3",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ sample            : chr  "S1" "S1" "S1" "S1" ...
```

3.1 ANOVA

Write the ANOVA table (or do the multilevel ANOVA from Gelman (2005) from day 4) including source of variability and degrees of freedom.

Source of variability	df
Block	2
Fertilizer	7
Whole plot error	14
Moment	3
Fertilizer × Moment	21
Split plot error	48
Error	96
Total	191

3.2 Marginal means

Get the marginal means, and explain what sources of uncertainty are contained in the standard errors of the means.

Response:

See model and results below. The variance components that go into the mean s.e. of these marginal means include σ^2_b, σ^2_{wp}, σ^2_{sp}, and σ^2_ε.

```
library(lme4)
library(emmeans)
library(ggpubr)

m <- lmer(grams_plant ~ fertilizer_source*application_moment +
            (1|block/fertilizer_source/application_moment),
          data = dat)

emmeans(m, ~fertilizer_source*application_moment)
```

```
##  fertilizer_source application_moment emmean   SE   df lower.CL upper.CL
##  Fert1             M1                  106.7 3.45 16.1     99.4    114.1
##  Fert2             M1                  105.6 3.45 16.1     98.3    112.9
##  Fert3             M1                  102.5 3.45 16.1     95.2    109.9
##  Fert4             M1                  101.3 3.45 16.1     94.0    108.6
##  Fert5             M1                  104.1 3.45 16.1     96.7    111.4
##  Fert6             M1                  101.7 3.45 16.1     94.3    109.0
##  Fert7             M1                  100.0 3.45 16.1     92.7    107.4
##  Fert8             M1                  102.0 3.45 16.1     94.7    109.3
##  Fert1             M2                  111.9 3.45 16.1    104.6    119.2
##  Fert2             M2                  105.4 3.45 16.1     98.1    112.7
##  Fert3             M2                  103.5 3.45 16.1     96.2    110.8
##  Fert4             M2                  110.3 3.45 16.1    102.9    117.6
##  Fert5             M2                  109.2 3.45 16.1    101.9    116.5
##  Fert6             M2                  107.8 3.45 16.1    100.5    115.2
##  Fert7             M2                   90.8 3.45 16.1     83.5     98.1
##  Fert8             M2                  108.2 3.45 16.1    100.9    115.6
##  Fert1             M3                  108.1 3.45 16.1    100.8    115.4
##  Fert2             M3                  104.4 3.45 16.1     97.1    111.8
##  Fert3             M3                  101.6 3.45 16.1     94.3    108.9
##  Fert4             M3                  105.5 3.45 16.1     98.2    112.8
##  Fert5             M3                  100.3 3.45 16.1     93.0    107.7
##  Fert6             M3                   95.5 3.45 16.1     88.2    102.8
##  Fert7             M3                   97.5 3.45 16.1     90.2    104.8
##  Fert8             M3                   99.1 3.45 16.1     91.8    106.4
```

```
##  Fert1             M4                  112.3 3.45 16.1    105.0    119.6
##  Fert2             M4                  103.6 3.45 16.1     96.3    110.9
##  Fert3             M4                  103.8 3.45 16.1     96.5    111.2
##  Fert4             M4                  112.3 3.45 16.1    105.0    119.6
##  Fert5             M4                  106.3 3.45 16.1     98.9    113.6
##  Fert6             M4                  102.9 3.45 16.1     95.6    110.2
##  Fert7             M4                   94.2 3.45 16.1     86.9    101.5
##  Fert8             M4                  101.5 3.45 16.1     94.2    108.8
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
```

3.3 Drawing conclusions

The scientists that ran this experiment were mostly interested in finding out the best Fertilizer-moment combination that maximizes yield, while understanding what makes that combination the best (i.e., is the effect of fertilizer source or application moment more important? Are they equally important?)

Response:

This is not something we can answer with a simple ANOVA. This is something one can tell by looking at the marginal effects and/or visualizing the results. See the figure below – Fertilizer source was one of the major factors affecting the response.

```
emmeans(m, ~fertilizer_source)
```

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

```
##  fertilizer_source emmean  SE   df lower.CL upper.CL
##  Fert1              109.8 2.6 5.49    103.3      116
##  Fert2              104.8 2.6 5.49     98.3      111
##  Fert3              102.9 2.6 5.49     96.4      109
##  Fert4              107.3 2.6 5.49    100.8      114
##  Fert5              105.0 2.6 5.49     98.5      111
##  Fert6              102.0 2.6 5.49     95.5      108
##  Fert7               95.6 2.6 5.49     89.1      102
##  Fert8              102.7 2.6 5.49     96.2      109
##
## Results are averaged over the levels of: application_moment
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
```

```
emmeans(m, ~application_moment)
```

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

```
##  application_moment emmean   SE   df lower.CL upper.CL
##  M1                    103 2.14 2.71     95.7      110
##  M2                    106 2.14 2.71     98.7      113
##  M3                    102 2.14 2.71     94.3      109
##  M4                    105 2.14 2.71     97.4      112
##
## Results are averaged over the levels of: fertilizer_source
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
```

```
as.data.frame(emmeans(m, ~fertilizer_source*application_moment)) |>
  ggplot(aes(emmean, fertilizer_source))+
```

```
  geom_errorbarh(aes(group = application_moment, color = application_moment,
                     xmin = lower.CL, xmax = upper.CL),
                 position = position_dodge(.4))+
  geom_point(aes(group = application_moment, color = application_moment),
                 position = position_dodge(.4))+
  theme_pubclean()+
  labs(color = "Application moment",
       y = "Fertilizer Source")
```

```
## Warning: `geom_errorbarh()` was deprecated in ggplot2 4.0.0.
## i Please use the `orientation` argument of `geom_errorbar()` instead.
## This warning is displayed once per session.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```