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?

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.

```r
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.

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").

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

```
##   block grams_plant fertilizer_source application_moment sample
## 1     1   109.04000             Fert1                 M1     S1
## 2     1   108.84399             Fert2                 M1     S1
## 3     1    97.29945             Fert3                 M1     S1
## 4     1   108.13468             Fert4                 M1     S1
## 5     1   102.49419             Fert5                 M1     S1
## 6     1   108.05124             Fert6                 M1     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.

3.2 Marginal means

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

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?)