Correlated Data

Rebecca C. Steorts (slide adaption from Maria Tacket) and material from Chapter 7 of Roback and Legler text.

Computing set up

```r
library(tidyverse)
library(tidymodels)
library(knitr)
library(patchwork)
library(viridis)
library(kableExtra)

knitr::opts_chunk$set(fig.width = 8,
                      fig.asp = 0.618,
                      fig.retina = 3,
                      dpt = 300,
                      out.width = "70%",
                      fig.align = "center")

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Learning goals

- Recognize a potential for correlation in a data set
- Identify observational units at varying levels
- Understand issues correlated data may cause in modeling
- Understand how random effects models can be used to take correlation into account

Notes based on Chapter 7 of Roback and Legler (2021) unless noted otherwise.

Correlated observations

Examples of correlated data

- In an education study, test scores for students from a particular teacher are typically more similar than test scores of other students with a different teacher

Examples of correlated data

- In an education study, test scores for students from a particular teacher are typically more similar than test scores of other students with a different teacher
- In a study measuring depression indices weekly over a month, the four measures for the same patient tend to be more similar than depression indices from other patients

Examples of correlated data

- In an education study, test scores for students from a particular teacher are typically more similar than test scores of other students with a different teacher
- In a study measuring depression indices weekly over a month, the four measures for the same patient tend to be more similar than depression indices from other patients
- In political polling, opinions of members from the same household tend to be more similar than opinions of members from another household

Examples of correlated data

- In an education study, test scores for students from a particular teacher are typically more similar than test scores of other students with a different teacher
- In a study measuring depression indices weekly over a month, the four measures for the same patient tend to be more similar than depression indices from other patients
- In political polling, opinions of members from the same household tend to be more similar than opinions of members from another household

Examples of correlated data

- In an education study, test scores for students from a particular teacher are typically more similar than test scores of other students with a different teacher
- In a study measuring depression indices weekly over a month, the four measures for the same patient tend to be more similar than depression indices from other patients
- In political polling, opinions of members from the same household tend to be more similar than opinions of members from another household

Correlation among outcomes within the same group (teacher, patient, household) is called **intraclass correlation**

Multilevel data

- We can think of correlated data as a multilevel structure

Multilevel data

- We can think of correlated data as a multilevel structure
    - Population elements are aggregated into groups

Multilevel data

- We can think of correlated data as a multilevel structure
  - Population elements are aggregated into groups
  - There are observational units and measurements at each level

Multilevel data

- We can think of correlated data as a multilevel structure
  - Population elements are aggregated into groups
  - There are observational units and measurements at each level
- For now we will focus on data with two levels:

Multilevel data

- We can think of correlated data as a multilevel structure
  - Population elements are aggregated into groups
  - There are observational units and measurements at each level
- For now we will focus on data with two levels:
  - **Level one**: Most basic level of observation

Multilevel data

- We can think of correlated data as a multilevel structure
  - Population elements are aggregated into groups
  - There are observational units and measurements at each level
- For now we will focus on data with two levels:
  - **Level one**: Most basic level of observation
  - **Level two**: Groups formed from aggregated level-one observations

Multilevel data example

Example: education study

- **Level one**
    - **Observational units**: students
    - **Level-one covariates**: test scores (response), year in school, demographics
- **Level two**
    - **Observational units**: teachers
    - **Level-two covariates**: years of experience

Two types of effects in model

- **Fixed effects**: Effects that are of interest in the study
  - Can think of these as effects whose interpretations would be included in a write up of the study

Two types of effects in model

- **Fixed effects**: Effects that are of interest in the study
  - Can think of these as effects whose interpretations would be included in a write up of the study
- **Random effects**: Not interested in studying effects of specific values in the data but we want to understand the variability
  - Can think of these as effects whose interpretations would not necessarily be included in a write up of the study

Example

Researchers are interested in understanding the effect social media has on opinions about a proposed economic plan. They randomly select 1000 households. They ask each adult in the household how many minutes they spend on social media daily and whether they support the proposed economic plan.

Example

Researchers are interested in understanding the effect social media has on opinions about a proposed economic plan. They randomly select 1000 households. They ask each adult in the household how many minutes they spend on social media daily and whether they support the proposed economic plan.

- daily minutes on social media is the fixed effect
- household is the random effect

Practice

Four wards were randomly selected at 25 hospitals and randomly assigned to offer a stress reduction program for nurses on the ward or serve as a control. At the end of the study period, a random sample of 10 nurses from each ward completed a test to measure job-related stress. Factors assumed to be related include experience, age, hospital size, and type of ward.

From Ex. 1 (a.) in Section 7.10.1 of BMLR

Practice

Four wards were randomly selected at 25 hospitals and randomly assigned to offer a stress reduction program for nurses on the ward or serve as a control. At the end of the study period, a random sample of 10 nurses from each ward completed a test to measure job-related stress. Factors assumed to be related include experience, age, hospital size, and type of ward.

From Ex. 1 (a.) in Section 7.10.1 of BMLR

1. What are the level one and level two observational units?
2. What is the response variable and what is its type (normal, Poisson, etc.)?
3. Describe the within-group correlation.
4. What are the fixed effects? What are the random effects?

Solution Practice

Four wards were randomly selected at 25 hospitals and randomly assigned to offer a stress reduction program for nurses on the ward or serve as a control. At the end of the study period, a random sample of 10 nurses from each ward completed a test to measure job-related stress. Factors assumed to be related include experience, age, hospital size, and type of ward.

1. What are the level one and level two observational units?

- level-one units (basic-level): the nurses
- level-two units (could be multiple-levels of grouping): ward and hospital

Solution Practice

Four wards were randomly selected at 25 hospitals and randomly assigned to offer a stress reduction program for nurses on the ward or serve as a control. At the end of the study period, a random sample of 10 nurses from each ward completed a test to measure job-related stress. Factors assumed to be related include experience, age, hospital size, and type of ward.

2. What is the response variable and what is its type (normal, Poisson, etc.)?

The response variable is test score of job-related stress. Normal.

Solution Practice

Four wards were randomly selected at 25 hospitals and randomly assigned to offer a stress reduction program for nurses on the ward or serve as a control. At the end of the study period, a random sample of 10 nurses from each ward completed a test to measure job-related stress. Factors assumed to be related include experience, age, hospital size, and type of ward.

3. Describe the within-group correlation.

We expect correlation of stress levels between nurses on the ward. In addition, we expect correlation in stress levels between wards at the same hospital.

Solution Practice

Four wards were randomly selected at 25 hospitals and randomly assigned to offer a stress reduction program for nurses on the ward or serve as a control. At the end of the study period, a random sample of 10 nurses from each ward completed a test to measure job-related stress. Factors assumed to be related include experience, age, hospital size, and type of ward.

4. What are the fixed effects? What are the random effects?

The fixed effects are experience, age (nurse level), type (ward level), size (hospital level).

The random effects are ward and hospital.

Teratogen and rat pups

Data: Teratogen and rat pups

Today's data are simulated results of an experiment with 24 dams (mother rats) randomly divided into four groups that received different doses of teratogen, a substance that could potentially cause harm to developing fetuses. The four groups are

- High dose (3 mg)
- Medium dose (2 mg)
- Low dose (1 mg)
- No dose (Control)

Each dam produced 10 rat pups and the presence of a deformity was noted for each pup.

Data: Teratogen and rat pups

Today's data are simulated results of an experiment with 24 dams (mother rats) randomly divided into four groups that received different doses of teratogen, a substance that could potentially cause harm to developing fetuses. The four groups are

- High dose (3 mg)
- Medium dose (2 mg)
- Low dose (1 mg)
- No dose (Control)

Each dam produced 10 rat pups and the presence of a deformity was noted for each pup.

Goal: Understand the association between teratogen exposure and the probability a pup is born with a deformity.

Sources of variation

- **Dose effect:** Studying whether different dose levels are associated with different probabilities of birth defects in the pups.

Sources of variation

- **Dose effect:** Studying whether different dose levels are associated with different probabilities of birth defects in the pups.
  - Dose is a **fixed effect**. Study is interested in defect rate at specific dose levels.

Sources of variation

- **Dose effect:** Studying whether different dose levels are associated with different probabilities of birth defects in the pups.
    - Dose is a **fixed effect**. Study is interested in defect rate at specific dose levels.
- **Dam (litter) effect:** Different dams may have different propensity to produce pups with defects, i.e. pups from same litter are more likely to be similar than pups from different litters.

Sources of variation

- **Dose effect:** Studying whether different dose levels are associated with different probabilities of birth defects in the pups.
  - Dose is a **fixed effect**. Study is interested in defect rate at specific dose levels.
- **Dam (litter) effect:** Different dams may have different propensity to produce pups with defects, i.e. pups from same litter are more likely to be similar than pups from different litters.
  - Dam is a **random effect.** Study is not interested in defect rate for each specific dam in the study but is interested in variability between litters.

Sources of variation

- **Dose effect:** Studying whether different dose levels are associated with different probabilities of birth defects in the pups.
    - Dose is a **fixed effect**. Study is interested in defect rate at specific dose levels.
- **Dam (litter) effect:** Different dams may have different propensity to produce pups with defects, i.e. pups from same litter are more likely to be similar than pups from different litters.
    - Dam is a **random effect.** Study is not interested in defect rate for each specific dam in the study but is interested in variability between litters.
- **Pup-to-pup variability**: Within litter pup differences. This is the unexplained variability.

Scenario 1: No dose effect

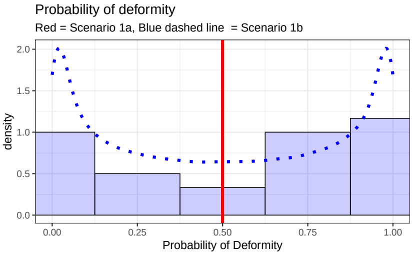
Assume dose has no effect on, p, the probability of a pup born with a deformity.

- **Scenario 1a.**: $p = 0.5$ for each dam
- **Scenario 1b.**: $p \sim Beta(0.5, 0.5)$ (expected value $= 0.5$)

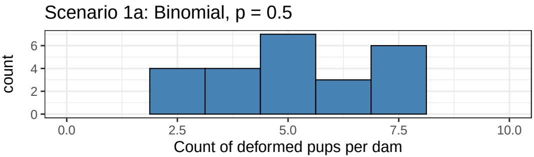
Scenario 1: No dose effect

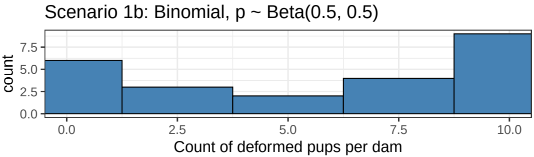
Assume dose has no effect on, p, the probability of a pup born with a deformity.

- **Scenario 1a.**: $p = 0.5$ for each dam
- **Scenario 1b.**: $p \sim Beta(0.5, 0.5)$ (expected value $= 0.5$)



1. Would you expect the number of pups with a deformity for dams in Scenario 1a to follow a distribution similar to the binomial distribution with $n = 10$ and $p = 0.5$? Why or why not?
2. Would you expect the number of pups with a deformity for dams in Scenario 1b to follow a distribution similar to the binomial distribution with $n = 10$ and $p = 0.5$? Why or why not?
3. Which scenario do you think is more realistic - Scenario 1a or 1b?


Scenario 1a: Binomial, p = 0.5


Scenario 1b: Binomial, p ~ Beta(0.5, 0.5)

mean_1a	sd_1a	mean_1b	sd_1b
5.166667	1.493949	5.666667	4.103727

Solutions

1. Would you expect the number of pups with a deformity for dams in Scenario 1a to follow a distribution similar to the binomial distribution with $n = 10$ and $p = 0.5$? Why or why not?

Yes. Every dam is assumed to have exactly 10 pups, and each pup is assumed to have a probability of exactly 0.5 of being deformed, regardless of their dam.

Solutions

2. Would you expect the number of pups with a deformity for dams in Scenario 1b to follow a distribution similar to the binomial distribution with $n = 10$ and $p = 0.5$? Why or why not?

No. Although every dam has exactly 10 pups, each dam has a unique probability of having deformed pups. So some pups have a much higher chance of being deformed and some much lower, depending on their dam. The probabilities across all dams just happen to average out to 0.5. In terms of coins, we could envision this as each dam having a unique weighted coin, so that each pup from a specific dam has the same weighted coin flipped for them.

Solutions

3. Which scenario do you think is more realistic - Scenario 1a or 1b?

1b because it seems more realistic that some dams might be prone to having deformed pups based on genetics, diet, environment, or other factors.

Scenario 2: Dose effect

Scenario 2: Dose effect

Now we will consider the effect of the dose of teratogen on the probability of a pup born with a deformity. The 24 pups have been randomly divided into four groups:

- High dose (`dose = 3`)
- Medium dose (`dose = 2`)
- Low dose (`dose = 1`)
- No dose (`dose = 0`)

Scenario 2: Dose effect

Now we will consider the effect of the dose of teratogen on the probability of a pup born with a deformity. The 24 pups have been randomly divided into four groups:

- High dose (`dose = 3`)
- Medium dose (`dose = 2`)
- Low dose (`dose = 1`)
- No dose (`dose = 0`)

We will assume the true relationship between p and dose is the following:

$$\log\left(\frac{p}{1-p}\right) = -2 + 1.33\ dose$$

Scenario 2

Scenario 2a.

$$p = \frac{e^{-2+1.33\ \mathit{dose}}}{1 + e^{-2+1.33\ \mathit{dose}}}$$

Scenario 2

Scenario 2a.

$$p = \frac{e^{-2+1.33\ \ dose}}{1 + e^{-2+1.33\ \ dose}}$$

Scenario 2b.

$$p \sim Beta\Big(\frac{2p}{(1-p)}, 2\Big)$$

Scenario 2

Scenario 2a.

$$p = \frac{e^{-2+1.33\ \mathit{dose}}}{1 + e^{-2+1.33\ \mathit{dose}}}$$

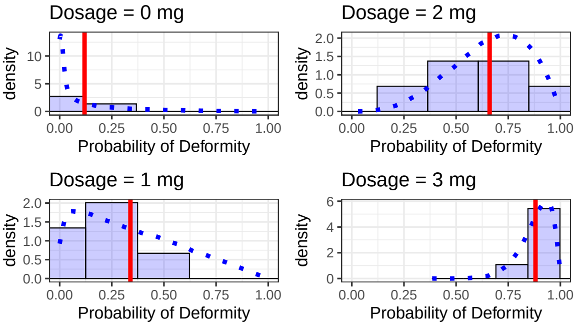
Scenario 2b.

$$p \sim \mathit{Beta}\Big(\frac{2p}{(1-p)}, 2\Big)$$

On average, dams who receive dose x have the same probability of pup born with deformity as dams with dose x under Scenario 2a.

- e.g., If `dose = 1`, the probability of a pup born with deformity is 0.338 in Scenario 2a and the mean is 0.338 in Scenario 2b.

Distributions under Scenario 2



Reproduced from Figure 7.3 in BMLR

Scenario 2 summary statistics

Table 2: Summary statistics of Scenario 2 by dose.

	Scenario 2a				Scenario 2b			
Dosage	Mean p	SD p	Mean Count	SD Count	Mean p	SD p	Mean Count	SD Count
0	0.119	0	1.333	1.366	0.061	0.069	0.500	0.837
1	0.339	0	3.167	1.835	0.239	0.208	3.500	2.881
2	0.661	0	5.833	1.472	0.615	0.195	5.833	1.941
3	0.881	0	8.833	1.169	0.872	0.079	8.833	1.169

From Table 7.2 in BMLR

1. In Scenario 2a, dams produced 4.79 deformed pups on average, with standard deviation 3.20. Scenario 2b saw an average of 4.67 with standard deviation 3.58. Why are comparisons by dose more meaningful than these overall comparisons?
2. We will use binomial and quasibinomial regression to model the relationship between dose and probability of pup born with a deformity. What can you say about the center and the width of the confidence intervals under Scenarios 2a and 2b?
   2.1 Which will be similar and why?
   2.2 Which will be different and how?

Scenario 2: Estimated odds ratio

The estimated effect of dose and the 95% CI from the binomial and quasibinomial models are below:

Scenario 2a

	Odds Ratio	95% CI
Binomial	3.536	(2.604, 4.958)
Quasibinomial	3.536	(2.512, 5.186)

Scenario 2b

	Odds Ratio	95% CI
Binomial	4.311	(3.086, 6.271)
Quasibinomial	4.311	(2.735, 7.352)

1. Describe how the quasibinomial analysis of Scenario 2b differs from the binomial analysis of the same simulated data. Do confidence intervals contain the true model parameters? Is this what you expected? Why?
2. Why does Scenario 2b contain correlated data that we must account for, while Scenario 2a does not?

Solutions

1. Describe how the quasibinomial analysis of Scenario 2b differs from the binomial analysis of the same simulated data. Do confidence intervals contain the true model parameters? Is this what you expected? Why? (See Table 7.1 in the book).

Both analyses give the same estimated coefficient and odds ratio, but the standard error for the quasi-binomial analysis is larger, leading to a lower t-statistic, higher p-value, and wider CI.

Both CI contain the true odds ratio (3.79), but, we would expect across many simulations the quasi-binomial would have closer 95 percent coverage than the binomial CI.

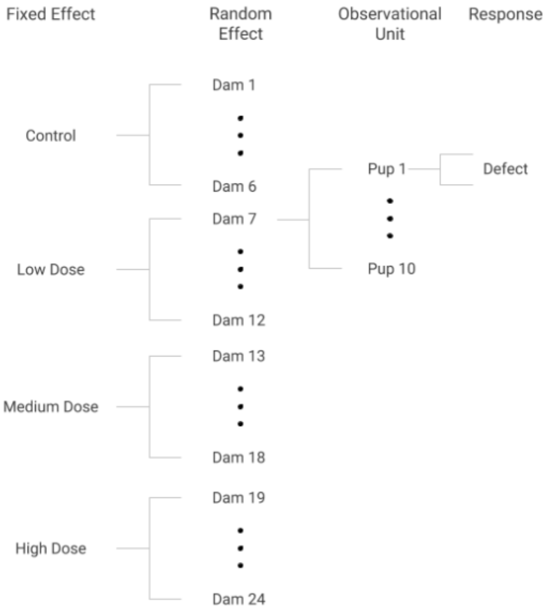
Solutions

2. Why does Scenario 2b contain correlated data that we must account for, while Scenario 2a does not?

There is structurally no extra-binomial variation to adjust for since all pups at a single dose behave similarly, regardless of dam (all dams at a single dos have the same probability of deformity).

In scenario 2b, there is extra binomial variation to adjust for since the results from pups at a single does depend on their dam and the specific probability associated with that dam.

Data structure



Preview: Fit model with random effect

```
## Rows: 240 Columns: 3
## -- Column specification -------------------------------
## Delimiter: ","
## dbl (3): dam, dose, deformity
##
## i Use `spec()` to retrieve the full column specificatio
## i Specify the column types or set `show_col_types = FALS
```

Preview: Fit model with random effect

Preview: Fit model with random effect

effect	group	term	estimate	std.error	statistic	p.value	conf.low	conf.high
fixed	NA	(Intercept)	-2.819	0.528	-5.343	0	-3.853	-1.785
fixed	NA	dose	1.691	0.282	5.992	0	1.138	2.244
ran_pars	dam	sd__(Intercept)	0.834	NA	NA	NA	NA	NA

Summary

- The structure of the data set may imply correlation between observations.
- Correlated observations provide less information than independent observations; we need to account for this reduction in information.
- Failing to account for this reduction could result in underestimating standard error, thus resulting in overstating significance and the precision of the estimates.
- We showed how we can account for this by incorporating the dispersion parameter or a random effect.

References

Roback, Paul, and Julie Legler. 2021. *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. CRC Press.