Correlated data

02.23.22

Click here for PDF of slides

Announcements

- HW 03 **due Mon, Feb 28 at 11:59pm**
- No office hours Thursday.
    - Office hours will resume Friday 2 - 3pm
    - Schedule an appointment

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

Correlated observations

Examples of correlated data

- In an education study, scores for students from a particular teacher are typically more similar than 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
  - 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
- Example: political polling
  - Level one: individual members of household
  - Level two: household

Two types of effects

- **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**: Effects we're not interested in studying but whose variability we want to understand
    - 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.

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

Practice

Researchers conducted a randomized controlled study where patients were randomly assigned to either an anti-epileptic drug or a placebo. For each patient, the number of seizures at baseline was measured over a 2-week period. For four consecutive visits the number of seizures were determined over the past 2-week period. Patient age and sex along with visit number were recorded.

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

Click here to submit your response.

Ex. 1 from Section 7.10.1 in BMLR

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)

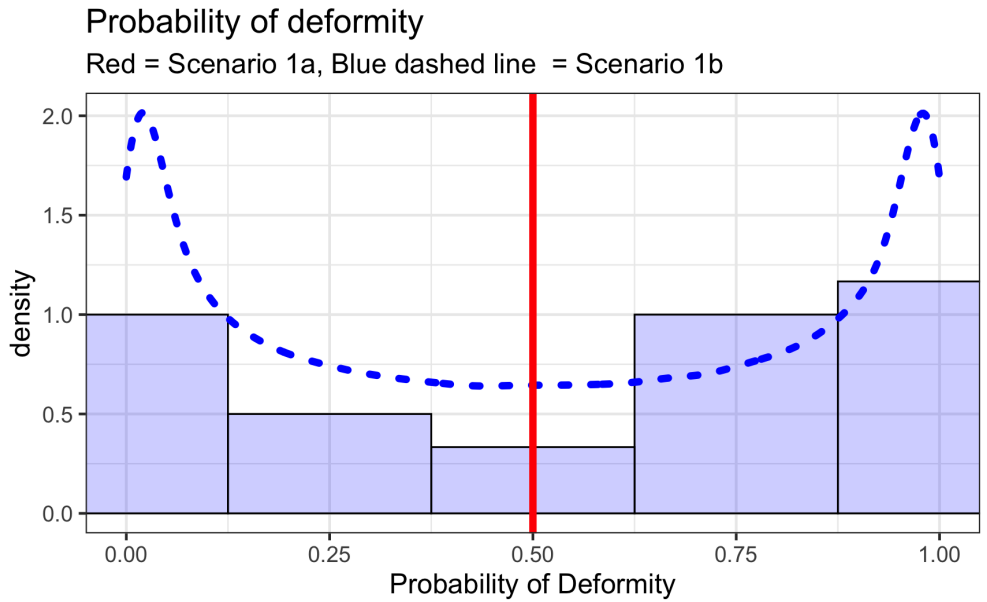
Each dam produced 10 rat pups and the presence of a deformity was noted.

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

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)

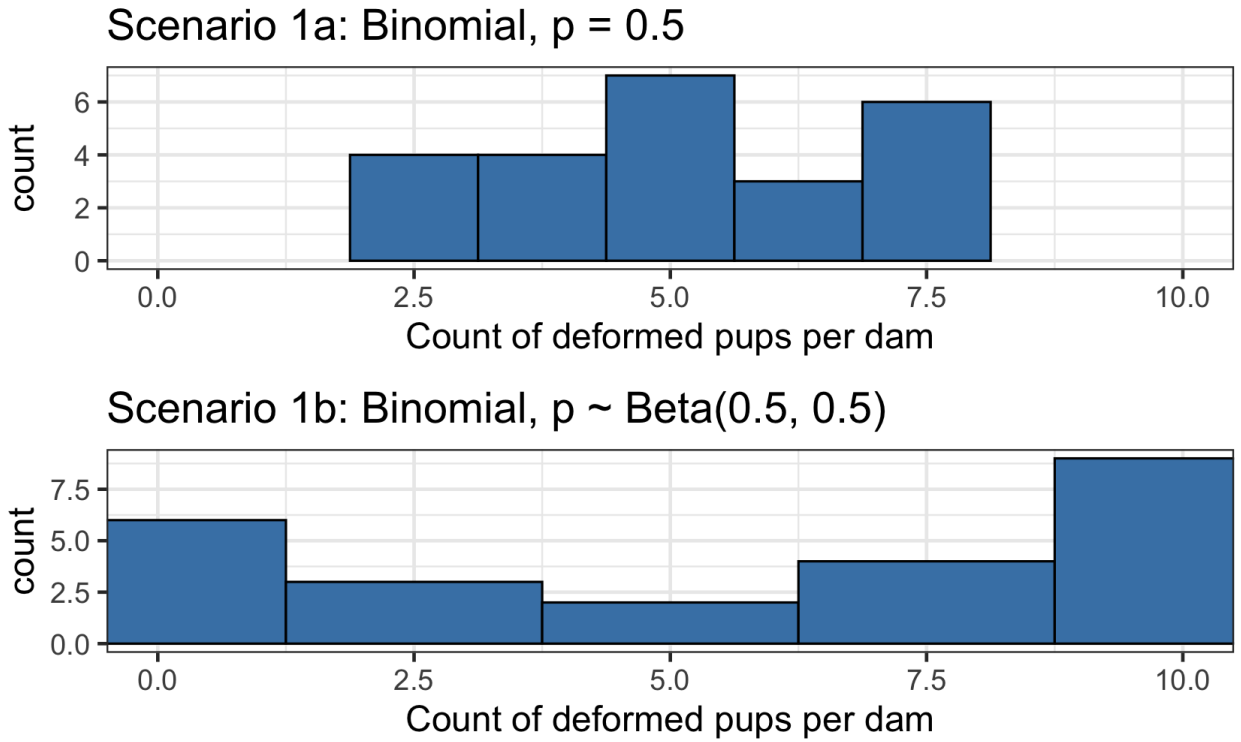


From Figure 7.1 in BMLR

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?



mean_1a	sd_1a	mean_1b	sd_1b
5.166667	1.493949	5.666667	4.103727

Let's take a look at a binomial and quasibinomial model for Scenarios 1a and 1b.

Complete Scenario 1 of `lecture-14.Rmd`

Click here to submit your response.

10:00

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`**)

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.

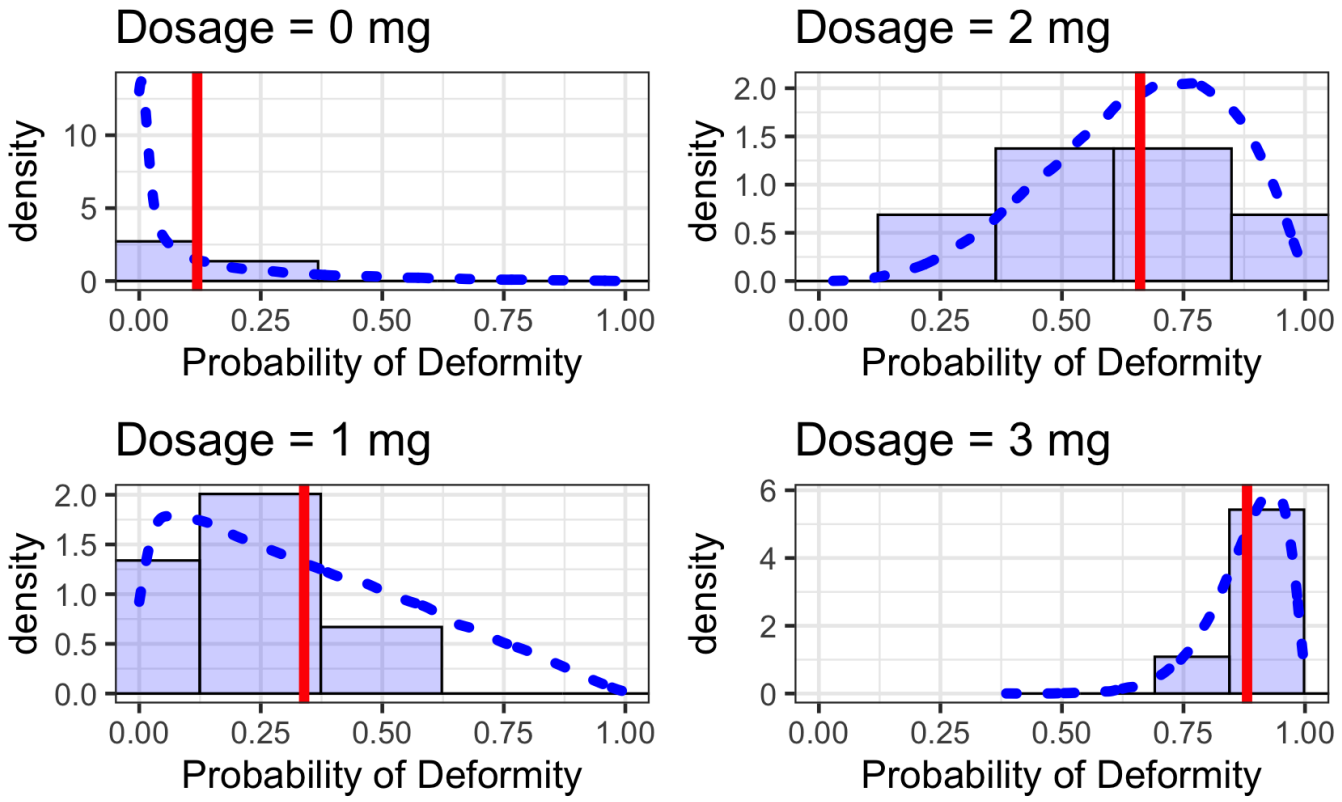
$$p = \frac{e^{-2+1.33\ dose}}{1 + e^{-2+1.33\ dose}}$$

Scenario 2b.:

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

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

Distributions under Scenario 2



Replicated from Figure 7.3 in BMLR

Summary statistics under Scenario 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?

    - Which will be similar and why?
    - 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 are differences between quasibinomial and binomial models of Scenario 2a less noticeable than the differences in Scenario 2b?

Preview: Add random effect to model

```
library(lme4)
random_effect_model <- glmer(deformity ~ dose + (1|dam),
                             family = binomial, data = scenario2_raw)
random_effect_model
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: deformity ~ dose + (1 | dam)
##    Data: scenario2_raw
##       AIC       BIC    logLik  deviance  df.resid
##  228.2567  238.6986 -111.1284  222.2567       237
## Random effects:
##  Groups Name        Std.Dev.
##  dam    (Intercept) 0.8335
## Number of obs: 240, groups:  dam, 24
## Fixed Effects:
## (Intercept)         dose
##      -2.819        1.691
```

Preview: Add random effect

```
confint(random_effect_model)
```

```
##                   2.5 %    97.5 %
## .sig01        0.3215584  1.536878
## (Intercept) -4.0762551 -1.899790
## dose         1.1988973  2.359106
```

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.

Acknowledgements

The content in the slides is from BMLR: Chapter 7 - Correlated data