Categorical: Repeated measures

Table of contents

1 Goals

1.1 Goals

1.1.1 Goals of this section

- Models for repeated measures
  - Mixed models framework
- First continuous outcomes

    - Then categorical outcomes

- Two types of mixed model
    - Marginal mixed model
    - Conditional mixed model

1.1.2 Goals of this lecture

- What's the problem with repeated measures?
- Mixed model for continuous outcomes
    - Marginal model: **R** matrix
    - Conditional model: **G** matrix

2 Repeated measures

2.1 Repeated measures

2.1.1 Assumptions of GLM

- GLM (ANOVA / linear regression) assumes
    - Conditional normality of residuals
    - Constant variance of residuals
    - **Independence** of residuals

2.1.2 Assumptions of GLiM

- GLiM relaxes the assumptions of
    - Conditional normality of residuals
    - Constant variance of residuals
- But GLiM still assumes **independence** of residuals

2.1.3 Independence

- **Independent observations**: Information about one observation doesn't provide any information about other observations
  - *Lack of independence implies correlation* (but the reverse is not true)
    * If observations are **not independent**, they will be correlated
    * If observations are **correlated**, they are not independent
    * If observations are **not correlated**, we don't know if they're independent or not

2.1.4 Repeated measures = non-independence

- Repeated measures from the same person are **not independent**
  - An observation from a person provides information about other observations from that person
  - Observations from the same person are more like one another than observations from different people
  - Observations from the same person are **correlated**

2.1.5 Violation of independence

- Does not impact regression coefficients
  - Impacts **standard errors**
  - Impacts **statistical significance**
- Will it be too *large* or too *small*? **It depends**
  - Hu, Goldberg, Hedeker, Flay, Pentz (1998)
  - Predictors about person: Standard errors are **too small**
  - Predictors about occasion: Standard errors are **too large**
  - Also depends on other things

2.1.6 Individual effects

- In models so far, there is **an** effect of a predictor
  - **Individual differences** in terms of *variables*
- But what if the effect of a predictor varied depending on the person?
  - **Individual differences** in terms of *effects or slopes*
- Mixed models can estimate person-specific effects

2.2 Example

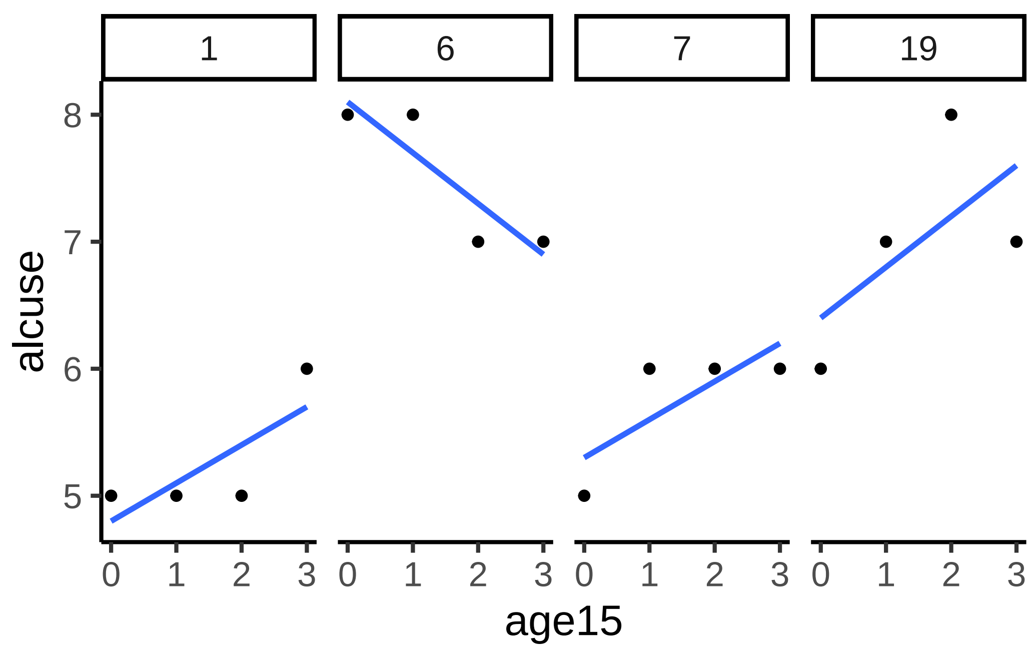
2.2.1 Example data: Substance use in adolescence

- `id`: ID variable
- `age`: Age in years
- `alcuse`: Alcohol use
- `ciguse`: Cigarette use
- `potuse`: Marijuana use
- Demographics: gender, family structure, other variables

2.2.2 Tall or univariate data

id	female	twopars	peerenc	parconf	paruse	age	alcuse	ciguse	potuse
1	0	1	7	11	13	15	5	6	5
1	0	1	7	11	13	16	5	6	5
1	0	1	7	11	13	17	5	5	4
1	0	1	7	11	13	18	6	7	5
6	1	0	6	14	16	15	8	9	6
6	1	0	6	14	16	16	8	8	5

2.2.3 Plot: Individual effects



2.2.4 Two types of mixed model

- Marginal models
  - Population-averaged models or generalized estimating equations (GEE)
  - Treat the person as a *nuisance* and adjust standard errors
- Conditional models
  - Generalized linear mixed models (GLMM)
  - Explicitly model the person (and variability among people) to get person-specific effects

2.3 Repeated-measures ANOVA

2.3.1 Review: Repeated-measures ANOVA assumptions

- Covariance matrix of outcomes

$$\mathbf{S}_{YY} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_3^2 & \sigma_{34} \\ & & & \sigma_4^2 \end{bmatrix}$$

- σ_1^2 = variance of outcome at time 1
- σ_{12} = covariance between outcome at time 1 and outcome at time 2

2.3.2 Review: Repeated-measures ANOVA assumptions

- **Compound symmetry** of the covariance matrix of outcomes
  - Homogeneity of variances (i.e., variances are all the same):
    - $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
  - Homogeneity of covariances (i.e., covariances are all the same):
    - $\sigma_{12} = \sigma_{13} = \sigma_{14} = \sigma_{23} = \sigma_{24} = \sigma_{34}$
- Actual assumption: **Sphericity**
  - Compound symmetry holds for **differences** between pairs of scores
  - Slightly weaker assumption

3 Linear mixed model

3.1 Linear mixed model

3.1.1 Linear mixed model

- Also known as: random coefficient model, multi-level model, nested model, hierarchical linear model, random effects model
- Many names because they were developed in parallel in different disciplines
  - Multi-level models and hierarchical linear models from education
  - Random coefficient from statistics

3.1.2 Linear mixed model

- Extension of GLM that allows for non-independence
  - Partitions variation, just like ANOVA, linear regression
  - More control over the form of non-independence
    - Linear regression: Independence
    - Repeated-measures ANOVA: Compound symmetry
- Two approaches
  - Correlated residuals: **R** matrix
  - Random effects: **G** matrix

3.1.3 Linear mixed model: Equations

The linear mixed model:

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

- where $\gamma \sim N(0, \mathbf{G})$ and $\epsilon \sim N(0, \mathbf{R})$
- $variance(Y) = V = \mathbf{ZGZ}' + \mathbf{R}$

3.1.4 Linear mixed model: Simpler

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

- $\mathbf{X}\beta$ are the **fixed effects**
  - Average effects
  - Think predictors ($\mathbf{X}$) and regression coefficients (β)
- $\mathbf{Z}\gamma$ and ϵ are *special residuals* that let us include correlations among the repeated observations
  - Specifically, among their **residuals**

3.1.5 Two approaches

- $\mathbf{Z}\gamma$ are the **random effects**
  - γ has mean = 0 and variance given by *covariance matrix* $\mathbf{G}$
  - Generalized linear mixed models (GLMM) use this
- ϵ is the **error** or **residual term**
  - ϵ has mean = 0 and variance given by *covariance matrix* $\mathbf{R}$
  - Generalized estimating equations (GEE) and related methods use this

3.1.6 Continuous vs categorical outcomes

- **Continuous outcomes**
  - $\mathbf{G}$ and $\mathbf{R}$ portions of the model are *independent*
- **Categorical outcomes**
  - $\mathbf{G}$ and $\mathbf{R}$ portions of the model are **NOT independent**
- This week: $\mathbf{G}$ and $\mathbf{R}$ for the continuous model
  - Understand how they function
- Next week: Categorical outcomes

3.2 R matrix

3.2.1 R matrix

- $\mathbf{R}$ is the covariance matrix among the repeated outcomes / timepoints

$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_3^2 & \sigma_{34} \\ & & & \sigma_4^2 \end{bmatrix}$$

3.2.2 How are timepoints related?

- Linear mixed model is really flexible about what this matrix can look like
  - Unstructured
  - Compound symmetry
  - Autoregressive
  - Diagonal
  - Many others: See Kinkaid (2005)
- Repeated-measures ANOVA is a special case of LMM
  - Not flexible: Compound symmetry only

3.2.3 Unstructured R

$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_3^2 & \sigma_{34} \\ & & & \sigma_4^2 \end{bmatrix}$$

- Estimate every value in the matrix
- $\frac{t(t+1)}{2}$ values: Here, 10 values

3.2.4 Compound symmetry R

$$\mathbf{R} = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ & \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ & & \sigma^2 + \sigma_1^2 & \sigma_1^2 \\ & & & \sigma^2 + \sigma_1^2 \end{bmatrix}$$

- One value for all variances

- One value for all covariances
- 2 values: Here, 2 values

3.2.5 Autoregressive R

$$\mathbf{R} = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 \\ & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ & & \sigma^2 & \sigma^2\rho \\ & & & \sigma^2 \end{bmatrix}$$

- One value for variances
- Covariance decreases as time between points increases (ρ)
- 2 values: Here, 2 values

3.2.6 Diagonal (independence) R

$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ & \sigma_2^2 & 0 & 0 \\ & & \sigma_3^2 & 0 \\ & & & \sigma_4^2 \end{bmatrix}$$

- One value for variance for each time point
- t values: Here, 4 values

3.2.7 Why do we care about R?

- **R** is the residual variance matrix
  - Residual variance impacts the **standard errors of the fixed effects**
  - **R** impacts the standard error (and therefore the significance) of the fixed effects
- Variance structure you choose affects what is significant
  - Choose the variance structure that most closely reflects reality

3.2.8 Which form of R to use?

- Run models with different versions of **R** matrix
  - Compare using AIC and likelihood
    * AIC: smaller is better
    * Likelihood: likelihood ratio test

3.2.9 Which form of R to use?

- Unstructured: Most information, but also most parameters
  - Difficult with more than a handful of timepoints
  - First try to get an idea of what the covariance matrix looks like
- Diagonal: Independence
  - All timepoints are uncorrelated
  - Unlikely given our discussion of repeated measures
- Compound symmetry: In between
- Autoregressive: In between

3.2.10 Example: Unstructured

```
Generalized least squares fit by maximum likelihood
  Model: alcuse ~ 1 + age15
  Data: alcuse_tall
       AIC      BIC    logLik
  981.0951 1017.108 -481.5476

Correlation Structure: General
 Formula: ~1 | id
 Parameter estimate(s):
 Correlation:
  1     2     3
2 0.580
3 0.421 0.605
4 0.324 0.366 0.340

Coefficients:
               Value  Std.Error  t-value p-value
(Intercept) 5.901857 0.08824313 66.88177  0.0000
age15       0.105372 0.03342207  3.15278  0.0017

 Correlation:
      (Intr)
age15 -0.633

Standardized residuals:
         Min           Q1          Med           Q3          Max
-3.418749537 -0.239414053 -0.007940977  0.858943288  2.304510243
```

```
Residual standard error: 0.9104505
Degrees of freedom: 404 total; 402 residual
```

3.2.11 Example: Unstructured

- Working correlation matrix ($\mathbf{R}$)

```
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.5797687 0.4208941 0.3236446
[2,] 0.5797687 1.0000000 0.6053009 0.3663275
[3,] 0.4208941 0.6053009 1.0000000 0.3404776
[4,] 0.3236446 0.3663275 0.3404776 1.0000000
```

3.2.12 Example: Compound symmetry

```
Generalized least squares fit by maximum likelihood
  Model: alcuse ~ 1 + age15
  Data: alcuse_tall
       AIC      BIC    logLik
  988.1413 1004.147 -490.0706

Correlation Structure: Compound symmetry
 Formula: ~1 | id
 Parameter estimate(s):
      Rho
0.4430272

Coefficients:
               Value  Std.Error  t-value p-value
(Intercept) 5.890099 0.08302555 70.94321  0.0000
age15       0.107921 0.03036304  3.55435  0.0004

 Correlation:
      (Intr)
age15 -0.549

Standardized residuals:
         Min           Q1          Med           Q3          Max
-3.405625651 -0.234496379  0.002171263  0.861991319  2.313480479
```

```
Residual standard error: 0.9120029
Degrees of freedom: 404 total; 402 residual
```

3.2.13 Example: Compound symmetry

- Working correlation matrix (**R**)

```
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.4430272 0.4430272 0.4430272
[2,] 0.4430272 1.0000000 0.4430272 0.4430272
[3,] 0.4430272 0.4430272 1.0000000 0.4430272
[4,] 0.4430272 0.4430272 0.4430272 1.0000000
```

3.2.14 Compare

Model	AIC	-2LL	# parameters
Unstructured	981.095	963.095	10
Compound symmetry	988.141	980.141	2

- Difference in -2LL: $980.141 - 963.095 = 17.046$
- Degrees of freedom: $10 - 2 = 8$
- Critical $\chi(8)^2 = 15.507$
- Test is significant: More complex model fits better
  - Unstructured

3.2.15 Interpretation

- This is a **marginal** model
  - Treat the person as a *nuisance* and adjust standard errors
- Use **R** to account for additional correlation of repeated measures
  - But don't care about it
  - Just want to get rid of it

3.3 G matrix

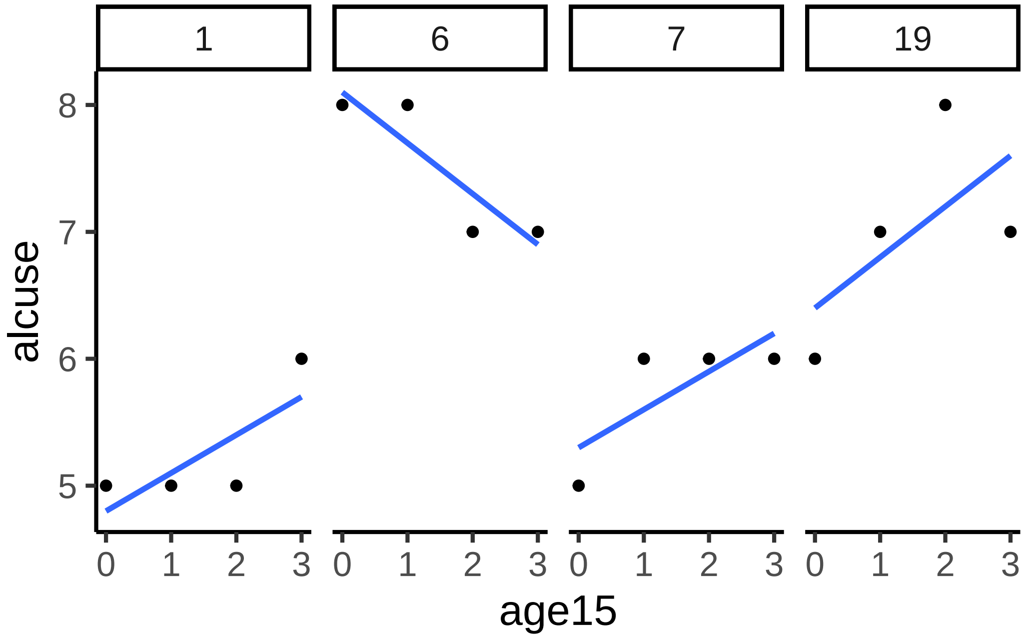
3.3.1 Individual effects

- Before:
  - There is **an** effect of a predictor
    - **Individual differences** in terms of *variables*
- Now:
  - Effect of a predictor varies depending on the person
    - **Individual differences** in terms of *effects or slopes*

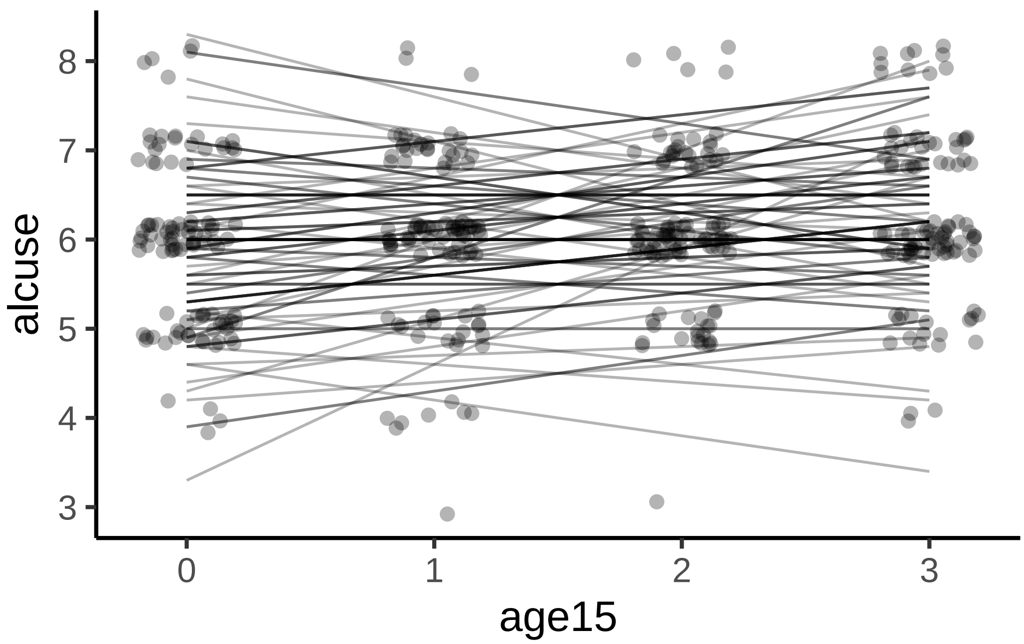
3.3.2 Conceptually

- Perform a regression on each person's data
  - Predictor: **time**
  - Outcome: Outcome
- Separate regression for **each person** in the study
  - Each person has an intercept
  - Each person has a slope

3.3.3 Time vs outcome for the first four participants



3.3.4 Time vs outcome for all participants



3.3.5 Assumptions

- Figures are a good way to think about the model
  - But still violate assumptions of linear regression
- Observations for each line are **not independent**
  - Multiple observations from the same person
- Good news!
  - Lack of independence only impacts **standard errors**
  - Estimates of the **intercepts** and **slopes** are good

3.3.6 That's a lot of lines!

- Intercept and slope for every single person in the sample
  - Can't report all of those
  - Summarize the intercepts and slopes in some way
- Variances and covariances
  - Variance of intercepts
  - Variance of slopes
  - Covariances between intercepts and slopes

3.3.7 G

- **G** is the variance-covariance matrix of intercepts and slopes

$$\mathbf{G} = \begin{bmatrix} \sigma^2_{int} & \sigma_{int-slope} \\ & \sigma^2_{slope} \end{bmatrix}$$

- σ^2_{int} is the variance of the intercepts
- σ^2_{slope} is the variance of the slopes
- $\sigma_{int-slope}$ is the covariance between intercepts and slopes

3.3.8 Why do we care about G?

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

- $\mathbf{Z}\gamma$ is where the random effects ($\mathbf{G}$) are
  - It accounts for some variation in scores
- Remaining variation ends up in ϵ, the residual variance
  - Residual variance impacts the **standard errors of the fixed effects**
  - **G** impacts s.e.s (and therefore significance) of fixed effects
- Random effects affect what is significant

3.4 Example: Random intercept and slope

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
  method [lmerModLmerTest]
Formula: alcuse ~ 1 + age15 + (1 + age15 | id)
   Data: alcuse_tall

     AIC      BIC   logLik deviance df.resid
   985.7   1009.7   -486.8    973.7      398

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.71801 -0.60424 -0.07987  0.54069  2.97288

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.55174  0.7428
          age15       0.03706  0.1925   -0.59
 Residual             0.40148  0.6336
Number of obs: 404, groups:  id, 101

Fixed effects:
             Estimate Std. Error        df t value             Pr(>|t|)
(Intercept)   5.89010    0.09080 100.99824  64.866 < 0.0000000000000002 ***
age15         0.10792    0.03409 100.99947   3.166              0.00204 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
```

```
      (Intr)
age15 -0.653
```

3.4.1 Example: Random intercept and slope

- Random effects (and residual variance)

```
 Groups   Name        Std.Dev. Corr
 id       (Intercept) 0.74279
          age15       0.19252  -0.586
 Residual             0.63363
```

3.4.2 Interpretation

- This is a **conditional** model
  - Individual variability in intercept and slope are of interest
- Explicitly model this using **G**
  - Person-specific intercepts and slopes

3.5 Comparing both sets of models

3.5.1 Compare fixed effects

Model	Intercept	Slope
R unstructured	5.902	0.105
G with random intercept and slope	5.89	0.108

- **R** unstructured: Correlated residuals
- **G** with random intercept and slope: Random effects

3.5.2 Continuous models vs categorical models

- Continuous outcome (here)
  - Interpretations are different
  - But numerical results are basically identical
    - Assuming some things

- Categorical outcome
  - Interpretations are different
  - Numerical results are **very** different
    - Nonlinearity for categorical outcomes

4 Summary

4.1 Summary

4.1.1 Summary of this week

- Repeated measures violate the assumption of independence
- Mixed model (/ multilevel model / hierarchical linear model)
  - Marginal and conditional models
    - **R** and **G** matrix approaches
    - Equivalent (numerically) for continuous outcomes

4.1.2 Next weeks

- Extend mixed models to categorical outcomes
  - Marginal: **R** matrix, population averaged, GEE, cluster robust
  - Conditional: **G** matrix, generalized linear mixed models (GLMM)
- Different interpretations, different numbers