PSY 5939: Longitudinal Data Analysis

Contents

1 Linear mixed model

1.1 Linear mixed model

1.1.1 Linear mixed model

Model for **non-independent** observations

- These can be longitudinal or cross-sectional

For example

- Several children from the same family
- Multiple schoolchildren with the same teacher
- Employees who work in teams or workgroups
- **Multiple observations from the same individual over time**

Observations from the same family / class / team / person are more similar to one another than observations from different families / classes / teams / persons

1.1.2 Linear mixed models

Non-independent observations means that there is some redundancy (or correlation) between observations

The effective sample size is smaller that the actual sample size

- There are 100 subjects but we only have 72 subjects' worth of information, due to correlations between observations

The standard error is **underestimated** if you ignore non-independence

- How much the standard errors are underestimated depends on how much the observations are related to one another

1.1.3 Linear mixed model: motivation

The linear mixed model (LMM) is an extension of the general linear model (GLM)

Partitions variation, just like ANOVA and regression

But **more ways** to partition and **more control** over the form

- Between-subjects ANOVA only uses "independence"
- Repeated-measures ANOVA only uses "compound symmetry"

The linear mixed model has **two different approaches** to sorting out the additional variation due to repeated measures

1.1.4 Linear mixed model: equations

The linear mixed model: $Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$

$\mathbf{X}\beta$ is a linear combination of predictors (matrix $\mathbf{X}$) and regression coefficients (vector β)

The last two terms are the two places where additional variation is modeled

- $\mathbf{Z}\gamma$ is used to random effects (we alredy talked about this)
- ϵ is used for correlated residuals
    - $\epsilon \sim N(0, \mathbf{R})$

2 The R matrix

2.1 The R matrix

2.1.1 Linear mixed model: correlated residuals

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

$\mathbf{X}\beta$ are the **fixed effects**

- $\mathbf{X}$ is the matrix of predictors
- β is the vector of regression coefficients or weights

ϵ is the error or residual term

- ϵ has a mean of 0 and variance given by **covariance matrix R**

2.1.2 Correlated residuals

In linear regression, a **single** residual for each person and a **single** variance of the residuals

With repeated measures, there is a single residual for each person **at each time point**

- Now there is a variance for *each time point*

In addition, due to the **repeated measures** and **non-independence**, residuals from different time points are related to each other

- There are *covariances* between the residuals at each time point

This results in a **covariance matrix** among the residuals at each time point

2.1.3 Residual matrix R

R is a $t \times t$ **covariance** matrix where t is the number of repeated measures

- Values on the main diagonal are *variances* at each time point
- Values off the main diagonal reflect the relationships between time points
- Matrix can take several different forms: auto-regressive, compound symmetry, unstructured, diagonal, many others

Sometimes, we convert **R** to a correlation matrix to aid interpretation

2.1.4 Residual matrix R

When observations are not independent, standard errors are underestimated if you ignore non-independence

How much do we adjust the standard errors up?

- It depends on how much observations are related to one another
- Which we find out by looking at the **R** matrix

2.1.5 How might timepoints be related?

Observations at different timepoints are related to one another

Summarize all those various relationships in a covariance matrix

Specifically, the **R** matrix

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

This should look familiar from repeated measures ANOVA

2.1.6 Mixed models with repeated measures

Repeated measures ANOVA is a special case of a mixed model where **R** has compound symmetry form

However, the linear mixed model is really flexible (while repeated measures ANOVA is not)

The **R** matrix can take on many different forms

- Unstructured
- Compound symmetry (we already know this one)
- Autoregressive
- Diagonal (a.k.a. independence)

2.1.7 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

In general, the unstructured matrix estimates $\frac{t \times (t+1)}{2}$ values

This 4×4 matrix would estimate 10 values

2.1.8 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 (main diagonal)

One value for all covariances (off diagonal)

In general, the CS matrix estimates 2 values

This 4×4 matrix would estimate 2 values

2.1.9 Auto-regressive 1 R

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

Estimate one value for the variances (main diagonal)

Covariances decrease as time between points increases

In general, the AR matrix estimates 2 values

This 4×4 matrix would estimate 2 values

2.1.10 Diagonal 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 at each time point (main diagonal)

One value for all covariances (off diagonal) = 0

In general, the diagonal matrix estimates t values

This 4×4 matrix would estimate 4 values

(This is independence)

2.1.11 Which form of R to use?

Run models with different versions of the R matrix and compare

- AIC: smaller is better
- Likelihood ratio test
  - Difference between -2LL (in output) for two models

  - Degrees of freedom for a model = # parameters in **R**
  - The difference in -2LL values is distributed as χ^2 with df = difference # parameters

2.1.12 Which form of R to use?

Unstructured gives you the most information, but generally requires you to estimate the most parameters

- does not work well with more than a handful of timepoints
- try this to get an idea of what the covariance matrix looks like

Diagonal assumes that all timepoints are uncorrelated

- unlikely given our discussion of how people are like themselves

Compound symmetry and **autoregressive** are somewhat in between

- fewer estimated parameters than unstructured
- captures the correlations among time points (unlike diagonal)

3 Inference

3.1 Inference and interpretation

3.1.1 What does R do?

R is the residual variance matrix

In the linear mixed model (as in GLM), the residual variance impacts the standard errors of the fixed effects

So our estimate of the matrix **R** will impact the standard errors (and therefore the significance) of the fixed effects

- but generally does not change the **estimates** of the fixed effects themselves

The variance structure you choose affects what is significant → choose the variance structure that most closely reflects reality to have the most accurate tests of significance

3.1.2 Interpreting fixed effects

Interpret fixed effects just like linear regression effects

Slightly complicated by the "time" variable

- SAS and SPSS want you to enter "time" as a categorical variable, which makes things hard to interpret but there is a work-around
- Have 2 "time" variables, include 1 as a categorical variable and 1 as a continuous variable
- Need to consider the zero-point of "time" and center the "time" variable as appropriate

Note: fixed effects don't really change as you change the R matrix

3.1.3 Interpreting correlated residuals

You can request the estimated **R** matrix

Report the *type* of R matrix you used

- e.g., compound symmetry, unstructured, etc.

Report the values that were estimated

- No tests of significance
- You don't typically interpret these values

R matrix is important because it influences the standard errors of the fixed effects

3.1.4 Stacked dataset

SAS, SPSS, and R require the data to be in **stacked** or **tall** format

- 1 line per subject, per occasion
- Subjects have multiple lines of data

Remember that you have syntax to convert data from wide to stacked

- Don't do it by hand

3.1.5 Software

Both correlated residuals and random effects are **mixed models**

SPSS and SAS: Both types of mixed models use the MIXED procedure

- Random effects (G matrix) with the "random" statement
- Correlated residuals (R matrix) with the "repeated" statement

R: Two different packages and procedures

- Random effects (G matrix) with the `lmer()` function in **lme4**
- Correlated residuals (R matrix) with the `gls()` function in **nlme**

3.1.6 Degrees of freedom

Degrees of freedom how much information is left over after we calculate quantities that we need

- how much information you're really basing your analysis on
- You've already used up some of the information (data) to calculate things, and you can't re-use it

We care about this because it determines the critical values for our statistical tests

3.1.7 Degrees of freedom with correlated observations: hard!

Basic premise for mixed models:

Observations are clustered in some way that makes them not independent

Now we don't have N (total # observations) independent pieces of information

We have fewer – but how many fewer???

Somewhere between the number of clusters (here, people) and the total number of observations

There are a number of different ways to calculate that

Different programs have different methods and different defaults

3.1.8 Significance tests

SPSS

- Satterthwaite degrees of freedom are the only option
- Bootstrapping (used before) **does not work** for this type of model

SAS

- Default is between-within (but other options are available)
- Bootstrapping is an option with additional code

R

- No degrees of freedom are reported
- Bootstrapping is an option

4 Summary and comparison

4.1 Advantages

4.1.1 Advantages

Versus linear regression

- Correctly accounts for non-independence
- Standard errors are (appropriately) inflated

Versus repeated measures ANOVA

- Can use other R forms besides compound symmetry
- No assumptions of sphericity, so no need for adjustments

Versus random effects mixed models

- Complex models converge when random effects models don't

4.1.2 Advantages

Uses all observations

- Doesn't drop participants with fewer observations
- Improved power compared to RM ANOVA

Can conceive of "time" as a continuous variable

- Observations don't need to be equally spaced
- Observations don't need to be the same for everyone

4.2 Shortcomings

4.2.1 Shorcomings

Participants do not need to have the same number of observations across time, but it's *better* if they do (approximately)

Consider: people measured 4 times across some timespan
Person 1: month 0, 1, 2, 3
Person 2: month 0, 2, 4, 5
Person 3: month 0, 2, 3, 4

R is a 6×6 matrix: 0, 1, 2, 3, 4, 5
You can run this model, but the estimates of the correlations among some time points are not very accurate
In this example, only one person has an observation at month 5

- How accurate can correlations that include time point 5 be?

4.2.2 Not a shortcoming necessarily

The R matrix approach frames the correlated observations across time (i.e., the individual people) as a nuisance

- Yes, observations from the same person are correlated and can affect standard errors, but I don't really care about the individual people - I just want accurate standard errors

This is a completely legitimate perspective

But what if you are actually interested in how the individual people differ from one another?

- Individual variability in trajectories with random effects mixed models

4.3 Comparing random effects versus correlated residuals

4.3.1 Continuous, normally distributed outcomes

With continuous and normally distributed outcomes, the fixed effects will be the same for both types of models

- Assuming no missing data, equal number of observations per person, etc.
- As you diverge from that, they will also diverge

Even though they are **different models**, the results are basically the same (numerically)

4.3.2 Non-continuous / non-normal outcomes

With non-normal outcomes (e.g., binary, count) that are modeled using appropriate models for them, the fixed effects will be different

The main reason is that they are **different models**

- They were different models with normally distributed outcomes too, but since the numbers are the same, it's not obvious
- But this only becomes apparent when you have non-normal outcomes

4.3.3 Conditional versus marginal models

Random effects mixed model is a **conditional** model

- Individual level is of interest
- **Conditional on** a person, what is the effect?
- For **this person**, what is the effect of 1 additional hour of studying on test score?
- Estimate individual trajectories, then average those

Correlated residuals model is a **marginal** model

- Group level is of interest
- **Averaging across** all people (i.e., looking at the marginal value), what is the effect?
- For **people**, what is the effect of 1 additional hour of studying on test score?
- Average across all people (ignore individual)

4.3.4 Which to use?

Use random effects models if:

- Individuals may be very different from one another, even within a group (and that's of interest)
- You have predictors at multiple levels (this model has "multiple levels")
- You are interested in "contextual effects"

Use correlated residuals models if:

- Individuals will be relatively similar in their effects (or you **really** don't care how they might differ)
- All predictors are at level 1 (there aren't "multiple levels")
- You have a vary complicated model that's not working as a random effects model

ALWAYS:
Remember that marginal and conditional effects are different