Beyond MLR Lab 6: Longitudinal data analysis

In this lab, we will explore the analysis of longitudinal data using linear mixed-effects models. Longitudinal data involve repeated measurements taken on the same subjects over time, allowing us to study changes in outcomes within individuals.

Adolescent alcohol use data

The dataset we will use comes from a study of adolescent alcohol use, featured in Singer and Willett's *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (2003). This study tracked 82 adolescents over three years (ages 14, 15, and 16) to examine how their alcohol use changed over time.

Dataset description

The dataset contains the following variables:

- **Outcome variable**:
  - `alcuse`: Continuous measure of alcohol use based on various survey items.
- **Grouping variable**:
  - `id`: Unique identifier for each adolescent in the study.
- **Covariates**:
  - `coa`: Dichotomous variable indicating parental alcoholism (1 = yes, 0 = no).
  - `male`: Dichotomous variable indicating male gender (1 = yes, 0 = no).
  - `peer`: Continuous measure of peer alcohol use, assessed at age 14.
  - `age`: Numerical variable representing the age of the adolescent at each time point (14, 15, 16).

Research question

In this lab, we are going to address the following research question:

> Does being a child of an alcoholic parent influence the level and change in alcohol use from ages 14 to 16?

Data import, cleaning, and inspection

```
# Load the required libraries
library(dplyr)
library(ggplot2)
library(lmerTest)
library(haven)

# Load the adolescent alcohol use data. Note that the call below assumes the dataset is place
alcohol_data <- read_sav("data/alcoholpp.sav")

# Convert the dummy coded dichotomous variables into factors
alcohol_data <- alcohol_data %>%
  mutate(
    coa = as.factor(coa),
    male = as.factor(male),
  )

# Inspect the structure of the dataset
str(alcohol_data)
```

Exploratory data analysis

Individual trajectories of alcohol use

We will start by examining the individual trajectories of alcohol use over time. One way to achieve this is a facet plot, which displays each individual's trajectory in a separate subplot. In this case, however, the dataset consists of 82 individuals, making it impractical to display all trajectories. We therefore take a random sample of 16 individuals for visualization.

```r
set.seed(123)  # Set seed for reproducibility

# Sample 16 unique individuals
sampled_ids <- sample(1:82, 16)

# Filter dataset for the sampled IDs and plot
alcohol_data %>%
  filter(id %in% sampled_ids) %>%
  ggplot(aes(x = age, y = alcuse)) +
  geom_line() +                                  # Add trajectories (lines)
  geom_point(size = 2, alpha = 0.7) +            # Add individual data points (dots)
  facet_wrap(~ id) +                             # Create a subplot for each individual
  labs(
    title = "Individual Alcohol Use Trajectories with Data Points",
    x = "Age",
    y = "Alcohol Use"
  ) +
  theme_minimal()
```

Observations:

- Many adolescents report no alcohol use (`alcuse = 0`) across all time points.
- For adolescents who do report alcohol use, the trajectories vary, with some increasing, decreasing, or remaining relatively stable over time.
- This variability underscores the need for statistical models, such as linear mixed-effects models, to account for differences both within and between individuals.

> ⚠ Disclaimer
>
> The dataset used in this lab is zero-inflated, with a large number of zero alcohol use values. While linear mixed-effects models are not necessarily the best approach for analyzing such data, we will use them in this lab to focus on the methodology. The results derived in this lab are intended for educational purposes only.

Mean trajectories by parental alcoholism

Next, we will examine the mean trajectories of alcohol use by parental alcoholism (`COA`). This will provide an overview of how this variable relates to alcohol use over time.

```r
# Compute the mean alcohol use by age, and COA status
mean_alcuse <- alcohol_data %>%
  group_by(age, coa) %>%
  summarise(mean_alcuse = mean(alcuse, na.rm = TRUE))

# Plot the mean trajectories by COA
ggplot(mean_alcuse, aes(x = age, y = mean_alcuse, color = coa)) +
  geom_line() +
  labs(
    title = "Mean Alcohol Use Trajectories by Parental Alcoholism",
    x = "Age",
    y = "Mean Alcohol Use"
  ) +
  scale_color_manual(values = c("0" = "blue", "1" = "red")) +
  ylim(0, 3) +
  theme_minimal()
```

> Question
>
> Based on the plot, do you expect there to be a main effect of parental alcoholism (COA) or age on alcohol use?

> Question
>
> Do you expect there to be an interaction effect between COA and age? Explain your reasoning based on the trends shown in the plot.

Random intercept model

We will start by fitting a random intercept model to the data. This model assumes that each individual has a unique baseline level of alcohol use at age 14, which varies randomly around the population mean. The model can be specified as follows:

```r
# Center age at 14 for interpretability
alcohol_data$age_centered <- alcohol_data$age - 14

# Fit a random intercept model with age centered at 14 and coa
random_intercept_model <- lmer(alcuse ~ age_centered*coa + (1 | id), data = alcohol_data)
summary(random_intercept_model)
```

Question

Calculate the ICC based on the estimated variance components. What proportion of variance in alcohol use is due to differences between individuals?

Random slope model

To extend the previous model to include a random slope, we specify `(1 + age_centered | id)` in the model formula:

```
# Fit the random slope model
random_slope_model <- lmer(alcuse ~ age_centered*coa + (1 + age_centered | id), data = alcoh

# Display the summary of the model
summary(random_slope_model)
```

We begin by assessing the variance components of the model. In addition to including a random intercept for the grouping variable `id`, the model now incorporates a random slope for the `age_centered` variable. This allows the rate of change in alcohol use with age to vary across individuals. Furthermore, the model includes a term for the correlation between the random intercept and the random slope, which captures the relationship between an individual's baseline level of alcohol use and their rate of change over time. This correlation provides insight into whether individuals with higher initial levels of alcohol use are more likely to increase or decrease their usage as they age.

Question

What does the estimated correlation between the random intercept and random slope tell us about the relationship between baseline alcohol use and the rate of change over time?

The inclusion of the correlation between the random intercept and the random slope arises from how we specified the formula object in the model. By default, the random effects are modeled as a multivariate normal distribution, allowing for the estimation of a correlation between the intercept and slope. However, an alternative specification is also possible, where the random intercept and random slope are assumed to be independent. This can be achieved by using a formula that separates the random intercept and slope terms, such as `(1 | id) + (0 + age_centered | id)`. Such a model may be appropriate if there is no theoretical or empirical justification for expecting a relationship between baseline levels and rates of change, or if simplifying the model does not significantly worsen model fit, which can be tested for using a Likelihood ratio test.

Here, we focus on comparing the random slope model against the previously fitted random intercept model. This approach allows us to evaluate whether modeling individual trajectories with these additional random effects provides a significantly better fit to the data compared to a simpler model that accounts only for variability in baseline alcohol use.

```
# Perform the likelihood ratio test
anova(random_intercept_model, random_slope_model, refit = FALSE)
```

> Question
>
> What does the likelihood ratio test tell us about the comparison between the random intercept and random slope models?

Assessing the fixed effects structure

Now that we have determined an appropriate structure for the random effects, we turn our attention to evaluating the fixed effects. The fixed effects represent the average relationships between the predictors and the outcome variable across all individuals in the study. Specifically, this includes the main effects of `age_centered` and `coa` as well as their interaction. By assessing the fixed effects, we aim to determine whether these predictors significantly contribute to explaining variation in alcohol use and whether their inclusion aligns with our theoretical expectations.

To determine whether the relationship between age and alcohol use differs depending on `coa` status, we start by testing the interaction effect between `age_centered` and `coa`:

```
# Obtain the ANOVA table for the random slope model
anova(random_slope_model)
```

> Question
>
> Based on the ANOVA table, what can you conclude about the interaction effect between `age_centered` and `coa` on alcohol use?

Because the interaction effect is insignificant, we can simplify the model by removing this term. We then refit the model and assess the main effects of `age_centered` and `coa`:

```
# Fit the simplified model without the interaction term
simplified_model <- lmer(alcuse ~ age_centered + coa + (1 + age_centered | id), data = alcoho
summary(simplified_model)
```

```r
# Obtain the ANOVA table for the simplified model
anova(simplified_model, refit=TRUE)
```

> Question
>
> What can you conclude about the main effects of age and parental alcoholism on adolescent alcohol use based on the simplified model?

Model building principles for longitudinal data analysis

Model building for longitudinal data analysis follows a systematic approach to ensure that the final model is both parsimonious and adequately represents the data. The process typically involves the following steps:

1. **Start with a saturated fixed-effects structure and a complex random-effects structure**

   - A **saturated fixed-effects structure** includes all plausible predictors and their interactions based on theoretical considerations and prior knowledge. This ensures that no potentially important effect is excluded at the outset.
   - A **complex random-effects structure** allows for a flexible representation of variability in the data. For example, random intercepts and random slopes are included to account for between-subject variability and other grouping factors.

   This initial specification provides a robust starting point for model refinement, ensuring the inclusion of all meaningful variability in the model.

2. **Determine the appropriate random-effects structure**

   - Simplify the random-effects structure by comparing models with different random-effects terms using likelihood ratio tests (LRTs).
   - Importantly, these comparisons are made without refitting the models using maximum likelihood (ML). Restricted maximum likelihood (REML) is retained during this step to ensure accurate estimation of variance components.
   - Remove unnecessary random-effects terms to avoid overfitting while retaining terms that account for substantial variability in the data.

3. **Simplify the fixed-effects structure**

   - Starting from the saturated fixed-effects structure combined with the final reduced random-effects structure, iteratively remove non-significant fixed effects.

- For fixed-effects structure refinement, models must be refitted using ML rather than REML. This ensures correct inference for hypothesis testing and comparison of nested models.
- Once the final fixed-effects structure is determined, the model can be refitted using REML for final parameter estimation.

Additional considerations for model comparison

The model building principles described above apply specifically to the comparison of **nested models**, where one model is a special case of another (e.g., a simpler model can be obtained by constraining parameters in the more complex model). In such cases, likelihood ratio tests (LRTs) provide a robust framework for selecting between models.

However, when comparing **non-nested models**, where one model cannot be derived from the other by parameter constraints, LRTs are not appropriate. Instead, model selection criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) should be used. These criteria provide a balance between model fit and complexity, aiding in the selection of the most parsimonious model that adequately represents the data.