Beyond MLR Lab 6: Longitudinal data analysis

Preliminary setup

In this lab, we will use the R packages `haven`, `ggplot2`, `dplyr`, `emmeans`, `lmerTest`. You can use the script below to automatically install and load them using the `pacman` package.

```
# Check whether pacman is available and install if needed
options(repos = c(CRAN = "https://cloud.r-project.org"))
if (!requireNamespace("pacman", quietly = TRUE)) install.packages("pacman")

# Use pacman to install (if needed) and load the required packages
pacman::p_load(haven, dplyr, ggplot2, emmeans, lmerTest)
```

Adolescent alcohol use data

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.

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).
  - `sex`: Dichotomous variable indicating sex.
  - `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:

> What factors predict adolescent alcohol use and its change over time? Specifically, do parental alcoholism, sex, and peer alcohol use influence baseline levels and trajectories of alcohol use from ages 14 to 16?

Data import, cleaning, and inspection

```
# Load the adolescent alcohol use data
# Note: the call below assumes the dataset is placed in the 'data' folder
# directly above the root folder of the project. Update the path as needed.
alcohol_data <- read_sav("data/alcoholpp.sav")

# Convert labelled variables (dbl+lbl) to factors using their value labels
alcohol_data <- alcohol_data |> mutate(across(where(is.labelled), as_factor))

# Use effects coding for the categorical variables
options(contrasts = c("contr.sum", "contr.poly"))

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

Exploratory data analysis

Overall pattern and individual variation

We begin by examining both the overall mean trajectory and individual trajectories simultaneously. This "spaghetti plot" shows all individual trajectories (thin gray lines) overlaid with the

overall mean trajectory (thick blue line), providing an immediate sense of both the average pattern and the variability around it.

```
# Compute the overall mean alcohol use by age
mean_trajectory <- alcohol_data |>
  group_by(age) |>
  summarise(mean_alcuse = mean(alcuse, na.rm = TRUE), .groups = "drop")

# Create spaghetti plot: individual trajectories + mean trajectory
ggplot(alcohol_data, aes(x = age, y = alcuse)) +
  geom_line(aes(group = id), alpha = 0.3, color = "gray60") +  # Individual trajectories
  geom_line(data = mean_trajectory, aes(x = age, y = mean_alcuse),
            color = "steelblue", linewidth = 1.5) +             # Mean trajectory
  geom_point(data = mean_trajectory, aes(x = age, y = mean_alcuse),
             color = "steelblue", size = 3) +                   # Mean points
  labs(
    title = "Individual and Mean Alcohol Use Trajectories",
    subtitle = "Gray lines = individual trajectories; Blue line = overall mean",
    x = "Age",
    y = "Alcohol Use"
  ) +
  theme_minimal()
```

Question

Based on the spaghetti plot:

1. Does the change in alcohol use over time appear to be linear, or would a quadratic trend be more appropriate?
2. How much variability is there in individual trajectories around the mean? Does this variability suggest the need for random effects in our models?

Examining individual trajectories in detail

While the spaghetti plot gives us a good overview of the overall variability, it can be difficult to distinguish specific individual patterns when all trajectories are overlaid. To get a better sense of the different types of individual trajectories, we'll examine a random sample of 16 individuals using a facet plot.

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

```r
# 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 = "Sample of Individual Alcohol Use Trajectories",
    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 show diverse patterns: some increasing, some decreasing, and some remaining relatively stable over time.
- The heterogeneity in both baseline levels and rates of change motivates the use of random intercepts and random slopes in our models.

> ⚠ 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), .groups = "drop")
```

```r
# 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"
  ) +
  ylim(0, 3) +
  theme_minimal()
```

Question

Based on the plot, do you expect there to be a effect of parental alcoholism on alcohol use? Do you expect there to be an interaction effect between parental alcoholism and age? Explain your reasoning based on the trends shown in the plot.

Coding exercise

Create similar plots to explore mean trajectories by `sex` and by `peer` alcohol use.
Hint for peer alcohol use: Since `peer` is a continuous variable, you can categorize it into tertiles (low, medium, high) for visualization purposes using the following code:

```r
alcohol_data <- alcohol_data |>
  mutate(peer_cat = cut(peer,
                        breaks = quantile(peer, probs = c(0, 1/3, 2/3, 1), na.rm = TRUE),
                        labels = c("Low", "Medium", "High"),
                        include.lowest = TRUE))
```

Then compute mean trajectories by `age` and `peer_cat`, and create a plot similar to the one above.

Modeling the time structure

Before adding covariates, we first establish the appropriate model for time. This "time structure first" approach ensures that we correctly capture how the outcome changes over time before investigating what factors might explain individual differences.

Step 1: Fit a flexible model for time

We start with a model that includes:

- **Saturated fixed effects for time**: linear and quadratic terms (with 3 time points, this perfectly captures the mean trajectory)
- **Random intercept and random slope**: allowing individuals to differ in both their baseline level and their rate of change

```
# Center age at 15 (middle time point)
alcohol_data$age_centered <- alcohol_data$age - 15

# Create quadratic term
alcohol_data$age_centered_sq <- alcohol_data$age_centered^2

# Fit the full model with random intercept and random slope
time_model_full <- lmer(alcuse ~ age_centered + age_centered_sq + (age_centered | id),
                        data = alcohol_data,
                        control = lmerControl(optimizer = "bobyqa"))
summary(time_model_full)
```

Using the bobyqa optimizer

We use `control = lmerControl(optimizer = "bobyqa")` to specify the BOBYQA (Bound Optimization BY Quadratic Approximation) optimizer. This derivative-free optimizer is often more robust than the default, especially for models with random slopes or when predictor variables are on different scales. It can help avoid convergence warnings that sometimes occur with the default optimizer.

Understanding the random effects syntax

The term `(age_centered | id)` specifies the random effects structure. This is shorthand for `(1 + age_centered | id)`, which includes:

- **Random intercept** (`1`): each individual has their own baseline level
- **Random slope** (`age_centered`): each individual has their own rate of change over time
- **Correlation**: the intercept and slope are allowed to be correlated (e.g., do individuals who start higher also change faster?)

Compare this to `(1 | id)` used in previous labs, which only included a random intercept. The random slope is the key new element for longitudinal data, capturing the **subject × time interaction** discussed in the lecture.

Uncorrelated random effects: If you want to assume the intercept and slope are uncorrelated, use `(1 | id) + (0 + age_centered | id)` or equivalently `(age_centered || id)`. This estimates separate variances but fixes the correlation to zero, reducing the number of parameters by one. This is sometimes used when the model with correlated random effects fails to converge, or when there is no theoretical reason to expect a correlation between baseline level and rate of change.

Why not random quadratic slopes?

With only 3 observations per person, we cannot reliably estimate individual-specific quadratic trajectories (syntax: `(age_centered + age_centered_sq | id)`). The issue is one of **identifiability**: with 3 time points, we cannot distinguish between "person i has a different curvature" and "person i has random measurement error."
General guideline:

- 3 time points: Random intercept + random linear slope is feasible
- 4+ time points: Could consider random quadratic slopes (though rarely needed)
- The quadratic term remains as a **fixed effect** to model the average curvature

Step 2: Test the random slope

We compare the full model (with random slope) against a model with only a random intercept using a likelihood ratio test. Since we are comparing random effects, we use `refit = FALSE` to retain REML estimation.

```
# Fit a model with only random intercept
time_model_ri <- lmer(alcuse ~ age_centered + age_centered_sq + (1 | id),
                      data = alcohol_data,
                      control = lmerControl(optimizer = "bobyqa"))

# Compare models using likelihood ratio test
anova(time_model_ri, time_model_full, refit = FALSE)
```

Question

Based on the likelihood ratio test, is the random slope for time necessary? What does this tell us about whether individuals differ in their rate of change in alcohol use?

Step 3: Calculate the ICC

Regardless of the random slope test result, we calculate the ICC from the random intercept model. This tells us what proportion of the variance in alcohol use (after accounting for time trends) is due to stable differences between individuals.

```
# Extract variance components from the random intercept model
vc <- as.data.frame(VarCorr(time_model_ri))
var_intercept <- vc$vcov[1]
var_residual <- vc$vcov[2]

# Calculate ICC
icc <- var_intercept / (var_intercept + var_residual)
cat("ICC:", round(icc, 3), "\n")
```

> Question
>
> Interpret the ICC. What does it tell you about the relative importance of between-person versus within-person variation in alcohol use?

Step 4: Test the fixed effects for time

Now we test whether the quadratic and linear time components are needed. We use likelihood ratio tests with `refit = TRUE` because we are comparing models with different fixed effects. We keep the random slope structure throughout.

```
# Test quadratic term: compare model with vs without age_centered_sq
time_model_linear <- lmer(alcuse ~ age_centered + (age_centered | id), data = alcohol_data,
                          control = lmerControl(optimizer = "bobyqa"))
anova(time_model_linear, time_model_full, refit = TRUE)
```

> Question
>
> Is the quadratic term significant? What does this tell you about the shape of the average trajectory?

```
# Test linear term: compare model with vs without age_centered
time_model_null <- lmer(alcuse ~ 1 + (age_centered | id), data = alcohol_data,
                        control = lmerControl(optimizer = "bobyqa"))
anova(time_model_null, time_model_linear, refit = TRUE)
```

Question

Is the linear term significant? What does this tell you about whether alcohol use changes over time on average?

Step 5: Residual diagnostics

After determining the time structure, we check whether the model adequately captures the time trend by examining the residuals.

```
# Use the final time model (based on tests above, linear model with random slopes)
final_time_model <- time_model_linear

# Create a data frame for plotting
diag_data <- data.frame(
  fitted = fitted(final_time_model),
  residuals = resid(final_time_model)
)

# Residuals vs Fitted using ggplot2
ggplot(diag_data, aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linewidth = 1) +
  geom_smooth(method = "loess", se = TRUE, color = "blue", fill = "lightblue") +
  labs(
    title = "Residuals vs Fitted",
    x = "Fitted values",
    y = "Conditional Residuals"
  ) +
  theme_minimal()
```

Question

Does the loess curve (blue line) lie close to zero across the range of fitted values?

⚠ Interpreting the diagnostic plot

The residual plot reveals patterns that are largely driven by the **zero-inflated nature** of the data. The clustering of points at low fitted values reflects the many adolescents with zero alcohol use. The loess curve dipping below zero at low fitted values and rising above zero at higher values, along with the increasing spread of residuals (heteroscedasticity),

are typical of zero-inflated outcomes. While a linear mixed model provides a useful approximation for teaching purposes, more appropriate models for these data might include two-part (hurdle) models or zero-inflated models.

Adding covariates

Now that we have established the appropriate time structure, we can investigate which covariates predict alcohol use and whether they modify the effect of time (cross-level interactions).

Step 1: Fit a full model with all covariates

We start with a model that includes all predictors and their interactions with time, combined with the random-effects structure and fixed time structure determined above (linear time with random intercept and random slope).

```
# Center peer at its mean for interpretability
alcohol_data$peer_centered <- alcohol_data$peer - mean(alcohol_data$peer, na.rm = TRUE)

# Fit the full model with all predictors and their interactions with time
# Fixed effects: linear time only (quadratic was not significant in Step 4)
# Random effects: intercept and slope for age_centered (based on the test in Step 2)
full_model <- lmer(
  alcuse ~ age_centered * (coa + sex + peer_centered) + (age_centered | id),
  data = alcohol_data,
  control = lmerControl(optimizer = "bobyqa")
)
summary(full_model)
```

Step 2: Simplify the fixed-effects structure

We use the `step()` function to perform backward elimination of non-significant fixed effects.

```
# Use step() to simplify fixed effects
step_result <- step(full_model, reduce.fixed = TRUE, reduce.random = FALSE)
print(step_result)

# Extract the final model
final_model <- get_model(step_result)
```

The `step()` function performs backward elimination of fixed effects, adhering to important model building principles such as the marginality principle (lower-order terms are retained if higher-order terms involving them remain in the model) and uses F-tests with Satterthwaite approximation for hypothesis testing.

> Question
>
> Based on the output from `step()`, which fixed effects terms were removed from the model? Does the final model include any interaction effects with time (cross-level interactions), or only main effects?

Step 3: Final model and interpretation

Based on the model building process, we arrive at our final model. We can now interpret the results in detail:

```
# Display the final model summary
summary(final_model)
```

> Question
>
> Based on the final model, what can you conclude about the factors that predict adolescent alcohol use and its change over time? Consider:
>
> - Which predictors have significant effects on alcohol use?
> - Are there differences between groups (e.g., children of alcoholic parents vs. not, sex differences)?
> - Do any predictors interact with time (age), suggesting different trajectories for different groups?