Beyond MLR Lab 3: Two-Way Factorial ANOVA

Preliminary setup

In this lab, we will use the R packages `ggplot2`, `dplyr`, and `emmeans`. 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(dplyr, ggplot2, emmeans)
```

Example: Effect of Rehabilitation Therapy Type and Disease Severity on Post-Stroke Recovery

In this lab, we will examine how the combination of rehabilitation therapy type and disease severity influences patient recovery rates after stroke. The data for this example are hypothetical and will be created through simulation:

```
# Generate some hypothetical data
set.seed(567)
Therapy <- factor(rep(c("Standard", "Enhanced", "Control"), each = 20),
                  levels = c("Standard", "Enhanced", "Control"))
Severity <- factor(rep(rep(c("Mild", "Severe"), each = 10), 3),
                   levels = c("Mild", "Severe"))

mu <- expand.grid(Therapy = levels(Therapy), Severity = levels(Severity))
mu$mean <- c(45, 42, 72, 58, 20, 18)
```

```r
recovery_rate <- rnorm(60, mean = rep(mu$mean, each = 10), sd = 5)

# Combining data into a data frame
data_twoway <- data.frame(
  Therapy = Therapy,
  Severity = Severity,
  RecoveryRate = recovery_rate
)
```

The R chunk above simulates patient recovery rate observations for 60 stroke patients in a two-way factorial design with the following characteristics:

- `Therapy`: Factor variable with three levels (Standard, Enhanced, Control) representing the type of rehabilitation therapy
- `Severity`: Factor variable with two levels (Mild, Severe) representing disease severity
- `RecoveryRate`: Numeric variable representing the percentage improvement in functional status after 12 weeks of rehabilitation

The design consists of 6 treatment combinations (3 therapy types $\times$ 2 severity levels), with 10 patients randomly assigned to each combination.

Exploratory Data Analysis

To understand how recovery rates vary across therapy type and disease severity, we start by creating an **interaction plot**. An interaction plot is a graphical tool that helps visualize whether the effect of one factor (in this case, therapy) depends on the levels of another factor (in this case, disease severity). If the lines on the interaction plot are parallel, this suggests there is no interaction, meaning the effect of therapy is the same for both disease severities. If the lines are not parallel, this indicates a potential interaction, meaning the effect of therapy differs depending on the severity of the disease.

```r
# Visualizing the results
ggplot(data_twoway, aes(x = Therapy, y = RecoveryRate, color = Severity, group = Severity)) +
  stat_summary(fun = mean, geom = "point", size = 3) +
  stat_summary(fun = mean, geom = "line") +
  labs(title = NULL,
       x = "Therapy Type",
       y = "Mean recovery rate (%)",
       color = "Disease Severity") +
  theme_minimal()
```

We also calculate summary statistics for each combination of therapy type and disease severity:

```r
# Generating summary statistics by Therapy and Severity
summary_stats <- data_twoway |>
  group_by(Therapy, Severity) |>
  summarise(
    n = n(),
    Mean_RecoveryRate = mean(RecoveryRate),
    SD_RecoveryRate = sd(RecoveryRate),
    .groups = "drop"
  )
summary_stats
```

> Question
>
> Based on the plot and summary statistics, does there appear to be an interaction between therapy type and disease severity?

Two-Way Factorial ANOVA Model

Model Specification

Using effects coding, the two-way factorial Analysis of Variance (ANOVA) model can be specified as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where:

- Y_{ijk}: The recovery rate for the k-th patient with therapy type i and disease severity level j.
- μ: The grand mean (overall average recovery rate across all groups).
- α_i: The main effect of therapy type i, representing the deviation from the grand mean. Constraint: $\sum_i \alpha_i = 0$.
- β_j: The main effect of disease severity level j, representing the deviation from the grand mean. Constraint: $\sum_j \beta_j = 0$.
- $(\alpha\beta)_{ij}$: The interaction effect between therapy type i and severity level j, representing the deviation from additivity. Constraint: $\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$.
- ϵ_{ijk}: The residual error term, assumed to be independent and normally distributed with mean zero and constant variance.

Model Estimation

We can fit the two-way ANOVA model using the `lm()` function. The key feature of the model formula is the interaction term (specified using `*`), which includes both main effects and their interaction:

```
options(contrasts = c("contr.sum", "contr.poly"))  # Effects coding

# Fit the two-way ANOVA model with interaction
model_twoway <- lm(RecoveryRate ~ Therapy * Severity, data = data_twoway)
summary(model_twoway)
```

ANOVA Table and Global Hypothesis Tests

To test whether the main effects and interaction are statistically significant, we use the `anova()` function to obtain the ANOVA table:

```
# ANOVA table
anova_results <- anova(model_twoway)
anova_results
```

The ANOVA table tests three global hypotheses:

1. No interaction: $H_0 : (\alpha\beta)_{ij} = 0$ for all i, j
2. No main effect of Therapy: $H_0 : \alpha_i = 0$ for all i
3. No main effect of Severity: $H_0 : \beta_j = 0$ for all j

> Question
>
> What do the F-tests and p-values in the ANOVA table tell us about the significance of the interaction and main effects?

Simple Effects Analysis

Because the interaction effect is significant, we proceed with examining the simple effects: the effect of one factor within each level of the other factor. When analyzing an interaction, we have a choice in how to decompose it. We could examine either the effect of therapy type (Standard vs Enhanced vs Control) separately for each disease severity level, or alternatively, we could examine the effect of disease severity (Mild vs Severe) separately for each therapy type.

In this clinical context, we choose to examine the effect of therapy type conditional on disease severity level. This choice is clinically motivated: from a treatment planning perspective, clinicians need to know which therapies work best for patients presenting with different disease severity levels. Understanding how therapy effectiveness varies across severity groups directly informs treatment decisions and allows for severity-stratified recommendations. In other words, the question "which therapy works best for mild cases versus severe cases?" is more clinically actionable than asking "how does severity affect outcomes for each therapy type?"

```
# Compute simple effects: effect of therapy type within each severity level
se_severity <- emmeans(model_twoway, ~ Therapy | Severity)
se_severity
```

To test which simple effects are significant, we perform pairwise comparisons within each severity level (using the Bonferroni correction to adjust for multiplicity):

```
# Pairwise comparisons (Therapy types) within each disease severity level
se_pairs <- pairs(se_severity, adjust = "bonferroni")
se_pairs
```

> Question
>
> For which disease severity level(s) does therapy type have a significant effect on recovery rate? What is the practical interpretation of this finding?

Alternative Decomposition: Simple Effects Conditional on Therapy Type

As mentioned earlier, we could also decompose the interaction by examining the effect of disease severity separately for each therapy type. While this perspective is less clinically motivated than the severity-stratified approach above, it can still provide useful insights.

> Exercise
>
> Perform a simple effects analysis conditional on therapy type (i.e., examine the effect of disease severity (Mild vs Severe) within each therapy type).

> Question
>
> For which therapy type(s) does disease severity have a significant effect on recovery rate?

Question

How do these findings complement or differ from the severity-stratified analysis we performed above?

Model Diagnostics

To assess the adequacy of the fitted model, we create two diagnostic plots:

- **Residuals vs Fitted Plot**: Used to assess homoscedasticity (constant variance). A random scatter around zero suggests equal variance.
- **Normal Q-Q Plot**: Used to assess normality of the errors. Points following the reference line suggest normally distributed residuals.

```
# Residuals vs Fitted
plot(model_twoway, which = 1, main = "Residuals vs Fitted")

# Normal Q-Q
plot(model_twoway, which = 2, main = "Normal Q-Q")
```

Question

Do the diagnostic plots suggest any violations of the model assumptions?