Beyond MLR Lab 1: One-way ANOVA

Welcome to lab 1 of the Beyond MLR course. 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.

```
# Install pacman if not already available
if (!require("pacman")) install.packages("pacman")

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

Example: Effect of Three Different Drug Treatments on Blood Pressure Reduction

In this lab, we will examine the effect of three different drug treatments (Control, Drug A, and Drug B) on the change in systolic blood pressure (SBP) from baseline to week 52 in patients with hypertension. The data for this example are hypothetical and will be created through simulation:

```
# Generating data for the single-factor experimental design
set.seed(123)
control <- rnorm(40, mean = -2, sd = 10)
drugA <- rnorm(40, mean = -20, sd = 10)
drugB <- rnorm(40, mean = -5, sd = 10)

# Combining data into a data frame
data_oneway <- data.frame(
  Treatment = factor(rep(c("Control", "DrugA", "DrugB"), each = 40),
                     levels = c("Control", "DrugA", "DrugB")),
  BP_change = c(control, drugA, drugB)
)
```

The above R chunk simulates hypothetical results for 120 patients (40 per treatment group) and stores the results in the data frame `data_oneway` that consists of the following two columns:

- `Treatment`: Factor variable indicating the treatment group.
- `BP_change`: Numeric variable representing the change in SBP (in mm Hg).

Exploratory Data Analysis

To understand the distribution of blood pressure reduction across treatment groups, we start by creating a boxplot:

```
# Visualizing the results
ggplot(data_oneway, aes(x = Treatment, y = BP_change, fill = Treatment)) +
  geom_boxplot() +
  labs(title = NULL,
       x = "Treatment", y = "Change in SBP (mm Hg)") +
  theme_minimal() +
  theme(legend.position = "none")
```

We also calculate some summary statistics:

```
# Generating summary statistics
summary_stats <- data_oneway |>
  group_by(Treatment) |>
  summarise(
    Mean_BP_change = mean(BP_change),
    SD_BP_change = sd(BP_change)
  )
summary_stats
```

> Question
>
> How do the mean changes in SBP compare among the treatment groups?

Performing a One-Way ANOVA

Model Specification

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

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where:

- Y_{ij}: The change in SBP for the j-th patient in the i-th treatment group.
- μ: The mean change in SBP across all treatment groups.
- τ_i: The effect of the i-th treatment (deviation from the overall mean).
- ϵ_{ij}: The error term, assumed to be independently and normally distributed with mean zero and constant variance.

> **Note 1: Coding schemes for categorical variables**
>
> In regression models, categorical variables are represented using coding schemes that allow their inclusion as explanatory variables. In R, dummy coding is the default coding scheme for nominal variables, where each level is compared to a reference category. For ordinal variables (i.e., ordered factors), R uses orthogonal polynomials as the default coding scheme, which can capture linear or nonlinear trends across ordered levels.
> While dummy coding is widely used in multiple linear regression models, effects coding is more common in ANOVA models as it enables interpretation of group differences relative to the grand mean rather than a single reference category. This interpretative difference makes effects coding particularly useful for examining group-level effects in experimental designs.
> To change the coding scheme globally to effects coding:
>
> ```
> # Set contrasts to effects coding globally
> options(contrasts = c("contr.sum", "contr.poly"))  # Effects coding
> ```
>
> To reset the coding scheme to the default dummy coding:
>
> ```
> # Reset contrasts to default dummy coding
> options(contrasts = c("contr.treatment", "contr.poly"))  # Dummy coding
> ```

Model Estimation

We can fit the one-way ANOVA model using the `lm()` function:

```
options(contrasts = c("contr.sum", "contr.poly"))  # Effects coding
contrasts(data_oneway$Treatment) # Inspect contrasts for Treatment

# Fitting the model with effects coding
```

```r
model_oneway <- lm(BP_change ~ Treatment, data = data_oneway)
summary(model_oneway)
```

> Question
>
> How do the estimated coefficients relate to the group means?

ANOVA Table

To test whether the effect of treatment is statistically significant, we use the `anova()` function to obtain the ANOVA table:

```r
# ANOVA table
anova_results <- anova(model_oneway)
anova_results
```

> Question
>
> What does the F-test tell us about the treatment effect?

Estimated Marginal Means

Estimated marginal means, also known as least-squares means, are model-based means that represent the predicted (or expected) response at each level of a factor, averaged over the levels of other variables in the model.

In the context of a one-way ANOVA, where there is only one factor (e.g., treatment group), the estimated marginal means are equal to the observed group means. In more complex models with multiple factors, estimated marginal means provide predicted group means that are averaged over the levels of these additional factors. For example, in a model with both treatment and age as factors, the estimated marginal means for the treatment groups show the treatment means averaged over age, illustrating what these group means are expected to be at specific values of age (or averaged over a grid of age values).

We can calculate the estimated marginal means using the `emmeans()` function from the `emmeans` package:

```r
# Obtain estimated marginal means
emms <- emmeans(model_oneway, ~ Treatment)
emms  # displays EMMs with standard errors and 95% confidence intervals
```

Question

How do the estimated marginal means compare to the observed group means calculated during the exploratory data analysis?

Pairwise comparisons

Since we found a significant treatment effect, we will conduct pairwise comparisons of the estimated marginal means to identify which specific treatment groups differ significantly from one another. We perform these pairwise comparisons using the `pairs()` function from the `emmeans` package. To account for the increased risk of Type I error due to multiple comparisons, we apply the Bonferroni correction to adjust the p-values, ensuring that the overall significance level remains controlled.

```
# Performing post-hoc analysis with emmeans
pairs(emms, adjust="Bonferroni")
```

Question

Based on the results of the pairwise comparisons, which treatment groups differ significantly?

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) for the errors. A random scatter of residuals around zero indicates equal variance.
- **Normal Q-Q Plot**: Used to assess normality of the errors. Residuals following the reference line suggest normally distributed errors.

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

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

Question

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

Reporting

A one-way ANOVA showed a statistically significant effect of treatment on the change in systolic blood pressure, $F(2, 117) = 43.86$, $p < 0.001$. Estimated marginal means (95% CI) were -20.1 (-22.9, -17.2) mm Hg for Drug A, -4.9 (-7.8, -2.1) mm Hg for Drug B, and -1.6 (-4.4, 1.3) mm Hg for the Control group. Pairwise comparisons (Bonferroni-adjusted) indicated statistically significant differences between Drug A and both Drug B ($p < 0.001$) and the Control group ($p < 0.001$), but not between Drug B and the Control group ($p = 0.291$).