Analysis of Variance (ANOVA)

ANOVA using R

Kamarul Imran Musa

2025-08-10

Table of contents

1 Introduction 1
1.1 Learning Objectives 1
2 Definition 2
2.1 Mathematical Foundation 2
2.2 Motivation for Using ANOVA 2
2.2.a Why Not Multiple t-tests? 2
2.2.b Advantages of ANOVA 2
3 Types of ANOVA 2
3.1 1. One-Way ANOVA 2
3.2 2. Two-Way ANOVA 5
3.3 3. One-Way ANCOVA (Analysis of Covariance) 9
3.4 4. Two-Way ANCOVA 11
4 ANOVA Assumptions 12
4.1 The Four Key Assumptions 12
4.2 Diagnostic Procedures 12
4.3 Dealing with Assumption Violations 15
5 Post-Hoc Analysis 16
5.1 Why Post-Hoc Tests Are Important 16
5.2 Common Post-Hoc Tests 16
6 Practical Example: Complete Analysis 19
7 Summary and Best Practices 23
7.1 When to Use Each Type of ANOVA 23
7.2 Recommendations 23
7.3 R Code Best Practices 23
8 Conclusion 24

1 Introduction

1.1 Learning Objectives

By the end of this session, students will be able to:

1. **Define** Analysis of Variance (ANOVA) and explain its purpose in statistical analysis
2. **Identify** appropriate situations for using different types of ANOVA in epidemiological studies
3. **Distinguish** between one-way, two-way ANOVA, and ANCOVA designs
4. **Apply** ANOVA techniques using R programming with tidyverse principles
5. **Evaluate** ANOVA assumptions and implement appropriate diagnostic procedures
6. **Interpret** ANOVA results including F-statistics, p-values, and effect sizes
7. **Conduct** post-hoc analyses when significant differences are detected
8. **Address** violations of ANOVA assumptions using appropriate statistical methods

2 Definition

Analysis of Variance (ANOVA) is a statistical method used to test whether there are statistically significant differences between the means of three or more groups. Despite its name suggesting analysis of variance, ANOVA actually compares means by analyzing the variance within and between groups.

2.1 Mathematical Foundation

ANOVA partitions the total variation in the data into components:

$$\text{Total Variation} = \text{Between-Group Variation} + \text{Within-Group Variation}$$

The F-statistic is calculated as:

$$F = \frac{\text{Mean Square Between Groups}}{\text{Mean Square Within Groups}} = \frac{MSB}{MSW}$$

2.2 Motivation for Using ANOVA

2.2.a Why Not Multiple t-tests?

When comparing multiple groups, conducting multiple pairwise t-tests leads to:

1. **Inflated Type I Error Rate**: With k groups, we need $\binom{k}{2}$ comparisons
   - For 3 groups: 3 comparisons, $\alpha = 1 - (1\text{-}0.05)^3 \approx 0.14$
   - For 5 groups: 10 comparisons, $\alpha \approx 0.40$
2. **Loss of Statistical Power**: Multiple testing corrections reduce power

2.2.b Advantages of ANOVA

- Controls Type I error rate at the specified α level
- More efficient use of data
- Can examine multiple factors simultaneously
- Provides omnibus test before specific comparisons

3 Types of ANOVA

3.1 1. One-Way ANOVA

Purpose: Compare means across three or more independent groups for one factor.

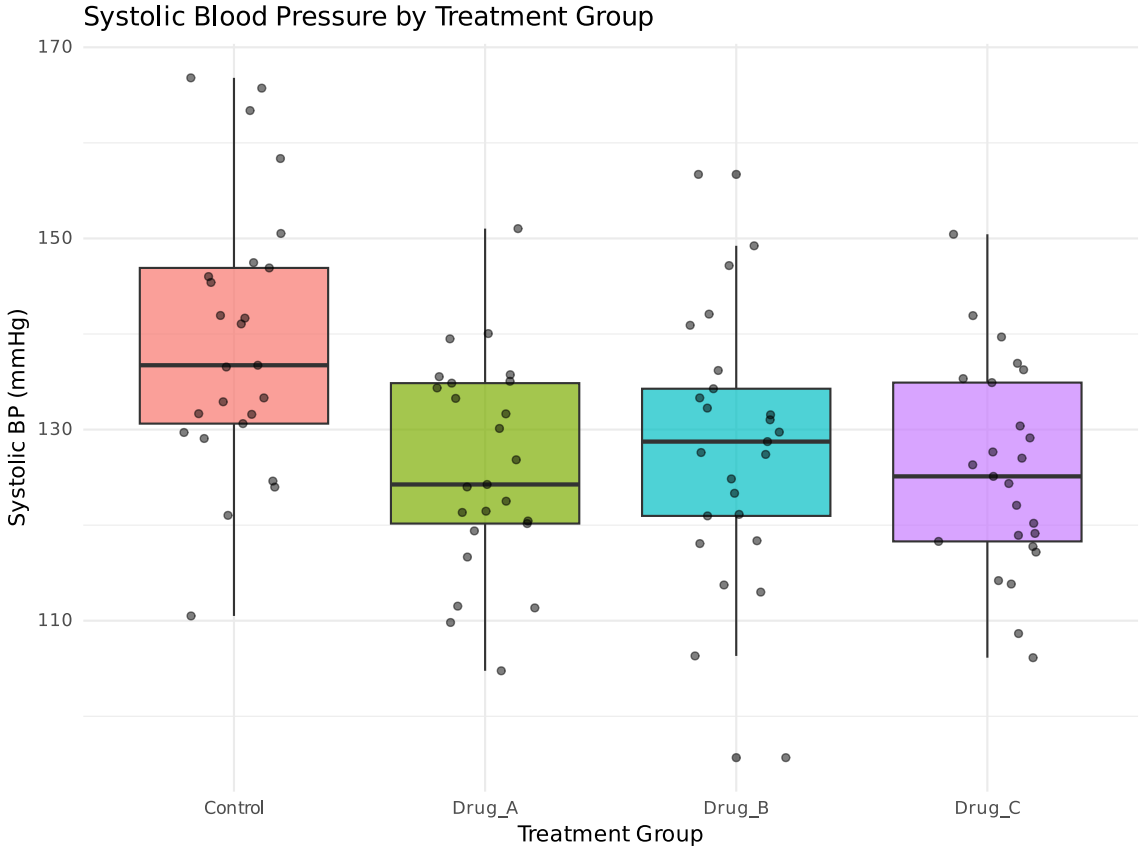
Research Question: "Do mean blood pressure levels differ significantly across different treatment groups?"

```r
# Simulate data: Blood pressure across 4 treatment groups
set.seed(123)
bp_data <- tibble(
  treatment = rep(c("Control", "Drug_A", "Drug_B", "Drug_C"), each = 25),
  systolic_bp = c(
    rnorm(25, 140, 15),  # Control
    rnorm(25, 125, 12),  # Drug A
    rnorm(25, 128, 14),  # Drug B
    rnorm(25, 122, 13)   # Drug C
  ),
  patient_id = 1:100
)

# Exploratory Data Analysis
bp_summary <- bp_data %>%
  group_by(treatment) %>%
  summarise(
    n = n(),
    mean_bp = mean(systolic_bp),
    sd_bp = sd(systolic_bp),
    se_bp = sd_bp / sqrt(n),
    .groups = "drop"
  )

print(bp_summary)
```

```
# A tibble: 4 × 5
  treatment     n mean_bp sd_bp se_bp
  <chr>     <int>   <dbl> <dbl> <dbl>
1 Control      25    140.  14.2  2.84
2 Drug_A       25    126.  11.0  2.21
3 Drug_B       25    128.  13.6  2.72
4 Drug_C       25    126.  10.8  2.16
```

```r
# Visualization
bp_data %>%
  ggplot(aes(x = treatment, y = systolic_bp, fill = treatment)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  labs(
    title = "Systolic Blood Pressure by Treatment Group",
    x = "Treatment Group",
    y = "Systolic BP (mmHg)",
```

```
    fill = "Treatment"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```




```
# One-way ANOVA
anova_model <- aov(systolic_bp ~ treatment, data = bp_data)
anova_summary <- summary(anova_model)
print(anova_summary)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
treatment    3   3165  1055.1   6.751 0.000352 ***
Residuals   96  15003   156.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Using broom for tidy output
anova_tidy <- tidy(anova_model)
print(anova_tidy)
```

```
# A tibble: 2 × 6
  term         df  sumsq meansq statistic   p.value
  <chr>     <dbl>  <dbl>  <dbl>     <dbl>     <dbl>
1 treatment     3  3165.  1055.      6.75  0.000352
2 Residuals    96 15003.   156.     NA    NA
```

```r
# Effect size (eta-squared)
eta_squared <- anova_tidy$sumsq[1] / sum(anova_tidy$sumsq)
cat("Eta-squared (effect size):", round(eta_squared, 3), "\n")
```

```
Eta-squared (effect size): 0.174
```

Interpretation: - F-statistic tests the null hypothesis that all group means are equal - A significant p-value indicates at least one group differs from others - Effect size (η^2) indicates practical significance

3.2 2. Two-Way ANOVA

Purpose: Examine the effects of two factors and their interaction.

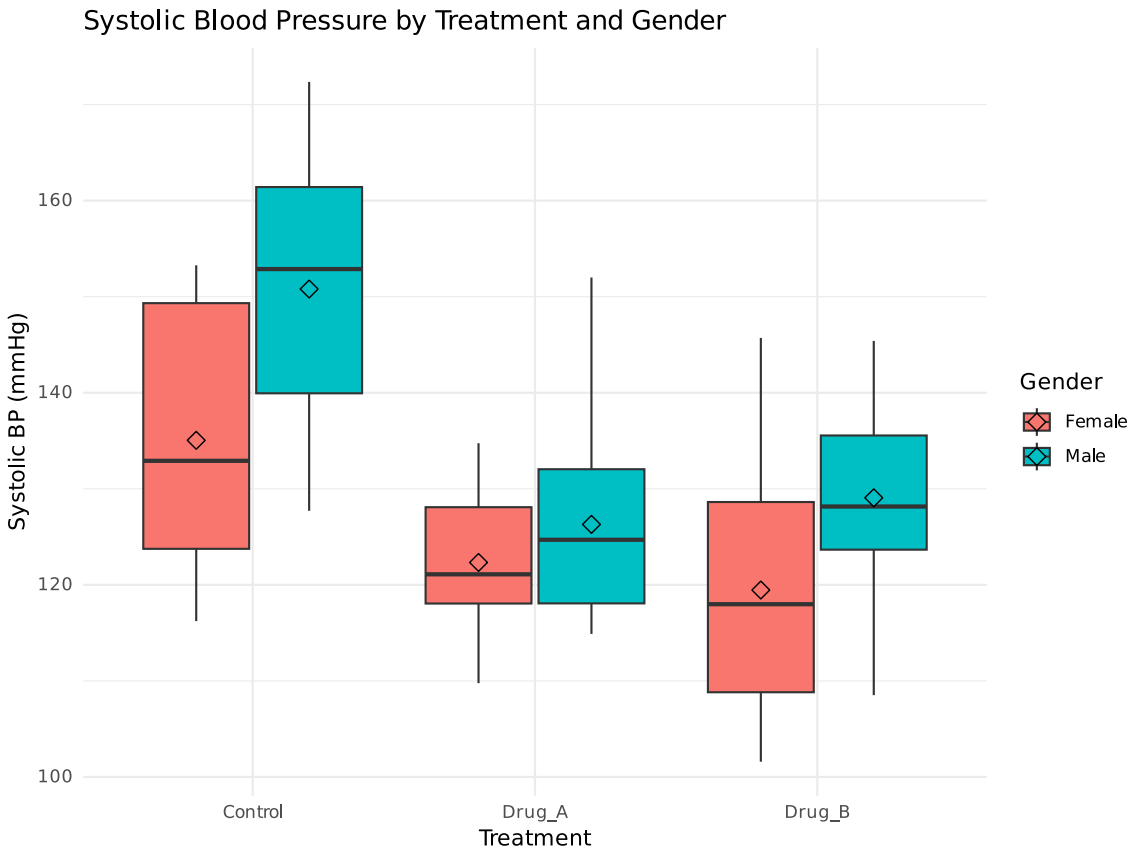
Research Question: "Do blood pressure levels differ by treatment and gender, and is there an interaction?"

```r
# Simulate data with two factors
set.seed(456)
bp_data_2way <- expand_grid(
  treatment = c("Control", "Drug_A", "Drug_B"),
  gender = c("Male", "Female")
) %>%
  slice(rep(1:n(), each = 20)) %>%
  mutate(
    patient_id = 1:n(),
    # Main effects and interaction
    systolic_bp = case_when(
      treatment == "Control" & gender == "Male" ~ rnorm(n(), 145, 12),
      treatment == "Control" & gender == "Female" ~ rnorm(n(), 135, 12),
      treatment == "Drug_A" & gender == "Male" ~ rnorm(n(), 125, 10),
      treatment == "Drug_A" & gender == "Female" ~ rnorm(n(), 120, 10),
      treatment == "Drug_B" & gender == "Male" ~ rnorm(n(), 130, 11),
      treatment == "Drug_B" & gender == "Female" ~ rnorm(n(), 118, 11)
    )
```

```r
  )

# Summary statistics
bp_summary_2way <- bp_data_2way %>%
  group_by(treatment, gender) %>%
  summarise(
    n = n(),
    mean_bp = mean(systolic_bp),
    sd_bp = sd(systolic_bp),
    .groups = "drop"
  )

print(bp_summary_2way)
```

```
# A tibble: 6 × 5
  treatment gender     n mean_bp sd_bp
  <chr>     <chr>  <int>   <dbl> <dbl>
1 Control   Female    20    135. 12.9
2 Control   Male      20    151. 13.9
3 Drug_A    Female    20    122.  6.77
4 Drug_A    Male      20    126.  9.81
5 Drug_B    Female    20    119. 12.2
6 Drug_B    Male      20    129.  9.33
```

```r
# Visualization
bp_data_2way %>%
  ggplot(aes(x = treatment, y = systolic_bp, fill = gender)) +
  geom_boxplot(position = position_dodge(0.8)) +
  stat_summary(fun = mean, geom = "point",
               position = position_dodge(0.8), size = 3, shape = 23) +
  labs(
    title = "Systolic Blood Pressure by Treatment and Gender",
    x = "Treatment",
    y = "Systolic BP (mmHg)",
    fill = "Gender"
  ) +
  theme_minimal()
```




```r
# Two-way ANOVA
anova_2way <- aov(systolic_bp ~ treatment * gender, data = bp_data_2way)
summary(anova_2way)
```

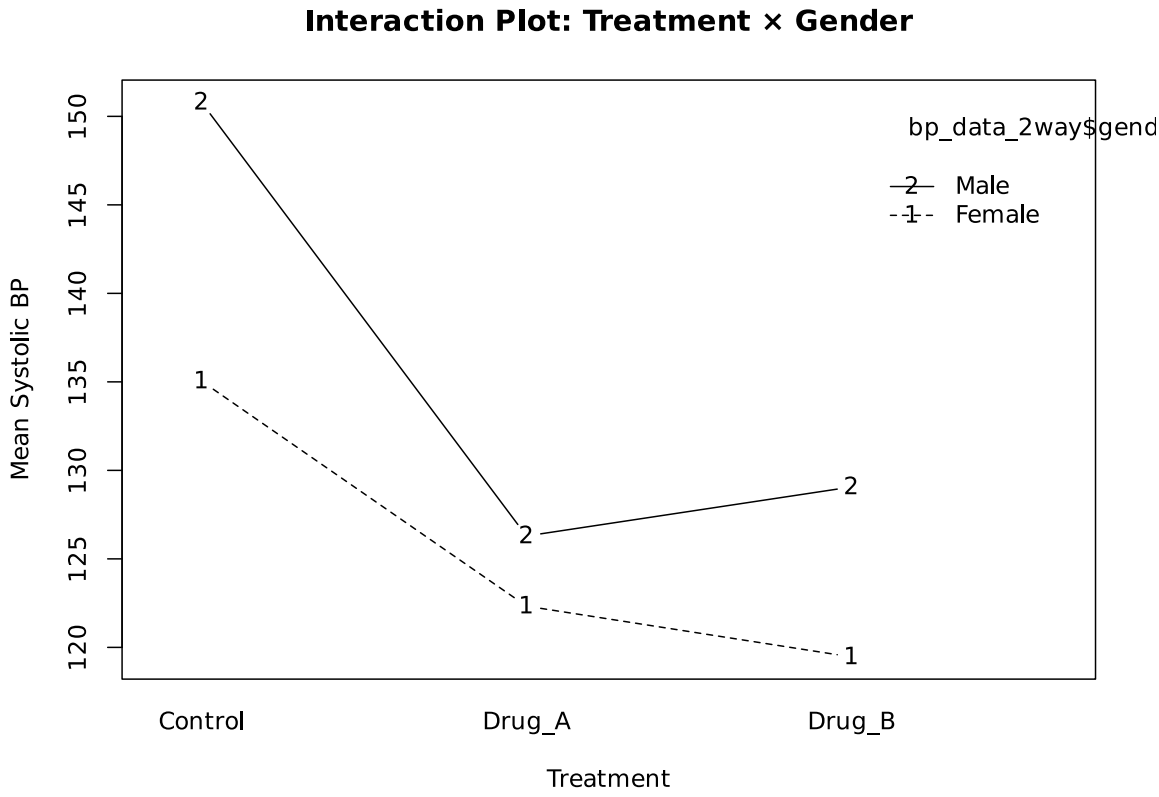
```
                 Df Sum Sq Mean Sq F value   Pr(>F)
treatment         2   9260    4630  37.658 2.78e-13 ***
gender            1   2865    2865  23.299 4.34e-06 ***
treatment:gender  2    695     348   2.828   0.0633 .
Residuals       114  14016     123
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Type III ANOVA (balanced design)
library(car)
Anova(anova_2way, type = "III")
```

```
Anova Table (Type III tests)
```

```
Response: systolic_bp
                 Sum Sq  Df   F value    Pr(>F)
(Intercept)      364748   1 2966.7208 < 2.2e-16 ***
treatment          2752   2   11.1924 3.647e-05 ***
gender             2481   1   20.1775 1.701e-05 ***
treatment:gender    695   2    2.8276   0.06331 .
Residuals         14016 114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Interaction plot
interaction.plot(
  x.factor = bp_data_2way$treatment,
  trace.factor = bp_data_2way$gender,
  response = bp_data_2way$systolic_bp,
  type = "b",
  legend = TRUE,
  xlab = "Treatment",
  ylab = "Mean Systolic BP",
  main = "Interaction Plot: Treatment × Gender"
)
```



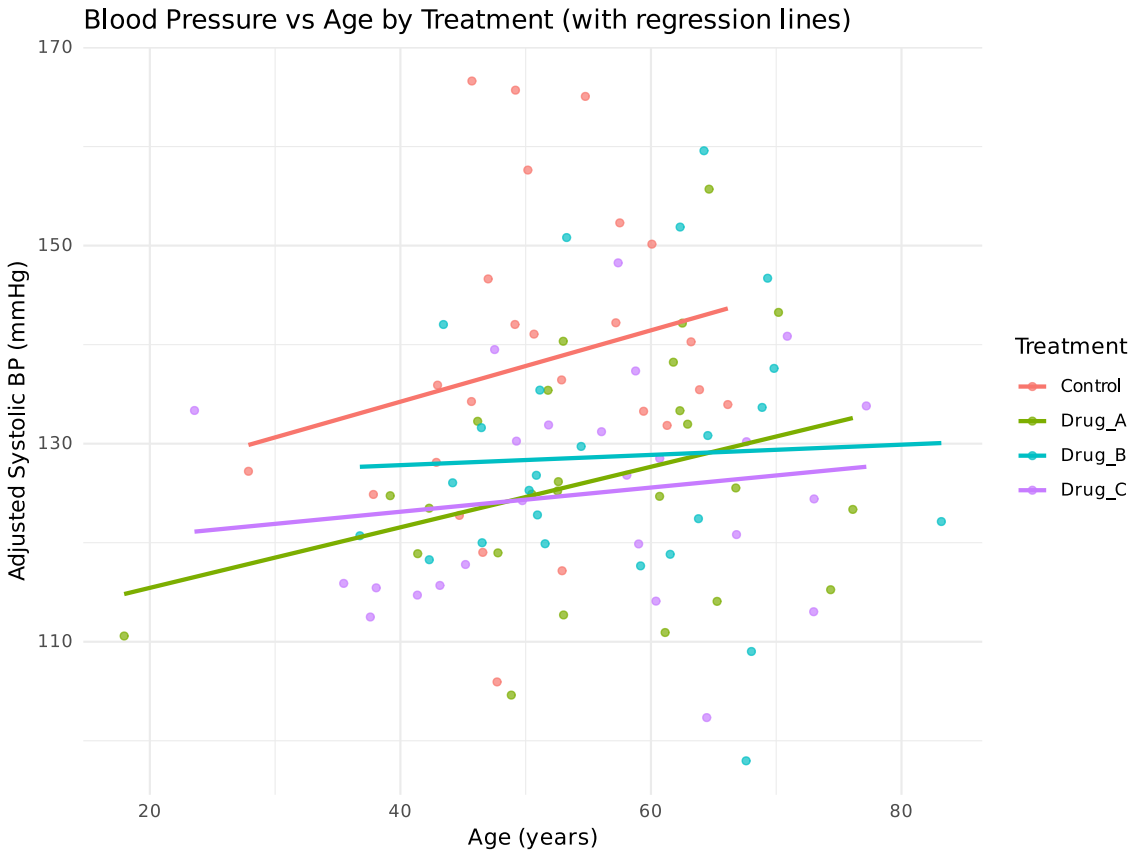
3.3 3. One-Way ANCOVA (Analysis of Covariance)

Purpose: Compare group means while controlling for a continuous covariate.

Research Question: "Do treatment effects on blood pressure remain significant after controlling for age?"

```
# Add age as covariate to original data
set.seed(789)
bp_data_ancova <- bp_data %>%
  mutate(
    age = rnorm(n(), 55, 12),
    # Adjust BP based on age (positive correlation)
    systolic_bp_adj = systolic_bp + 0.3 * (age - 55) + rnorm(n(), 0, 3)
  )

# Visualize relationship
bp_data_ancova %>%
  ggplot(aes(x = age, y = systolic_bp_adj, color = treatment)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
```

```
    title = "Blood Pressure vs Age by Treatment (with regression lines)",
    x = "Age (years)",
    y = "Adjusted Systolic BP (mmHg)",
    color = "Treatment"
  ) +
  theme_minimal()
```




```
# ANCOVA model
ancova_model <- aov(systolic_bp_adj ~ treatment + age, data = bp_data_ancova)
summary(ancova_model)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    3   2710   903.3   5.288 0.00205 **
age          1    496   496.4   2.906 0.09151 .
Residuals   95  16227   170.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Check for homogeneity of slopes (interaction test)
ancova_interaction <- aov(systolic_bp_adj ~ treatment * age, data =
bp_data_ancova)
summary(ancova_interaction)
```

```
              Df Sum Sq Mean Sq F value  Pr(>F)
treatment      3   2710   903.3   5.180 0.00237 **
age            1    496   496.4   2.847 0.09496 .
treatment:age  3    184    61.2   0.351 0.78845
Residuals     92  16043   174.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Adjusted means
emmeans_result <- emmeans(ancova_model, "treatment")
print(emmeans_result)
```

```
 treatment emmean   SE df lower.CL upper.CL
 Control      139 2.64 95      134      144
 Drug_A       126 2.62 95      121      131
 Drug_B       128 2.63 95      123      133
 Drug_C       125 2.61 95      120      130

Confidence level used: 0.95
```

3.4 4. Two-Way ANCOVA

Purpose: Examine two factors while controlling for covariates.

```r
# Add age to two-way data
bp_data_2way_ancova <- bp_data_2way %>%
  mutate(
    age = rnorm(n(), 55, 12),
    systolic_bp_adj = systolic_bp + 0.25 * (age - 55) + rnorm(n(), 0, 2)
  )

# Two-way ANCOVA
ancova_2way <- aov(systolic_bp_adj ~ treatment * gender + age,
                   data = bp_data_2way_ancova)
summary(ancova_2way)
```

```
                Df Sum Sq Mean Sq F value   Pr(>F)
treatment        2   9706    4853  39.150 1.21e-13 ***
gender           1   2563    2563  20.676 1.37e-05 ***
```

```
age                1    515     515   4.159   0.0438 *
treatment:gender   2    670     335   2.702   0.0714 .
Residuals        113  14007     124
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Adjusted means
emmeans_2way <- emmeans(ancova_2way, ~ treatment | gender)
print(emmeans_2way)
```

```
gender = Female:
 treatment emmean   SE  df lower.CL upper.CL
 Control      136 2.49 113      131      141
 Drug_A       122 2.49 113      117      127
 Drug_B       120 2.49 113      115      125

gender = Male:
 treatment emmean   SE  df lower.CL upper.CL
 Control      151 2.49 113      146      156
 Drug_A       126 2.49 113      121      131
 Drug_B       128 2.51 113      123      133

Confidence level used: 0.95
```

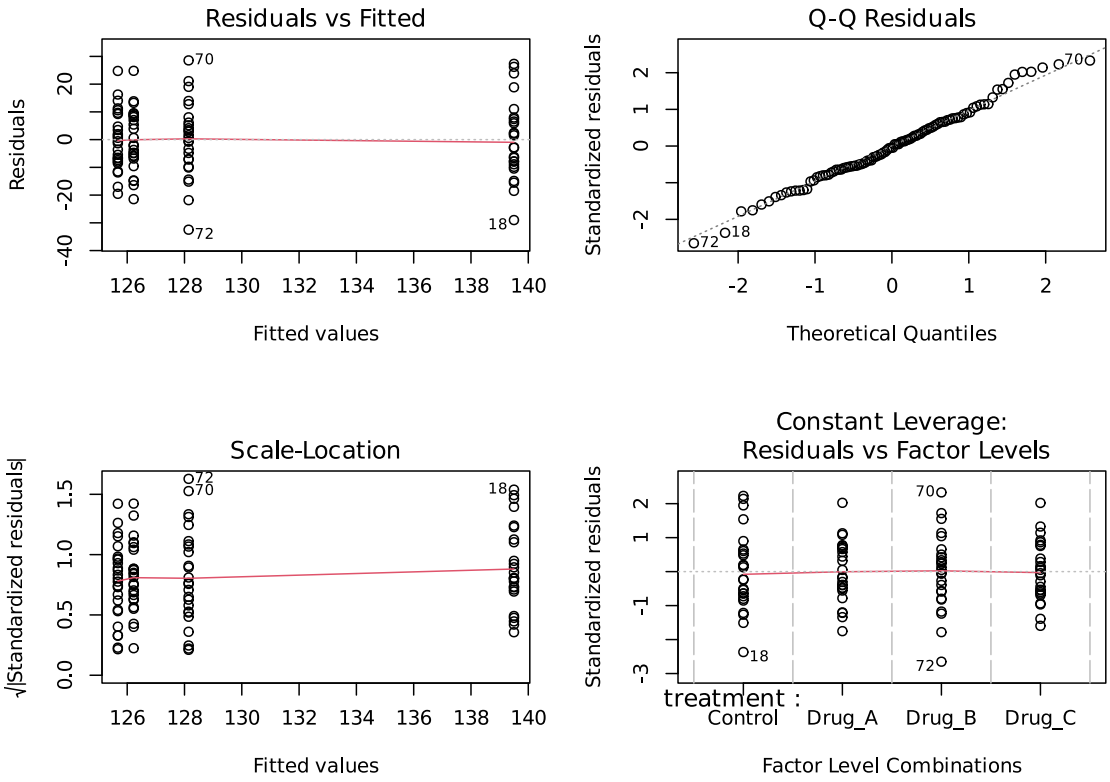
4 ANOVA Assumptions

4.1 The Four Key Assumptions

1. **Independence**: Observations are independent
2. **Normality**: Residuals are normally distributed
3. **Homogeneity of Variance**: Equal variances across groups
4. **Linearity** (for ANCOVA): Linear relationship between covariate and outcome

4.2 Diagnostic Procedures

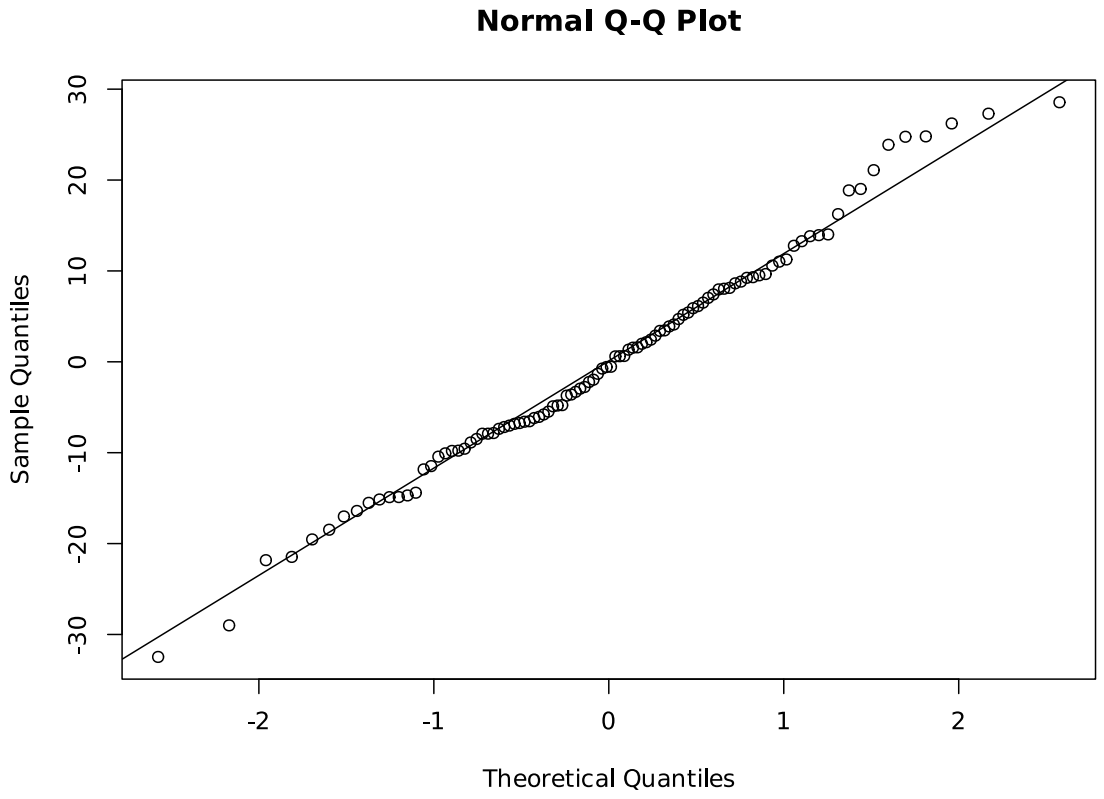
```r
# Using the one-way ANOVA model for diagnostics
model <- aov(systolic_bp ~ treatment, data = bp_data)

# 1. Residual plots
par(mfrow = c(2, 2))
plot(model)
```




```
par(mfrow = c(1, 1))

# 2. Normality tests
# Shapiro-Wilk test on residuals
shapiro.test(residuals(model))
```

```

    Shapiro-Wilk normality test

data:  residuals(model)
W = 0.99017, p-value = 0.6782
```

```
# QQ plot
qqnorm(residuals(model))
qqline(residuals(model))
```



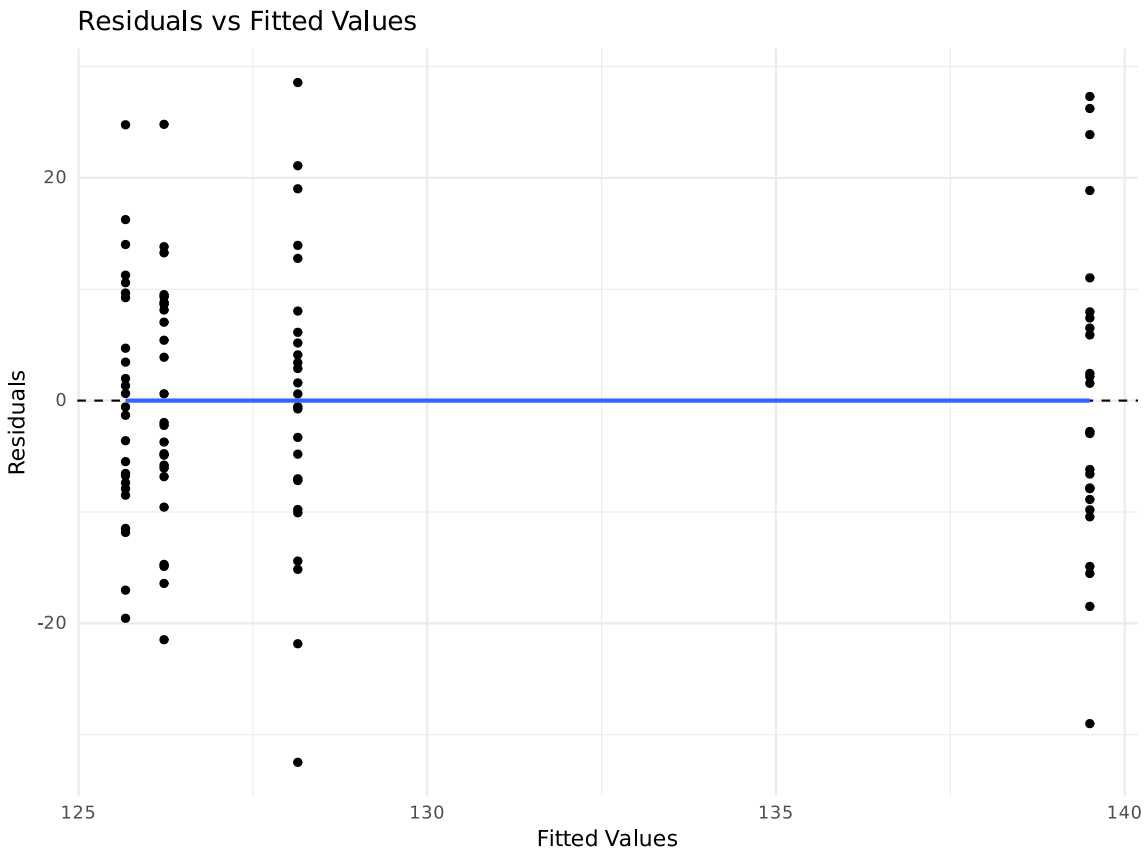

```r
# 3. Homogeneity of variance
# Levene's test
leveneTest(systolic_bp ~ treatment, data = bp_data)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.5761 0.6321
      96
```

```r
# Bartlett's test (more sensitive to non-normality)
bartlett.test(systolic_bp ~ treatment, data = bp_data)
```

```

    Bartlett test of homogeneity of variances

data:  systolic_bp by treatment
Bartlett's K-squared = 2.8076, df = 3, p-value = 0.4222
```

```r
# 4. Visual inspection
bp_data %>%
  mutate(residuals = residuals(model),
         fitted = fitted(model)) %>%
  ggplot(aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_smooth(se = FALSE) +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()
```



4.3 Dealing with Assumption Violations

```r
# 1. Non-normality: Transformations
# Log transformation (for positive skewed data)
bp_data_log <- bp_data %>%
  mutate(log_bp = log(systolic_bp))
```

```r
# Square root transformation
bp_data_sqrt <- bp_data %>%
  mutate(sqrt_bp = sqrt(systolic_bp))

# 2. Unequal variances: Welch's ANOVA
oneway.test(systolic_bp ~ treatment, data = bp_data, var.equal = FALSE)
```

```
    One-way analysis of means (not assuming equal variances)

data:  systolic_bp and treatment
F = 5.758, num df = 3.000, denom df = 52.996, p-value = 0.001743
```

```r
# 3. Non-parametric alternative: Kruskal-Wallis test
kruskal.test(systolic_bp ~ treatment, data = bp_data)
```

```
    Kruskal-Wallis rank sum test

data:  systolic_bp by treatment
Kruskal-Wallis chi-squared = 15.265, df = 3, p-value = 0.001604
```

```r
# 4. Robust ANOVA (using WRS2 package if available)
# install.packages("WRS2")
# library(WRS2)
# t1way(systolic_bp ~ treatment, data = bp_data)
```

5 Post-Hoc Analysis

5.1 Why Post-Hoc Tests Are Important

When ANOVA indicates significant differences, post-hoc tests help identify: - Which specific groups differ from each other - The magnitude of differences - Control family-wise error rate

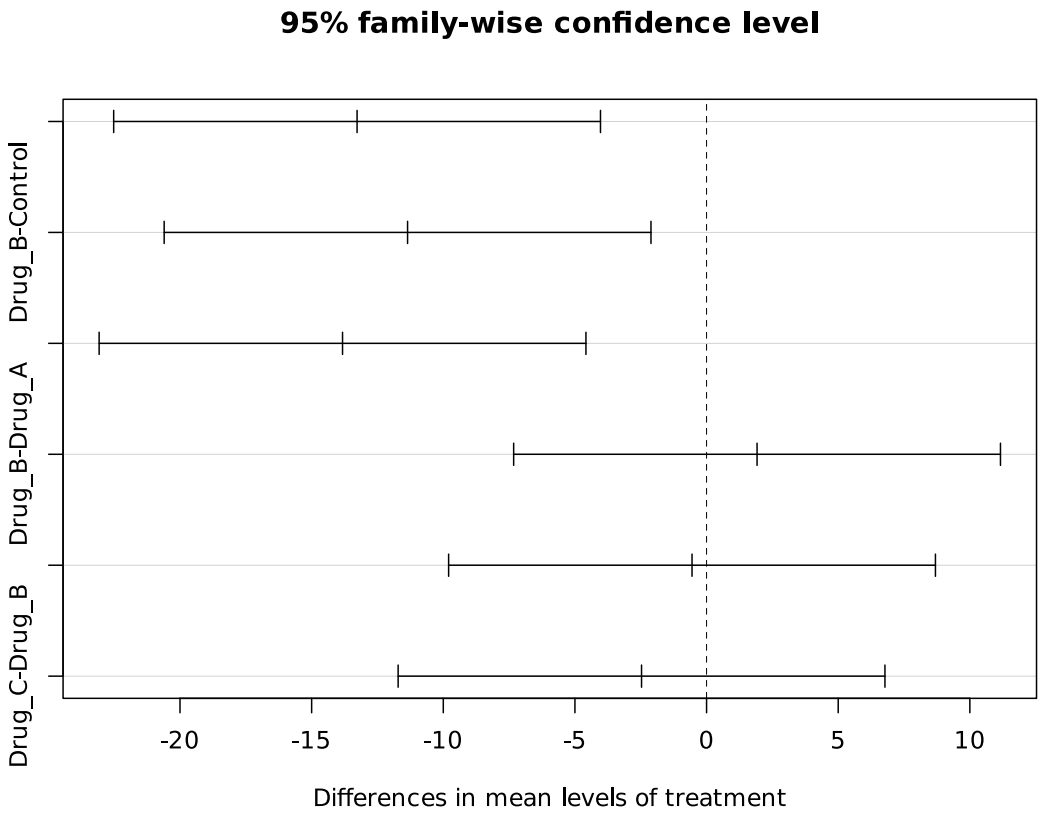
5.2 Common Post-Hoc Tests

```r
# 1. Tukey's HSD (Honest Significant Difference)
tukey_result <- TukeyHSD(anova_model)
print(tukey_result)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

```
Fit: aov(formula = systolic_bp ~ treatment, data = bp_data)

$treatment
                     diff        lwr       upr     p adj
Drug_A-Control -13.2743967 -22.519400 -4.029393 0.0016715
Drug_B-Control -11.3566710 -20.601674 -2.111668 0.0095608
Drug_C-Control -13.8265640 -23.071567 -4.581561 0.0009743
Drug_B-Drug_A    1.9177257  -7.327278 11.162729 0.9483822
Drug_C-Drug_A   -0.5521673  -9.797171  8.692836 0.9986345
Drug_C-Drug_B   -2.4698930 -11.714896  6.775110 0.8974118
```

```r
# Visualization of Tukey results
plot(tukey_result)
```




```r
# 2. Using emmeans for more flexibility
pairwise_emmeans <- emmeans(anova_model, "treatment")
pairs(pairwise_emmeans, adjust = "tukey")
```

```
 contrast         estimate   SE df t.ratio p.value
 Control - Drug_A   13.274 3.54 96   3.754  0.0017
 Control - Drug_B   11.357 3.54 96   3.212  0.0096
 Control - Drug_C   13.827 3.54 96   3.910  0.0010
 Drug_A - Drug_B    -1.918 3.54 96  -0.542  0.9484
 Drug_A - Drug_C     0.552 3.54 96   0.156  0.9986
 Drug_B - Drug_C     2.470 3.54 96   0.699  0.8974

P value adjustment: tukey method for comparing a family of 4 estimates
```

```r
# 3. Bonferroni correction
pairs(pairwise_emmeans, adjust = "bonferroni")
```

```
 contrast         estimate   SE df t.ratio p.value
 Control - Drug_A   13.274 3.54 96   3.754  0.0018
 Control - Drug_B   11.357 3.54 96   3.212  0.0108
 Control - Drug_C   13.827 3.54 96   3.910  0.0010
 Drug_A - Drug_B    -1.918 3.54 96  -0.542  1.0000
 Drug_A - Drug_C     0.552 3.54 96   0.156  1.0000
 Drug_B - Drug_C     2.470 3.54 96   0.699  1.0000

P value adjustment: bonferroni method for 6 tests
```

```r
# 4. False Discovery Rate (FDR) control
pairs(pairwise_emmeans, adjust = "fdr")
```

```
 contrast         estimate   SE df t.ratio p.value
 Control - Drug_A   13.274 3.54 96   3.754  0.0009
 Control - Drug_B   11.357 3.54 96   3.212  0.0036
 Control - Drug_C   13.827 3.54 96   3.910  0.0009
 Drug_A - Drug_B    -1.918 3.54 96  -0.542  0.7066
 Drug_A - Drug_C     0.552 3.54 96   0.156  0.8762
 Drug_B - Drug_C     2.470 3.54 96   0.699  0.7066

P value adjustment: fdr method for 6 tests
```

```r
# 5. Custom contrasts
# Example: Compare Drug treatments vs Control
contrast_matrix <- list(
  "Drugs_vs_Control" = c(-3, 1, 1, 1),  # Contrast coefficients
  "Drug_A_vs_B_C" = c(0, 2, -1, -1)
)
```

```
custom_contrasts <- contrast(pairwise_emmeans, contrast_matrix)
print(custom_contrasts)
```

```
 contrast         estimate   SE df t.ratio p.value
 Drugs_vs_Control   -38.46 8.66 96  -4.440  <.0001
 Drug_A_vs_B_C       -1.37 6.12 96  -0.223  0.8240
```

6 Practical Example: Complete Analysis

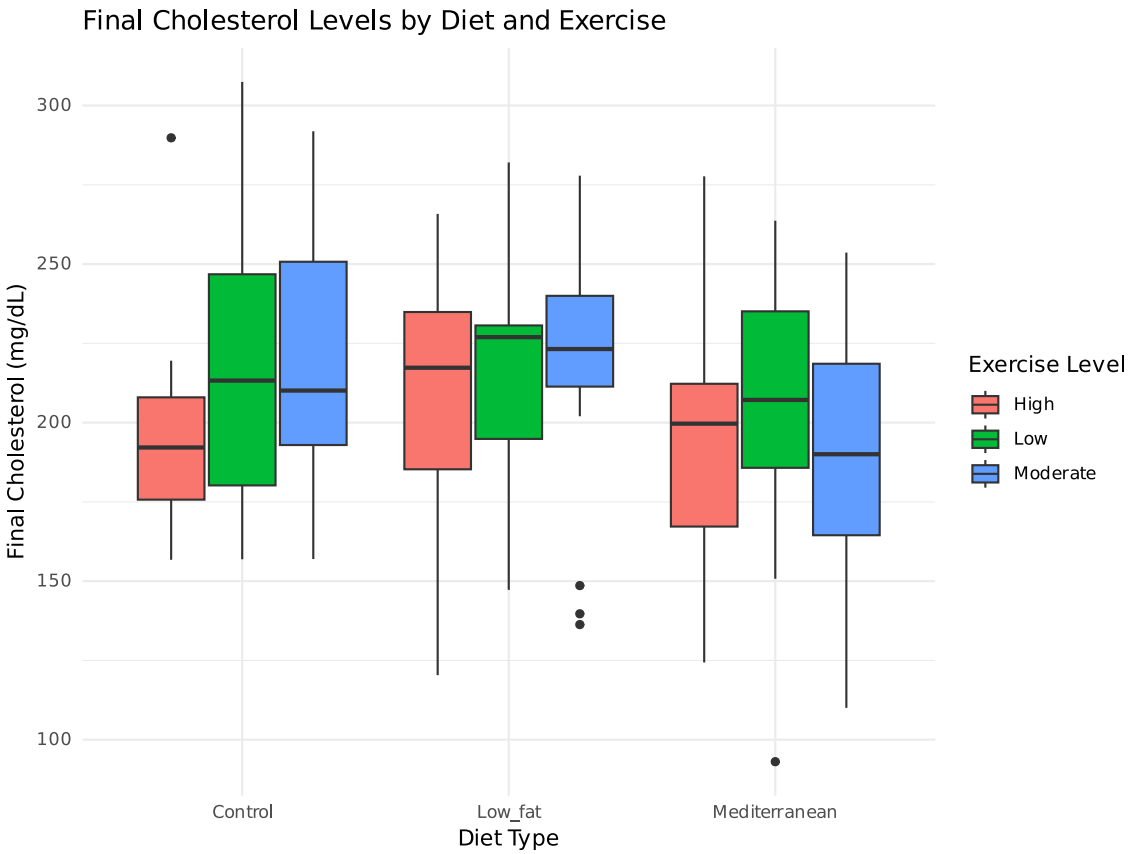
Let's work through a complete epidemiological example:

```
# Simulate a study: Effect of dietary intervention on cholesterol levels
# Factor 1: Diet type (Mediterranean, Low-fat, Control)
# Factor 2: Exercise level (Low, Moderate, High)
# Covariate: Baseline cholesterol

set.seed(2024)
chol_study <- expand_grid(
  diet = c("Mediterranean", "Low_fat", "Control"),
  exercise = c("Low", "Moderate", "High")
) %>%
  slice(rep(1:n(), each = 15)) %>%
  mutate(
    participant_id = 1:n(),
    baseline_chol = rnorm(n(), 220, 30),
    # Treatment effects
    diet_effect = case_when(
      diet == "Mediterranean" ~ -15,
      diet == "Low_fat" ~ -8,
      diet == "Control" ~ 0
    ),
    exercise_effect = case_when(
      exercise == "High" ~ -10,
      exercise == "Moderate" ~ -5,
      exercise == "Low" ~ 0
    ),
    # Interaction effect (Mediterranean diet works better with high exercise)
    interaction_effect = ifelse(diet == "Mediterranean" & exercise == "High",
-5, 0),
    # Final cholesterol level
    final_chol = baseline_chol + diet_effect + exercise_effect +
                 interaction_effect + 0.3 * (baseline_chol - 220) + rnorm(n(),
0, 15)
  )

# 1. Descriptive statistics
chol_summary <- chol_study %>%
```

```r
  group_by(diet, exercise) %>%
  summarise(
    n = n(),
    mean_baseline = mean(baseline_chol),
    mean_final = mean(final_chol),
    change = mean_final - mean_baseline,
    sd_change = sd(final_chol - baseline_chol),
    .groups = "drop"
  )

print(chol_summary)
```

```
# A tibble: 9 × 7
  diet          exercise     n mean_baseline mean_final  change sd_change
  <chr>         <chr>    <int>         <dbl>      <dbl>   <dbl>     <dbl>
1 Control       High        15          210.       194. -16.0         15.9
2 Control       Low         15          216.       216.  -0.131       11.1
3 Control       Moderate    15          219.       220.   1.05        14.7
4 Low_fat       High        15          222.       205. -17.2         19.5
5 Low_fat       Low         15          220.       216.  -4.82        13.6
6 Low_fat       Moderate    15          225.       216.  -8.20        15.8
7 Mediterranean High        15          218.       192. -25.9         18.8
8 Mediterranean Low         15          217.       205. -11.9         13.6
9 Mediterranean Moderate    15          211.       189. -21.6         18.1
```

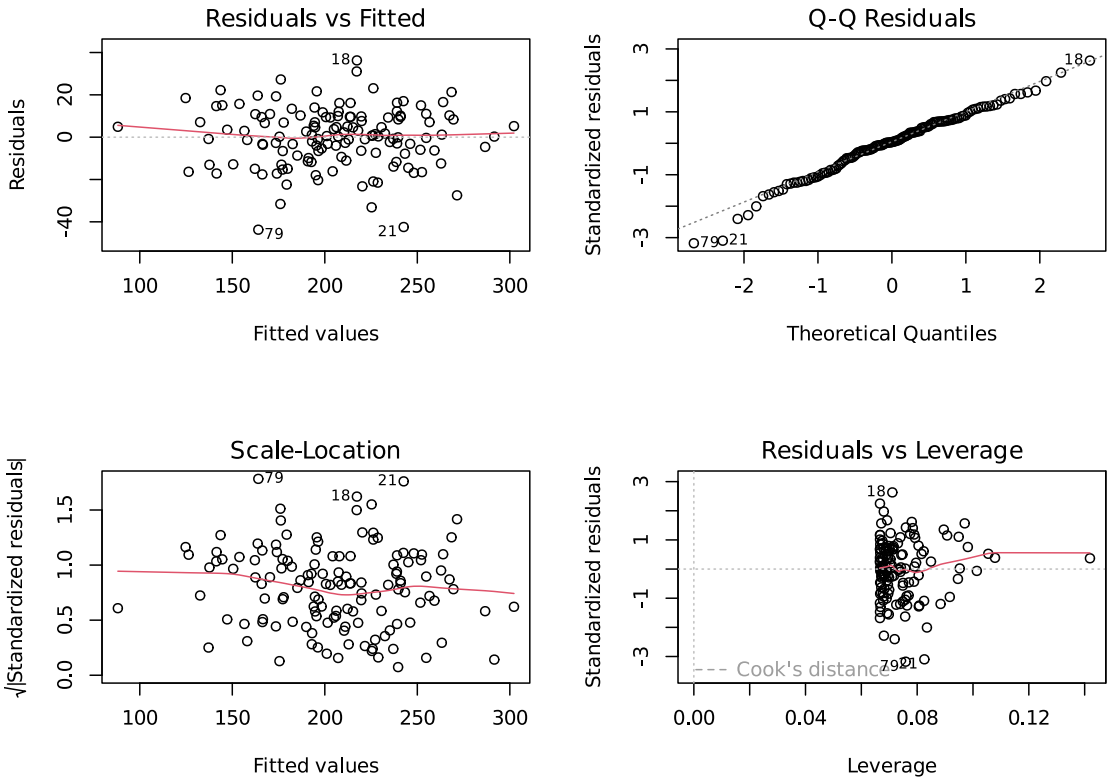
```r
# 2. Visualization
chol_study %>%
  ggplot(aes(x = diet, y = final_chol, fill = exercise)) +
  geom_boxplot(position = position_dodge(0.8)) +
  labs(
    title = "Final Cholesterol Levels by Diet and Exercise",
    x = "Diet Type",
    y = "Final Cholesterol (mg/dL)",
    fill = "Exercise Level"
  ) +
  theme_minimal()
```




```r
# 3. Two-way ANCOVA
final_model <- aov(final_chol ~ diet * exercise + baseline_chol, data =
chol_study)
summary(final_model)
```

```
               Df Sum Sq Mean Sq F value   Pr(>F)
diet            2   7724    3862  18.863 6.93e-08 ***
exercise        2   5487    2743  13.399 5.34e-06 ***
baseline_chol   1 183569  183569 896.607  < 2e-16 ***
diet:exercise   4    436     109   0.533    0.712
Residuals     125  25592     205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 4. Check assumptions
par(mfrow = c(2, 2))
plot(final_model)
```




```r
par(mfrow = c(1, 1))

# 5. Post-hoc analysis
emmeans_final <- emmeans(final_model, ~ diet | exercise)
pairs(emmeans_final, adjust = "tukey")
```

```
exercise = High:
 contrast               estimate   SE  df t.ratio p.value
 Control - Low_fat          3.76 5.25 125   0.717  0.7542
 Control - Mediterranean   11.52 5.23 125   2.201  0.0749
 Low_fat - Mediterranean    7.76 5.23 125   1.485  0.3017

exercise = Low:
 contrast               estimate   SE  df t.ratio p.value
 Control - Low_fat          5.68 5.23 125   1.087  0.5239
 Control - Mediterranean   11.85 5.22 125   2.269  0.0640
 Low_fat - Mediterranean    6.17 5.23 125   1.180  0.4670

exercise = Moderate:
 contrast               estimate   SE  df t.ratio p.value
```

```
 Control - Low_fat          10.58 5.23 125   2.023  0.1108
 Control - Mediterranean    20.75 5.24 125   3.964  0.0004
 Low_fat - Mediterranean    10.17 5.26 125   1.935  0.1332

P value adjustment: tukey method for comparing a family of 3 estimates
```

7 Summary and Best Practices

7.1 When to Use Each Type of ANOVA

Design	Use When	Example
One-way ANOVA	Comparing 3+ groups on one factor	Treatment efficacy across multiple drugs
Two-way ANOVA	Two factors, interested in main effects and interactions	Treatment × Gender effects
One-way ANCOVA	One factor + continuous covariate	Treatment effects controlling for age
Two-way ANCOVA	Two factors + continuous covariate	Treatment × Gender controlling for baseline

7.2 Recommendations

1. **Always check assumptions** before interpreting results
2. **Use post-hoc tests** only when overall ANOVA is significant
3. **Report effect sizes** alongside p-values
4. **Consider practical significance** not just statistical significance
5. **Use appropriate corrections** for multiple comparisons
6. **Visualize your data** before and after analysis

7.3 R Code Best Practices

```r
# 1. Use tidyverse for data manipulation
data %>%
  filter(condition) %>%
  group_by(factor) %>%
  summarise(mean_outcome = mean(outcome), .groups = "drop")

# 2. Use broom for tidy model outputs
model %>% tidy() %>% kable()

# 3. Use emmeans for post-hoc analysis
emmeans(model, "factor") %>% pairs(adjust = "tukey")

# 4. Always visualize
```

```r
ggplot(data, aes(x = factor, y = outcome)) +
  geom_boxplot() +
  theme_minimal()

# 5. Check assumptions systematically
# Normality
shapiro.test(residuals(model))
# Homogeneity
leveneTest(outcome ~ factor, data = data)
# Independence (by design)
```

8 Conclusion

ANOVA is a powerful statistical tool in epidemiology and public health research. Understanding when to apply different types of ANOVA, how to check assumptions, and how to interpret results is crucial for making valid statistical inferences. Always remember that statistical significance should be accompanied by practical significance and proper effect size reporting.

The combination of R programming with tidyverse principles provides an efficient and reproducible approach to conducting ANOVA analyses in epidemiological research.