Lecture 15 - Class Activity ANCOVA

Bill Perry

Lecture 15: Analysis of Covariance (ANCOVA)

What is ANCOVA?

ANCOVA (Analysis of Covariance) combines regression and ANOVA to: - Compare group means while adjusting for a continuous covariate - Increase statistical power by reducing residual error - Control for confounding variables

When to Use ANCOVA

Use ANCOVA when you have: - **Response variable**: Continuous - **Predictor variable**: Categorical (factor/groups) - **Covariate**: Continuous variable that affects the response

Key Assumptions of ANCOVA

1. **Independence** of observations
2. **Normality** of residuals
3. **Homogeneity of variances** across groups
4. **Linearity** between response and covariate within each group
5. **Homogeneity of slopes** (most critical!) - regression slopes must be equal across all groups

> ! Critical First Step
>
> Always test for **homogeneity of slopes** before proceeding with ANCOVA. If slopes differ significantly between groups, standard ANCOVA is inappropriate.

Part 1: Cricket Chirping Analysis

Data Overview

We want to compare chirping rate of two cricket species: - *Oecanthus exclamationis* - *Oecanthus niveus*

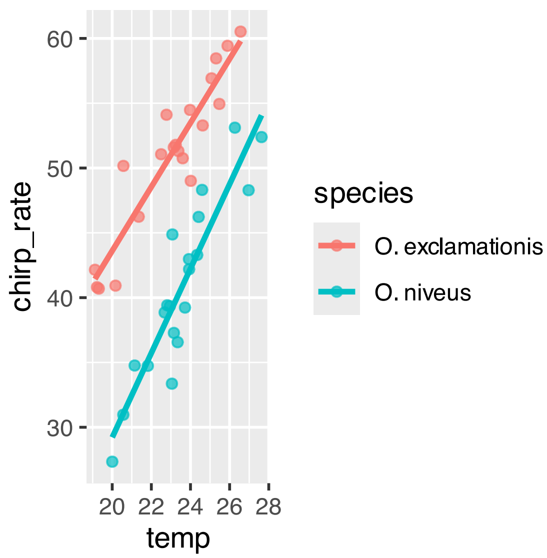
But we measured rates at different temperatures, and there's a relationship between pulse rate and temperature. ANCOVA lets us adjust for temperature effect to get a more powerful test!

```
# Create simulated cricket data based on lecture example
set.seed(456)
n <- 40
species <- rep(c("O. exclamationis", "O. niveus"), each = n/2)
temp <- c(rnorm(n/2, mean = 22, sd = 2), rnorm(n/2, mean = 24, sd = 2))
chirp_rate <- 40 + 2.5 * (temp - 23) + ifelse(species == "O. exclamationis", 10, 0) + rnorm(n,
sd = 3)
cricket_data <- data.frame(species = species, temp = temp, chirp_rate = chirp_rate)

# View data structure

head(cricket_data)
```

```
            species     temp chirp_rate
1 O. exclamationis 19.31296   40.69557
2 O. exclamationis 23.24355   51.78799
3 O. exclamationis 23.60175   50.75553
4 O. exclamationis 19.22222   40.80589
5 O. exclamationis 20.57129   50.16484
6 O. exclamationis 21.35188   46.24225
```

```
# Plot with regression lines by species
ggplot(cricket_data, aes(x = temp, y = chirp_rate, color = species)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Step 1: Test Homogeneity of Slopes

This is the most critical assumption! We test if the regression slopes are equal across all groups.

```
# Test for homogeneity of slopes by including interaction term
cricket_slopes_model <- lm(chirp_rate ~ temp * species, data = cricket_data)
Anova(cricket_slopes_model, type = 3)
```

```
Anova Table (Type III tests)

Response: chirp_rate
             Sum Sq Df F value           Pr(>F)
(Intercept)    6.32  1  0.9393         0.338915
temp         620.48  1 92.1572 0.00000000001828 ***
species       69.76  1 10.3617         0.002724 **
temp:species  26.08  1  3.8734         0.056796 .
Residuals    242.38 36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: If p > 0.05, slopes are homogeneous and we can proceed with ANCOVA. If p < 0.05, slopes differ and standard ANCOVA is inappropriate.

Step 2: Fit ANCOVA Model

Since slopes are homogeneous (p > 0.05), we can fit the ANCOVA model without the interaction term.

```
# Fit ANCOVA model (without interaction)
cricket_ancova <- lm(chirp_rate ~ temp + species, data = cricket_data)

# Get model summary
summary(cricket_ancova)
```

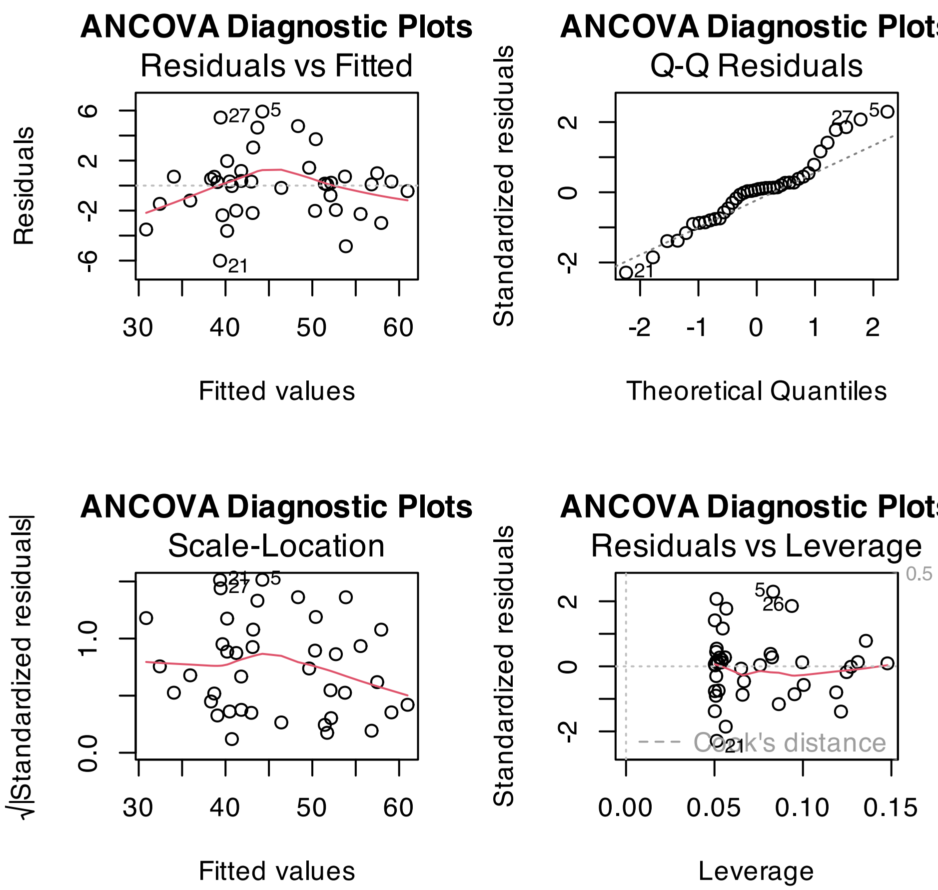
```

Call:
lm(formula = chirp_rate ~ temp + species, data = cricket_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0065 -1.9653  0.1923  0.7886  5.9192 

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)    
(Intercept)      -13.2012     4.7423  -2.784              0.00842 ** 
temp               2.7926     0.2048  13.634 0.000000000000000530 ***
speciesO. niveus -11.8005     0.8593 -13.733 0.000000000000000424 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.694 on 37 degrees of freedom
Multiple R-squared:  0.8994,	Adjusted R-squared:  0.894 
F-statistic: 165.4 on 2 and 37 DF,  p-value: < 0.00000000000000022
```

```
# View ANOVA table
Anova(cricket_ancova)
```

```
Anova Table (Type II tests)

Response: chirp_rate
           Sum Sq Df F value                Pr(>F)    
temp      1348.81  1  185.90 0.0000000000000005296 ***
species   1368.34  1  188.59 0.0000000000000004236 ***
Residuals  268.46 37                                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 3: Check Model Assumptions

```
# Create diagnostic plots
par(mfrow = c(2, 2))
plot(cricket_ancova, main = "ANCOVA Diagnostic Plots")
```




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

Step 4: Calculate Adjusted Means

ANCOVA compares adjusted means - what each group's mean would be at the overall mean of the covariate.

```r
# Calculate adjusted means using emmeans
cricket_adjusted_means <- emmeans(cricket_ancova, "species")

# Convert to dataframe for plotting
cricket_adj_means_df <- as.data.frame(cricket_adjusted_means)
cricket_adj_means_df
```

```
 species             emmean        SE df lower.CL upper.CL
 O. exclamationis 51.70513 0.6049702 37 50.47934 52.93091
 O. niveus        39.90462 0.6049702 37 38.67883 41.13040

Confidence level used: 0.95
```

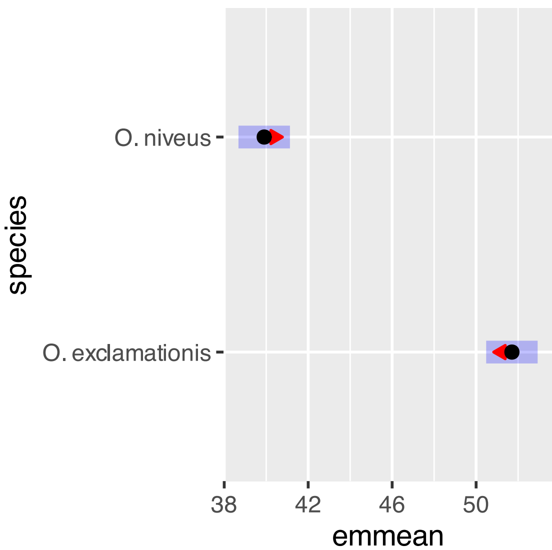
Step 5: Pairwise Comparisons

```r
# Pairwise comparisons of adjusted means
pairs(cricket_adjusted_means, adjust = "sidak")
```

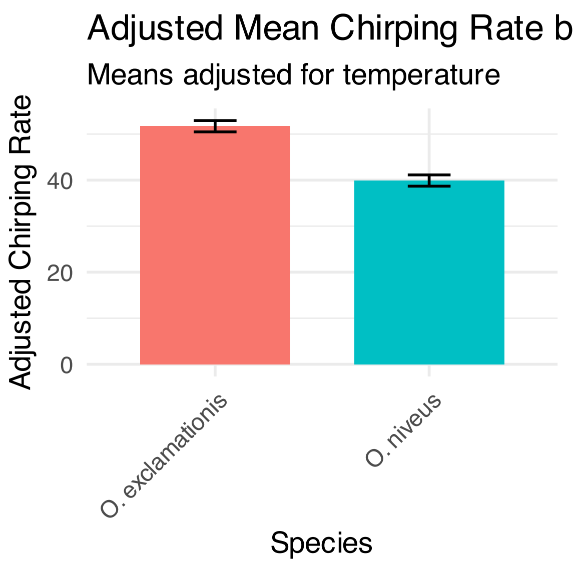
```
 contrast                          estimate    SE df t.ratio p.value
 O. exclamationis - O. niveus          11.8 0.859 37  13.733  <.0001
```

Step 6: Visualize Results

```r
# Plot adjusted means with confidence intervals
plot(cricket_adjusted_means, comparisons = TRUE)
```




```r
# Bar chart of adjusted means
ggplot(cricket_adj_means_df, aes(x = species, y = emmean, fill = species)) +
  geom_bar(stat = "identity", width = 0.7) +
  geom_errorbar(aes(ymin = lower.CL, ymax = upper.CL), width = 0.2) +
  labs(title = "Adjusted Mean Chirping Rate by Species",
       subtitle = "Means adjusted for temperature",
       x = "Species",
       y = "Adjusted Chirping Rate") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))
```



Part 2: Partridge Longevity Analysis

Data Overview

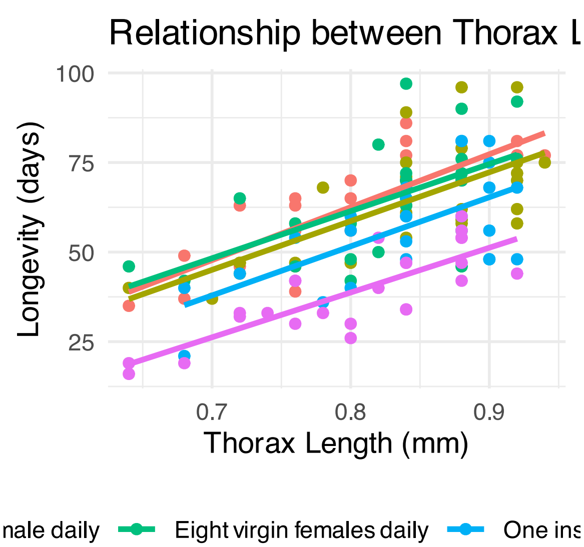
We'll analyze the effect of mating strategy on male fruitfly longevity, using thorax length as a covariate.

```
# Load the partridge dataset
partridge <- read.csv("data/partridge.csv")

# Create better treatment names
partridge$treatment <- factor(partridge$TREATMEN,
                              levels = 1:5,
                              labels = c("No females",
                                         "One virgin female daily",
                                         "Eight virgin females daily",
                                         "One inseminated female daily",
                                         "Eight inseminated females daily"))

# View data structure
head(partridge)
```

```
  PARTNERS TYPE TREATMEN LONGEV  LLONGEV THORAX     RESID1 PREDICT1      RESID2
1        8    0        1     35 1.544068   0.64  -5.868456 40.86846 -0.04743024
2        8    0        1     37 1.568202   0.68  -9.301196 46.30120 -0.07105067
3        8    0        1     49 1.690196   0.68   2.698804 46.30120  0.05094369
4        8    0        1     46 1.662758   0.72  -5.733936 51.73394 -0.02424867
5        8    0        1     63 1.799341   0.72  11.266064 51.73394  0.11233405
6        8    0        1     39 1.591065   0.76 -18.166676 57.16668 -0.14369601
  PREDICT2  treatment
1 1.591498 No females
2 1.639252 No females
3 1.639252 No females
4 1.687007 No females
5 1.687007 No females
6 1.734761 No females
```

```r
# Visualize the relationship between thorax length and longevity by treatment
ggplot(partridge, aes(x = THORAX, y = LONGEV, color = treatment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Relationship between Thorax Length and Longevity",
       x = "Thorax Length (mm)",
       y = "Longevity (days)",
       color = "Treatment") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Step 1: Test Homogeneity of Slopes

```r
# Test for homogeneity of slopes
homo_slopes_model <- lm(LONGEV ~ THORAX * treatment, data = partridge)
Anova(homo_slopes_model, type = 3)
```

```
Anova Table (Type III tests)

Response: LONGEV
                 Sum Sq  Df F value    Pr(>F)
(Intercept)       755.6   1  6.6320   0.01128 *
THORAX           3486.3   1 30.5999 2.017e-07 ***
treatment          36.9   4  0.0810   0.98805
THORAX:treatment   42.5   4  0.0933   0.98441
Residuals       13102.1 115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 2: Fit ANCOVA Model

```r
# Fit the ANCOVA model (without interaction)
ancova_model <- lm(LONGEV ~ THORAX + treatment, data = partridge)
```

```r
# Get more detailed summary
summary(ancova_model)
```

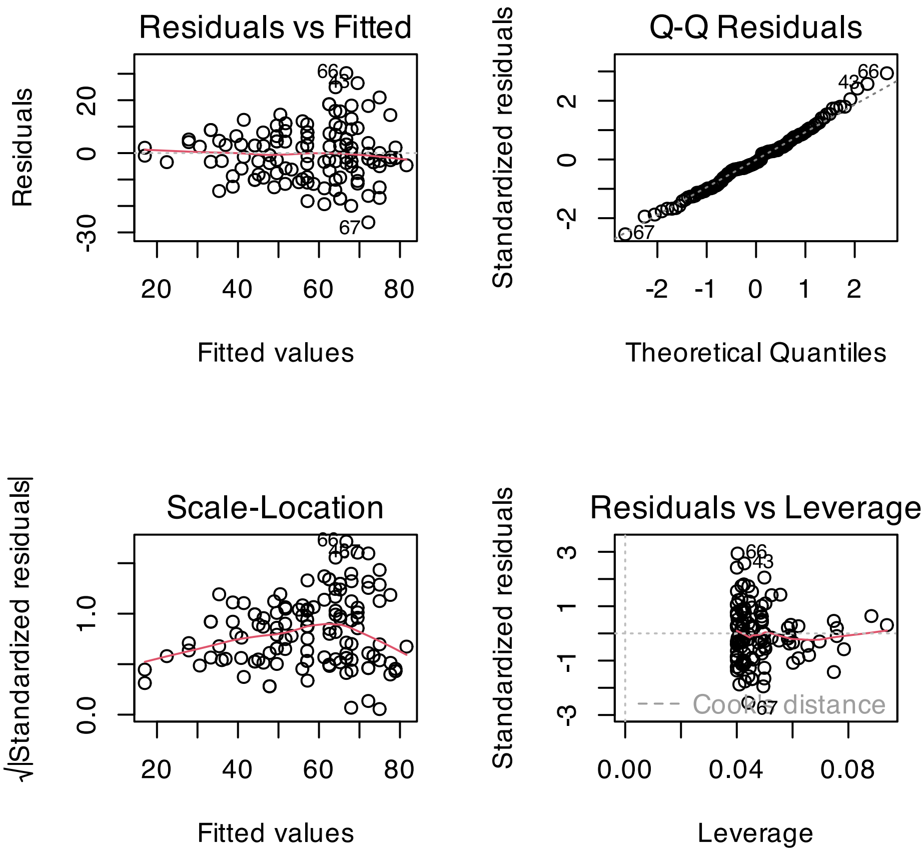
```

Call:
lm(formula = LONGEV ~ THORAX + treatment, data = partridge)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.189  -6.599  -0.989   6.408  30.244 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                               -46.055     10.239  -4.498 1.61e-05
THORAX                                    135.819     12.439  10.919  < 2e-16
treatmentOne virgin female daily           -3.929      2.997  -1.311 0.192347
treatmentEight virgin females daily        -1.276      2.983  -0.428 0.669517
treatmentOne inseminated female daily     -10.946      2.999  -3.650 0.000391
treatmentEight inseminated females daily  -23.879      2.973  -8.031 7.83e-13
                                            
(Intercept)                              ***
THORAX                                   ***
treatmentOne virgin female daily            
treatmentEight virgin females daily         
treatmentOne inseminated female daily    ***
treatmentEight inseminated females daily ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.51 on 119 degrees of freedom
Multiple R-squared:  0.6564,	Adjusted R-squared:  0.6419 
F-statistic: 45.46 on 5 and 119 DF,  p-value: < 2.2e-16
```

```r
# View ANOVA table
anova(ancova_model)
```

```
Analysis of Variance Table

Response: LONGEV
           Df  Sum Sq Mean Sq F value    Pr(>F)    
THORAX      1 15496.6 15496.6 140.293 < 2.2e-16 ***
treatment   4  9611.5  2402.9  21.753 1.719e-13 ***
Residuals 119 13144.7   110.5                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 3: Check Assumptions

```r
# Create diagnostic plots
par(mfrow = c(2, 2))
plot(ancova_model)
```



Step 4: Calculate Adjusted Means

```
# Get adjusted means using emmeans
adjusted_means <- emmeans(ancova_model, "treatment")
adjusted_means
```

```
 treatment                      emmean   SE  df lower.CL upper.CL
 No females                       65.4 2.11 119     61.3     69.6
 One virgin female daily          61.5 2.11 119     57.3     65.7
 Eight virgin females daily       64.2 2.10 119     60.0     68.3
 One inseminated female daily     54.5 2.11 119     50.3     58.7
 Eight inseminated females daily  41.6 2.12 119     37.4     45.8

Confidence level used: 0.95
```

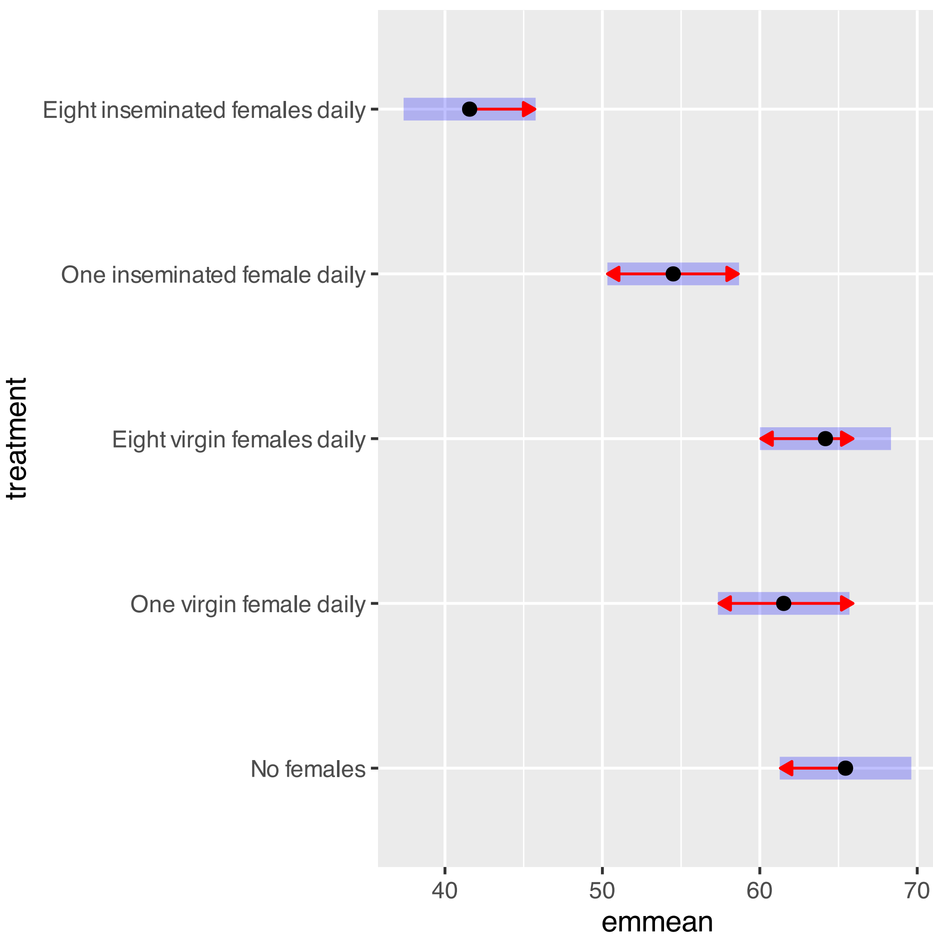
Step 5: Pairwise Comparisons

```
# Pairwise comparisons of adjusted means
pairs(adjusted_means, adjust = "tukey")
```

```
 contrast                                                  estimate   SE
 No females - One virgin female daily                          3.93 3.00
 No females - Eight virgin females daily                       1.28 2.98
 No females - One inseminated female daily                    10.95 3.00
 No females - Eight inseminated females daily                 23.88 2.97
```

```
 One virgin female daily - Eight virgin females daily              -2.65 2.98
 One virgin female daily - One inseminated female daily             7.02 2.97
 One virgin female daily - Eight inseminated females daily         19.95 3.01
 Eight virgin females daily - One inseminated female daily          9.67 2.98
 Eight virgin females daily - Eight inseminated females daily      22.60 2.99
 One inseminated female daily - Eight inseminated females daily    12.93 3.01
  df t.ratio p.value
 119   1.311  0.6849
 119   0.428  0.9929
 119   3.650  0.0035
 119   8.031  <.0001
 119  -0.891  0.8996
 119   2.361  0.1336
 119   6.636  <.0001
 119   3.249  0.0129
 119   7.560  <.0001
 119   4.298  0.0003

P value adjustment: tukey method for comparing a family of 5 estimates
```

```
# Plot adjusted means with confidence intervals
plot(adjusted_means, comparisons = TRUE)
```

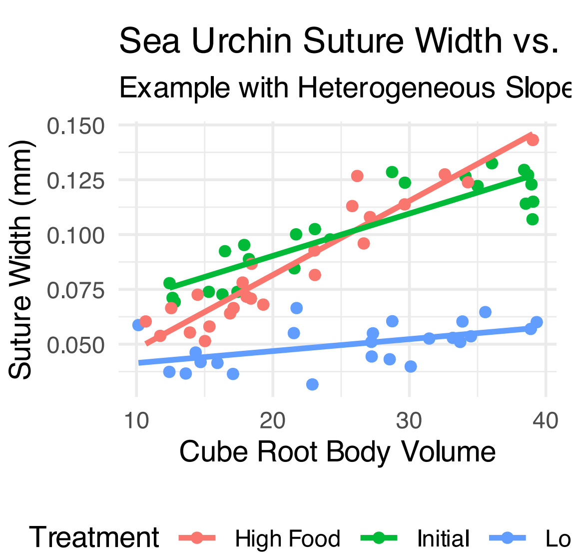


Part 3: Example with Heterogeneous Slopes

Let's look at an example where slopes are NOT homogeneous using sea urchin data.

```r
# Create simulated sea urchin data with heterogeneous slopes
set.seed(345)
n <- 72  # 24 urchins per group

# Create data frame
treatments <- rep(c("Initial", "Low Food", "High Food"), each = n/3)
volume <- c(
  runif(n/3, 10, 40),  # Initial
  runif(n/3, 10, 40),  # Low Food
  runif(n/3, 10, 40)   # High Food
)

# Create suture width with different slopes for each treatment
suture_width <- ifelse(
  treatments == "Initial", 0.05 + 0.002 * volume,
  ifelse(
    treatments == "Low Food", 0.04 + 0.0005 * volume,
    0.02 + 0.003 * volume  # High Food
  )
) + rnorm(n, 0, 0.01)

urchin_data <- data.frame(treatment = treatments, volume = volume, suture_width = suture_width)

# Plot the data with regression lines
ggplot(urchin_data, aes(x = volume, y = suture_width, color = treatment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Sea Urchin Suture Width vs. Volume",
       subtitle = "Example with Heterogeneous Slopes",
       x = "Cube Root Body Volume",
       y = "Suture Width (mm)",
       color = "Treatment") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Test for Homogeneity of Slopes

```r
# Fit model with interaction
urchin_model <- lm(suture_width ~ volume * treatment, data = urchin_data)
Anova(urchin_model, type = 3)
```

```
Anova Table (Type III tests)

Response: suture_width
                   Sum Sq Df F value    Pr(>F)    
(Intercept)      0.0005253  1    5.91   0.01778 *  
volume           0.0151663  1  170.64 < 2.2e-16 ***
treatment        0.0020070  2   11.29 6.064e-05 ***
volume:treatment 0.0062129  2   34.95 4.453e-11 ***
Residuals        0.0058662 66                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result: With $p < 0.05$, we have heterogeneous slopes! Standard ANCOVA would be inappropriate here.

What to do with Heterogeneous Slopes

When slopes are not homogeneous, you have several options:

```r
# Option: Analyze groups separately
initial_model <- lm(suture_width ~ volume, data = filter(urchin_data, treatment == "Initial"))
low_food_model <- lm(suture_width ~ volume, data = filter(urchin_data, treatment == "Low
Food"))
high_food_model <- lm(suture_width ~ volume, data = filter(urchin_data, treatment == "High
Food"))

# Summary for each group
initial_model
```

```

Call:
lm(formula = suture_width ~ volume, data = filter(urchin_data, 
    treatment == "Initial"))

Coefficients:
(Intercept)       volume  
   0.051785     0.001926  
```

```r
low_food_model
```

```

Call:
lm(formula = suture_width ~ volume, data = filter(urchin_data, 
    treatment == "Low Food"))

Coefficients:
(Intercept)       volume  
  0.0359532    0.0005453  
```

```
high_food_model
```

```

Call:
lm(formula = suture_width ~ volume, data = filter(urchin_data,
    treatment == "High Food"))

Coefficients:
(Intercept)       volume
   0.014077     0.003376
```

Summary Checklist for ANCOVA

When conducting ANCOVA, always follow these steps:

> ANCOVA Checklist
>
> 1. **Visualize your data** - plot response vs covariate, colored by groups
> 2. **Test homogeneity of slopes** - fit model with interaction term
>   - If $p > 0.05$: proceed with ANCOVA
>   - If $p < 0.05$: use alternative approaches
> 3. **Fit ANCOVA model** - response ~ covariate + factor
> 4. **Check assumptions** - use diagnostic plots
> 5. **Interpret results** - focus on adjusted means, not raw means
> 6. **Conduct post-hoc tests** - pairwise comparisons if needed
> 7. **Visualize results** - show adjusted means with confidence intervals

Key Points to Remember

- **ANCOVA increases power** by accounting for covariate variation
- **Adjusted means** are what we compare, not raw group means
- **Homogeneity of slopes** is the most critical assumption
- **Parallel lines** in your plot suggest homogeneous slopes
- **Non-parallel lines** indicate heterogeneous slopes - use alternative methods

> ! Key Points from ANCOVA Analysis
>
> 1. **Test homogeneity of slopes first** - this is the most critical assumption
> 2. **ANCOVA compares adjusted means** at the mean value of the covariate
> 3. **Increases statistical power** by removing variation due to the covariate
> 4. **Choose appropriate methods** based on whether slopes are homogeneous
> 5. **Visualize your results** clearly showing the relationship between variables
> 6. **Check all assumptions** using diagnostic plots
> 7. **Interpret in biological context** - what do the adjusted means tell us?
>
> Remember: The covariate should be measured independently of the treatment and should not be affected by the treatment itself!