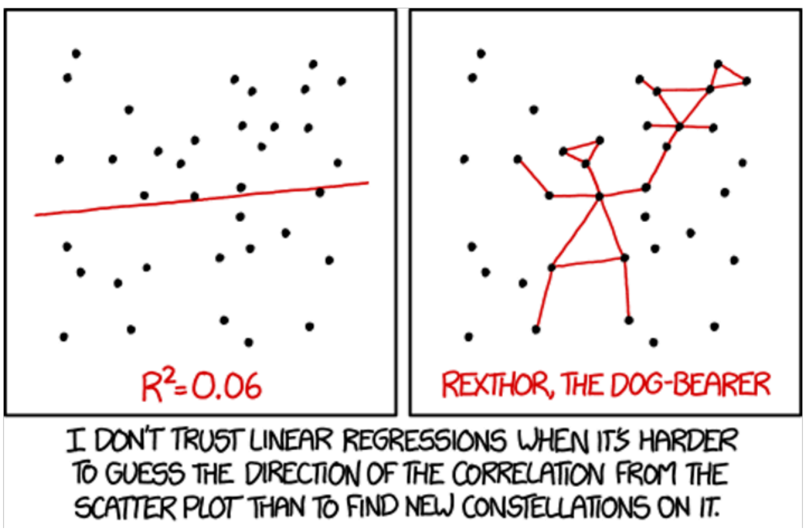
Lecture 11 - Single factor analysis of variance - ANOVA

Bill Perry

Lecture 11: Review

Multiple Regression

- MLR model
- Regression parameters
- Analysis of variance
- Null hypotheses
- Explained variance
- Assumptions and diagnostics
- Collinearity
- Interactions
- Dummy variables
- Model selection
- Importance of predictors



Lecture 12: Overview

ANOVA

Analysis of variance: single and multi-factor designs

- Examples: diatoms, circadian rhythms
- Predictor variables: fixed vs. random
- ANOVA model
- Analysis and partitioning of variance
- Null hypothesis
- Assumptions and diagnostics
- Post F Tests - Tukey and others
- Reporting the results

Lecture 12: ANOVA Introduction

What if response continuous and predictor(s) categorical?

	Independent variable	
Dependent variable	**Continuous**	**Categorical**
Continuous	Regression	ANOVA
Categorical	Logistic regression	Tabular

Lecture 12: ANOVA and Regression Connection

Remember

Key Insight

Both regression and ANOVA:

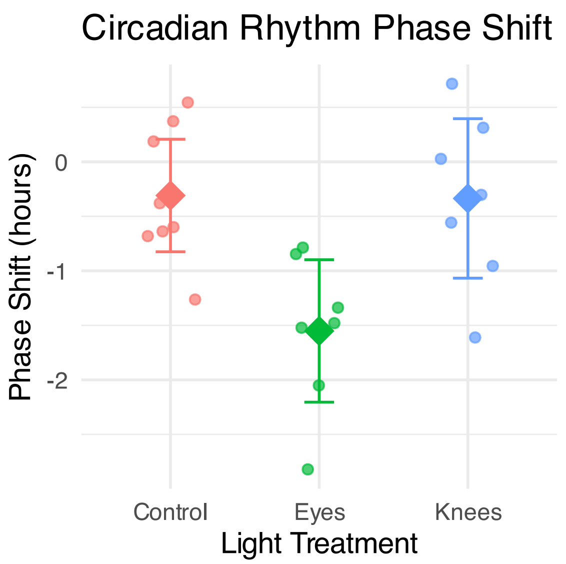
- Partition the total variation in Y
- Use F-tests for significance
- Are based on the General Linear Model
- Test if explanatory variables predict Y ANOVA is fundamentally connected to regression analysis - both are special cases of the General Linear Model.

```
# A tibble: 2 × 3
  Model      Form                 Tests
  <chr>      <chr>                <chr>
1 Regression Y = β₀ + β₁X + ε     H₀: β₁ = 0
2 ANOVA      Yᵢⱼ = μ + Aᵢ + εᵢⱼ   H₀: μ₁ = μ₂ = ... = μₖ
```

Lecture 12: ANOVA Partitioning

General method for partitioning variation in continuous dependent variable

- One or more continuous (and categorical) predictors:
  - regression
- One or more categorical predictors:
  - ANOVA
- Categorical predictor variables:
  - groups or experimental treatments

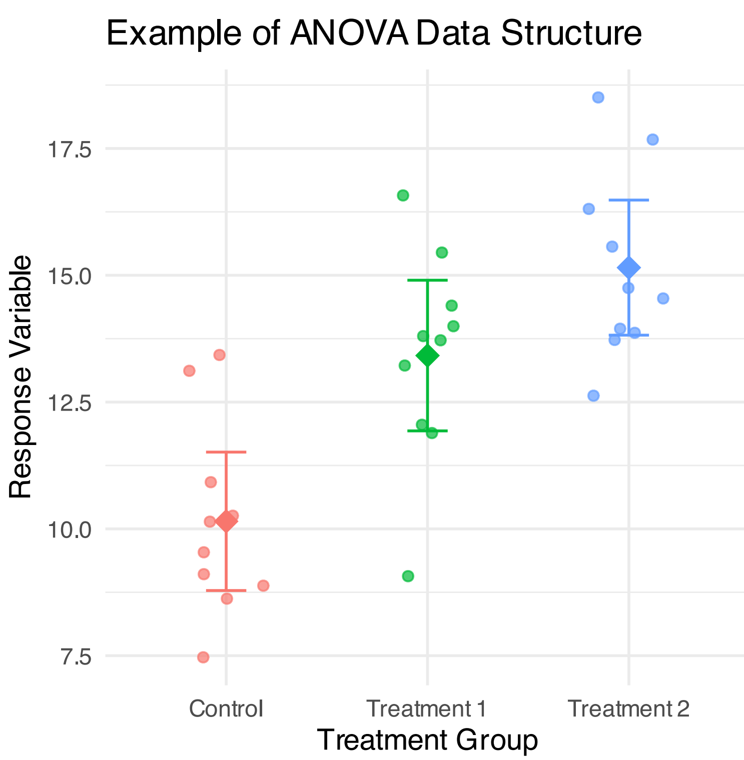


Lecture 12: ANOVA as Regression

ANOVA as Regression

With one categorical variable, ANOVA is equivalent to regression with dummy variables.

In fact when we will run ANOVAs we will use he smae code as for regression! See explanation on oher web page - Will link here



Lecture 12: ANOVA Goals

ANOVA aims to compare means of groups:

- Contribution of predictors + "error" to variability

- Test H_0 that population (random effects) or group (fixed effects) means are equal
- Single factor (1-way) and multifactor (2-, 3-way designs)
  - Single factor: one factor, more than two levels.
- Multifactor:
  - two or three factors, two or more levels.
  - Examines variation due to factors **AND** their interaction

The Analysis of Variance

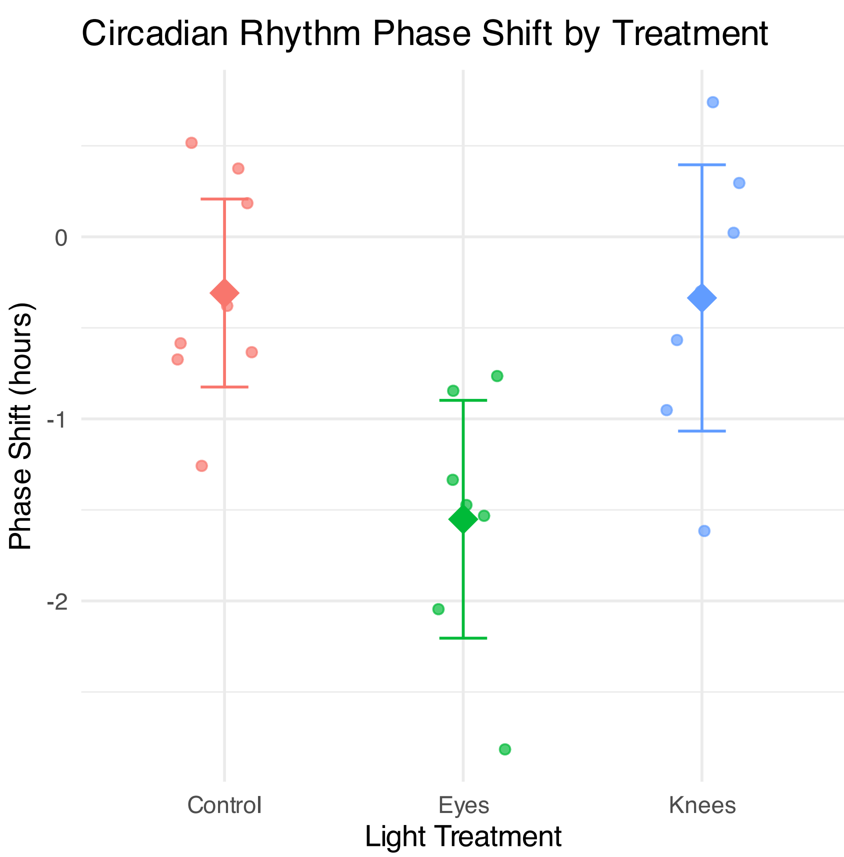
Analysis of variance is the most powerful approach known for simultaneously testing whether the means of k groups are equal. It works by assessing whether individuals chosen from different groups are, on average, more different than individuals chosen from the same group.

The null hypothesis of ANOVA is that the population means μ_i are the same for all treatments.

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1: At least one μ_i is different from the others.

> **i Note**
>
> Rejecting H_0 in ANOVA is evidence that the mean of at least one group is different from the others. It does not indicate *which* means differ.



Lecture 12: ANOVA Logic

Even if all groups had the same true mean, the data would likely show different sample means for each group due to sampling error.

The key insight of ANOVA is that we can estimate how much variation among group means ought to be present from sampling error alone if the null hypothesis is true.

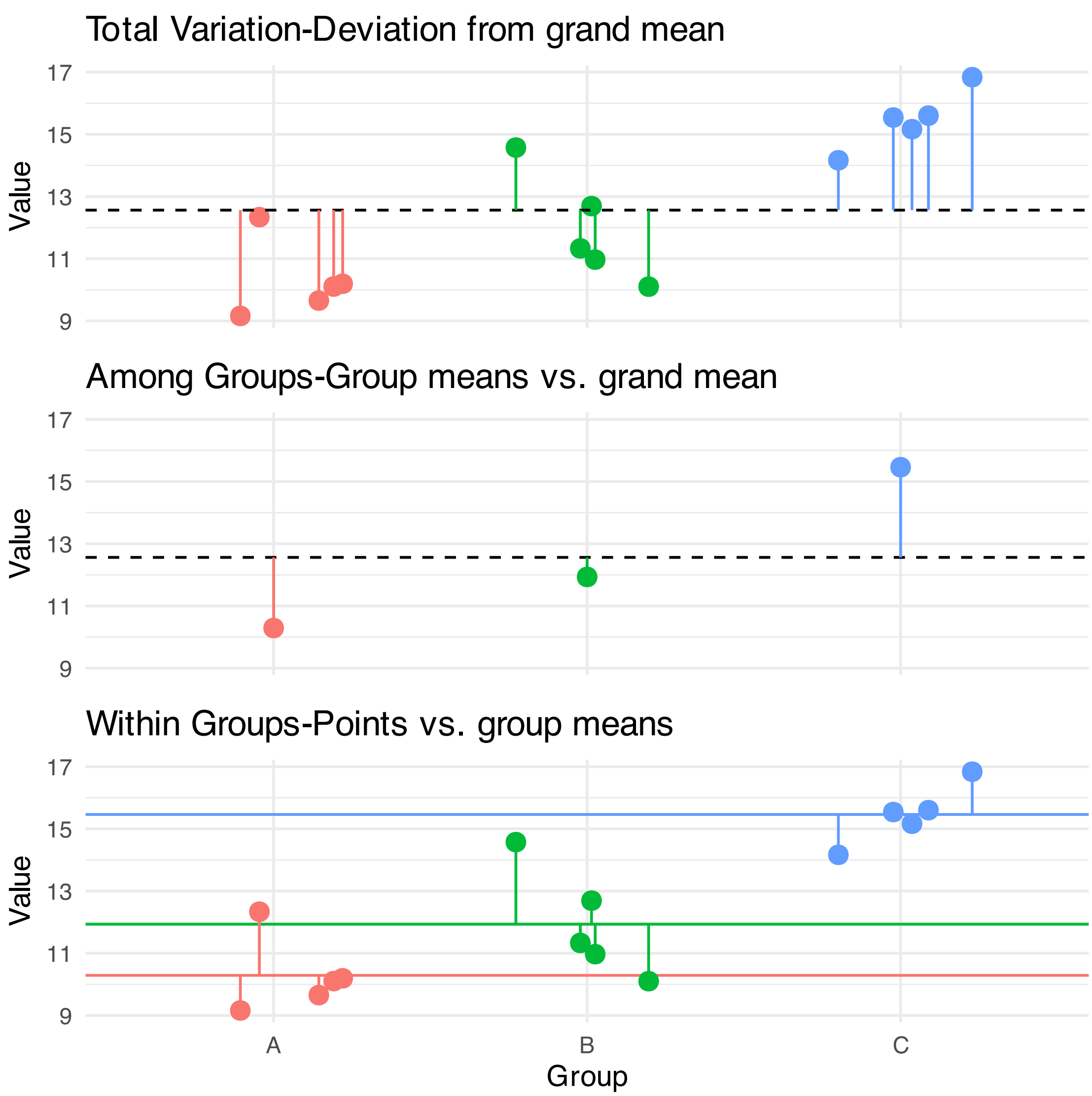
ANOVA lets us determine whether there is more variance among the sample means than we would expect by chance alone. If so, then we can infer that there are real differences among the population means.

Two key measures of variation are calculated and compared:

1. **Group mean square (MSgroups)** - variation among subjects from different groups
2. **Error mean square (MSerror)** - variation among subjects within the same group

The comparison is done with an F-ratio:

$$F = \frac{MS_{groups}}{MS_{error}}$$



Lecture 12: Partitioning the Sum of Squares

The total variation in Y can be expressed as a sum of squares:

$SS_{total} = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}\right)^2$

This can be partitioned into two components:

1. **Among Groups (Treatment)**: $SS_{among} = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(\bar{Y_i} - \bar{Y}\right)^2 = n \sum_{i=1}^{a} \left(\bar{Y_i} - \bar{Y}\right)^2$
2. **Within Groups (Error)**: $SS_{within} = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y_i}\right)^2$

These components are additive: $SS_{total} = SS_{among} + SS_{within}$

Lecture 12: Sum of Squares Example

```
Analysis of Variance Table

Response: phase_shift
          Df  Sum Sq Mean Sq F value   Pr(>F)
treatment  2 2.23686 1.11843   16.05 0.001076 **
Residuals  9 0.62714 0.06968
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# A tibble: 3 × 4
  Component     `Sum of Squares` `Degrees of Freedom` `Mean Square`
  <chr>                    <dbl>                <dbl>         <dbl>
1 Total                    2.86                    11       NA
2 Among Groups             2.24                     2        1.12
3 Within Groups            0.627                    9        0.0697
```

! Key Connection to Regression

This is the same partitioning we saw in regression analysis: $SS_{total} = SS_{regression} + SS_{residual}$

Where:

- SS_{among} in ANOVA = $SS_{regression}$ in regression
- SS_{within} in ANOVA = $SS_{residual}$ in regression

Both measure how much variation is explained by our model vs. unexplained (error).

Lecture 12: ANOVA Tables

The ANOVA table organizes all computations leading to a test of the null hypothesis of no differences among population means.

- **Source of variation**: What is being tested
- **Sum of squares**: Measure of total variation for each source
- **df**: Degrees of freedom for each source
- **Mean squares**: Sum of squares divided by df
- **F-ratio**: Ratio of mean squares, used to test significance
- **P-value**: Probability of observing our results if H_0 is true

Example: For a one-way ANOVA with 3 groups and 4 replicates per group:

- df for treatments = (a - 1) = 2
- df for error = a(n - 1) = 3(4 - 1) = 9
- df total = an - 1 = 11

Lecture 12: Circadian Rhythm Data Example

```
Analysis of Variance Table

Response: phase_shift
          Df Sum Sq Mean Sq F value   Pr(>F)
treatment  2 7.2245  3.6122  7.2894 0.004472 **
Residuals 19 9.4153  0.4955
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# A tibble: 3 × 4
  treatment   Mean    SD     N
  <fct>      <dbl> <dbl> <int>
1 Control   -0.309 0.618     8
2 Eyes      -1.55  0.706     7
3 Knees     -0.336 0.791     7
```

Lecture 12: ANOVA vs Regression Tables

! Comparing ANOVA and Regression Tables

An ANOVA table from an ANOVA model:

Source	df	SS	MS	F	p
Treatment	a-1	SS_treatment	MS_treatment	F	p
Error	a(n-1)	SS_error	MS_error		
Total	an-1	SS_total			

Is equivalent to an ANOVA table from a regression model:

Source	df	SS	MS	F	p
Regression	k	SS_regression	MS_regression	F	p
Error	n-k-1	SS_residual	MS_residual		
Total	n-1	SS_total			

where k = number of dummy variables = a-1

Lecture 12: F ratio

The F-ratio is calculated as:

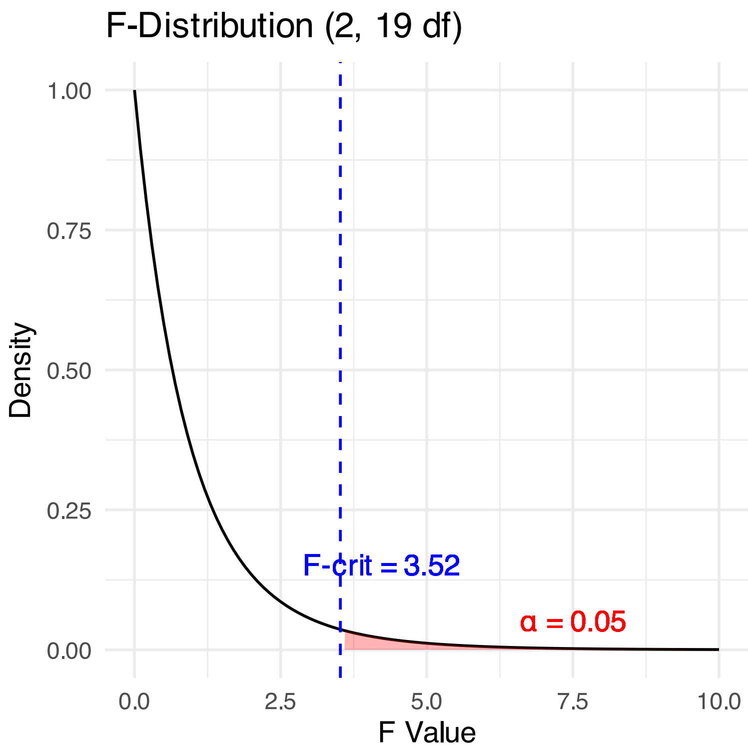
$$F = \frac{MS_{among}}{MS_{error}}$$

Under the null hypothesis (all means equal): - The F-ratio should be approximately 1 - Larger F-ratios suggest the among-group variance exceeds what would be expected by chance

With the circadian rhythm data: - F = 7.29 - p = 0.004 - We reject the null hypothesis

The F-ratio follows an F-distribution with (a - 1) and (a(n - 1)) degrees of freedom.

```
# A tibble: 2 × 2
  Metric                Value
  <chr>                 <dbl>
1 F-observed             7.29
2 F-critical (α = 0.05)  3.52
```



ℹ Connection to t-test

An ANOVA with two groups (a = 2) is equivalent to a t-test:

$$F = t^2$$

Lecture 12: F ratio Visualization

The F-ratio is calculated as:

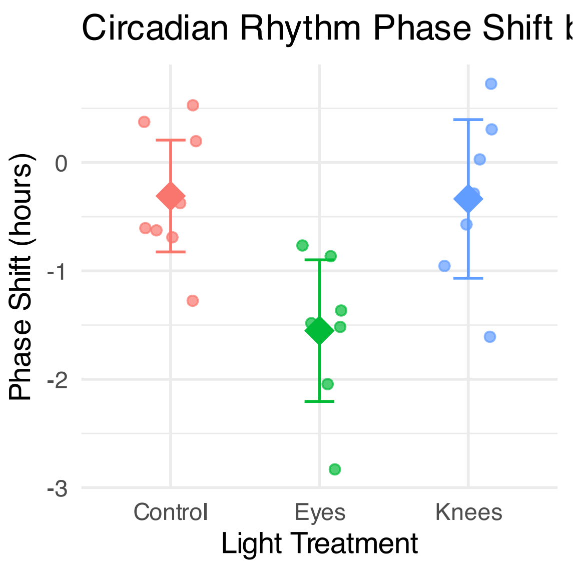
$$F = \frac{MS_{among}}{MS_{error}}$$

Under the null hypothesis (all means equal): - The F-ratio should be approximately 1 - Larger F-ratios suggest the among-group variance exceeds what would be expected by chance

With the circadian rhythm data: - F = 7.29 - p = 0.004 - We reject the null hypothesis

The F-ratio follows an F-distribution with (a - 1) and (a(n - 1)) degrees of freedom.

```
# A tibble: 2 × 2
  Metric                Value
  <chr>                 <dbl>
1 F-observed             7.29
2 F-critical (α = 0.05)  3.52
```



Lecture 12: Variation Explained: R2

R^2 summarizes the contribution of group differences to total variation:

$$R^2 = \frac{SS_{among}}{SS_{total}}$$

This is interpreted as the "fraction of the variation in Y that is explained by groups."

For the circadian rhythm data:

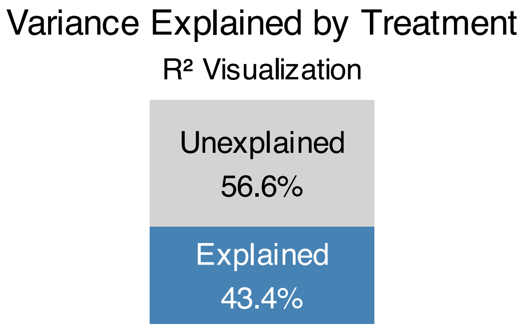
$$R^2 = \frac{7.224}{16.639} = 0.43$$

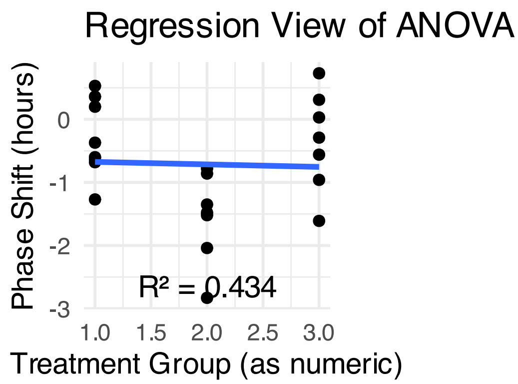
43% of the total variation in phase shift is explained by differences in light treatment, with the remaining 57% being unexplained variation.

Connection to Regression

This is exactly the same calculation as R^2 in regression:

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$





Lecture 12: ANOVA Assumptions

ANOVA has the same assumptions as the two-sample t-test, but applied to all k groups:

1. **Random samples** from corresponding populations
2. **Normality**: Y values are normally distributed in each population
3. **Homogeneity of variance**: variance is the same in all populations
4. **Independence**: observations are independent

Checking assumptions:

- Normality: Q-Q plots, histogram of residuals, Shapiro-Wilk test
- Homogeneity: plot residuals vs. predicted values or x-values
- Independence: examine experimental design

If assumptions are violated:

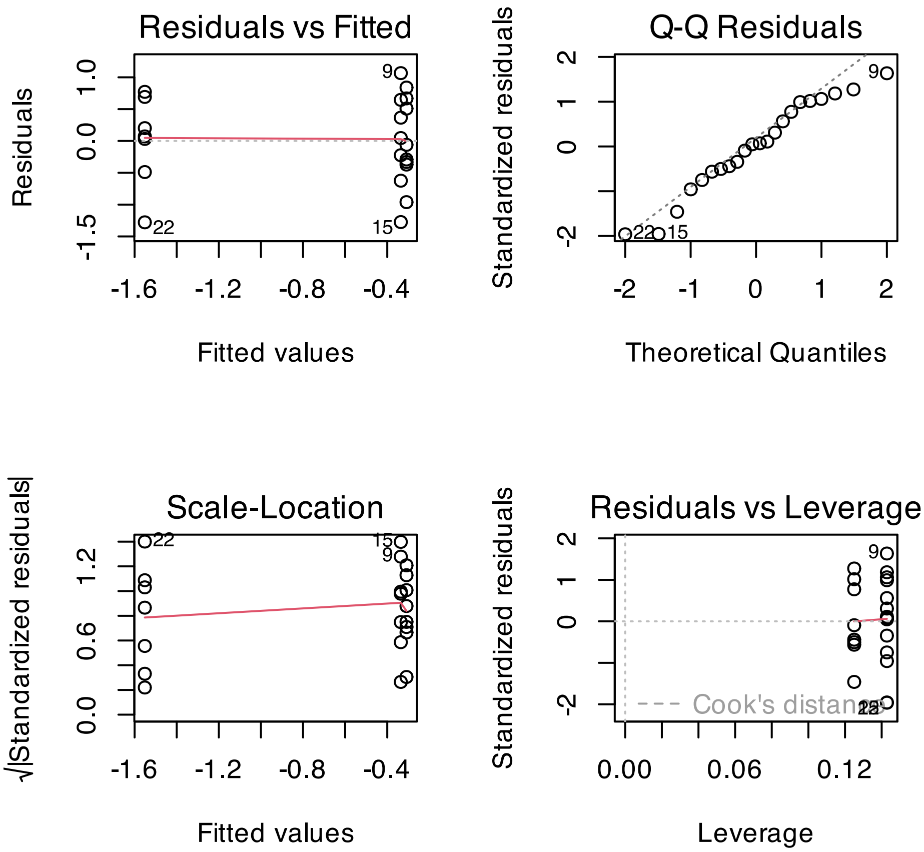
- Transform Y (e.g., log, square root)
- Use robust or non-parametric alternatives
- Use generalized linear models (GLMs)

Lecture 12: ANOVA diagnostics

This is the default output of base R

```
# Model diagnostics
par(mfrow = c(2, 2))
plot(circ_model)
dev.off() # This forces the plot to be written
```

```
null device
          1
```



A newer way to check with the performance library

```
# install.packages("performance")
library(performance)
check_model(circ_model)
```

Lecture 12: Levene's Test

Levene's test of homogeneity of variance Null Hypothesis is that they are homogeneous So you want a non significant result here

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.1586 0.8545
      19
```

Lecture 12: Shapiro-Wilk Test

Shapiro-Wilk Normality Test Null Hypothesis is that they are normally distributed So you want a non significant result here

```

	Shapiro-Wilk normality test
```

```
data:  residuals(circ_model)
W = 0.95893, p-value = 0.468
```

i Shared Assumptions with Regression

ANOVA and regression share virtually identical assumptions because they are both linear models:

Assumption	ANOVA	Regression
Linearity	Relationship between group membership and Y is additive	Relationship between X and Y is linear
Normality	Residuals within each group are normal	Residuals are normal
Equal variance	Variance is the same across all groups	Variance is the same across all X values
Independence	Observations are independent	Observations are independent

Lecture 12: ANOVA Post-Hoc Testing Overview

When ANOVA rejects H_0, we need to determine which groups differ.

Planned comparisons: - Identified during study design - Have strong prior justification - Use pooled variance from all groups - Have higher precision than separate t-tests

Unplanned (post hoc) comparisons: - Used when no specific comparisons were planned - Must adjust for multiple testing - Common methods: Tukey-Kramer, Bonferroni, Scheffé

Example: Using Tukey's HSD to compare all pairs of treatments in the circadian rhythm data.

```
 contrast        estimate    SE df t.ratio p.value
 Control - Eyes     1.243 0.364 19   3.411  0.0079
 Control - Knees    0.027 0.364 19   0.074  0.9970
 Eyes - Knees      -1.216 0.376 19  -3.231  0.0117

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

Lecture 12: Post-Hoc Testing Results

When ANOVA rejects H_0, we need to determine which groups differ.

Planned comparisons: - Identified during study design - Have strong prior justification - Use pooled variance from all groups - Have higher precision than separate t-tests

Unplanned (post hoc) comparisons: - Used when no specific comparisons were planned - Must adjust for multiple testing - Common methods: Tukey-Kramer, Bonferroni, Scheffé

Example: Using Tukey's HSD to compare all pairs of treatments in the circadian rhythm data.

```
 treatment emmean    SE df lower.CL upper.CL .group
 Eyes      -1.551 0.266 19   -2.108   -0.995  a
 Knees     -0.336 0.266 19   -0.893    0.221   b
 Control   -0.309 0.249 19   -0.830    0.212   b

Confidence level used: 0.95
```

```
P value adjustment: tukey method for comparing a family of 3 estimates
significance level used: alpha = 0.05
NOTE: If two or more means share the same grouping symbol,
      then we cannot show them to be different.
      But we also did not show them to be the same.
```

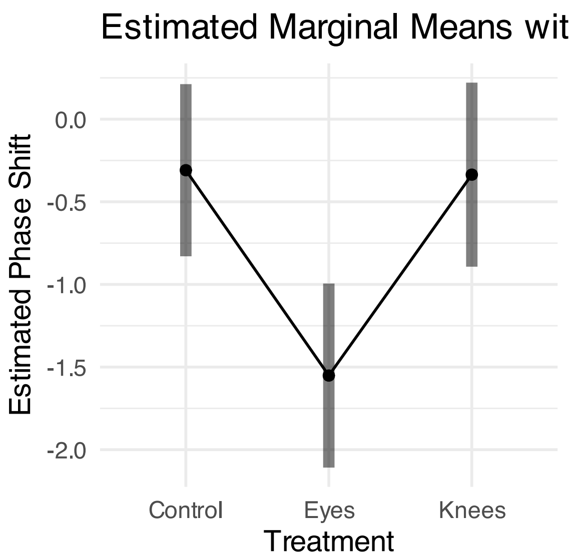
Lecture 12: Post-Hoc Visualization

When ANOVA rejects H_0, we need to determine which groups differ.

Planned comparisons: - Identified during study design - Have strong prior justification - Use pooled variance from all groups - Have higher precision than separate t-tests

Unplanned (post hoc) comparisons: - Used when no specific comparisons were planned - Must adjust for multiple testing - Common methods: Tukey-Kramer, Bonferroni, Scheffé

Example: Using Tukey's HSD to compare all pairs of treatments in the circadian rhythm data.



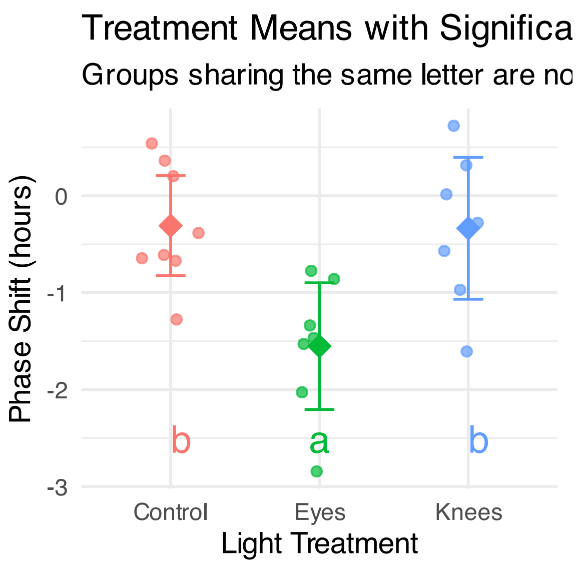
Lecture 12: Significance Groups Plot

When ANOVA rejects H_0, we need to determine which groups differ.

Planned comparisons: - Identified during study design - Have strong prior justification - Use pooled variance from all groups - Have higher precision than separate t-tests

Unplanned (post hoc) comparisons: - Used when no specific comparisons were planned - Must adjust for multiple testing - Common methods: Tukey-Kramer, Bonferroni, Scheffé

Example: Using Tukey's HSD to compare all pairs of treatments in the circadian rhythm data.



Lecture 12: Reporting ANOVA Results

Formal scientific writing example:

"The effect of light treatment on circadian rhythm phase shift was analyzed using a one-way ANOVA. There was a significant effect of treatment on phase shift ($F(2, 19) = 7.29$, $p = 0.004$, $\eta^2 = 0.43$). Post-hoc comparisons using Tukey's HSD test indicated that the mean phase shift for the Eyes treatment ($M = -1.55$ h, $SD = 0.71$) was significantly different from both the Control treatment ($M = -0.31$ h, $SD = 0.62$) and the Knees treatment ($M = -0.34$ h, $SD = 0.79$). However, the Control and Knees treatments did not significantly differ from each other. These results suggest that light exposure to the eyes, but not to the knees, impacts circadian rhythm phase shifts."

Lecture 12: ANOVA Summary

Key ANOVA Principles

1. **Purpose**: ANOVA (Analysis of Variance) compares means across multiple groups simultaneously
2. **Connection to Regression**:
   - Both are special cases of the General Linear Model
   - ANOVA with categorical predictors = Regression with dummy variables
   - Both partition variance into explained and unexplained components
3. **The Analysis of Variance**:
   - Partitions total variation into components
   - Tests whether differences among groups exceed what would be expected by chance
   - Uses F-tests to compare variance between groups to variance within groups
4. **Sum of Squares Partitioning**:
   - SS(Total) = SS(Between Groups) + SS(Within Groups)
   - Same as SS(Total) = SS(Regression) + SS(Error) in regression
5. **Fixed vs. Random Effects**:
   - Fixed effects: specific groups of interest (most common)
   - Random effects: sampling from a larger population

ANOVA Assumptions

1. Independence of observations

2. Normal distribution of residuals
3. Homogeneity of variances