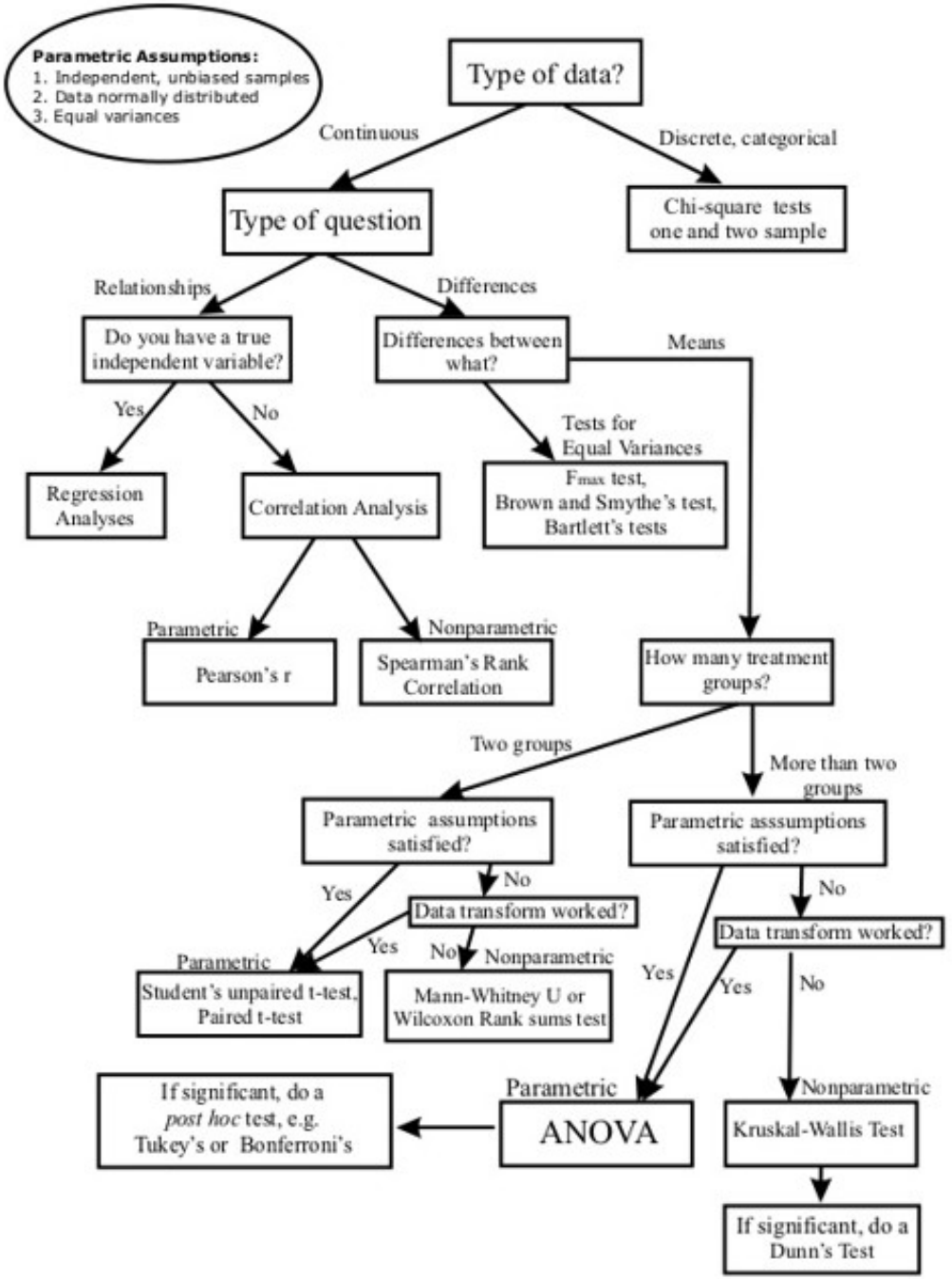
Day 4: Linear Models

A Unified Framework for Data Analysis

Flow Chart for Selecting Commonly Used Statistical Tests



Today's Journey: One Dataset, Many Questions

Our Running Example: Drosophila Weight Study

Research question: How does diet affect fly weight?

Data: 40 flies measured under two conditions

- **Diet:** Restricted vs. Control (between subjects)
- **Sex:** Male vs. Female (biological variable)
- **Response:** Weight (mg)

Questions we'll answer:

1. Does diet affect weight? (Simple linear regression)
2. Do males and females differ? (Adding a categorical predictor)
3. Does diet effect depend on sex? (Interactions)
4. What if we measured same flies on both diets? (Paired data)

All of these are LINEAR MODELS!

Overview

Part I: Introduction to Linear Models
The general framework and why it matters

Part II: Simple Linear Regression
One continuous predictor

Part III: Multiple Regression
Multiple predictors, interpretation of coefficients

Part IV: Categorical Predictors
t-tests and ANOVA as linear models

Part V: Interactions
When effects depend on other variables

Part VI: Repeated Measures / Paired Data
Multiple intercepts, partial pooling

Part VII: Model Diagnostics
Checking assumptions and improving models

Part I

Introduction to Linear Models

A unified framework

What is a Linear Model?

General Form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

where:

- Y = response variable (what we're trying to explain)
- X_i = predictor variables (explanatory variables)
- β_i = coefficients (parameters to estimate)
- ϵ = error term, $\epsilon \sim N(0, \sigma^2)$

Key properties:

1. **Linear in parameters** (β), not necessarily in X
2. Can include transformations: X^2, $\log(X)$, etc.
3. Can include categorical variables (via dummy coding)
4. Can include interactions: $X_1 \times X_2$

Why Linear Models Matter

Linear models unify many classical tests:

Classical Test	Linear Model Formulation
One-sample t-test	$Y = \beta_0 + \epsilon$
Two-sample t-test	$Y = \beta_0 + \beta_1 \text{Group} + \epsilon$
Paired t-test	$Y = \beta_0 + \beta_1 \text{Time} + \alpha_i + \epsilon$
ANOVA	$Y = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \cdots + \epsilon$
ANCOVA	$Y = \beta_0 + \beta_1 X + \beta_2 \text{Group} + \epsilon$
Correlation	$Y = \beta_0 + \beta_1 X + \epsilon$

Advantages of linear model framework:

- Handles complex designs naturally
- Clear interpretation via coefficients
- Easy to extend (GLMs, mixed models, etc.)
- Single computational approach

Assumptions of Linear Models

Key assumptions (remember: LINE):

1. **L**inearity: Relationship between X and Y is linear
2. **I**ndependence: Observations are independent
3. **N**ormality: Residuals ϵ are normally distributed
4. **E**qual variance (Homoscedasticity): Variance of residuals is constant

Important

- We assume normality of residuals, not raw data!
- For large n, normality less critical (CLT for coefficients)
- Independence is the most critical assumption

Part II

Simple Linear Regression

One continuous predictor

Simple Linear Regression Model

Model

$$\text{Weight}_i = \beta_0 + \beta_1 \times \text{Diet}_i + \epsilon_i$$

where Diet is coded numerically (e.g., 0 = Control, 1 = Restricted)

Parameters:

- β_0 = **intercept** (expected weight when Diet = 0)
- β_1 = **slope** (change in weight for 1-unit change in Diet)
- σ^2 = residual variance (variability not explained by model)

Goal: Estimate β_0, β_1, and σ^2 from data

Estimation method: Ordinary Least Squares (OLS) or Maximum Likelihood

Interpreting Coefficients

Example: Diet Effect on Weight

$$\widehat{\text{Weight}} = 60 + 15 \times \text{Diet}$$

Interpretation:

- $\hat{\beta}_0 = 60$: Flies on control diet (Diet=0) live 60 mg on average
- $\hat{\beta}_1 = 15$: Switching from control to restricted diet increases weight by 15 mg on average

Predictions:

- Control diet (Diet=0): $\hat{Y} = 60 + 15(0) = 60$ mg
- Restricted diet (Diet=1): $\hat{Y} = 60 + 15(1) = 75$ mg

Coefficients have units: slope is in (response units) / (predictor units)

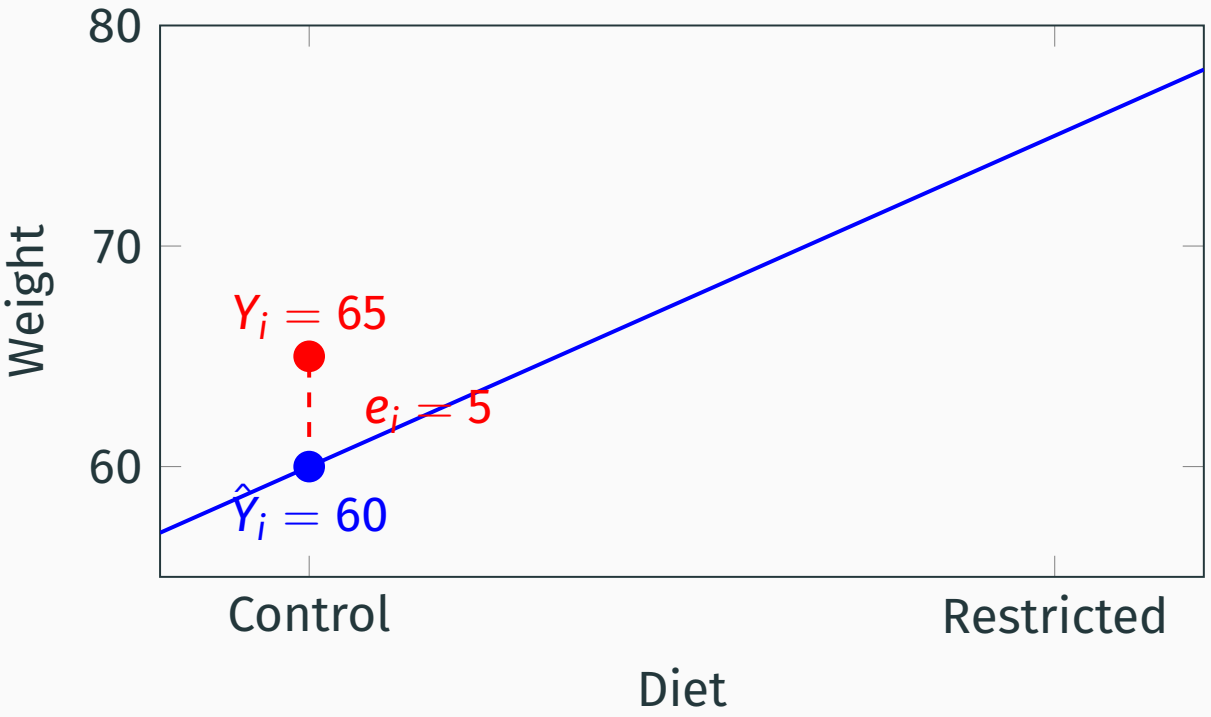
Visualizing the Regression Line

Residuals: Observed vs. Fitted

Residual

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

Difference between observed and predicted values



OLS minimizes: $\sum_{i=1}^{n} e_i^2$ (sum of squared residuals)

Model Fit: R^2

Coefficient of Determination

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Proportion of variance explained by the model

Interpretation:

- $R^2 = 0$: Model explains no variance (terrible fit)
- $R^2 = 1$: Model explains all variance (perfect fit)
- $R^2 = 0.6$: Model explains 60% of variance

Caution

- High R^2 doesn't mean the model is "correct" or "useful"
- Low R^2 doesn't mean the relationship isn't important
- In biology, R^2 often moderate (lots of uncontrolled variation)
- Focus on coefficient interpretation and inference

Inference for Coefficients

Question: Is the slope significantly different from zero?

Hypothesis Test

$H_0 : \beta_1 = 0$ (no effect of diet)
$H_A : \beta_1 \neq 0$ (diet has an effect)

Test statistic:

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

Calculate p-value from t-distribution

Confidence Interval

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \times \text{SE}(\hat{\beta}_1)$$

Quantifies uncertainty in the slope estimate

Part III

Multiple Regression

Multiple predictors

Adding a Second Predictor: Sex

Question: Do males and females have different weights?

Model with Two Predictors

$$\text{Weight}_i = \beta_0 + \beta_1 \times \text{Diet}_i + \beta_2 \times \text{Sex}_i + \epsilon_i$$

where Sex is coded: 0 = Female, 1 = Male

Now we have three parameters:

- β_0 = weight for females on control diet (baseline)
- β_1 = effect of restricted diet (holding sex constant)
- β_2 = difference between males and females (holding diet constant)

Key Concept

Each coefficient represents the effect holding other variables constant

Interpreting Multiple Regression Coefficients

Example: Diet + Sex Model

$$\widehat{\text{Weight}} = 65 + 12 \times \text{Diet} - 8 \times \text{Sex}$$

Interpretation:

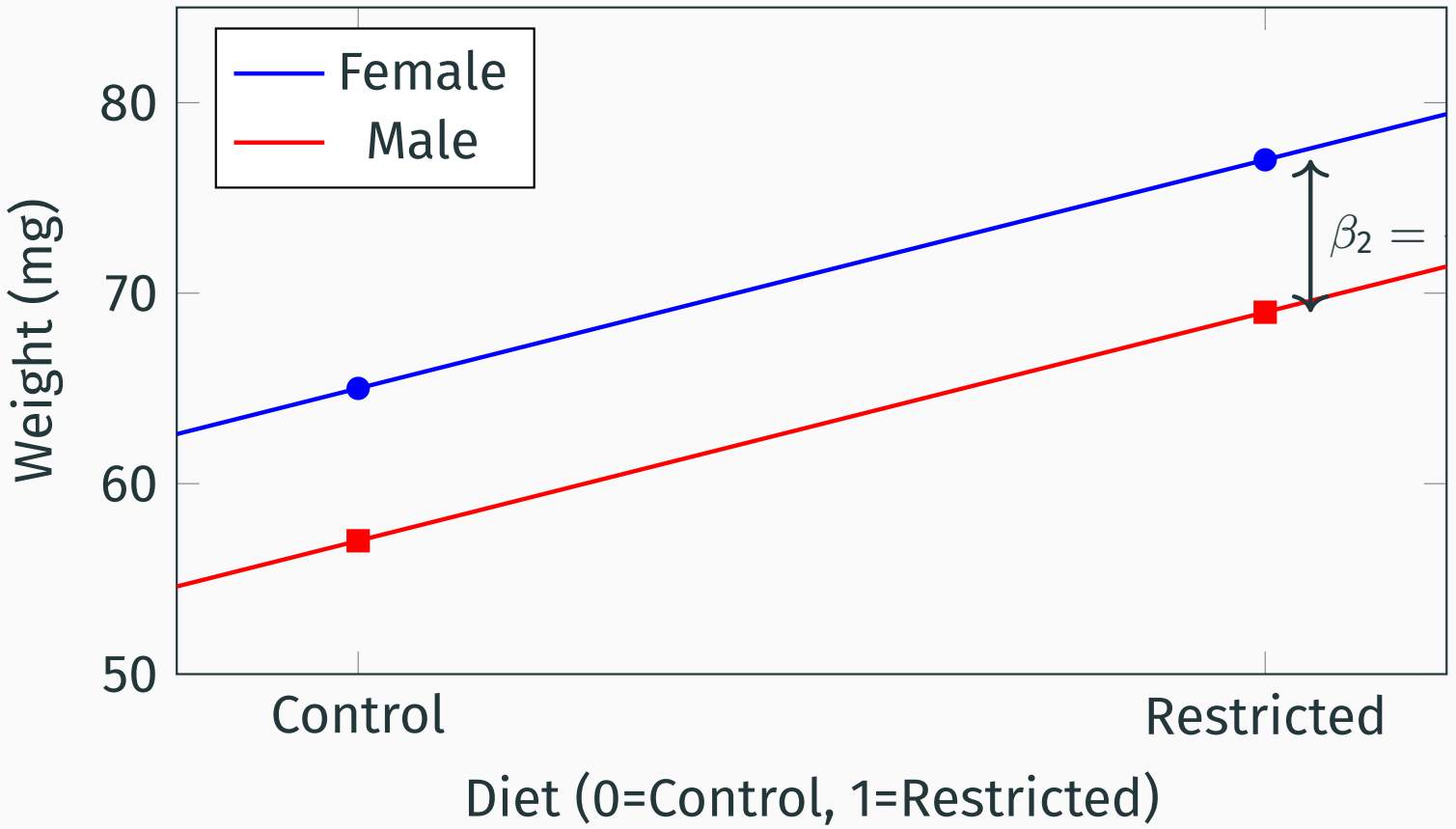
- $\hat{\beta}_0 = 65$: Baseline (female, control diet) weight = 65 mg
- $\hat{\beta}_1 = 12$: Restricted diet increases weight by 12 mg, for both sexes
- $\hat{\beta}_2 = -8$: Males live 8 mg less than females, on both diets

Predictions for all four groups:

- Female, Control: $65 + 12(0) - 8(0) = 65$ mg
- Female, Restricted: $65 + 12(1) - 8(0) = 77$ mg
- Male, Control: $65 + 12(0) - 8(1) = 57$ mg
- Male, Restricted: $65 + 12(1) - 8(1) = 69$ mg

Visualizing Multiple Regression

Model: Weight = 65 + 12×Diet - 8×Sex



Note: Parallel lines → no interaction (effect of diet same for both sexes)

Part IV

Categorical Predictors

t-tests and ANOVA as linear models

Dummy (Indicator) Variables

Problem: How to include categorical variables in linear models?

Dummy Coding

Create binary (0/1) variables for each category (except reference)

Example: Diet with 2 levels

Diet	Dummy variable
Control (reference)	0
Restricted	1

Model:

$$\text{Weight} = \beta_0 + \beta_1 \times I_{\text{Restricted}} + \epsilon$$

- β_0 = mean weight on control diet
- β_1 = difference (Restricted - Control)

This is exactly a two-sample t-test!

Categorical Variable with 3+ Levels

Example: Diet with 3 levels (Control, Restricted, High-protein)

Need 2 Dummy Variables

Diet	$I_{\text{Restricted}}$	$I_{\text{High-protein}}$
Control (reference)	0	0
Restricted	1	0
High-protein	0	1

Model:

$$\text{Weight} = \beta_0 + \beta_1 I_{\text{Restricted}} + \beta_2 I_{\text{High-protein}} + \epsilon$$

- β_0 = mean weight on control diet
- β_1 = difference (Restricted - Control)
- β_2 = difference (High-protein - Control)

This is one-way ANOVA!

Two-Sample t-test = Simple Linear Model

Traditional approach:

- Calculate means for both groups: $\bar{Y}_{\text{control}}$, $\bar{Y}_{\text{restricted}}$
- Calculate difference and SE
- Test H_0: means are equal

Linear model approach:

$$Y = \beta_0 + \beta_1 \times \text{Diet} + \epsilon$$

Test $H_0 : \beta_1 = 0$

They Give Identical Results!

- Same t-statistic
- Same p-value
- Same confidence interval

But linear model framework extends naturally to complex designs

ANOVA Table from Linear Model

Partition total variance:

Source	SS	df	MS	F
Model	$\sum(\hat{Y}_i - \bar{Y})^2$	p	SS/df	$\text{MS}_{\text{model}}/\text{MS}_{\text{resid}}$
Residual	$\sum(Y_i - \hat{Y}_i)^2$	$n-p-1$	SS/df	
Total	$\sum(Y_i - \bar{Y})^2$	$n-1$		

F-test:

$$F = \frac{\text{MS}_{\text{model}}}{\text{MS}_{\text{residual}}} \sim F_{p,n-p-1}$$

Tests H_0: all slope coefficients = 0 (model has no predictive value)

For simple regression: $F = t^2$ (F-test = squared t-test)

Part V

Interactions

When effects depend on other variables

What is an Interaction?

Interaction

The effect of one variable depends on the level of another variable

Synonyms: Effect modification, moderation, non-additivity

Biological examples:

- Drug effect differs by genotype
- Temperature effect differs by species
- Nutrient effect differs by developmental stage
- Diet effect differs by sex (our example!)

Two Scenarios

No interaction (additive): Diet increases weight by same amount in both sexes → parallel lines

Interaction: Diet effect larger in one sex than the other → non-parallel lines

Linear Model with Interaction

Model

$$\text{Weight} = \beta_0 + \beta_1 \text{Diet} + \beta_2 \text{Sex} + \beta_3 (\text{Diet} \times \text{Sex}) + \epsilon$$

where Diet and Sex are dummy coded (0/1)

Parameters:

- β_0 = baseline (female, control diet)
- β_1 = effect of diet in females (when Sex=0)
- β_2 = difference males-females on control diet (when Diet=0)
- β_3 = interaction term = how much diet effect differs between sexes

Diet effect by sex:

- In females (Sex=0): β_1
- In males (Sex=1): $\beta_1 + \beta_3$
- Difference in diet effects: β_3

Test for interaction: $H_0 : \beta_3 = 0$

Interpreting Interaction Coefficients

Example: Model with Interaction

$$\widehat{\text{Weight}} = 63 + 18 \times \text{Diet} - 6 \times \text{Sex} - 10 \times (\text{Diet} \times \text{Sex})$$

Predictions for all four groups:

- Female, Control: $63 + 18(0) - 6(0) - 10(0)(0) = 63$ mg
- Female, Restricted: $63 + 18(1) - 6(0) - 10(1)(0) = 81$ mg
- Male, Control: $63 + 18(0) - 6(1) - 10(0)(1) = 57$ mg
- Male, Restricted: $63 + 18(1) - 6(1) - 10(1)(1) = 65$ mg

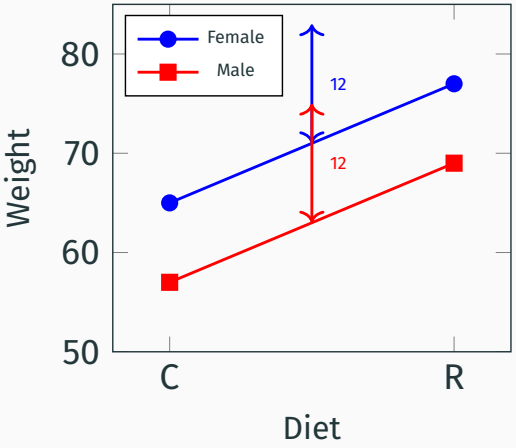
Diet effect by sex:

- In females: $81 - 63 = 18$ mg increase (β_1)
- In males: $65 - 57 = 8$ mg increase ($\beta_1 + \beta_3$)
- Interaction: $18 - 8 = 10$ mg (β_3)

Conclusion: Diet benefit is 10 mg larger in females than males

Visualizing Interactions

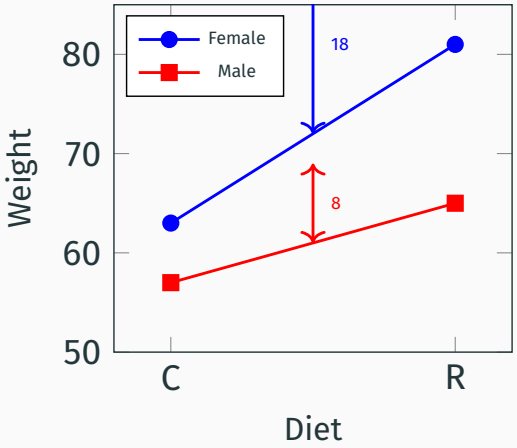
No Interaction (Parallel)



$\beta_3 = 0$

Same diet effect in both sexes

With Interaction (Non-parallel)



$\beta_3 \neq 0$

Diet effect differs by sex

Interaction = difference in slopes = non-parallelism

Building Models Hierarchically

Principle: Always include main effects when including interactions

Model Hierarchy

1. **Main effects only:**
   $Y = \beta_0 + \beta_1 \text{Diet} + \beta_2 \text{Sex} + \epsilon$

2. **Add interaction:**
   $Y = \beta_0 + \beta_1 \text{Diet} + \beta_2 \text{Sex} + \beta_3 (\text{Diet} \times \text{Sex}) + \epsilon$

Testing strategy:

1. Fit model with interaction
2. Test $H_0 : \beta_3 = 0$
3. If interaction not significant → use simpler model (main effects only)
4. If interaction significant → keep interaction, interpret carefully

Never Include Interaction Without Main Effects!

Coefficients become uninterpretable

Interactions: Continuous × Continuous

Interactions aren't just for categorical variables!

Example: Body Size and Temperature

$$\text{Metabolic Rate} = \beta_0 + \beta_1 \text{Mass} + \beta_2 \text{Temp} + \beta_3 (\text{Mass} \times \text{Temp}) + \epsilon$$

Interpretation:

- β_1 = effect of mass when Temp = 0 (usually not meaningful!)
- β_2 = effect of temperature when Mass = 0 (not meaningful!)
- β_3 = how much the effect of mass changes per unit increase in temperature

Better to center predictors first:

- Mass* = Mass - mean(Mass)
- Temp* = Temp - mean(Temp)
- Now β_1 = effect of mass at average temperature

Centering Predictors

Why Center?
Makes intercept and main effects interpretable when interactions present

Centering: $X_i^* = X_i - \bar{X}$

Example: Mass × Temperature
Uncentered:

$$Y = \beta_0 + \beta_1 \text{Mass} + \beta_2 \text{Temp} + \beta_3 (\text{Mass} \times \text{Temp})$$

β_1 = effect of mass when Temp = 0°C (often outside data range!)

Centered:

$$Y = \beta_0 + \beta_1 \text{Mass}^* + \beta_2 \text{Temp}^* + \beta_3 (\text{Mass}^* \times \text{Temp}^*)$$

β_1 = effect of mass at average temperature (more meaningful!)

Centering changes β_0, β_1, β_2 but NOT β_3 or predictions!

Higher-Order Interactions

Can have interactions among 3+ variables (but interpret carefully!)

Three-Way Interaction

$$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 (A \times B) + \beta_5 (A \times C) + \beta_6 (B \times C) + \beta_7 (A \times B \times C) + \epsilon$$

β_7 = how much the $A \times B$ interaction differs across levels of C

Biology Example

Does the diet-by-sex interaction differ between young and old flies?

Effect of diet may:

- Differ by sex (2-way: Diet × Sex)
- Differ by age (2-way: Diet × Age)
- AND the sex difference in diet effect may differ by age (3-way!)

Warning

Three-way interactions are hard to interpret and often underpowered. Use sparingly!

Testing for Interactions

Approach 1: t-test on interaction coefficient

Test $H_0 : \beta_3 = 0$ using t-test

Approach 2: F-test comparing models

- Model 1 (reduced): Main effects only
- Model 2 (full): Main effects + interaction
- Test: Does Model 2 fit significantly better?

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/(\text{df}_1 - \text{df}_2)}{\text{RSS}_2/\text{df}_2}$$

Power Consideration

Interactions require more power to detect than main effects

Rule of thumb: Need 4× the sample size to detect interaction vs. main effect of same magnitude

When to Include Interactions?

Include interactions when:

1. **A priori hypothesis:** Theory or prior research suggests interaction
2. **Exploratory finding:** Data show clear non-parallelism
3. **Biological plausibility:** Makes sense that effects would differ

Don't include interactions when:

1. No theoretical reason
2. "Fishing expedition" (testing all possible interactions)
3. Insufficient sample size

Caution

- Each interaction term uses degrees of freedom
- Multiple interactions → multiple testing problem
- Non-significant interaction doesn't prove effects are identical

Reporting Interactions

If interaction is significant:

1. Report the interaction test first
2. Present estimates for each group/level separately
3. Visualize with separate lines/groups
4. Don't over-interpret main effects in isolation

Good Example

"We found a significant diet-by-sex interaction ($F(1,36) = 5.2$, $p = 0.028$). The restricted diet increased weight by 18 mg in females ($t = 4.3$, $p < 0.001$) but only 8 days in males ($t = 2.1$, $p = 0.042$). This 10-day difference in diet effect between sexes was statistically significant."

Bad Example

"There was a significant effect of diet ($p < 0.05$) and a diet-by-sex interaction ($p < 0.05$)."

(Doesn't say what the interaction means!)

Common Mistakes with Interactions

1. **Including interaction without main effects**
   Solution: Always include main effects

2. **Interpreting main effects when interaction present**
   Solution: Main effects are ”conditional” when interaction present—report group-specific effects

3. **Concluding ”no interaction” from p > 0.05**
   Solution: Absence of evidence ≠ evidence of absence; may lack power

4. **Testing all possible interactions**
   Solution: Multiple testing problem; only test interactions with a priori justification

5. **Not centering continuous predictors**
   Solution: Center to make coefficients interpretable

6. **Overinterpreting three-way interactions**
   Solution: These are hard to interpret and often spurious; need strong justification

Summary: Interactions

Key Concepts

- **Interaction:** Effect of one variable depends on another
- **Visualization:** Non-parallel lines
- **Test:** Coefficient on product term
- **Interpretation:** Difference in slopes

Best Practices

- Always include main effects with interactions
- Center continuous predictors
- Visualize interactions (don't just report p-values)
- Report group-specific effects when interaction significant
- Test interactions only with theoretical justification
- Be cautious with three-way+ interactions

Interactions are common in biology—effects rarely identical across all contexts!

Part VI

Repeated Measures & Paired Data

When observations are not independent

The Problem: Non-Independence

Recall: Linear models assume observations are independent

But in biology, observations are often correlated:

- Same individual measured multiple times (repeated measures)
- Measurements from same family, litter, or colony
- Multiple cells from same organism
- Multiple samples from same location
- Before-after measurements (paired data)

Why this matters:

- Standard errors are underestimated (observations aren't truly independent)
- P-values are too small (false positives!)
- Confidence intervals are too narrow

Must account for correlation structure in the data

Revised Fly Study: Repeated Measures

New design: Measure the same flies under both diets

Study Protocol

- Start with 20 flies (10 male, 10 female)
- Measure weight on control diet (Period 1)
- *Somehow reset the flies...* (hypothetically!)
- Measure weight on restricted diet (Period 2)
- Each fly contributes two observations

Data structure:

FlyID	Sex	Diet	weight
1	F	Control	62
1	F	Restricted	78
2	F	Control	65
2	F	Restricted	80
...	...	...	...

Problem: Two observations from Fly 1 are more similar to each other than to observations from Fly 2!

Naive (Wrong) Analysis

Temptation: Just use the model we had before

$$\text{weight} = \beta_0 + \beta_1 \text{Diet} + \beta_2 \text{Sex} + \epsilon$$

Why This is Wrong

- Treats 40 observations as independent
- Actually only 20 independent units (flies)
- Inflates degrees of freedom
- Underestimates standard errors
- P-values too optimistic!

Consequence: More likely to claim ”significant” effect when there isn’t one (Type I error inflation)

We must account for the fact that observations are nested within flies

Solution 1: Paired t-test

Classic approach for paired data:

1. Calculate difference for each fly: $D_i = Y_{i,\text{Restricted}} - Y_{i,\text{Control}}$
2. Test if mean difference is zero: $H_0 : \mu_D = 0$
3. Use one-sample t-test on differences

Example

Fly	Control	Restricted	Difference
1	62	78	+16
2	65	80	+15
3	60	75	+15
...	...	...	...

Test: $\bar{D} = 15$ mg, $t = 6.5$, $p < 0.001$

Advantages: Simple, widely understood

Limitations: Can't include other predictors (like Sex), can't handle unbalanced data

Solution 2: Fixed Effects Model

Include fly as a predictor (dummy variables):

Model

$$\text{weight} = \beta_0 + \beta_1 \text{Diet} + \beta_2 \text{Sex} + \alpha_1 I_{\text{Fly2}} + \alpha_2 I_{\text{Fly3}} + \cdots + \alpha_{19} I_{\text{Fly20}} + \epsilon$$

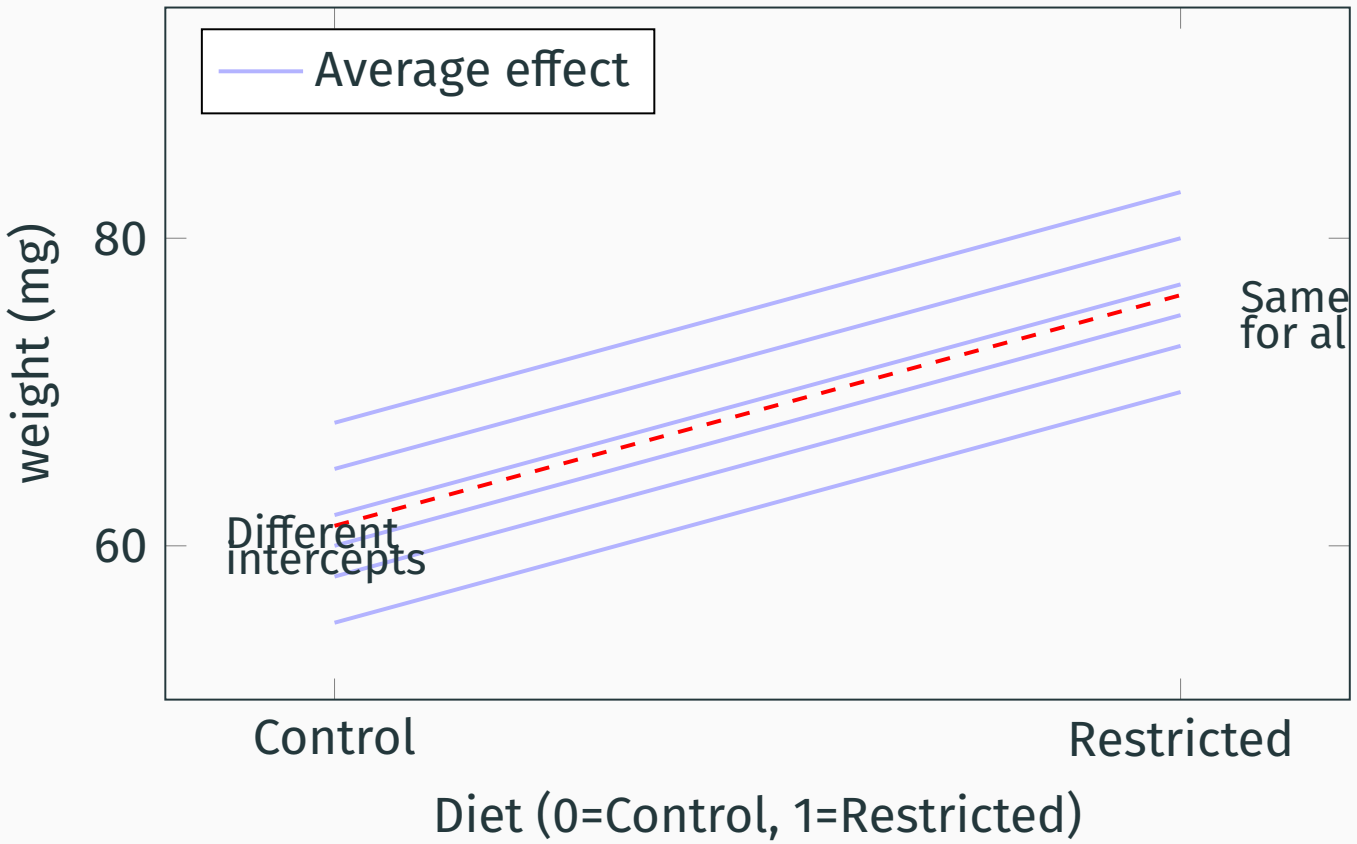
Interpretation:

- β_0 = baseline weight for Fly 1 (reference), female, control
- β_1 = effect of diet (comparing within same fly!)
- β_2 = difference between males and females
- α_i = deviation of Fly *i* from Fly 1 (fly-specific intercepts)

This accounts for correlation: Each fly has its own baseline

Limitation: Uses many degrees of freedom (19 fly parameters!), can't include fly-level predictors

Visualizing Fixed Effects: Multiple Intercepts



Each fly gets its own intercept (baseline) but shares the same diet effect

Solution 3: Random Effects (Mixed Models)

Problem with fixed effects: Treating flies as fixed uses too many parameters

Random Effects Approach

Instead of estimating each fly's intercept separately, model them as drawn from a distribution:

$$\alpha_i \sim N(0, \sigma^2_{\text{fly}})$$

Now we only estimate one parameter (σ^2_{fly}) instead of 19!

Mixed Model

$$\text{weight}_{ij} = \beta_0 + \beta_1 \text{Diet}_{ij} + \beta_2 \text{Sex}_i + \alpha_i + \epsilon_{ij}$$

where:

- $\alpha_i \sim N(0, \sigma^2_{\text{fly}})$ = random intercept for fly i
- $\epsilon_{ij} \sim N(0, \sigma^2)$ = residual error

Why Use Random Effects?

Advantages over fixed effects:

1. **More efficient:** Uses fewer parameters
2. **Generalizable:** Inferences apply to population of flies, not just these 20
3. **Partial pooling:** Borrows information across flies (shrinkage toward mean)
4. **Handles unbalanced data:** Works even if some flies measured only once
5. **Can include group-level predictors:** Can have fly-level AND observation-level predictors

When to Use Random Effects

Use random effects when:

- You have many groups (flies, individuals, sites)
- Groups are a random sample from a population
- You want to generalize beyond observed groups

Partial Pooling: Borrowing Information

Random effects provide "partial pooling" (shrinkage):

Three Approaches

1. **Complete pooling:** Ignore fly identity (naive model)
   Assumes all flies identical

2. **No pooling:** Fixed effects for each fly
   Treats each fly as completely unique

3. **Partial pooling:** Random effects
   Each fly has its own intercept, but shrunk toward overall mean

Result:

- Flies with more data → estimates closer to their own mean
- Flies with less data → estimates shrunk more toward overall mean
- Best of both worlds: Accounts for individuality while borrowing strength

Variance Components

Mixed model partitions total variance:

Total Variance

$$\text{Var}(Y) = \sigma^2_{\text{fly}} + \sigma^2$$

where:

- σ^2_{fly} = between-fly variance (how much flies differ)
- σ^2 = within-fly variance (residual, measurement error)

Intraclass Correlation (ICC)

$$\text{ICC} = \frac{\sigma^2_{\text{fly}}}{\sigma^2_{\text{fly}} + \sigma^2}$$

Proportion of variance due to fly identity

Example

If $\sigma^2_{\text{fly}} = 25$ and $\sigma^2 = 15$:

ICC = 25 / (25 + 15) = 0.625 (62.5% of variance is between flies)

Comparing the Three Approaches

Aspect	Complete Pooling	Fixed Effects	Random Effects
Model	Simple LM	LM with fly dummies	Mixed model
Parameters	Few	Many (n-1)	Few
SE of β_1	Too small	Correct	Correct
Generalizability	To these flies	To these flies	To population
Unbalanced data	OK	OK	OK
Computational	Easy	Easy	Moderate
Interpretation	Simple	Moderate	Moderate

Recommendation

For repeated measures with many groups: Use random effects (mixed models)

Adding Between-Subject Factors

Sex is a fly-level variable (doesn't change within fly):

Full Model

$$\text{weight}_{ij} = \beta_0 + \beta_1 \text{Diet}_{ij} + \beta_2 \text{Sex}_i + \beta_3 (\text{Diet} \times \text{Sex})_{ij} + \alpha_i + \epsilon_{ij}$$

where:

- β_1 = within-fly effect of diet (repeated measure)
- β_2 = between-fly effect of sex
- β_3 = interaction (does diet effect differ by sex?)
- α_i = random intercept for fly i

Interpretation:

- Diet is **within-subject** factor (varies within fly)
- Sex is **between-subject** factor (constant within fly)
- Can test both types of effects in one model!

Random Slopes: Letting Effects Vary

So far: Random intercepts only (flies differ in baseline)

Extension: Random slopes (flies differ in diet response)

Random Intercepts + Random Slopes

$$\text{weight}_{ij} = \beta_0 + \beta_1 \text{Diet}_{ij} + \alpha_{0i} + \alpha_{1i} \text{Diet}_{ij} + \epsilon_{ij}$$

where:

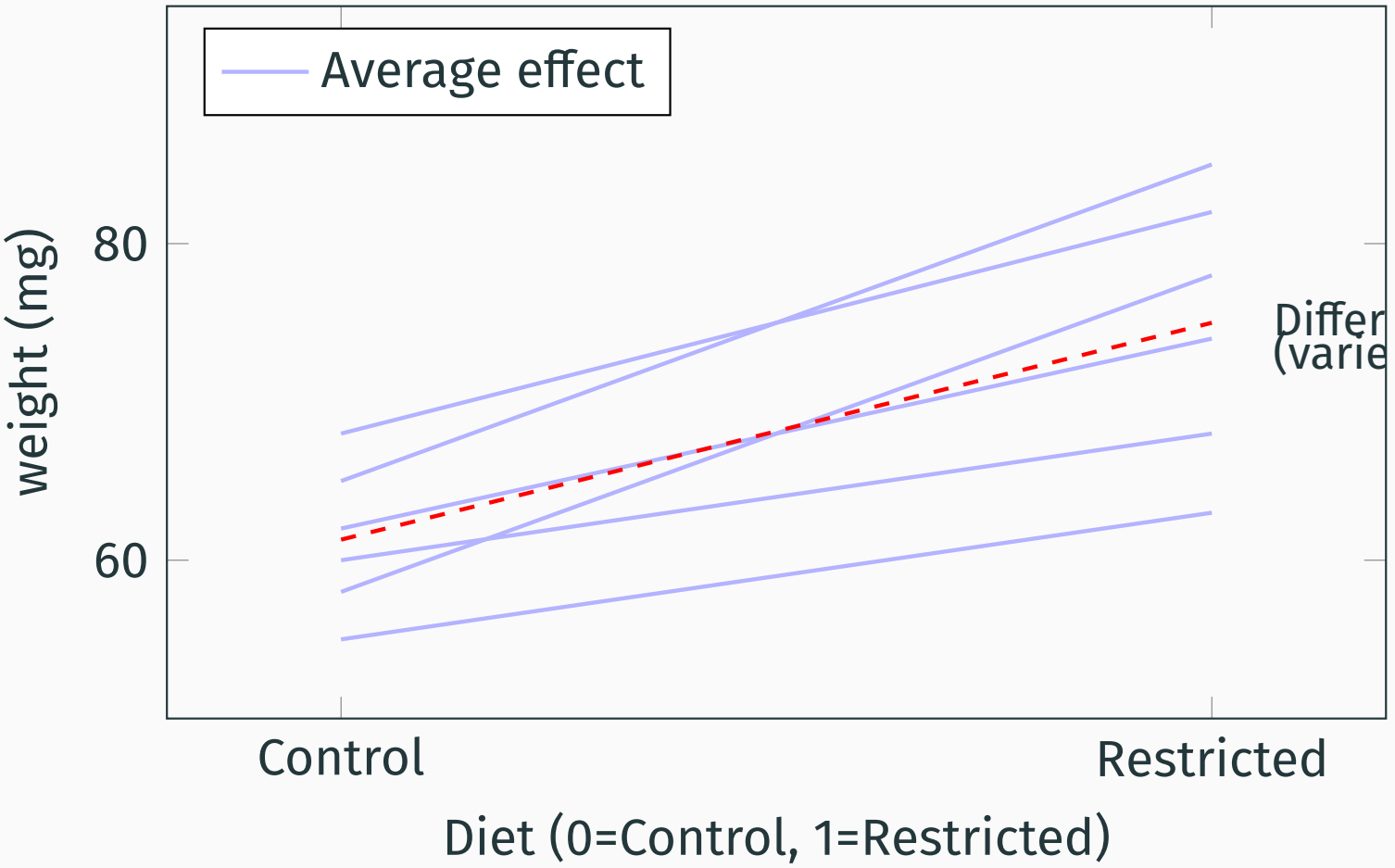
- $\alpha_{0i} \sim N(0, \sigma^2_{\text{int}})$ = random intercept (baseline varies)
- $\alpha_{1i} \sim N(0, \sigma^2_{\text{slope}})$ = random slope (diet effect varies)

When to use:

- Theoretical reason to expect individual variation in responses
- Enough data per individual (need $\geq$3 measurements)
- Want to model heterogeneity in treatment effects

Warning: More parameters → need more data; can lead to convergence issues

Visualizing Random Slopes



Each fly has its own intercept AND slope—captures individual variation in treatment response

Fitting Mixed Models in R

Package: *lme4* (most common)

Random Intercept Only

```
library(lme4)

model <- lmer(weight ~ Diet + Sex + (1 | FlyID),
              data = flies)
```

(1 | FlyID) = random intercept for each fly

Random Intercept + Random Slope

```
model <- lmer(weight ~ Diet + Sex +
              (1 + Diet | FlyID),
              data = flies)
```

(1 + Diet | FlyID) = random intercept and random slope for diet effect

summary(model) shows fixed effects and variance components

Assumptions for Mixed Models

Same as regular linear models, plus:

1. **Random effects are normally distributed:** $\alpha_i \sim N(0, \sigma^2_\alpha)$
2. **Random effects independent of residuals:** $\text{Cov}(\alpha_i, \epsilon_{ij}) = 0$
3. **Groups are a random sample:** Flies randomly sampled from population

Diagnostics:

- Check residual normality (Q-Q plot)
- Check homoscedasticity (residuals vs. fitted)
- Check random effects normality (Q-Q plot of BLUPs)
- Check for outlier groups

Mixed models are robust to moderate violations of normality, especially with large number of groups

Decision Guide: Which Approach?

Use Paired t-test when:

- Simple paired design (2 measurements per subject)
- No additional predictors
- Balanced data
- Want simplest analysis

Use Fixed Effects when:

- Few groups (<10)
- Only care about these specific groups
- All groups well-measured

Use Random Effects (Mixed Models) when:

- Many groups (≥ 10)
- Groups are random sample
- Want to generalize to population
- Unbalanced data
- Need to include both within- and between-subject factors

Extensions: More Complex Designs

Mixed models handle many complex designs:

1. **Nested designs:** Cells within organisms within populations

$$Y_{ijk} = \beta_0 + \alpha_i + \alpha_{ij} + \epsilon_{ijk}$$

2. **Crossed designs:** Multiple random factors (e.g., flies and time points)

$$Y_{ij} = \beta_0 + \alpha_i + \gamma_j + \epsilon_{ij}$$

3. **Longitudinal data:** Multiple time points with temporal correlation
4. **Spatial data:** Locations with spatial correlation
5. **Split-plot designs:** Different factors at different scales

Mixed models are extremely flexible—a whole course unto themselves!

Common Mistakes with Repeated Measures

1. **Ignoring non-independence**
   Solution: Always account for clustering/pairing

2. **Using complete pooling when you have repeated measures**
   Solution: At minimum use paired t-test; better yet, use mixed model

3. **Using fixed effects with many groups**
   Solution: Switch to random effects

4. **Treating between-subject variables as within-subject**
   Solution: Understand which variables vary within vs. between subjects

5. **Including random slopes when data are sparse**
   Solution: Need $\geq$3 measurements per subject for random slopes

6. **Not checking convergence warnings**
   Solution: Simplify model if convergence issues arise

Summary: Repeated Measures

Key Concepts

- **Non-independence:** Observations from same subject are correlated
- **Must account for it:** Otherwise SE too small, p-values too optimistic
- **Solutions:** Paired t-test, fixed effects, random effects
- **Random effects:** Most flexible and efficient approach

Variance Decomposition

- Between-subject variance (σ^2_α)
- Within-subject variance (σ^2)
- ICC quantifies proportion of variance between subjects

Ignoring non-independence is one of the most common mistakes in biological statistics!