Analysis of Environmental Data

Deck 7 Regression Modeling

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst
Michael France Nelson

Regression Concepts 1

What's in This Section?

Important take-home concepts

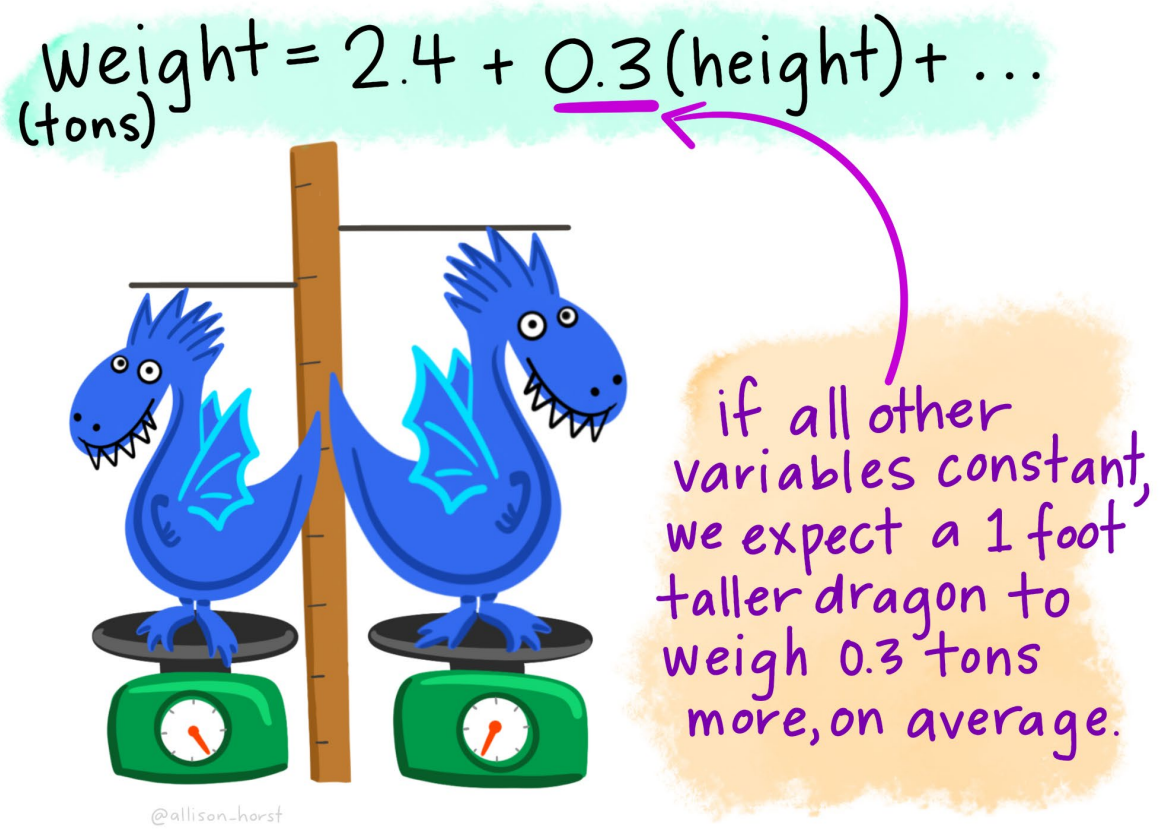
- What is a regression model?
- What is the constellation of methods?
- What are the 4 key assumptions?
  - Normality of the residuals
  - Homogeneity
  - Fixed x
  - Independent observations
- Residuals

What is a Regression?

Regressions embody the dual-model concept

Regression is a modeling paradigm in which we specify a mathematical relationship between independent and dependent variables.

- A regression includes a *deterministic model* to specify the average behavior.
- It specifies a *stochastic model* to describe the variability around the average behavior.

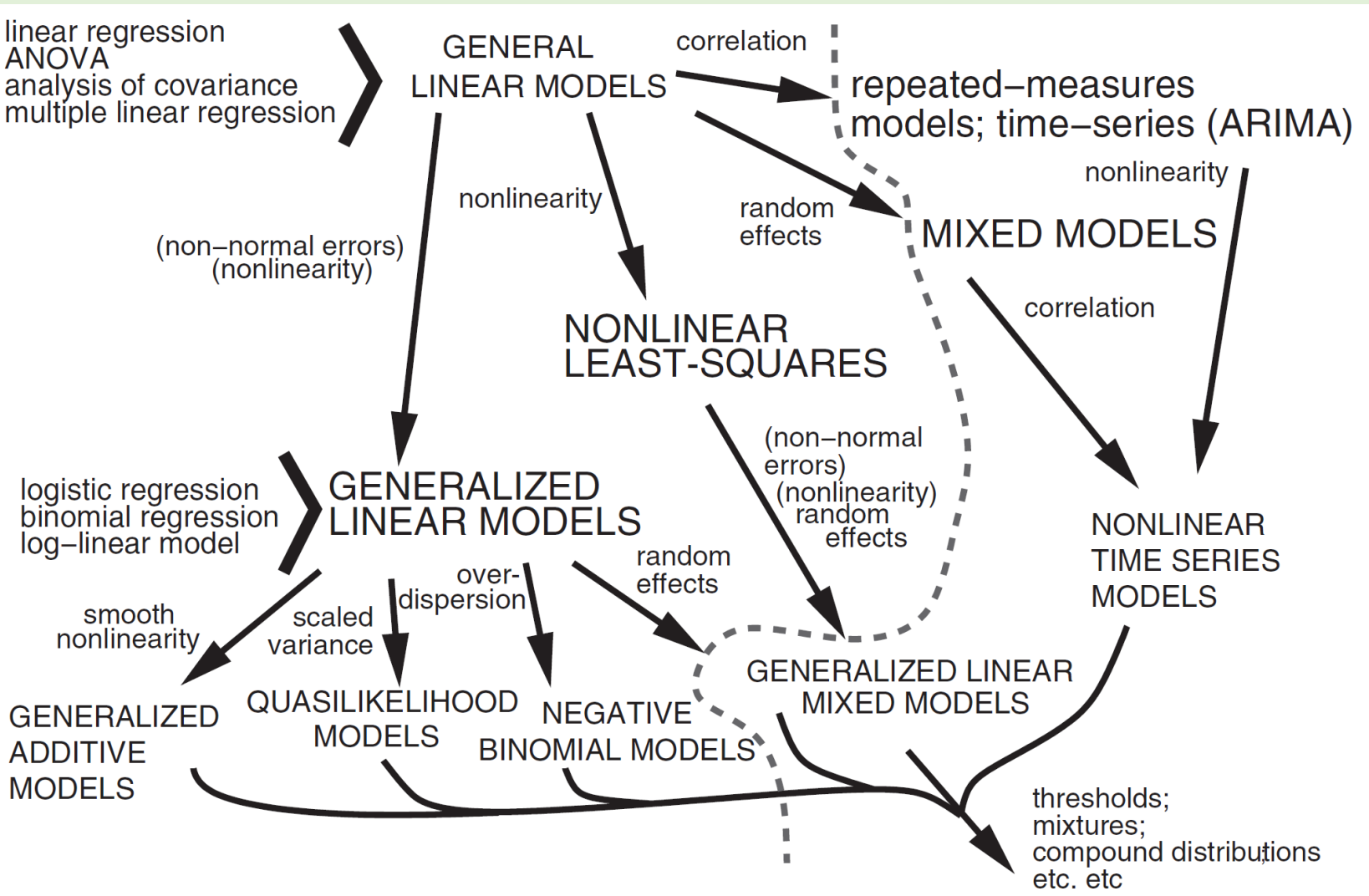


Artwork by @allison_horst

Regression Acronyms: The constellation

Bolker: Ecological Models and Data in R, Figure 9.2

- There are many types of regression models including:
  - General Linear Models
  - Generalized Linear Models
  - Mixed Models
- Think of the collection as a constellation of methods
- There are a lot of similar names and acronyms



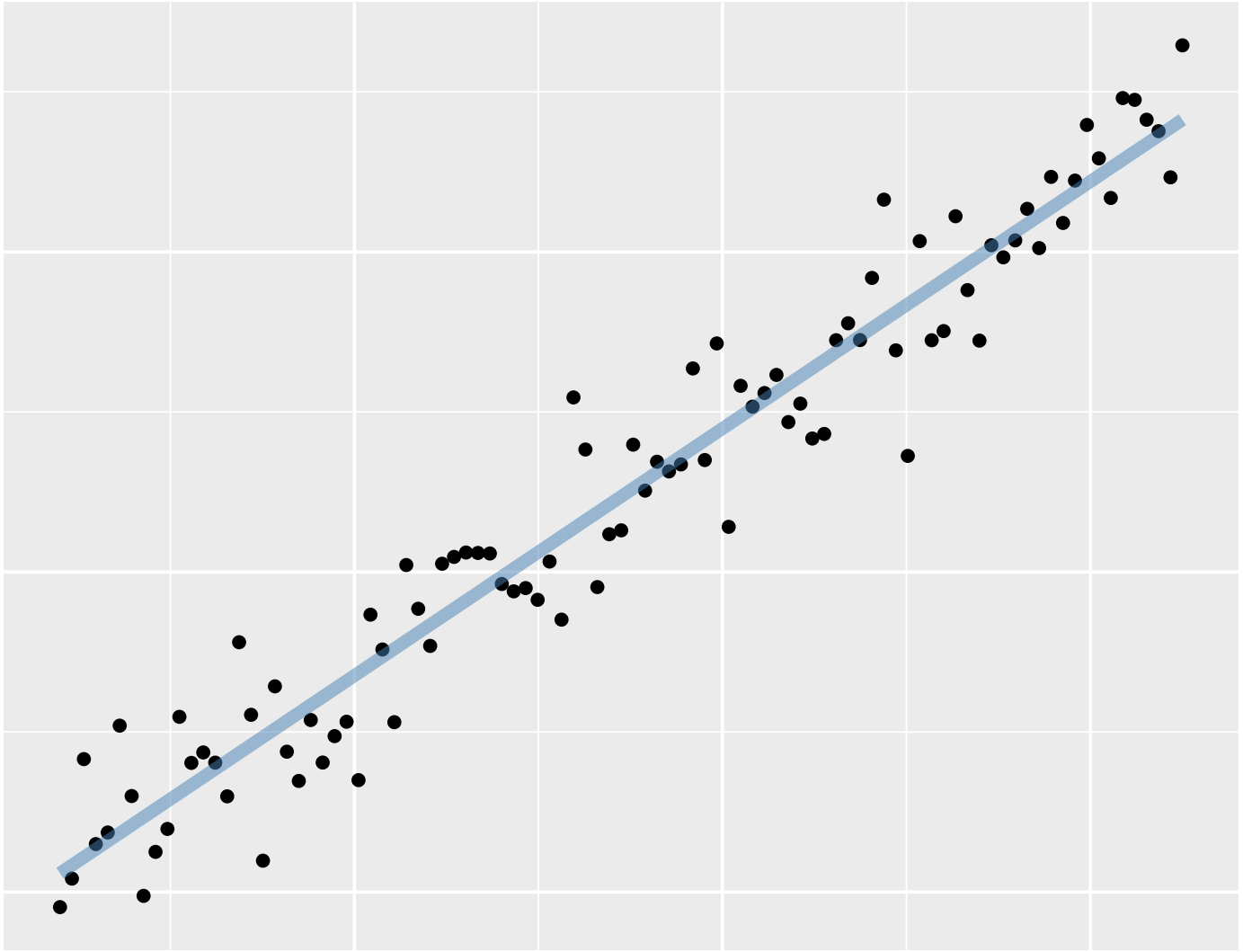
Group 1 Models

The simplest model we can fit is always a linear model!

General Linear Models form the core group of regression models.

- Other regression model paradigms are extensions of General Linear Models.

We'll spend a lot of time on this class of models, which I'll call *Group 1* models.



Group 1 Models – 4 Key Assumptions

Our Group 1 models carry some baggage… Specifically four key assumptions:

- Independent observations
- Constant variance a.k.a homoskedasticity, a.k.a. homogeneity
- Fixed x: no measurement error in our predictor variables
- Normality: normality refers to the model residuals

In addition, Group I requires that our models be *linear in the parameters* and have a response on a **continuous scale**.

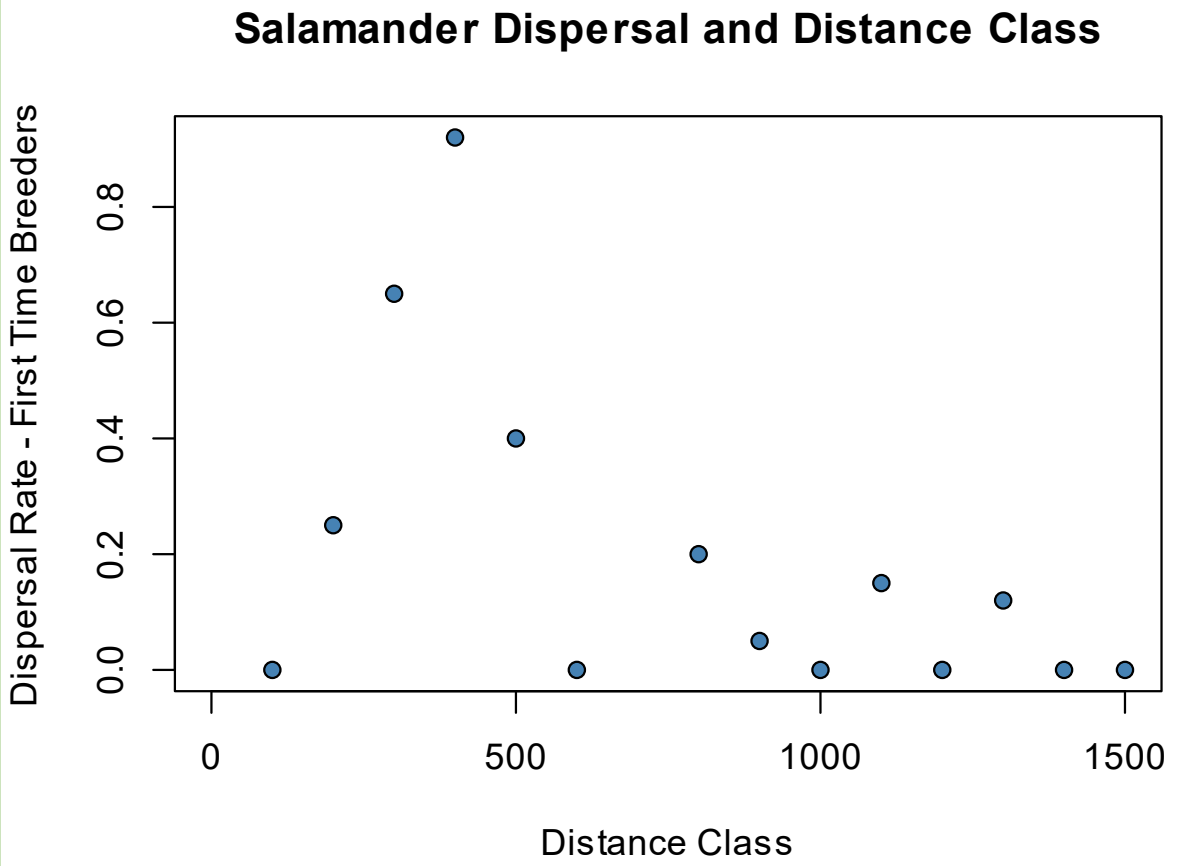
The extended models can deal with different violations of these assumptions and requirements.

[Residual] Normality Assumption

- Under repeated sampling, data would be normally distributed *at each x*.
- Normally distributed around each *predicted value* in the *deterministic model*.
- This assumption is often misunderstood to mean that the values for each variable in a data set must be normally-distributed by themselves.
- **But what is a residual?**
  - The differenced between a predicted and observed value

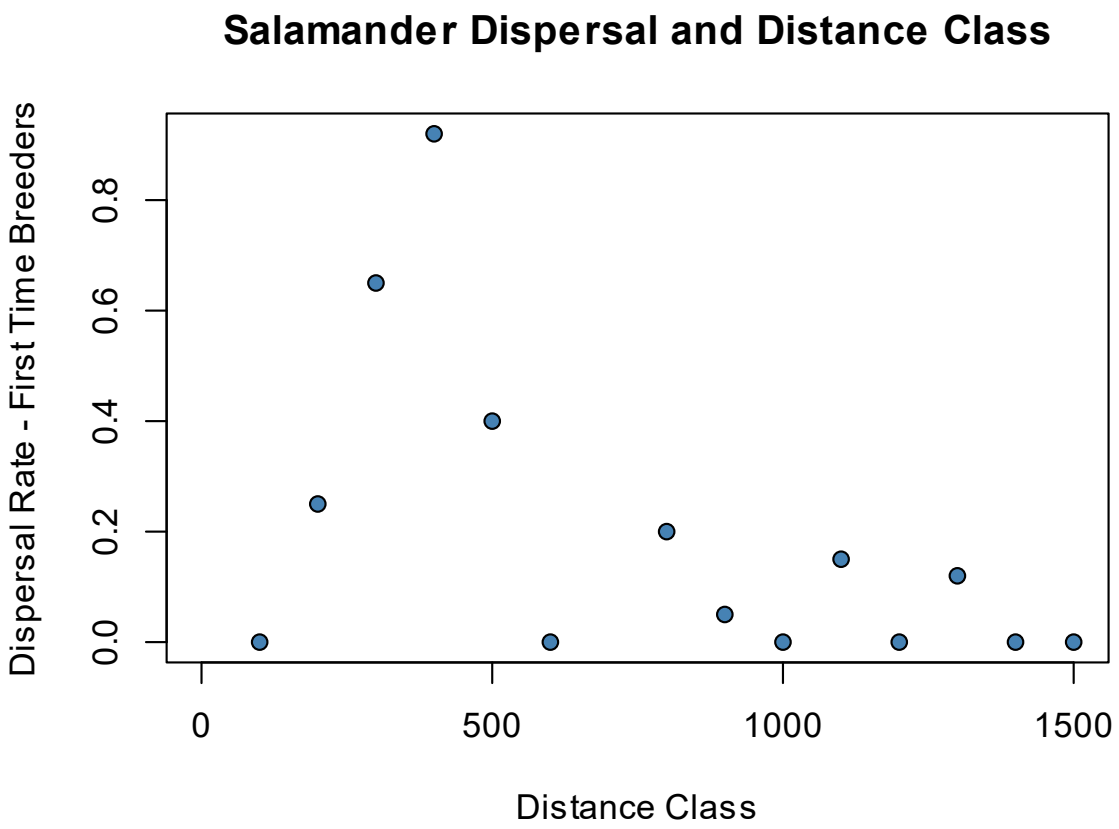
Model Residuals

- Salamander breeder dispersal data
- What kind of model should we fit?
- A Ricker curve might be a good choice.

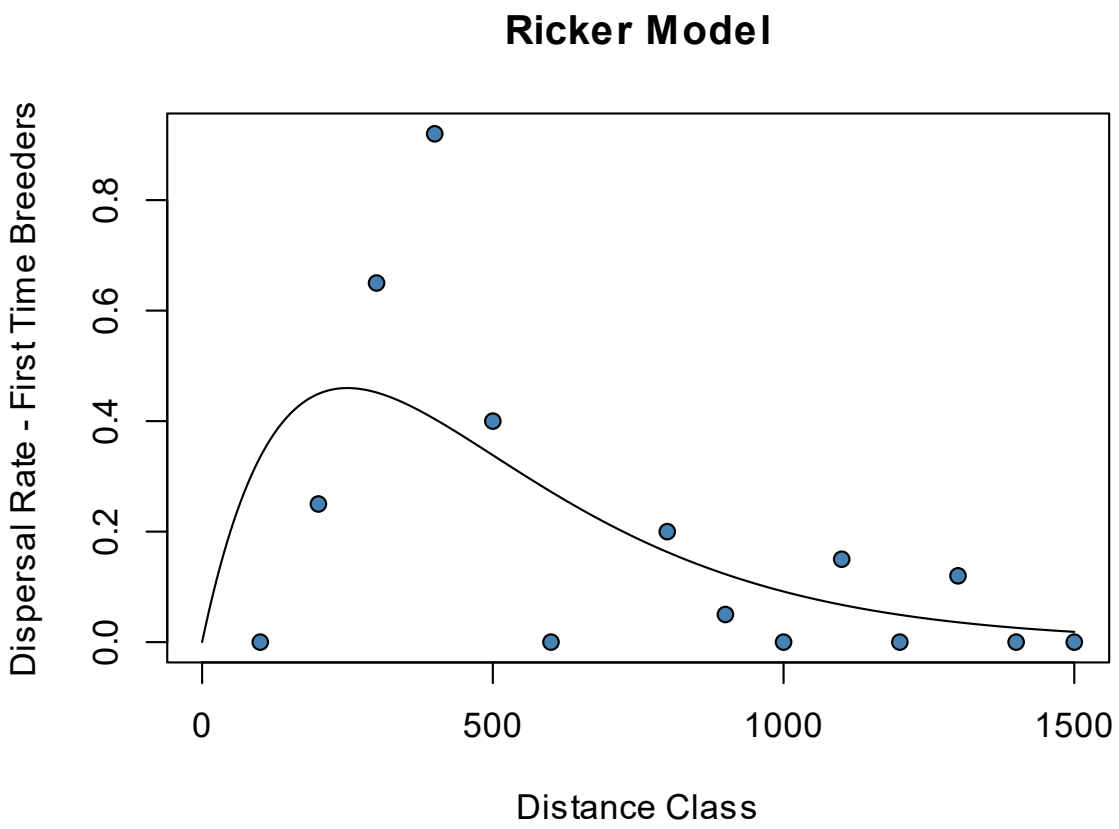


Model Residuals

The Data

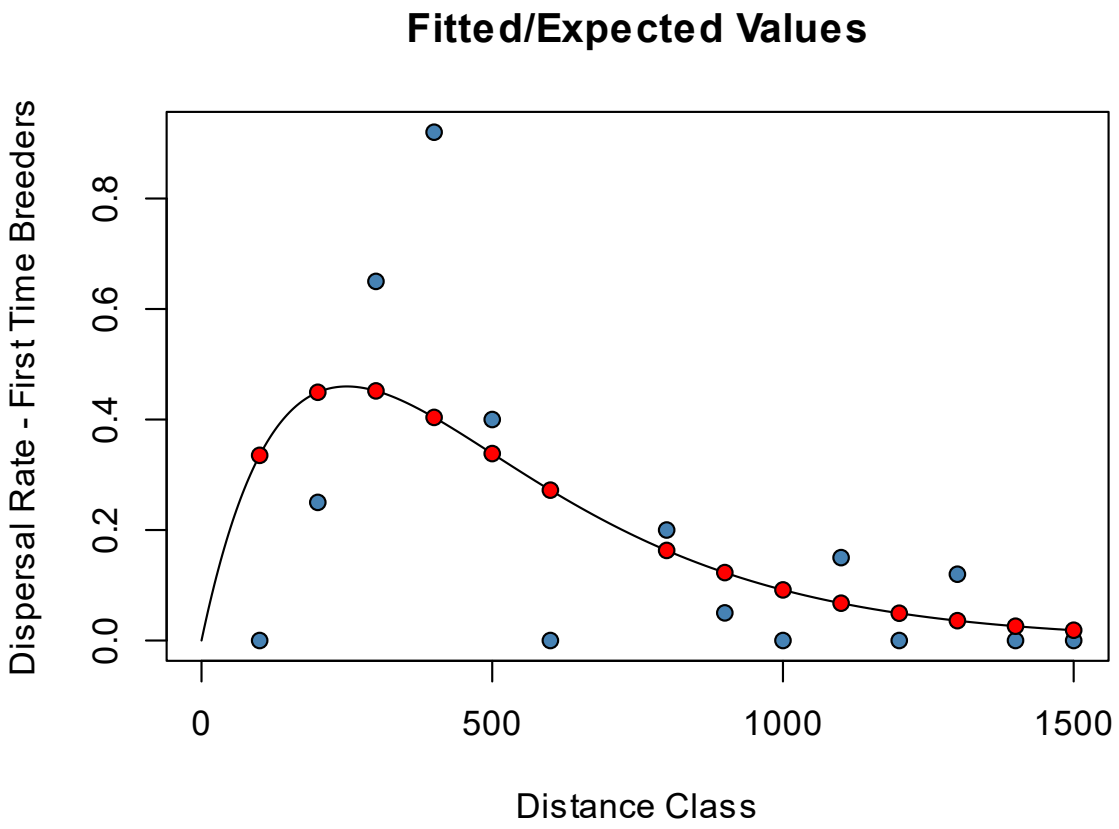


A Fitted Ricker Curve

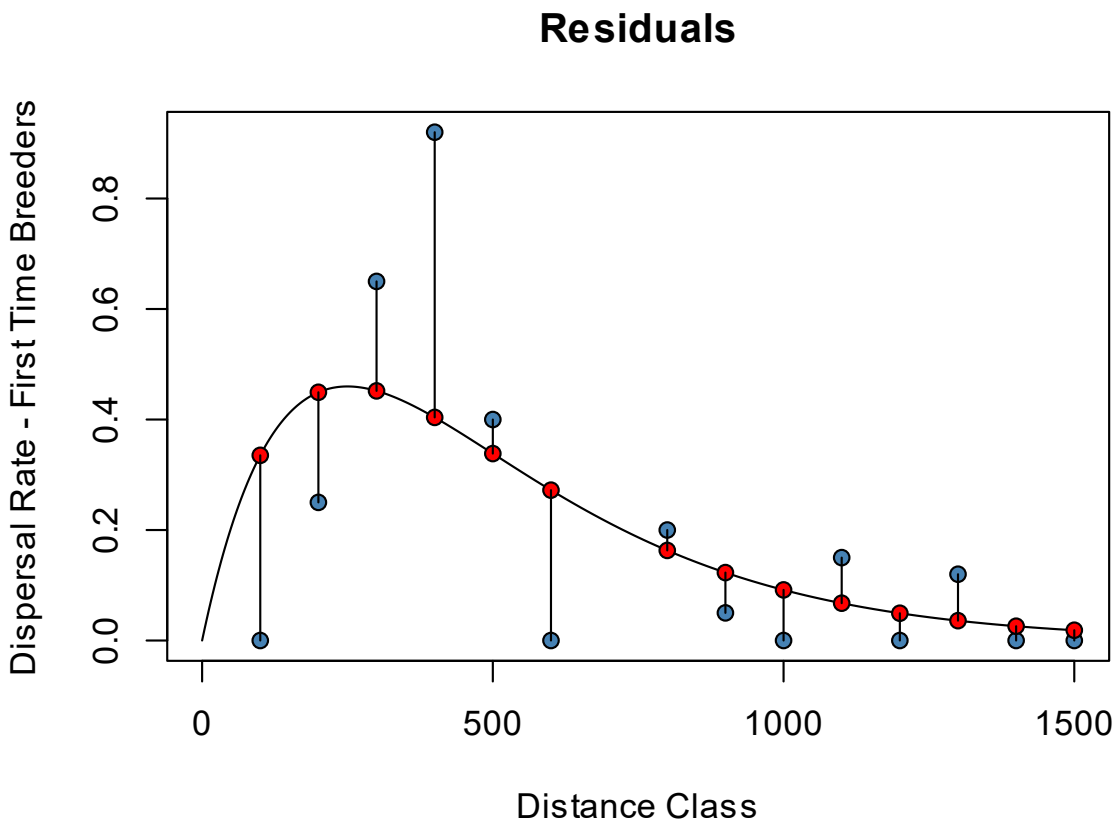


Model Residuals

The Predicted Values

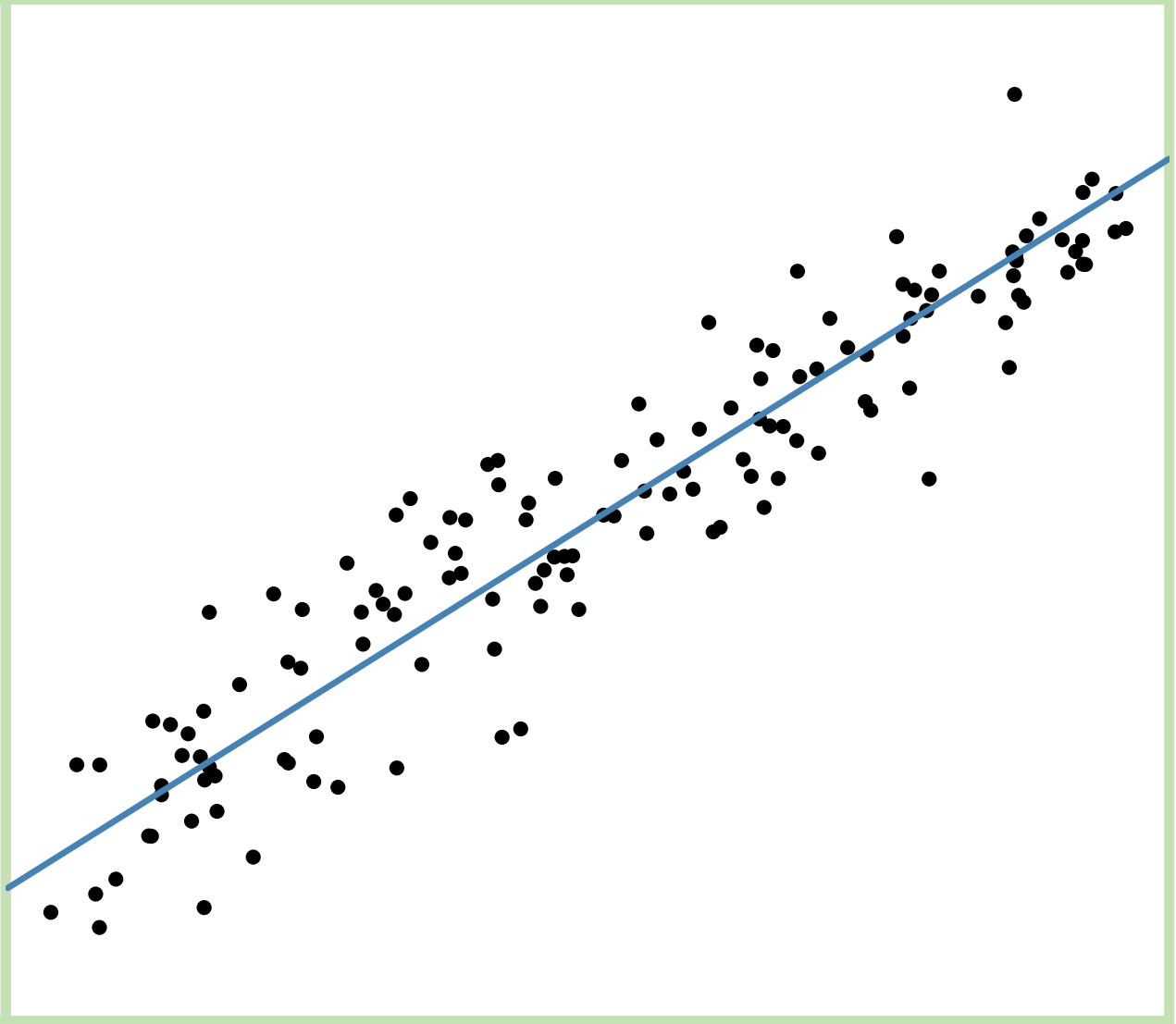


And…The Residuals!



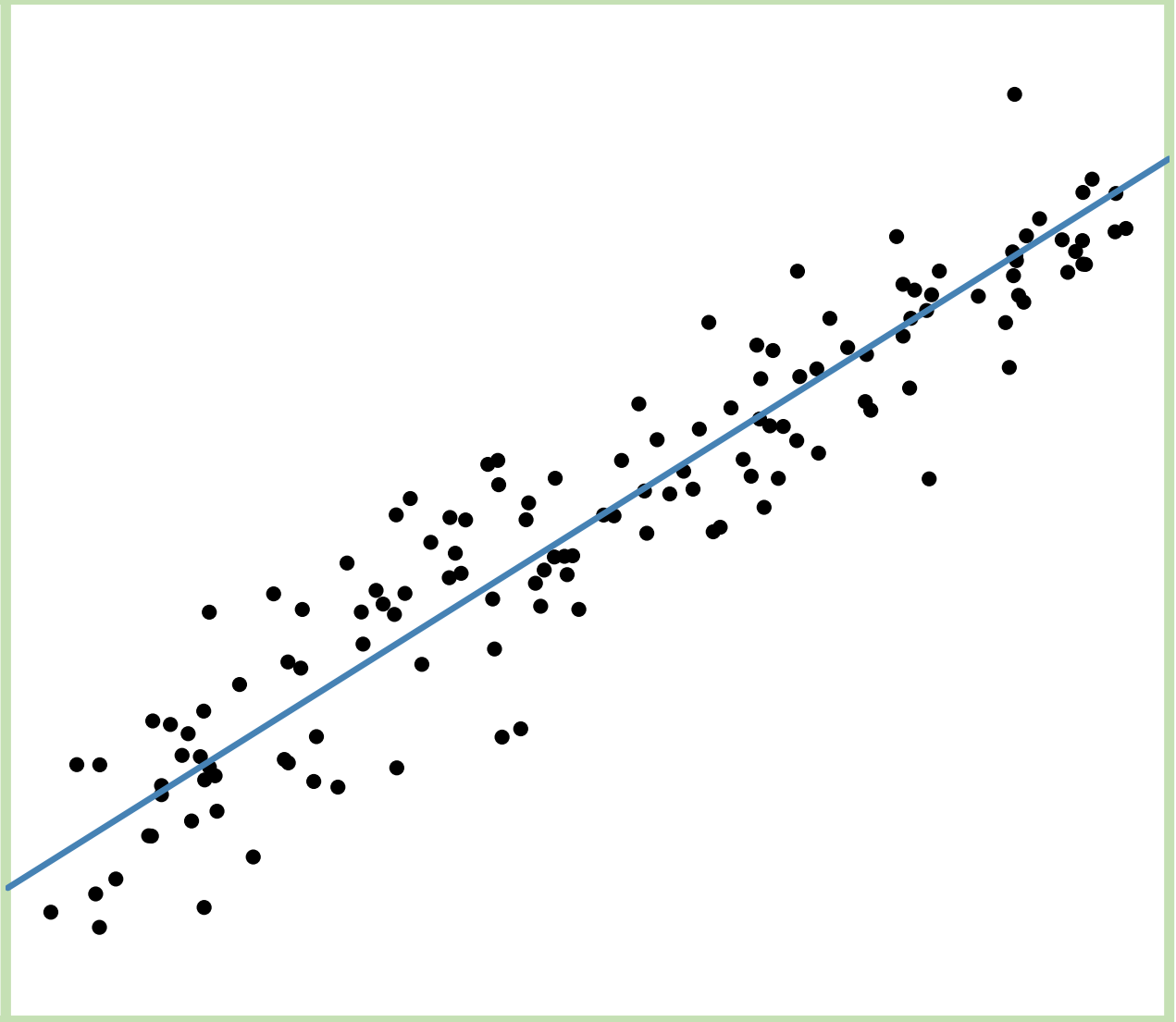
[Residual] Normality Assumption

- Group 1 models assume that residuals are Normally distributed



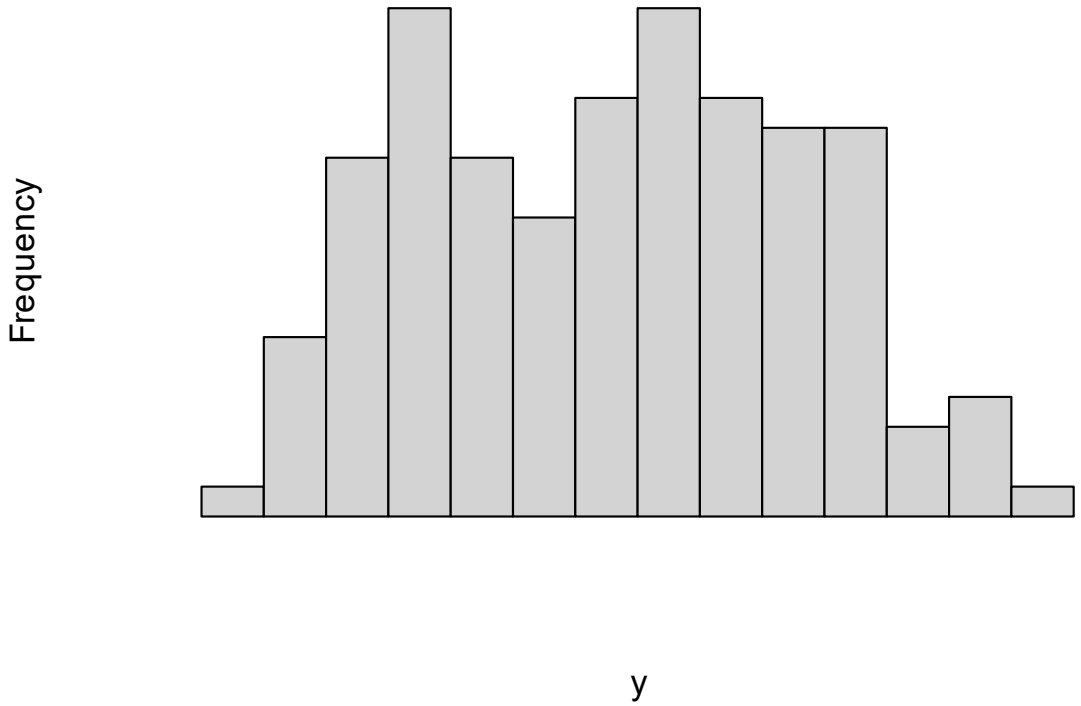
[Residual] Normality Assumption

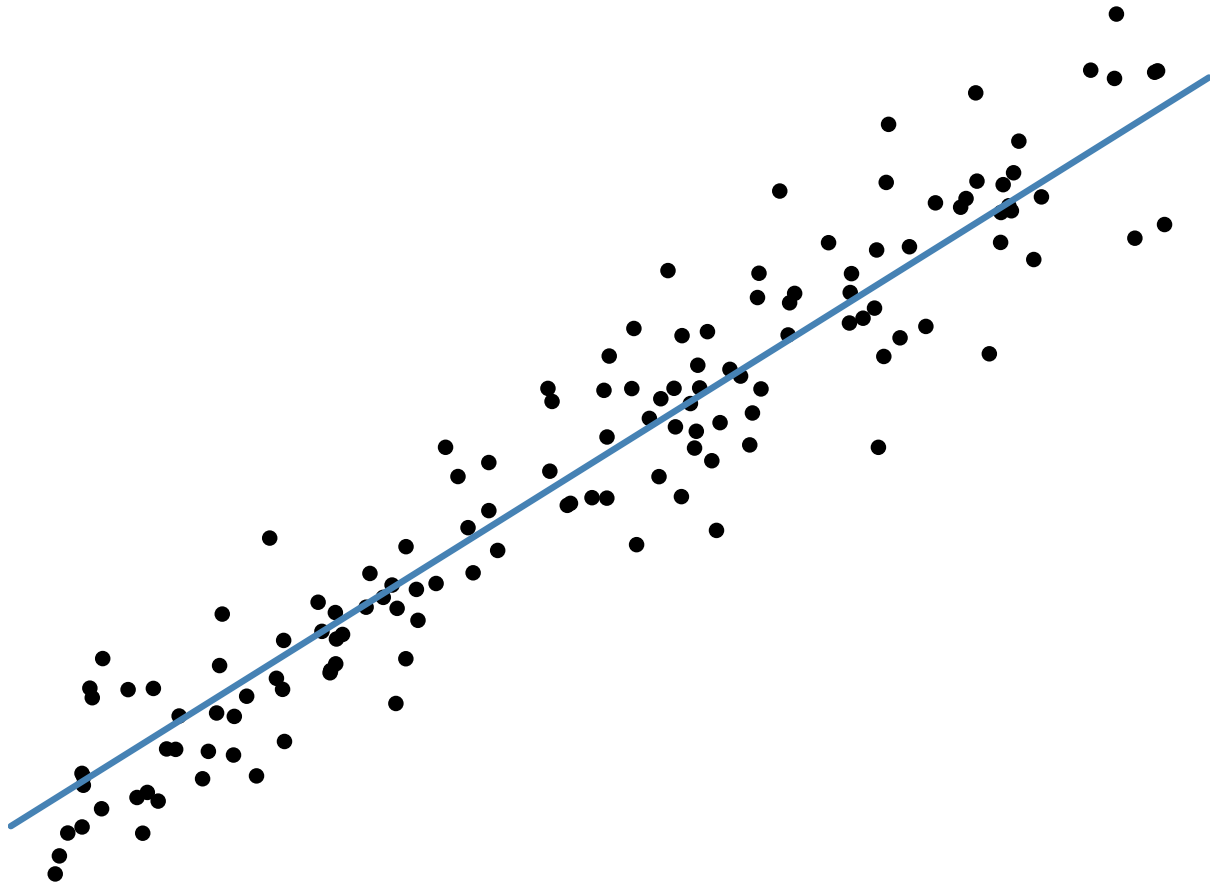
- This does not mean that ‘the data are normally distributed’.
  - Usually, the data points themselves aren’t Normally distributed.
  - This is a frequent point of confusion.



[Residual] Normality Assumption

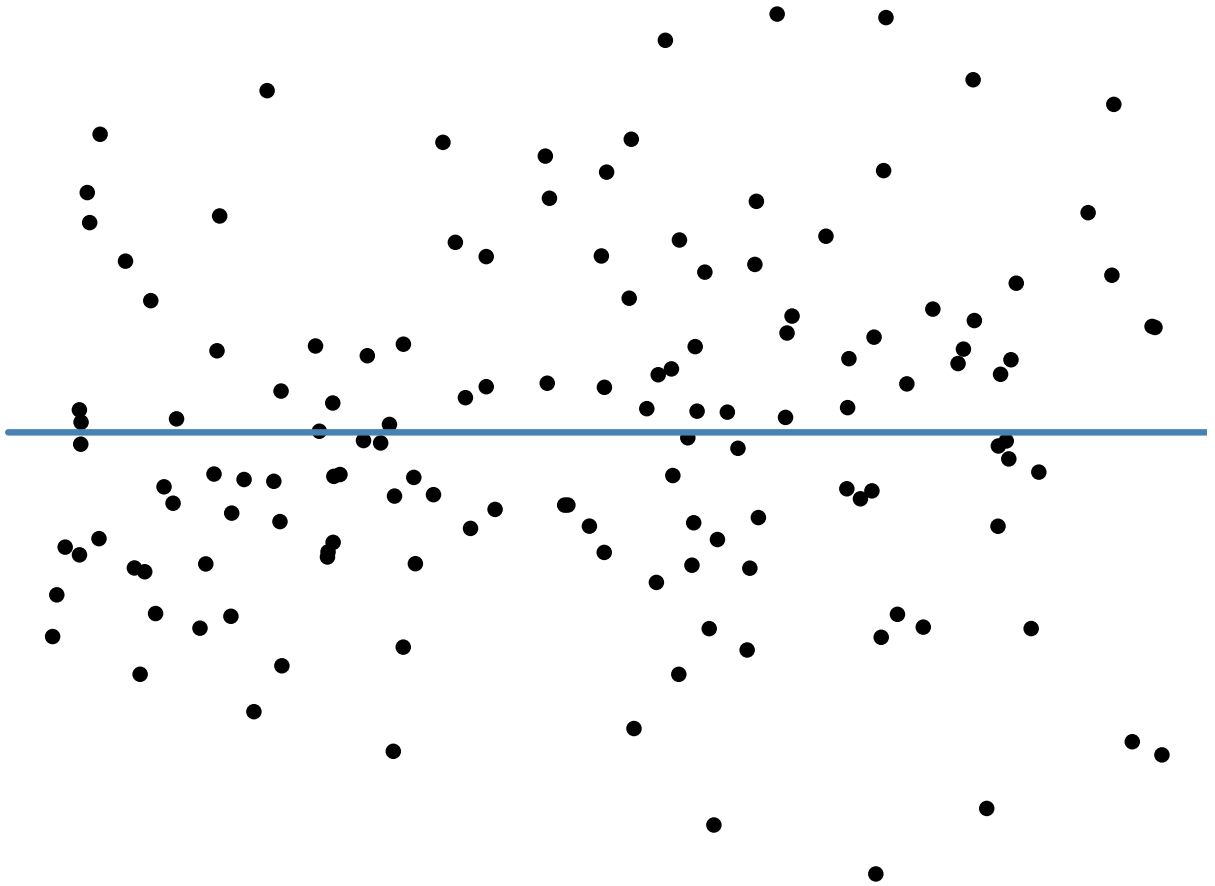
The following data look relatively well-behaved, but the histogram of the y-values suggests non-normality. A Shapiro test provides evidence of non-normality with p = 0.007.

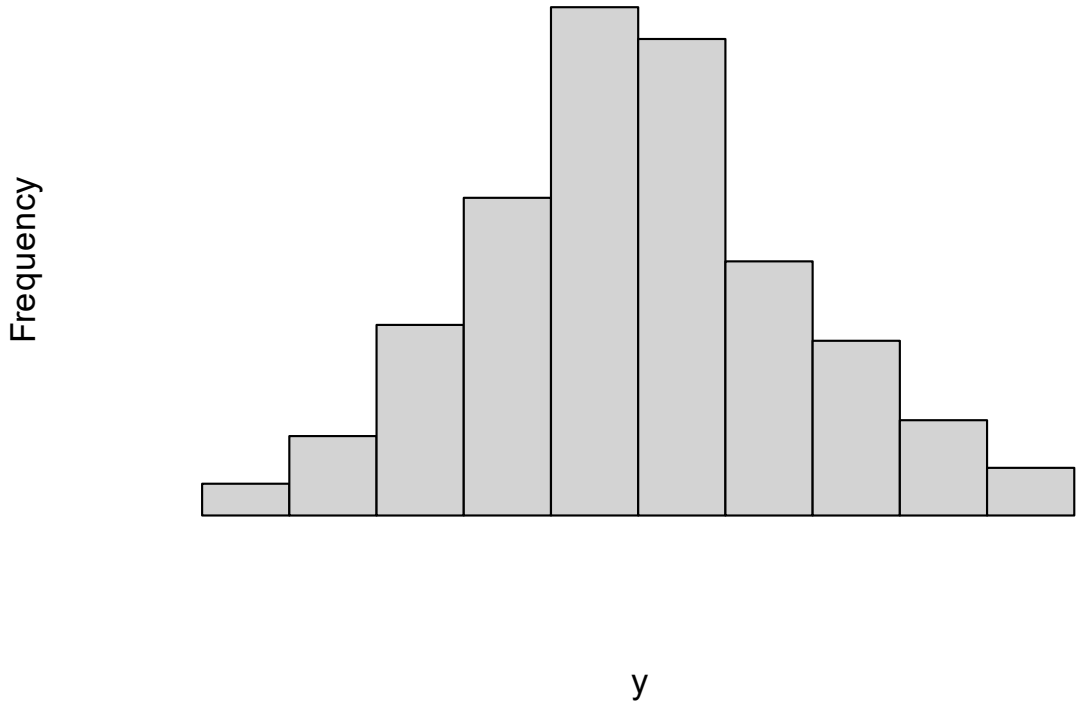




[Residual] Normality Assumption

We really care about the normality of the *residuals* from a model.
A Shapiro test on the residuals suggests normality with p = 0.833.



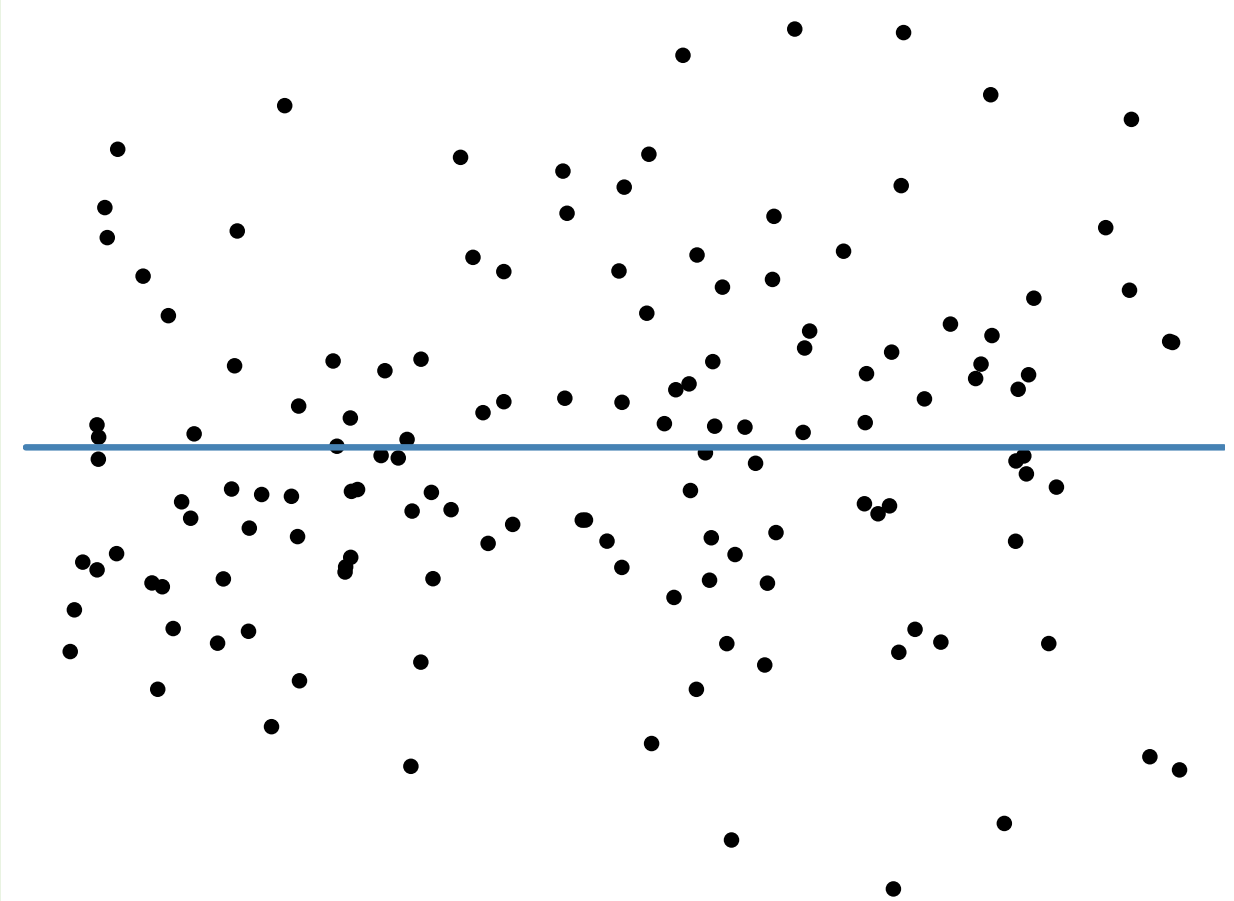


Homogeneity Assumption

The homogeneity assumption requires constant variance along the entire range of predictor values.

Key points of the assumption:

- The stochastic model is a Normal distribution.
- The spread parameter, σ is constant.
- In other words, the variability does not depend on the value of x

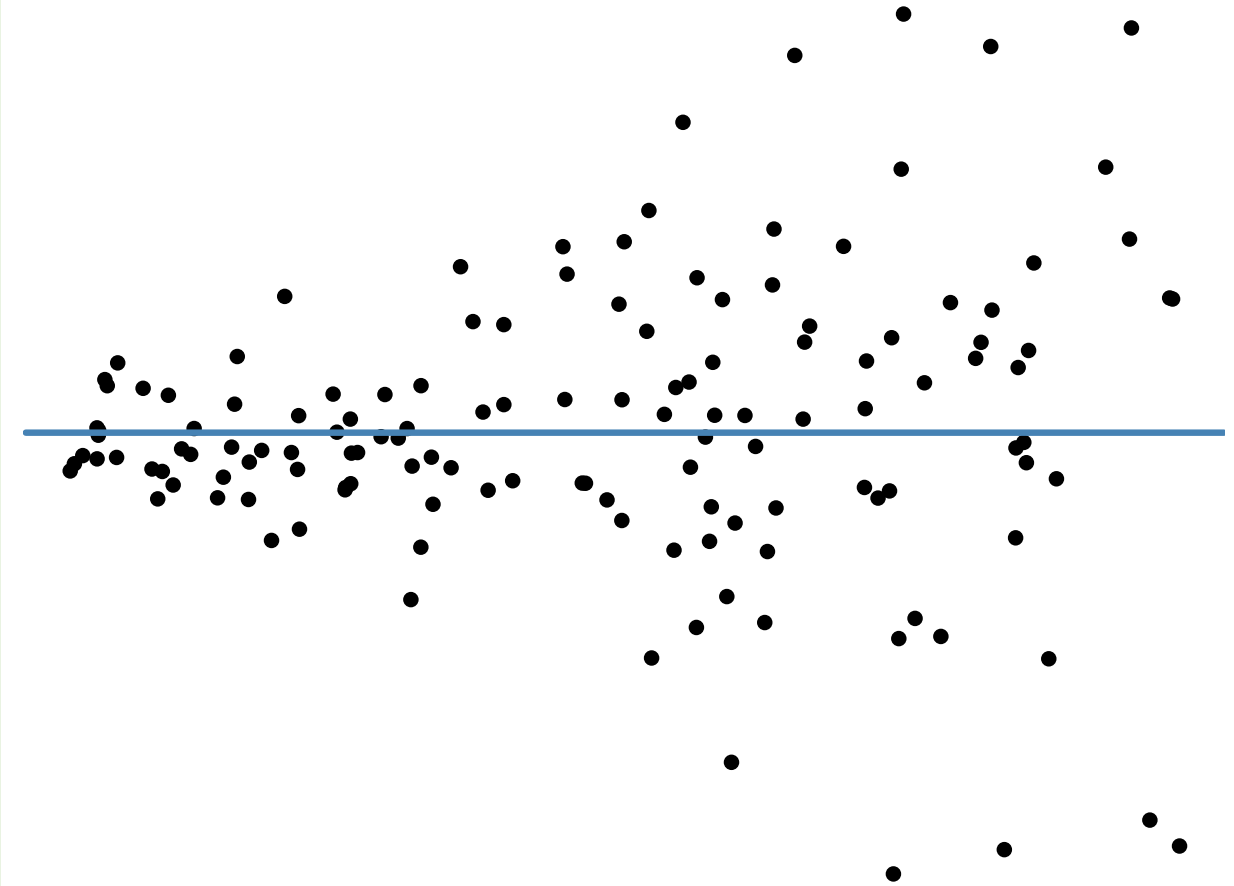


Heterogeneous REsiduals

The homogeneity assumption requires constant variance along the entire range of predictor values.

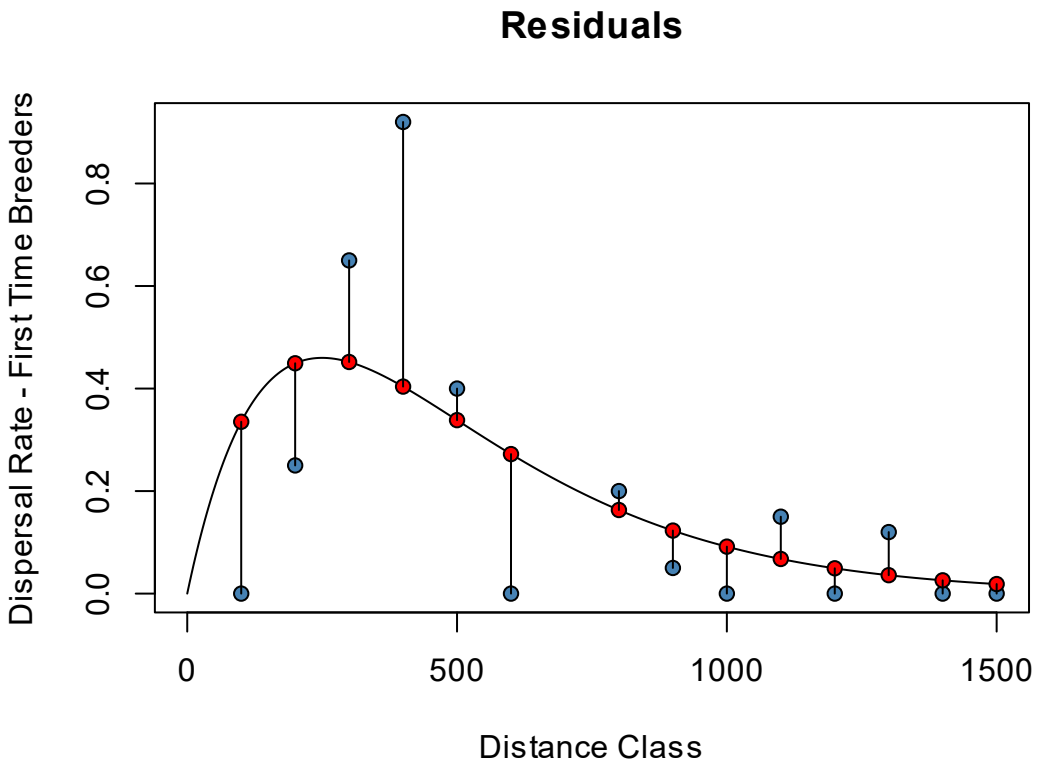
Heterogeneous residuals

- The spread parameter, σ is non-constant.
- In other words, the variability depends on the value of x

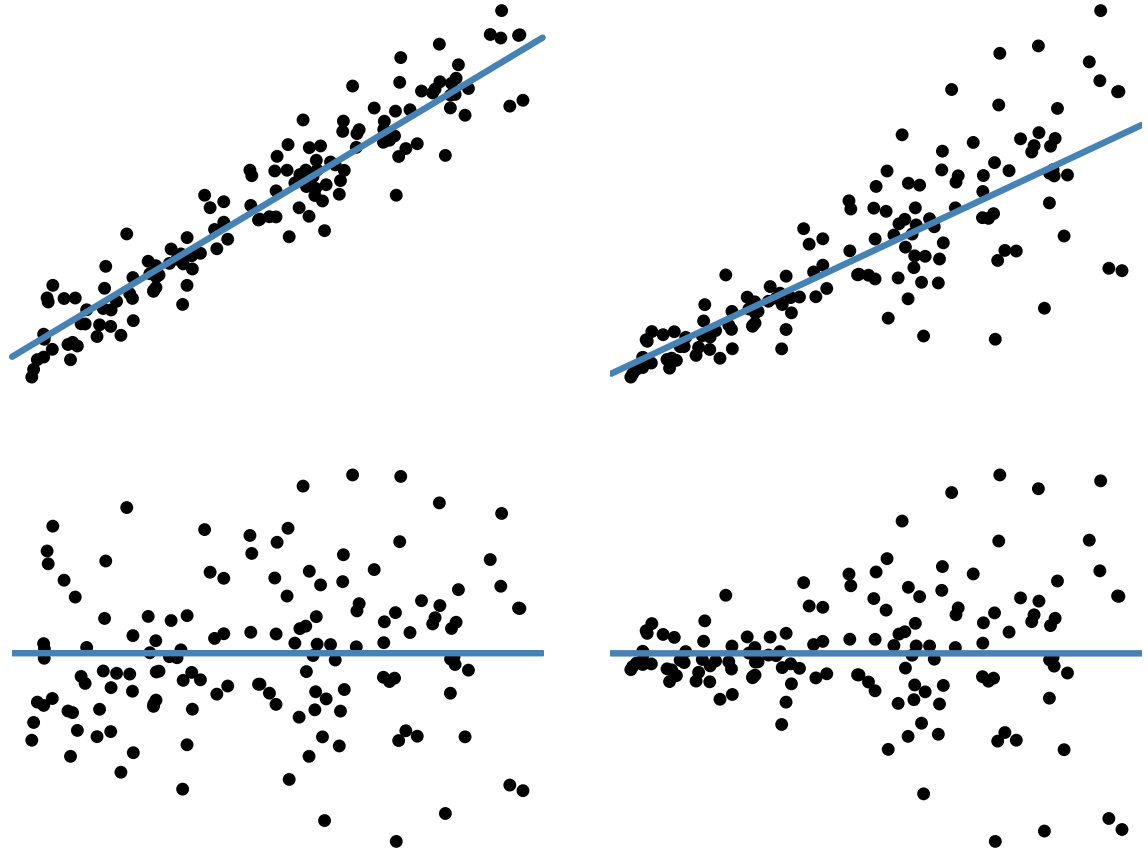


Homogeneity Assumption

Residuals of models of real data are often heterogeneous.



We don't like to see a megaphone shape



Independent Observations Assumption

Non-independence is one of the more challenging violations to deal with.

- Independent observation assumption key points:
  - Sampling is randomized.
  - Knowing something about observation x_1 gives us no information about observation x_2
  - The joint probability of independent events is the product of individual probabilities.
  - This is the basis for likelihood methods.

- Zuur, 2007:
  - "The independence assumption means that if an observed value is larger than the fitted value (positive residual) at a particular X value, then this should be independent of the Y value for neighboring X values."
- Non-independence can result from:
  - Proximity in space or time
  - Hierarchical structure

Fixed X Assumption

We often forget about the fixed-x assumption.

- Perfect accuracy in measurements of explanatory variables.
- This assumption is frequently violated
- It's OK-ish if the *noise* in the predictor variables' measurement is small relative to the noise in the response.

Regression Concepts 2

Key Concepts

- The regression equation
- Model coefficients and ANOVA (we'll talk much more about these)
- What is the constellation of methods?

Group 1 Models – 4 Key Assumptions Recap

- These assumptions apply to all of the Group 1 models we'll consider.

- Independent observations
- Constant variance a.k.a homoskedasticity, a.k.a. homogeneity
- Fixed x: no measurement error in our predictor variables
- Normality: normality refers to the model residuals

Regression Equation

We can express the dual model compactly with a regression equation.

- The basic regression equation can be expressed in several ways:

$$y_i = \alpha + \beta_1 x_1 + \epsilon_1$$

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_1$$

$$Y \sim Normal(\alpha + \beta X, \sigma)$$

Regression parameter interpretation

- Intercept: "The value of the response when the predictor is zero"
  - The intercept often occurs outside the range of our data: it is an exptrapolation.
- Slope parameters: "For each 1-unit change in x, we expect a β_1 change in the value of y (on average)."

Parameter Interpretation

A linear regression of penguin flipper length and body mass:

$$(Flipper\ length) = 136.7 + 0.015 \times (body\ mass)$$

```
lm(
  flipper_length_mm ~ body_mass_g,
  data = penguins)
```

```
Call:
lm(formula = flipper_length_mm ~ body_mass_g, data = penguins)

Coefficients:
(Intercept)  body_mass_g
  136.72956      0.01528
```

Overall Model Standard Deviation

Recall the basic regression equation:

$$y_i = \alpha + \beta_1 x_1 + \epsilon_1$$

We might ask: what is the overall model standard deviation?

- By that, we mean: what is the standard deviation of the residuals:

$$sd_{model} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}$$

Why n-2?

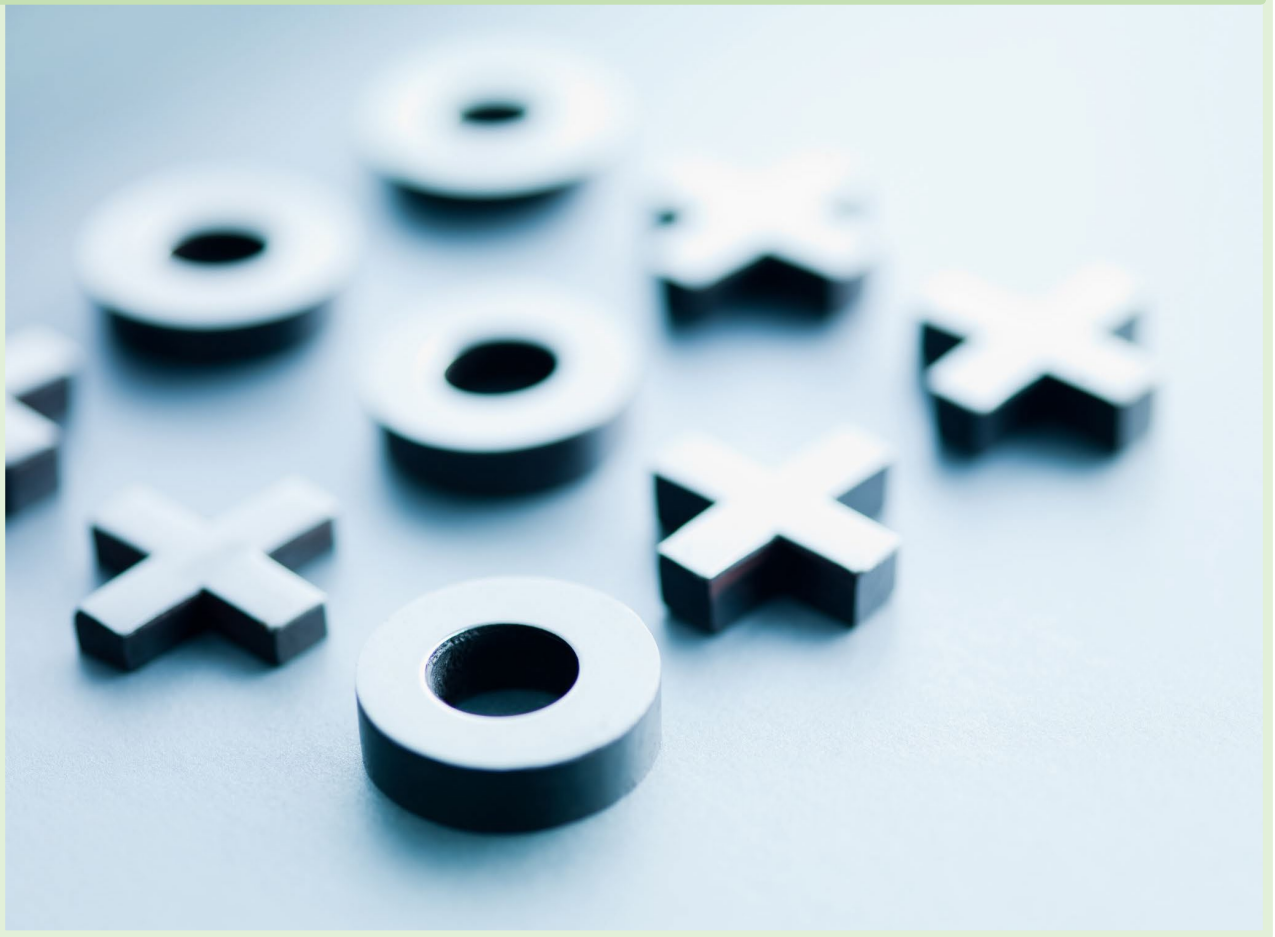
- We lose one degree of freedom for each parameter we estimate.
- We estimated two model parameters: α and β_1.

A Tale of Two Tables - Preview

Model Coefficients and the ANOVA Table

Two questions we might ask of a regression model:

1. What is the *magnitude* of the relationship between predictor x_1 and response y?
   - The model coefficient table tells us the direction and magnitude of the association between predictor and response.



A Tale of Two Tables - Preview

Model Coefficients and the ANOVA Table

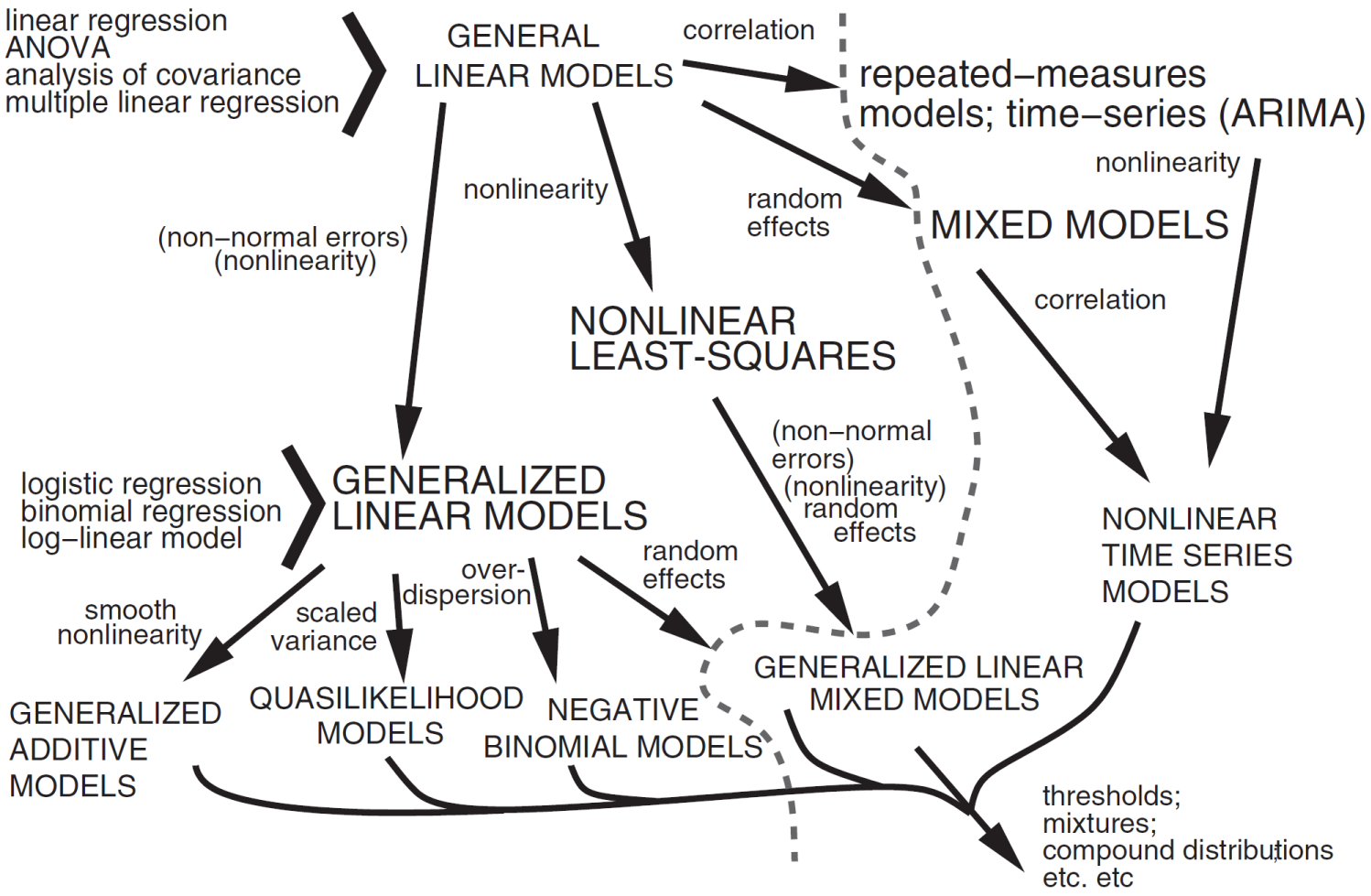
Two questions we might ask of a regression model:

2. How much of the variability in the model does predictor x_1 explain?
   - The Analysis of Variance (ANOVA) table tells us the relative importance of the various predictors to the overall model.

Model Diagnostics, Validation, and Selection

- How do we know that we have chosen the *best* model?
- Did we include the right predictors?
- Did our algorithm find the best parameter values?
- How well does our model fit the observed data?
- How well does our model predict new data?
- Does our data/model meet assumptions?
- Are the assumption violations *acceptably* small?

There are Many Types of Models: The Constellation



Bolker: Ecological Models and Data in R, Figure 9.2

There are Many Types of Models: The Constellation

- Many of the models beyond Group 1 were developed to handle violations of one or more of the Group 1 required assumptions.
- We'll spend most of our time on Group 1 models:
  - Easiest to understand, many principles transfer to other models.
  - Easiest to implement and interpret

Key Concepts

- The regression equation
- Model coefficients and ANOVA (we'll talk much more about these)
- What is the constellation of methods?

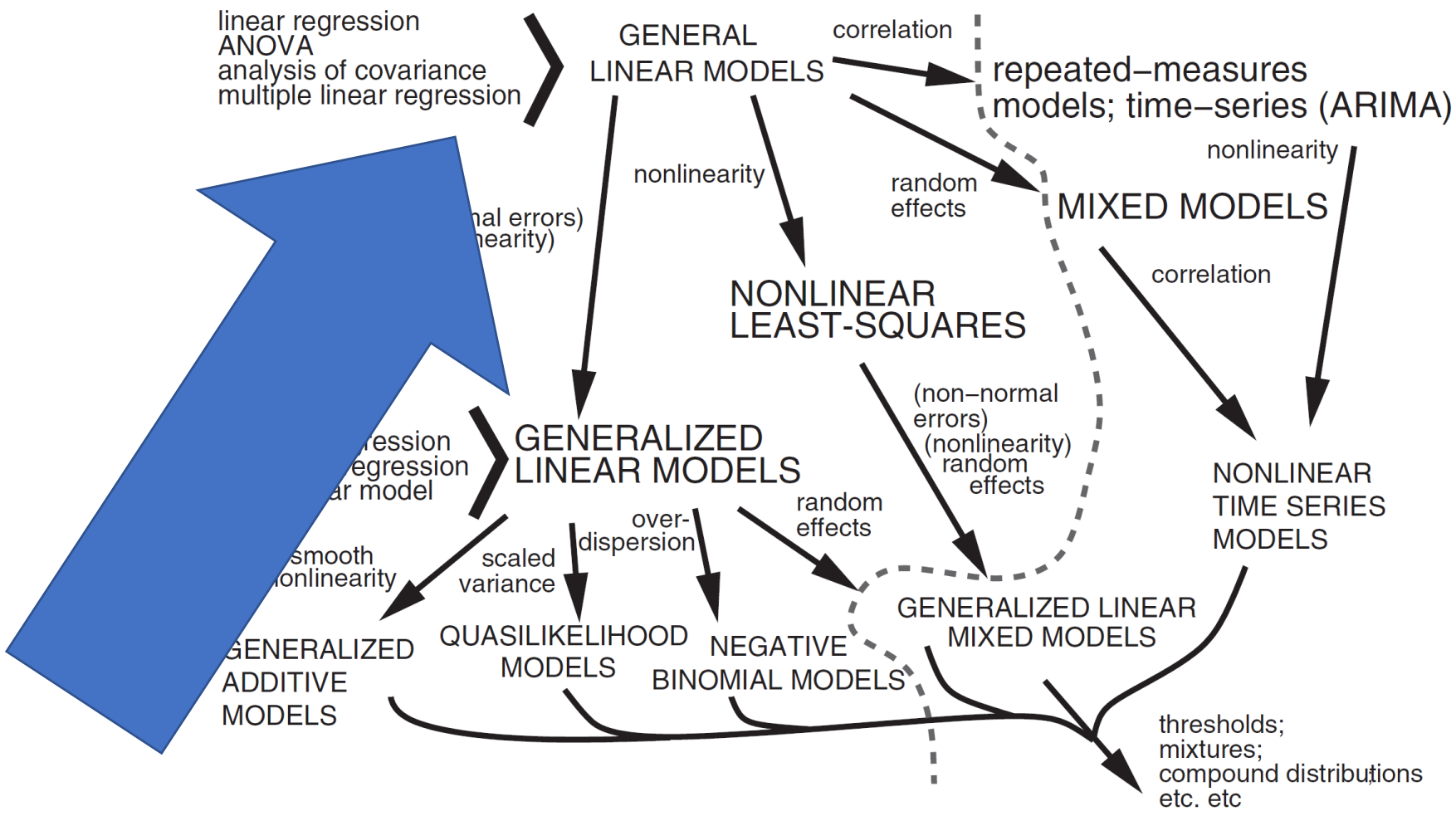
Group 1: General Linear Models

What's in This Section?

Take-Home Concepts

- What makes a model linear?
  - Linear in the parameters
- Categorical and continuous predictors.
- Group 1 responses are always continuous
- Key assumptions of general linear models.

There are Many Types of Models: The Constellation



Bolker: Ecological Models and Data in R, Figure 9.2

Group 1 Models – 4 Key Assumptions

Our Group 1 models carry some baggage… Specifically four key assumptions:

- Independent observations
- Constant variance a.k.a homoskedasticity, a.k.a. homogeneity
- Fixed x: no measurement error in our predictor variables
- Normality: normality refers to the model residuals

In addition, Group I requires that our models be *linear in the parameters* and have a response on a **continuous scale**.

The extended models can deal with different violations of these assumptions and requirements.

Group 1: General Linear Models

Four key assumptions:

- Normality: normality refers to the model residuals
- Constant variance a.k.a homoskedasticity, a.k.a. homogeneity
- Independent observations
- Fixed x: no measurement error in our predictor variables

Group 1 requirements:

- Group 1 models are linear in the parameters
- Group 1 models have a single continuous response variable

Terminology

- Response: Y
- Predictor(s): X
- Intercept: alpha
- Slope(s): beta

Group 1: Types of models

Group 1 methods are essentially variations on linear regression.

- T-Test Simple Linear Regression
- 1-Way ANOVA
- Multiple Linear Regression
- n-Way ANOVA
- ANCOVA

Group 1: general equation format

- Element-by-element form:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \epsilon_i$$

- Matrix/Vector form:

$$Y = \alpha + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \ \ldots + \beta_n \boldsymbol{X_n} + \epsilon$$

Group 1: Distribution Format

- We can also write the equations as:

$$y \sim Normal(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \sigma)$$

This format emphasizes the normality assumption of the residuals.

Linear in the Parameters

Linearity in parameters means that in the deterministic functions, the model coefficients can only have *multiplicative* relationships to the predictor variables.

- It will help to dissect some regression equations to identify variables, coefficients/parameters, and constants.

The classic simple linear regression equation:

$$y = \alpha + \beta x + \epsilon$$

Linear in the Parameters

This model is linear in the parameters: $y = \alpha + \beta x + \epsilon$

Things to note:

- x and y correspond to our *observations*. They are not estimated.
- α and β are the model coefficients, i.e. parameters. They are the quantities we want to estimate.
- β *multiplies* the predictor variable x.
- ϵ is the residuals, i.e. the stochastic model. For Group 1 this is the Normal distribution.

Linear in the Parameters

This model is also linear in the parameters:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

Things to note:

- x and y correspond to our *observations*. They are not estimated.
- β_1 ,β_2 ,and β_3 multiply the variables x_1 and x_2
- We used the product of the two predictors, x_1 and x_2 as a *third* predictor.

Linear in the Parameters

This model is *not* linear in the parameters:

$y = \alpha + \beta_1 x_1^2 + \alpha x_2^{\beta_2} + \epsilon$, Why not?

- The $\beta_1 x_1^2$ is ok. We've just used the square of the first predictor. It's like a modification of a predictor. Imagine that we could create another predictor column called 'sq' in our data that contained the squares of x_1.
  - Even though x^2 is not a linear function, the coefficient β_1 multiplies the term.

Linear in the Parameters

This model is *not* linear in the parameters:

$y = \alpha + \beta_1 x_1^2 + a x_2^{\beta_2} + \epsilon$, Why not?

- The term $a x_2^{\beta_2}$ is *not linear in the parameters*. Why?
  - The model coefficient β_2 does not *multiply* the predictor x_2 ,but rather it is an exponent.
  - The **constant** a multiplies x, but it is not a model coefficient estimated that is estimated from the data.

Linear in the Parameters

It seems weird that we can say $\beta_1 x_1^2$ is *linear* and $\alpha x_2^{\beta_2}$ is not.

- Both are nonlinear expressions.
- However, in the first term we are raising x_1 to a constant.
  - The *constant*, 2, is not estimated from the data therefore it is not a model *coefficient*.
- In the second term, we have specified a *model coefficient* as an exponent.
  - Since the coefficient does not *multiply* but rather *exponentiates* the predictor it is not *linear in the predictors*.

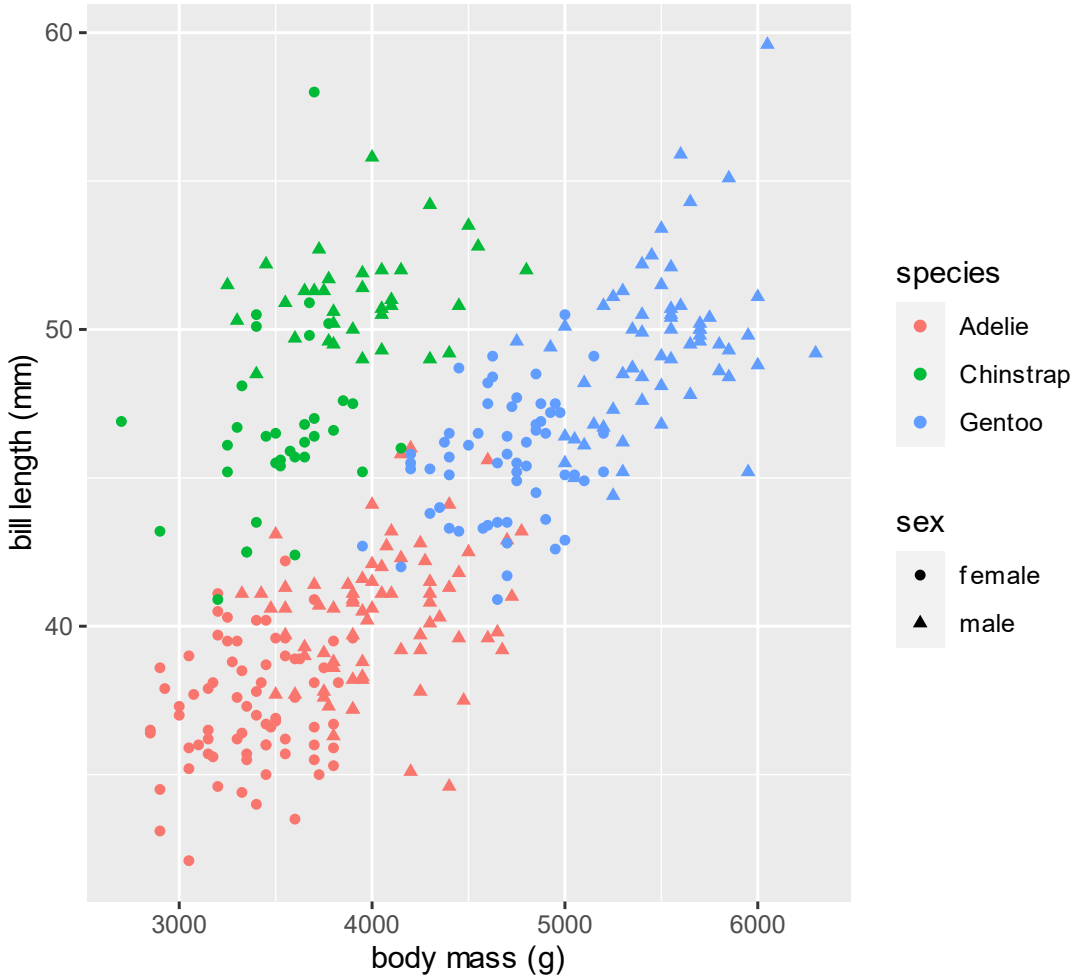
Palmer Penguin Data

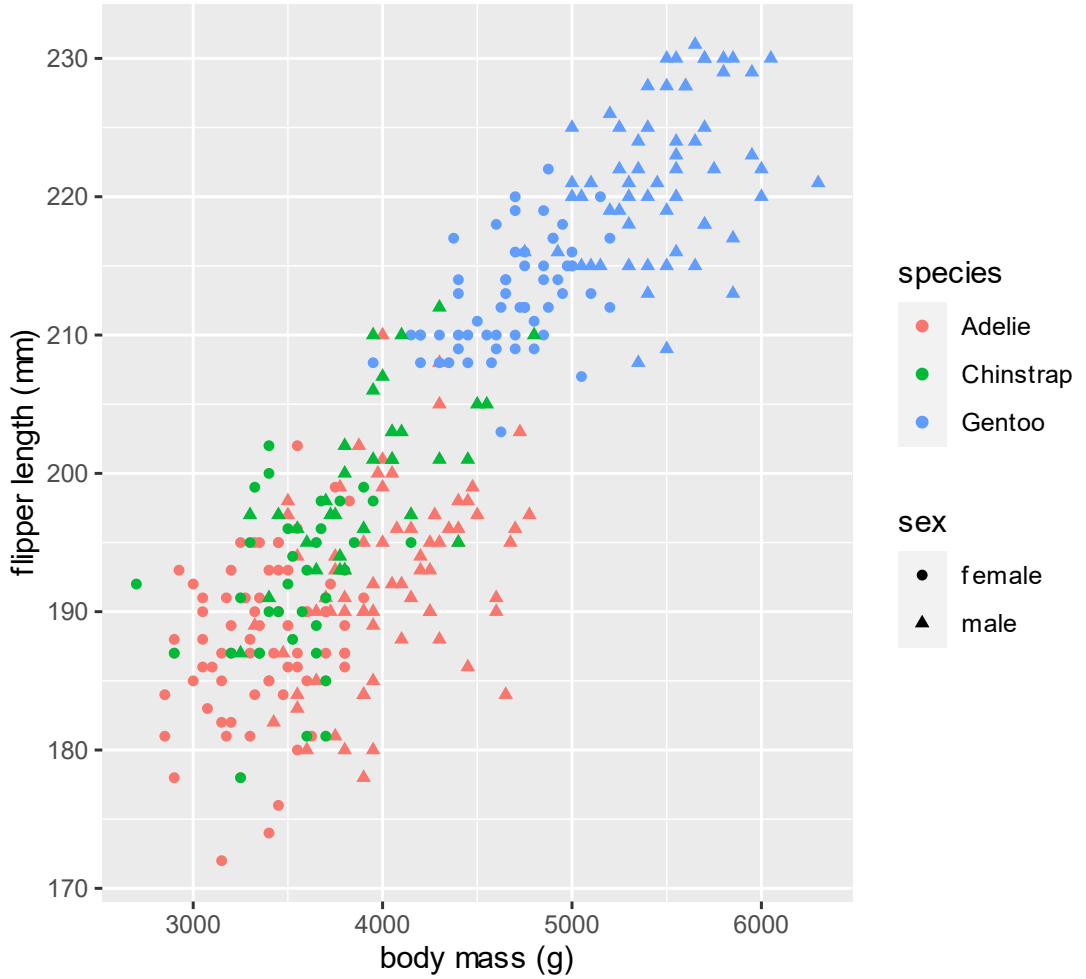
We'll use the Palmer Penguin dataset to illustrate group 1 methods

- Dr. Kristen Gorman and the Palmer Station Long Term Ecological Research (LTER) Program.
- 3 Penguin species in the Palmer Archipelago
  - size measurements: 4 continuous variables
  - species, island, and sex: categorical - nominal scale
- R package palmerpenguins

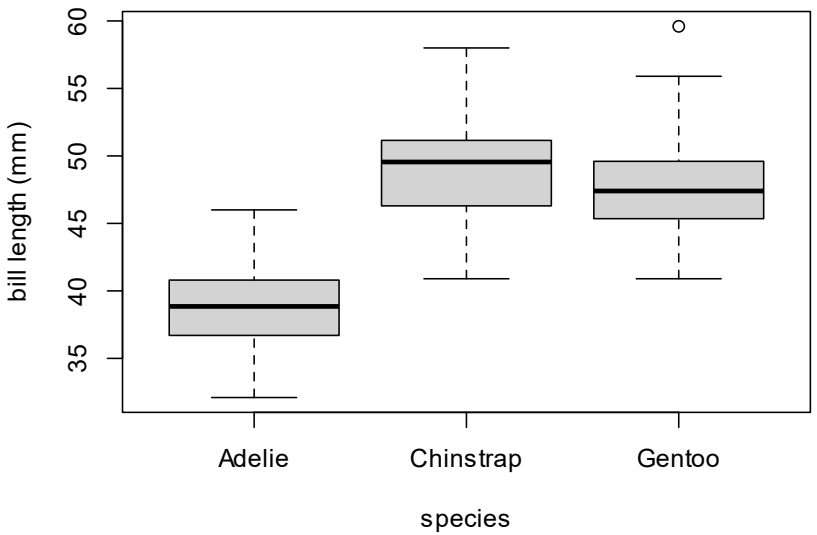
https://education.rstudio.com/blog/2020/07/palmerpenguins-cran/

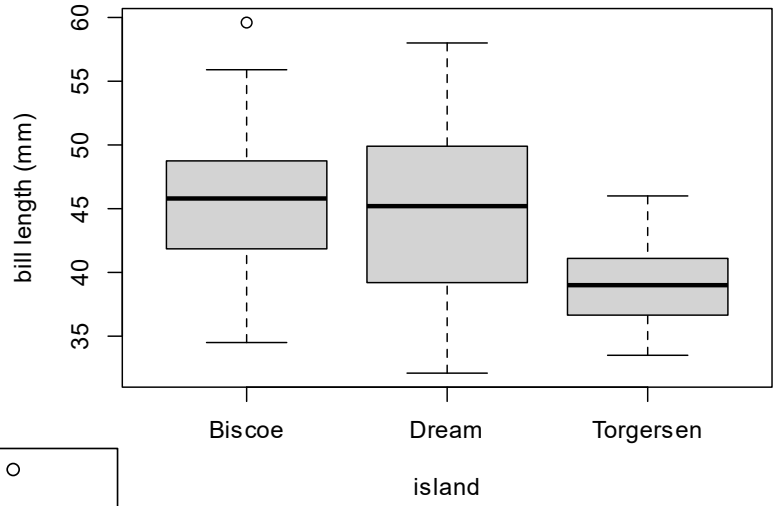
Palmer Penguins

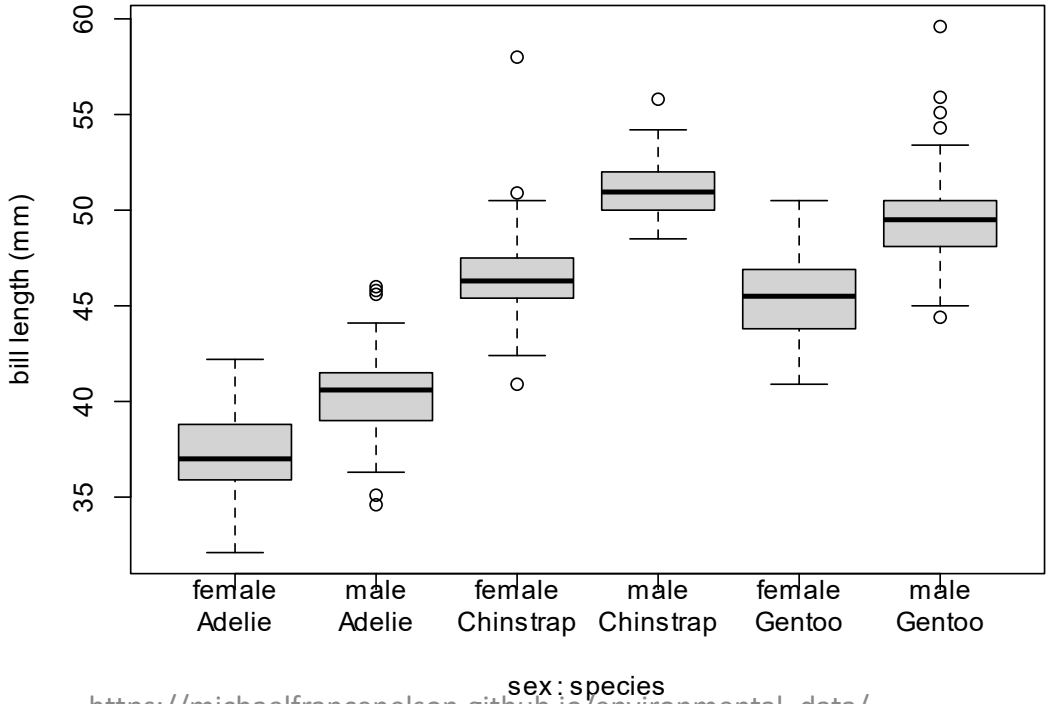




Palmer Penguins – Graphical Exploration







Tests For Differences: 2 Samples

Group 1: T-tests

- **t-test**
- Simple Linear Regression
- 1-Way ANOVA
- Multiple Linear Regression
- n-Way ANOVA
- ANCOVA

T-tests are appropriate with

- One categorical predictor with 1 or 2 levels
- One continuous response

T-tests analyze the following questions:

- Is the mean of one group different from a fixed value?
- Are the means of two groups different from each other ?

An elaboration of the t-test:

- 1-way ANOVA extends t-test to 3 or more groups.

What's a T-Test?

The problem: we want to know if the means of two groups of observations are different.

What could we do?

- Compare means the means of the two groups?
- How could we assess significance?

What's a T-Test?

A t-test tests the **null hypothesis** that the two groups of observations were drawn from the same population.

- The **alternative hypothesis** is that they were drawn from different populations.
- We use measures of center and spread to calculate a **t-statistic:**

For 1 sample: $\boldsymbol{t} = \frac{\bar{x}_1 - \boldsymbol{\mu}}{\boldsymbol{s}/\sqrt{\boldsymbol{n}}}$ For 2 samples: $\boldsymbol{t} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\boldsymbol{s_1^2}}{\boldsymbol{n_1}} + \frac{\boldsymbol{s_2^2}}{\boldsymbol{n_2}}}}$

What's a T-Test?

Large t-values support the alternative hypothesis

Small t-values support the null hypothesis

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

What factors contribute to the t-value?

- Difference in means: large difference = larger t-value
- Sample variances: small variance = larger t-value
- Sample sizes: larger sizes = larger t-value

T-test: Samples

There are 1- and 2-sample versions of the t-test:

- 1-sample compares the mean of a group of measurements to a fixed value.
- 2-sample compares the means of two groups of measurements

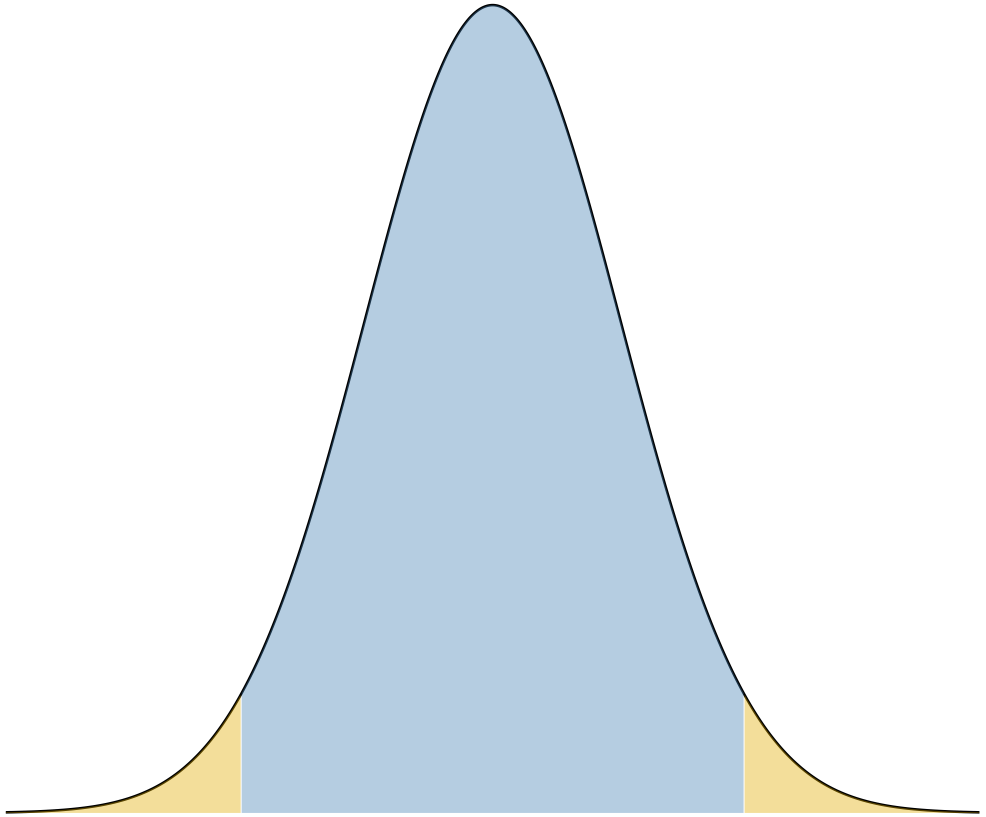
T-test: Tails

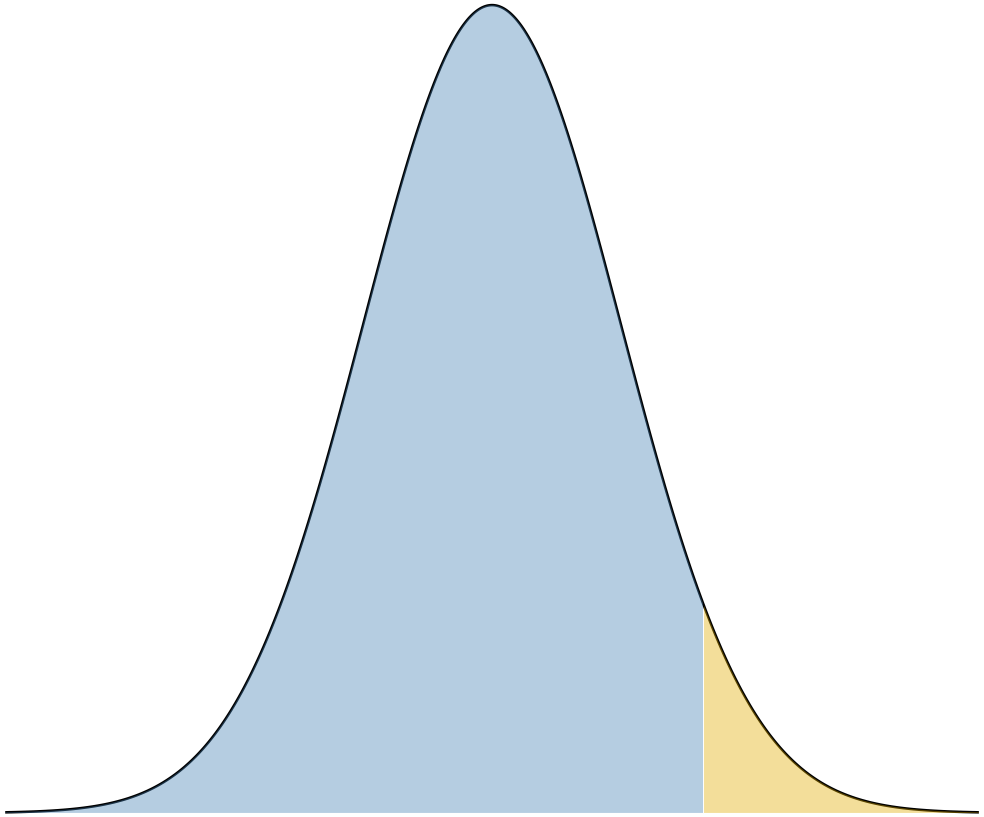
1-tailed	2-tailed
Specifies a directional alternative hypothesis: • "Chinstrap penguins weigh more than Adelie penguins." • You have to specify ahead of time. Usually requires prior knowledge or experience. • Smaller critical t-values	Specifies a non-directional alternative hypothesis: • "Chinstrap and Adelie penguins have different body masses." • More general than the 1-tailed, you don't need any prior knowledge. • Higher critical t-values.

Critical Values and Rejection Regions

Critical Value	Rejection Region
• Critical t-values are determined by the significance level (alpha) and the degrees of freedom. • Critical difference is the difference in means corresponding to the critical t-values.	• Rejection regions are in the tails of the distribution. • If the observed difference in means is greater than the critical difference, it falls within the rejection region.

Critical Values and Rejection Regions





Tell UMass to Pledge to Reduce Plastics!

Become a Plastics Reduction Partner

The Plastics Reduction Partner Pledge & Certification calls on institutions to reduce their dependence on plastics by making actionable commitments across four categories (awareness, behavior change, operational change, & demonstrating leadership). Tell UMass this is important to you!

Scan to **SIGN THE PETITION** for UMass to join the Plastics Reduction Partner Certification

Scan to **HELP DECIDE WHICH PRIORITIES** UMass should focus on for reducing plastics on campus

What could a t-test tell us about the penguins?

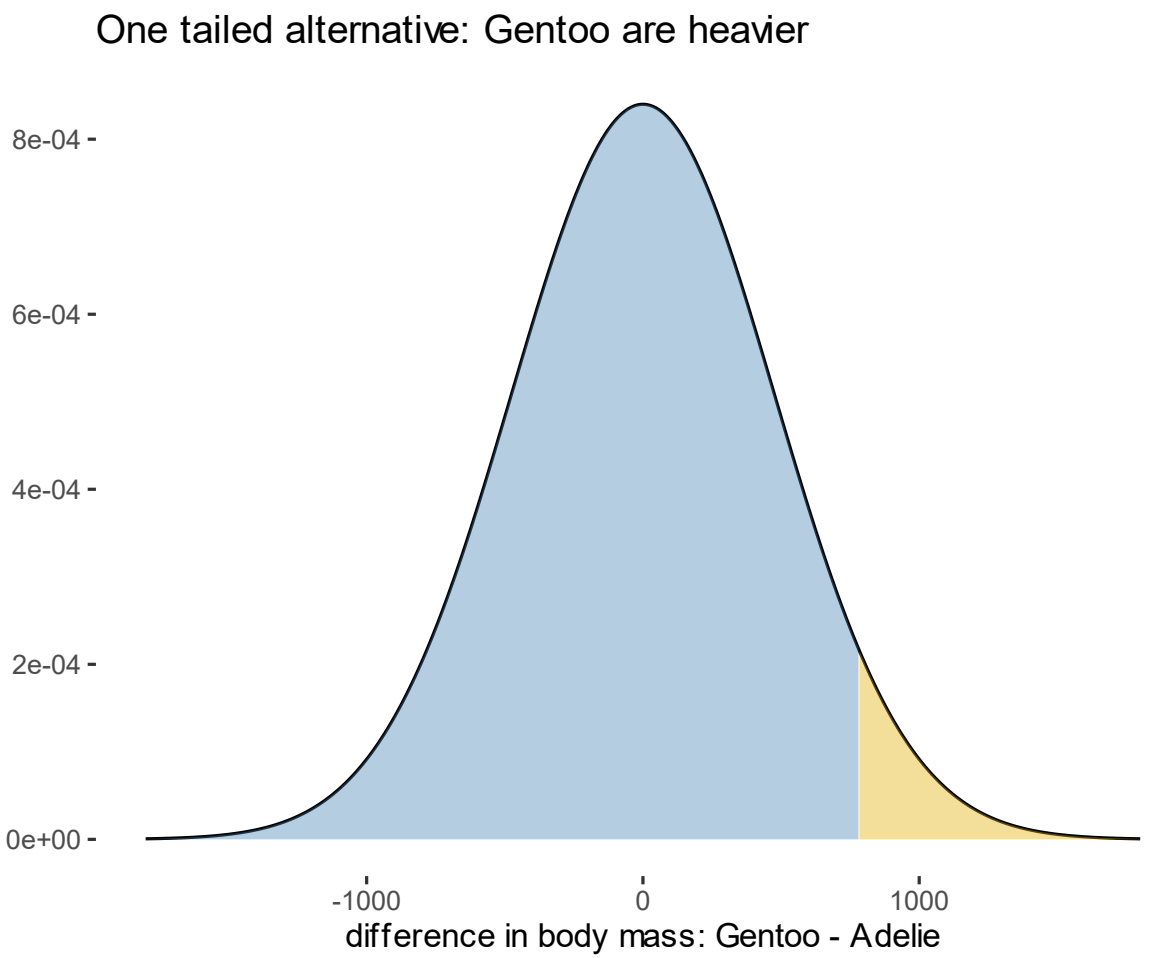
Hint: What are the categorical predictors?

Maximimal for two parameter!
Check actual standard deviation and mean values > actual measurements > Ausgangspunkt fuer guessing best fit

1-tailed Test: Gentoo are heavier

T.test
Compare the sex of penguins > body mass (continuous)
Predictor is categorical with two levels > response is continuous
Sex would be a predictor!!!!!
Critical difference is about 900g, 900g and heavier = significant > one tailed

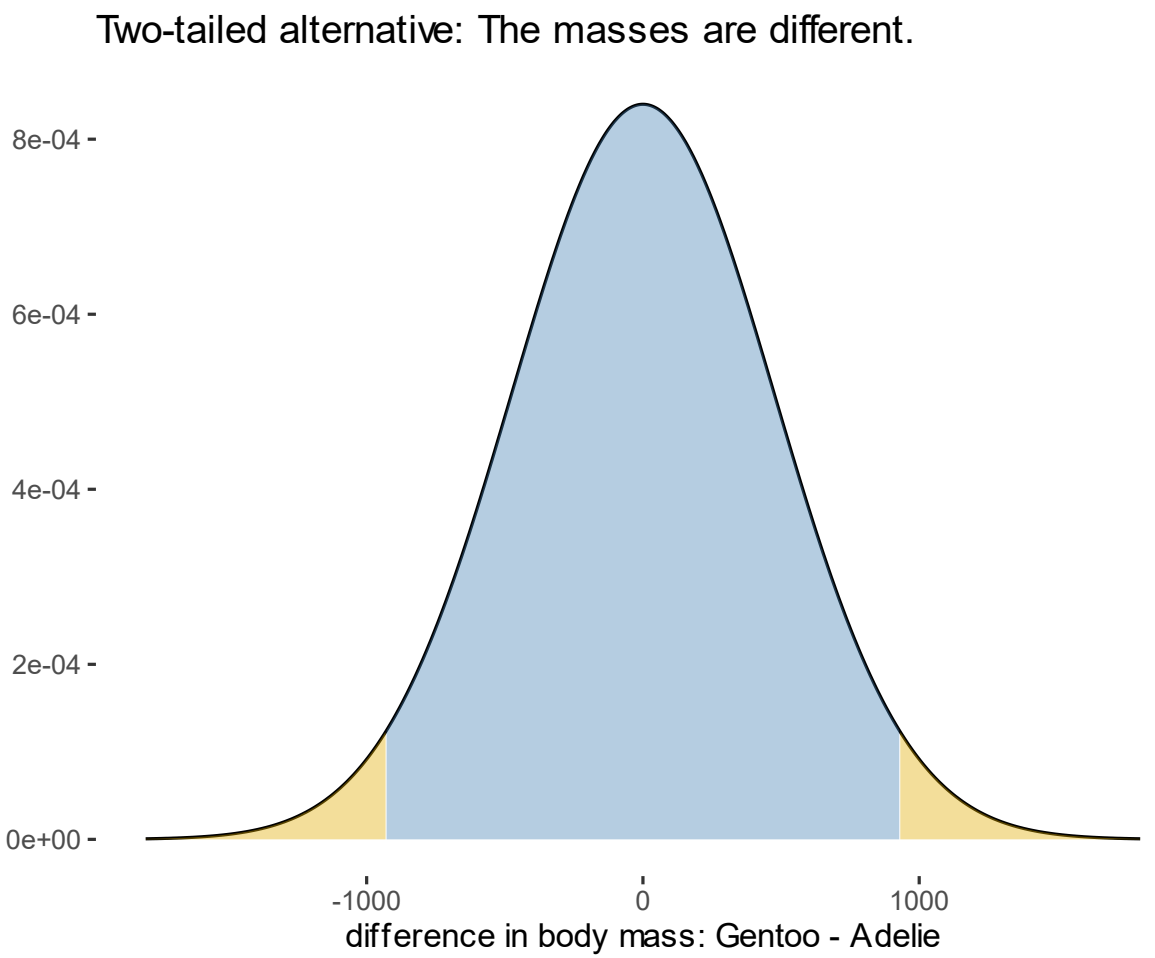
- Rejection region is a single tail.
- Critical difference is about 900g.



2-tails: masses are different

Two tailed test > differences are real > but don't know which is bigger
Critical difference is bigger for two tailed

- Rejection regions in both tails.
- Critical difference is about 950g.



Remember To Test Your Assumptions

4 Key assumptions:

For all general linear methods we have these assumptions
> normality of the residuals > difference of each observation from the mean value

1. Normality: normality refers to the model residuals
2. Constant variance a.k.a homoskedasticity, a.k.a. homogeneity
3. Independent observations
4. Fixed x: no measurement error in our predictor variables

We'll test the first 2

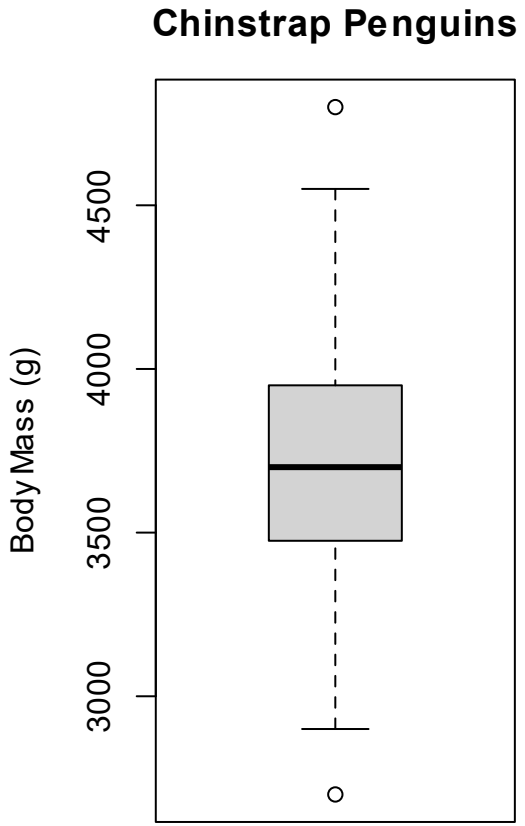
Testing Assumptions: Normality

Shapiro test: most common normality test.

- Null hypothesis: data are normal

In R: shapiro.test():

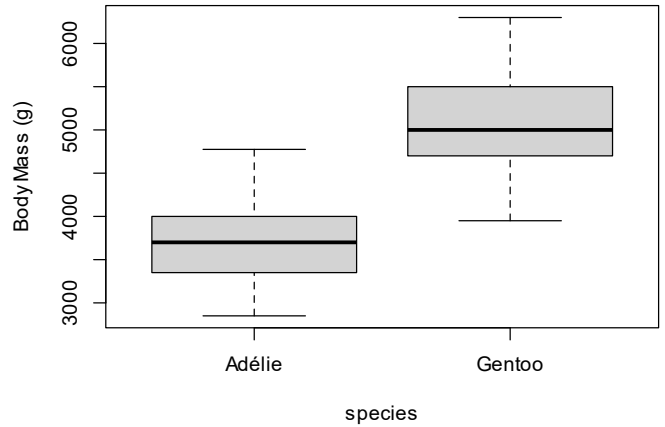
```
> dat_chinstrap = subset(
+ penguins, species == "Chinstrap")
> shapiro.test(dat_chinstrap$body_mass_g)

        Shapiro-Wilk normality test

data:  dat_chinstrap$body_mass_g
W = 0.98449, p-value = 0.5605
```



Testing Assumptions: Equal Variance

- We can use the Bartlett test

constant variance > when value higher
Bartlett test equality of variances/constant variance
> null hypothesis is that variances are the same > small value reject null > and say there are differences in variance



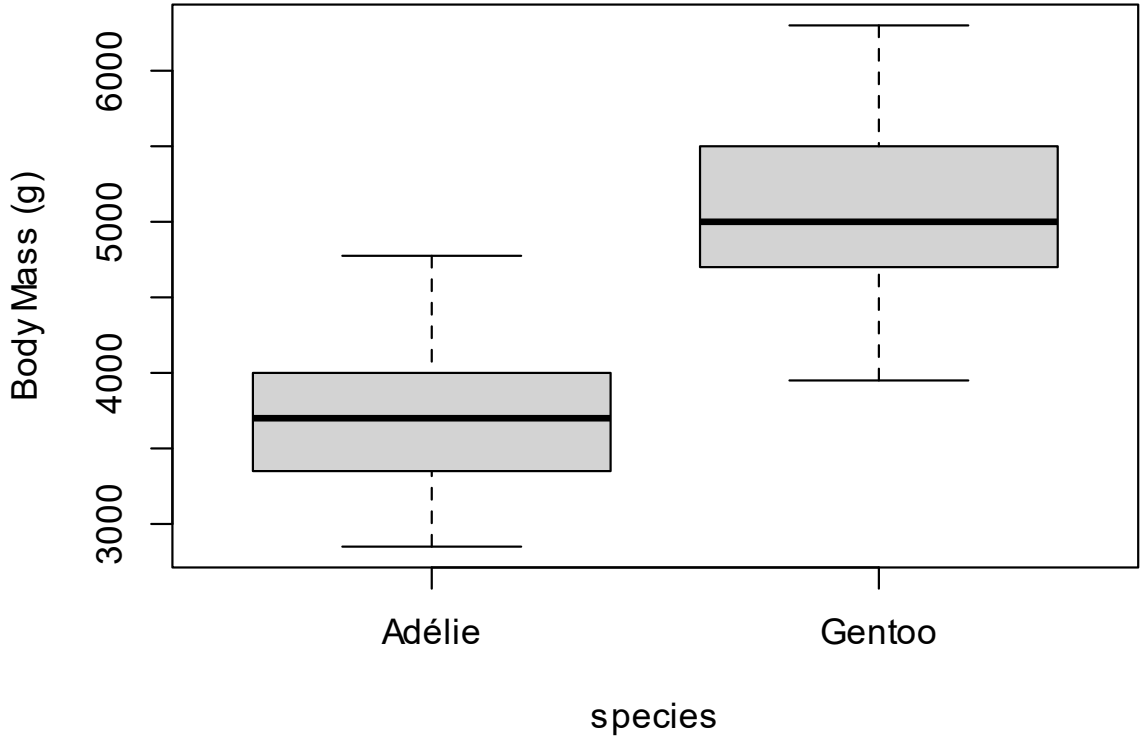

```
> dat_pen = subset(
+ penguins, species %in% c("Adelie", "Gentoo"))
> bartlett.test(body_mass_g ~ species, data = dat_pen)

        Bartlett test of homogeneity of variances

data:  body_mass_g by species
Bartlett's K-squared = 1.2084, df = 1, p-value = 0.2717
```

Testing Assumptions: Equal Variance

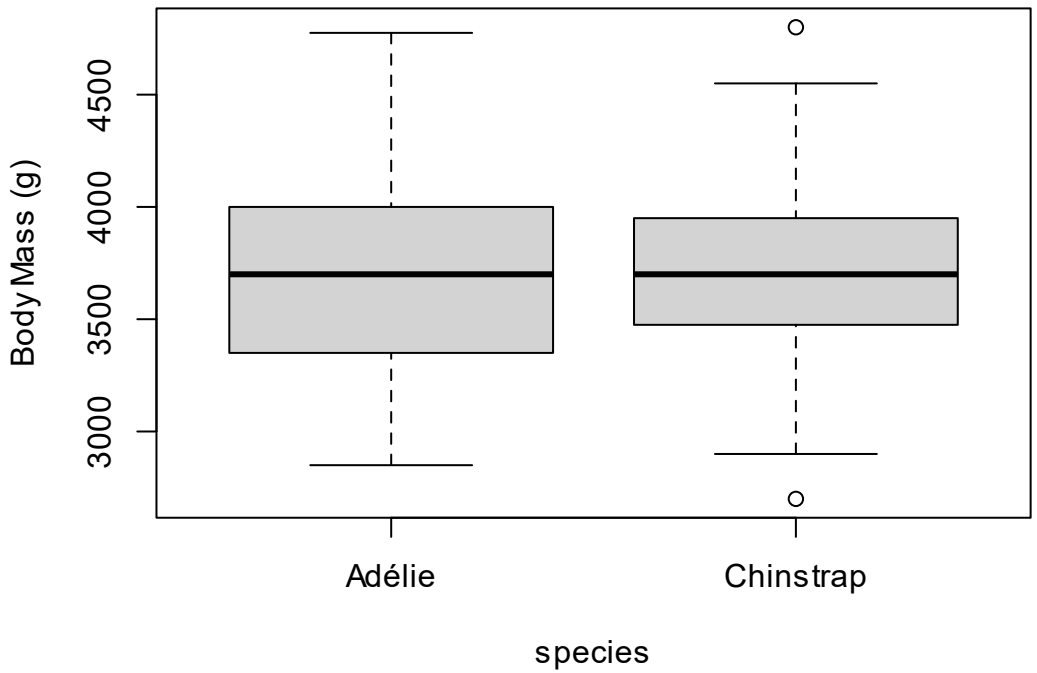
Anova can not deal with unequal variance
T test can

T-tests are robust to heterogeneity

- ANOVA is not!



T-test Adelie and Chinstrap Penguins



Boxplot best to see when doing exploration
> equal variance, equal mean

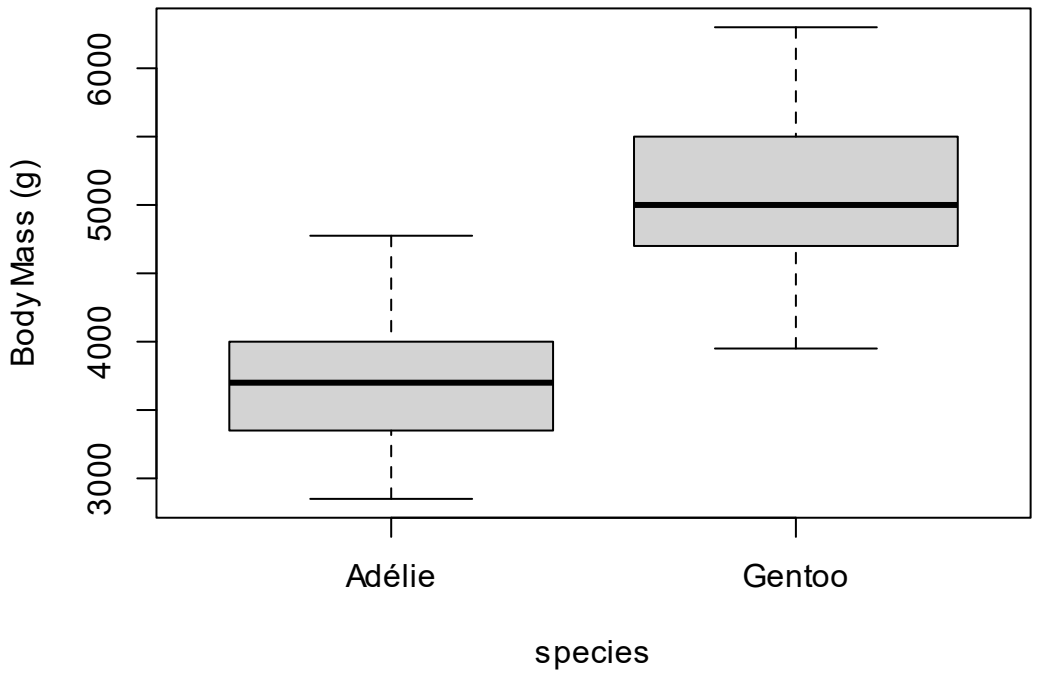
- Do the two groups seem different?

T-test Adelie and Chinstrap Penguins

```
> t.test(body_mass_g ~ species, data = p4)

        Welch Two Sample t-test

data:  body_mass_g by species
t = -0.54309, df = 152.45, p-value = 0.5879
alternative hypothesis: true difference in means between
 group Adélie and group Chinstrap is not equal to 0
95 percent confidence interval:
 -150.38481   85.53284
sample estimates:
   mean in group Adélie mean in group Chinstrap
               3700.662                3733.088
```

Variable on left explained by variable at the right
Confidence interval is t-value > contains > 0
indicates there is no difference

T-test Adelie and Gentoo Penguins



Look different
Equal variance? > yes because same vertical space > look symmetrical

- Do the two groups seem different?

T-test Adelie and Chinstrap Penguins

```
> t.test(body_mass_g ~ species, data = p3)

        Welch Two Sample t-test

data:  body_mass_g by species
t = -23.386, df = 249.64, p-value < 2.2e-16
alternative hypothesis: true difference in means between
 group Adélie and group Gentoo is not equal to 0
95 percent confidence interval:
 -1491.183 -1259.525
sample estimates:
mean in group Adélie mean in group Gentoo
            3700.662             5076.016
```

Confidence interval does not contain zero > different in means
Bottom values tell mean of each group

Nonparametric Alternative: Wilcoxon Test

Wilcoxon test > non-parametric test > don't have assumptions

- If our assumptions aren't met, we can use a non-parametric alternative: the Wilcoxon test.
  - Also known as the Mann-Whitney U test.
  - Syntax is very similar to t.test() in R
  - Function is wilcox.test()

Wilcoxon Test Syntax

Output: no confidence interval > tells us alternative hypotheses
W > test statistic

General Linear method > t test most easiest one

```
>     wilcox.test(body_mass_g ~ species, data = p4)

        wilcoxon rank sum test with continuity correction

data:  body_mass_g by species
W = 4831, p-value = 0.4855
alternative hypothesis: true location shift is not equal to 0
```

Group 1: Simple Linear Regression

- t-test
- **Simple Linear Regression**
- 1-Way ANOVA
- Multiple Linear Regression
- n-Way ANOVA
- ANCOVA

SLR Requires:

- One continuous response
- One continuous predictor
- What questions could we address in the penguin data?

Here have a continuous preceptor instead of categorical
> ask questions about continuous variables
> flipper length in terms of body mass

Simple Linear Regression elaborations

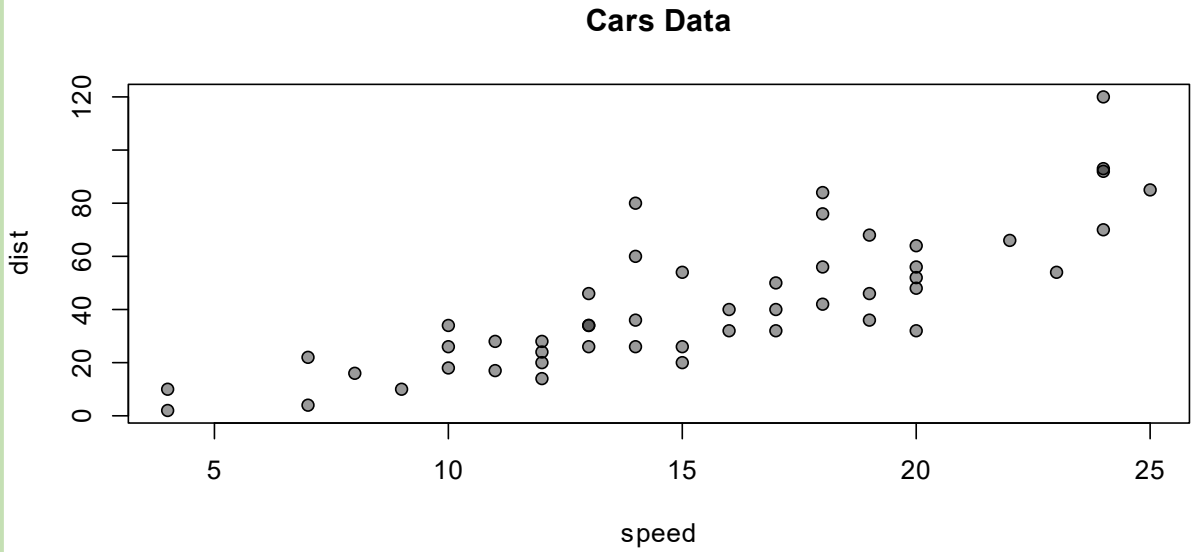
1. Multiple linear regression: More than one continuous predictors
2. ANOVA: One categorical predictor (instead of continuous)
3. ANCOVA: Mixture of categorical and continuous predictors

Simple Linear Regression: Example

The Cars Dataset

Example simple regression: Cars data

- The cars data set contains speed and stopping distance observations from 50 trials.
- The goal was to quantify the relationship between driving speed and stopping distance.



Cars: Fitting a Model

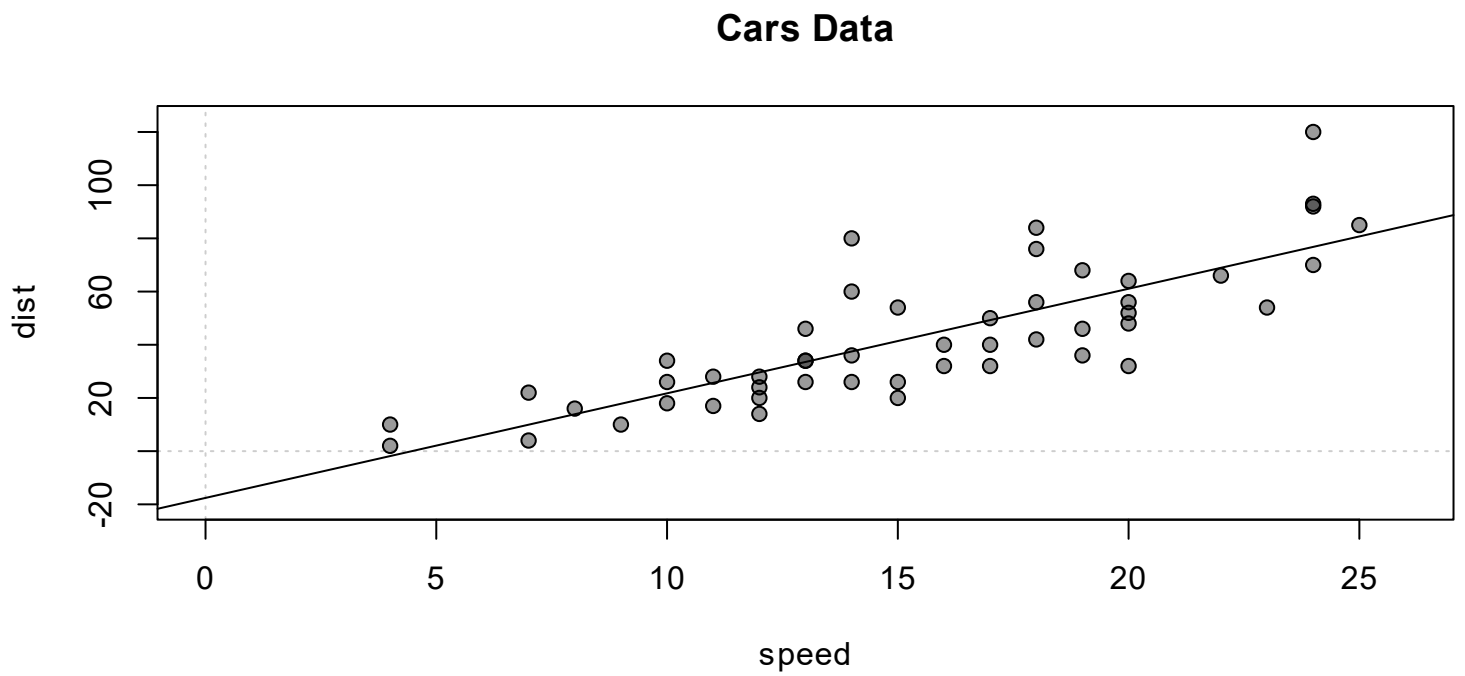
- Which variable is the predictor?
- Which is the response?
- We can fit a simple linear model:

looks linear
One predictor > speed
Response > stopping distance
Lm() > linear model
Distance in terms of speed
Lm creates linear model object
Abline > function to make line on top of the data

```
fit_cars = lm(dist ~ speed, data = cars)
```

```
plot(dist ~ speed, data = cars)
abline(fit_cars)
```

We can use abline() to plot the regression line.



Cars Model: What can we learn from the coefficient table?

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Model Formula

Residuals summary – we'll learn more about this later

>>Residuals: summary, want residual mean of zero > to be normally distributed

Model Coefficients

Amount of variation explained by the model

R-squared output > how well model fits the data > percentage of variability of the data explained by the model > R squared > how much the output is explained by predictor variable

> One measure of model fit

Example simple regression: Cars data

Let's translate the model coefficients to English. You can uses the `coefficients()` function in R to retrieve just the regression coefficients, without all of the information from the full summary:

```
coefficients(fit_cars)
(Intercept)       speed
-17.579095    3.932409
```

Intercept: linear equation > predictor coefficients is intercept = -17, and slope: 3 for speed
Intercept > Value that expect to observe when set all predictor variables to value of zero
Intercept >
R squared 64 percent
> not too bad in a noisy dataset

- The intercept is about -17.6.
  - What does that mean, is it sensible?
  - How could you translate the speed coefficient (3.9) into an English sentence?
- The R2 was 64%: That means the model explains about 64% of the variation in the response.
  - Is that a lot? Is it a good model? What is the other 36%?

Example simple regression: Cars data

Let's translate the model coefficients to English. You can uses the `coefficients()` function in R to retrieve just the regression coefficients, without all of the information from the full summary:

```
coefficients(fit_cars)
```

```
(Intercept)       speed
-17.579095    3.932409
```

Coefficients > intercepts of y axis > often outside of the range of data > ignore them
Meaningfull is slope coefficient > speed
If travel 10 mile it would take 40 feet longer to stop

- The intercept is about -17.6.
  - What does that mean, is it sensible?
- How could you translate the speed coefficient (3.9) into an English sentence?
  - "For each 1-mph increase in speed, it takes about 4 additional feet to stop".

Cars Model: Regression Equation

- We can build the regression equation from the model coefficient output from R:

$$distance = -17.6 + 3.9 \times speed$$

- How does this help us?
- One way we can use regression equations is for **prediction**.

- Given our regression equation
- $distance = -17.6 + 3.9 \times speed$
- we could calculate an **expected** stopping distance for any possible speed.

Cars Model: Regression Equation

We could do the calculation by hand, but fortunately R has a built-in function to use a model fit object to obtain predicted values.

- Perhaps counter intuitively, this function is called `predict()`.
- `predict` expects a data frame with a columns for each of the model predictors. In the cars model we had only one predictor: `speed`.
- We could calculate the expected stopping distance for a car travelling at 34 mph:

```
predict(fit_cars, newdata = data.frame(speed = 34))
```

```
       1
116.1228
```

Using slope can make predictions
> predict()
> give estimate of deterministic model > what is expected stopping distance if speed is 34 > 116 .1228

Regression Equation: Driving at 450 mph

Predictions for very high speeds > 1752 miles to stop
> we dopt trust the estimate, collect same data rather

What about a car traveling at 450 mph?

```
predict(fit_cars, newdata = data.frame(speed = 450))
       1
1752.005
```

The equation predicts that you would need about 1/3 mile to stop.

Regression Equation: Stopped Car

A stopped car (0 mph) would need:

```
predict(fit_cars, newdata = data.frame(speed =
0))
```

```
        1
-17.57909
```

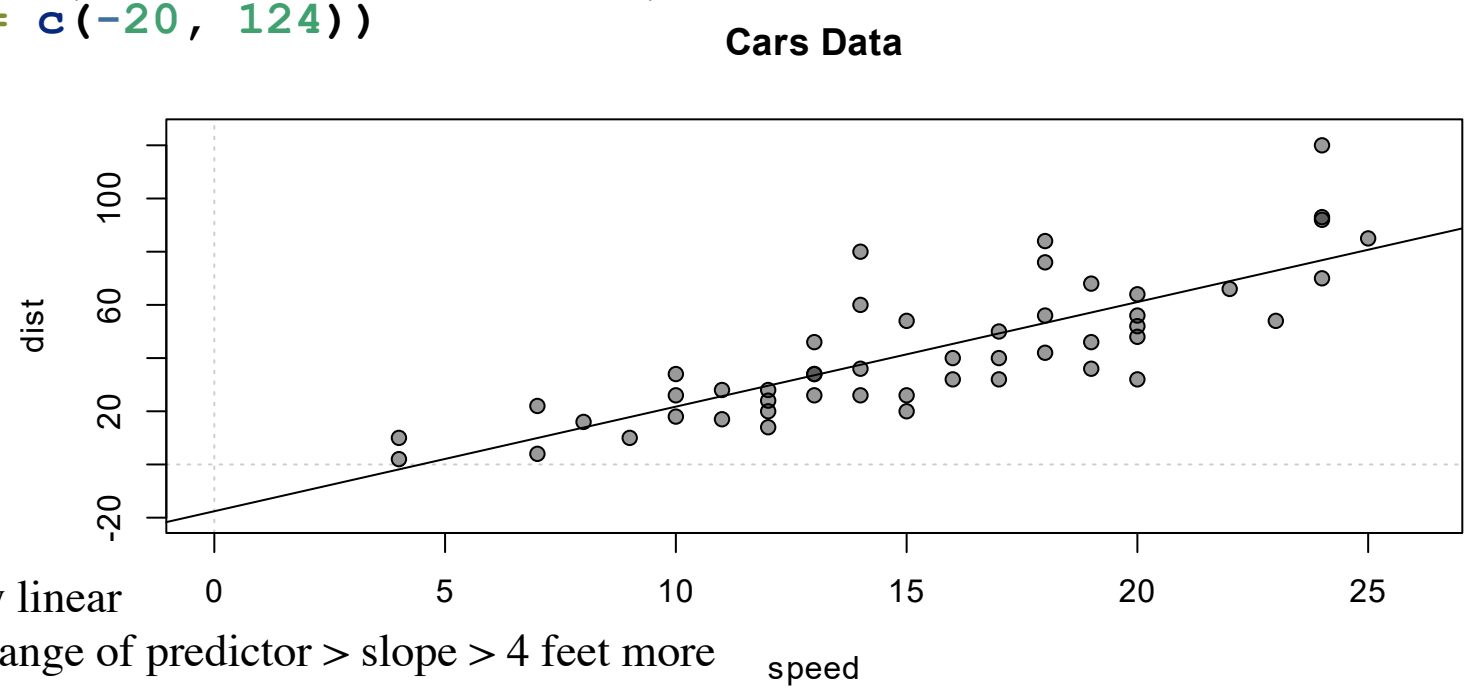
Intercept -17
Stopping distance expected to be when set all predictor variables set to zero > -17.57 > unrealistic > dont need

- Any potential issues with these predictions?
- Does that mean our model is *bad*?
- Is this value familiar to us from the model coefficient table?

Regression Equation: Stopped Car

Take another look at a plot of the data and the model fit:

```
plot(dist ~ speed, data = cars, main = "Cars Data",
     xlim = c(0, 26), ylim = c(-20, 124))
abline(fit_cars)
```



Even a curve could fit the data > when not really linear

Speed > rate of change of response per 1 unit change of predictor > slope > 4 feet more

0 speed is outside the range of our observations! It's an extrapolated value!

Group 1: 1-Way Analysis of Variance

- t-test
- Simple Linear Regression
- **1-Way ANOVA**
- Multiple Linear Regression
- n-Way ANOVA
- ANCOVA

ANOVA: Categorical predictor, 3 or more levels

- Continuous response
- Like an extended t-test

Analyzes the following questions:

1. Are the group means different from one another?
   - Note: ANOVA does not specify which pairs of groups are different from one another.

What could a 1-way ANOVA tell us about the penguins?

What were the categorical variables?

- Sex
- Species
- Island

Allows to test for categorical predictor that has 3 or more levels
Null hypothesis: null difference between the groups
Don't have option of specifying a directional hypothesis
>>> Body mass in terms of species or island

Single categorical predictor with 3 levels
Different body mass by species and also adding sex
When get significant result > doesn't tells which pairs are different
> post hoc testing which is turkey test

ANOVA elaborations

Two or more categorical predictors: multi-way ANOVA

Categorical and continuous predictors: Analysis of Covariance (ANCOVA)

Post ANOVA analysis: which groups are different from one another?

- Tukey Honest Significant Difference (HSD) test
  - Pairwise tests between all factor levels.
    - number of pairs gets large very quickly!
  - Correction for multiple testing: Bonferroni, etc.

Group 1: Multiple Linear Regression

Penguin flipper length > in terms of body mass and bill depth > three continues variables
Intraction btw body mass and flipper length, and bill depth and flipper length

- t-test
- Simple Linear Regression
- 1-Way ANOVA
- **Multiple Linear Regression**
- n-Way ANOVA
- ANCOVA

A multiple linear regression model has:

- One continuous response
- Two or more continuous predictors

The model attempts to quantify the pairwise relationships between each predictor and the response - combined effect of 2 or more predictors on the response

Multiple regression can fail with highly correlated predictors: collinearity and multicollinearity.

Group 1: Multiple Linear Regression

- t-test
- Simple Linear Regression
- 1-Way ANOVA
- **Multiple Linear Regression**
- n-Way ANOVA
- ANCOVA

Multiple regression elaborations:

- Mixture of categorical and continuous predictors:
  - Interaction terms: synergistic effects of two or more predictors.
  - Analysis of Covariance (ANCOVA)

Group 1: Multiple Linear Regression

- t-test
- Simple Linear Regression
- 1-Way ANOVA
- **Multiple Linear Regression**
- n-Way ANOVA
- ANCOVA

What can it tell us about the penguins?

- What were the continuous predictors?
  - flipper length, bill measurements, body mass.
- Could we use these three continuous variables to predict the species?
  - Hint: no! Group 1 methods require a continuous response!

Group 1: Multi-Way ANoVA

- t-test
- Simple Linear Regression
- 1-Way ANOVA
- Multiple Linear Regression**
- **n-Way ANOVA**
- ANCOVA

Categorical analogue of multiple regression

- Main effects
- Interactions

What could we ask with the penguin data?

- Categorical variables: island, sex

Elaboration: Mix of categorical and continuous variables Analysis of Covariance (ANCOVA)

Group 1: Analysis of Covariance

- t-test
- Simple Linear Regression
- 1-Way ANOVA
- Multiple Linear Regression**
- n-Way ANOVA
- **ANCOVA**

ANCOVA combines categorical and continuous data:

- A mix of categorical and continuous predictors
- Continuous response

What could we ask with the penguin data?

- Categorical variables: island, sex
- Continuous variables: flipper length, bill dimensions, body mass

ANCOVA
Effect of island as well as of sex and body mass when thinking of flipper length
Could all be predictors for bill length

When do group 1 methods start to fail?

We have to be able to multiply observations > independent
Megaphone shape when have no constant variance
> transforming variables or switch to other model > generalized liner model zum beispiel
Normality of residuals > least important when have decent sample sizes
Generalized > non constant variance

Violations of our key assumptions:

- Independent observations:
  - Problems with likelihood: inflated confidence
- Constant variance a.k.a homoskedasticity, a.k.a. homogeneity.
  - Inaccurate measures of confidence/significance
- Fixed x: no measurement error in our predictor variables
- Normality: normality refers to the model residuals

When do group 1 methods start to fail?

> complicated correlation structures in predictors > correlated with one another
Bill length and body mass are correlated > issues if include both as predictors
Want response variable to be correlated with predictor
But not both predictors

[Multi]Collinearity

- If two predictors are correlated they contain redundant information.
  - How does a model know which predictor should get the credit?
- Detecting collinearity between two variables is easy: just calculate the correlation coefficients

When do group 1 methods start to fail?

Pearson and spearman don't help if have correlations between three or more variables

[Multi]Collinearity

- Multi-collinearity: complex correlational structures can exist among 3 or more variables.
  - Pearson/Spearman correlation coefficient is only for 2 variables.
  - Multicollinearity is hard to detect.
  - It causes 'unstable' coefficients: coefficients can change drastically when one observation is removed.

Key Concepts

- What makes a model linear? Linearity in the parameters.
- Categorical and continuous predictors.
- Key assumptions of general linear models.
  - When can they fail?
- Classes of Group 1 models

Correlated Predictors: Collinearity

Correlated Predicctors

Whitebark Pine: Background

Richard Sniezko, US Forest Service - Forest Service Dorena lab

The whitebark pine, *Pinus albicaulis* is a high-altitude tree that grows in montane habitats in Western North America.

Whitebark Pine: Modeling

Warmer winters in recent decades are associated with many plant and animal species shifting their ranges to higher altitudes.

The seeds of whitebark pine is an important food source for many animals, including black bears.

Plant pests, including the Mountain Pine Beetle *Dendroctonus ponderosae* are also shifting their ranges, making novel (and susceptible) hosts available.

There is great interest in understanding, and predicting, how whitebark pine growth varies with altitude and temperature.

We could probably learn a lot by creating a regression model of pine growth predicted by average winter temperatures and altitude!

Whitebark Pine: Data

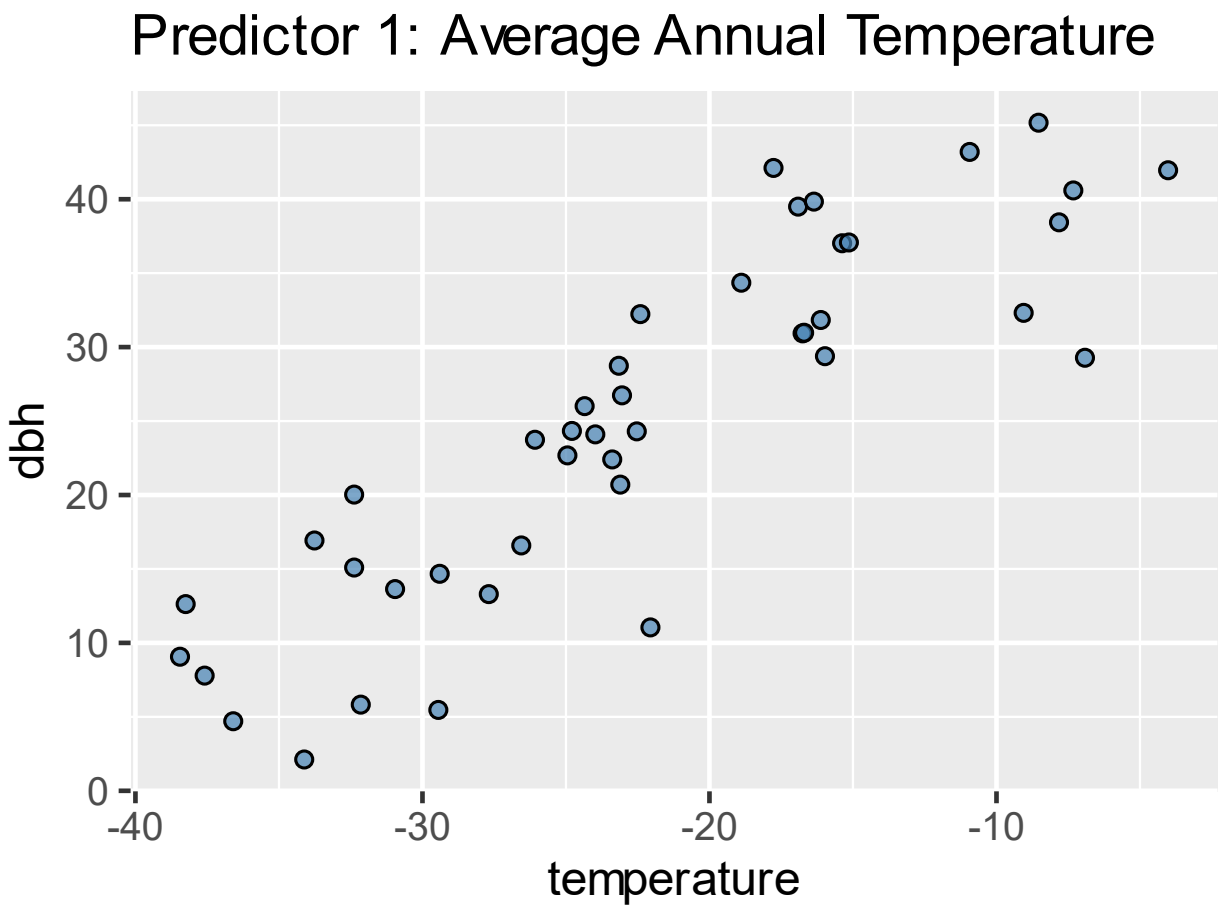
Suppose you have a dataset containing information about the size of individual trees `dbh`.

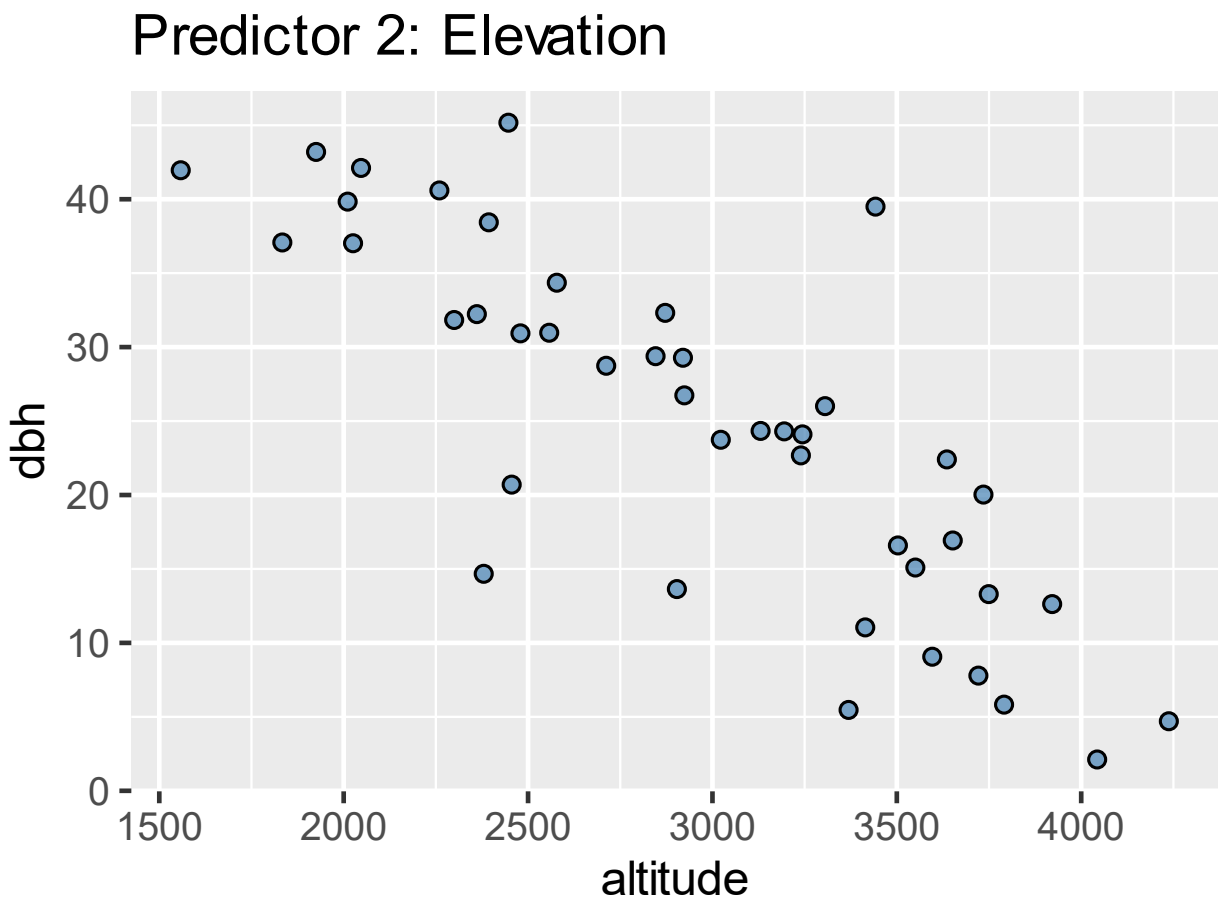
For each tree, you also know the altitude at which it grows, the man annual precipitation, and the mean annual temperature:

Do you think altitude and temperature are related?

```
head(dat_whitebark)
  X rainfall altitude      dbh
1 1 20.44596 2493.433 47.08338
2 2 33.23555 2301.189 52.09458
3 3 31.15216 2328.082 51.69528
4 4 25.24166 2390.111 49.53393
5 5 29.35912 2227.444 51.96837
6 6 28.51941 2421.535 49.85675
```

Whitebark Pine: Scatterplots



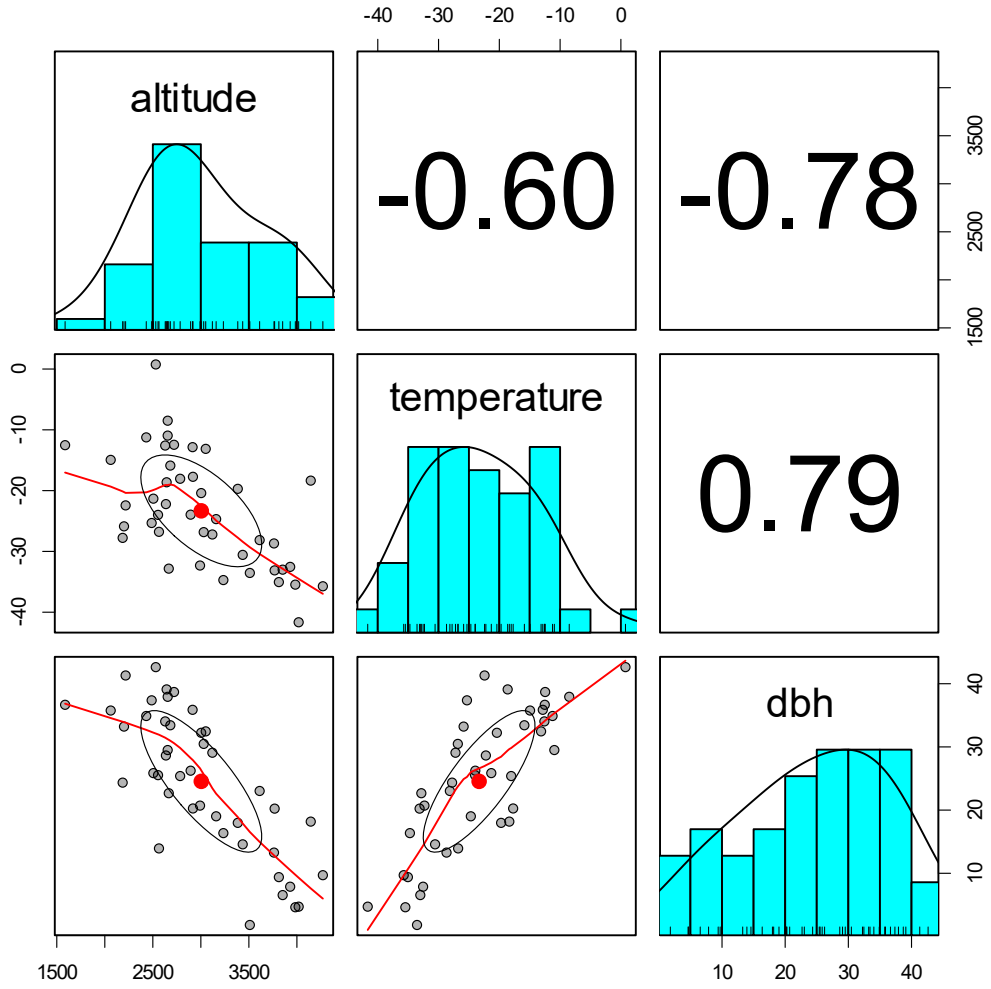


Whitebark Pine: Correlated Predictors

Dbh = stems of the trees
Mean annual temperature, rainfall, altitude
Temp and altitude may correlate
Correlation of -.60 > strong correlation
Tenperature of 0.8 > strong correlation to dbh
And altitudee -0.8 correlation

We know that they grow larger in warmer areas and at lower altitudes.

We can also see that altitude and temperature are strongly correlated in this **pair plot**:



Whitebark Pine: Simple Models

Fit two linear regression models
> look at r squared and coefficients
when mean temperature 0, intercept at 47

```
Call:
lm(formula = dbh ~ temperature, data =
dat_whitebark_collinear)


Residuals:
    Min      1Q  Median      3Q     Max
-14.014  -3.183   1.095   3.618  10.786


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.08033    2.05861   22.87  < 2e-16 ***
temperature  1.02641    0.08136   12.62 1.06e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1


Residual standard error: 5.881 on 41 degrees of freedom
Multiple R-squared:  0.7952,    Adjusted R-squared:
0.7902
F-statistic: 159.2 on 1 and 41 DF,  p-value: 1.057e-15
```

```
Call:
lm(formula = dbh ~ altitude, data =
dat_whitebark_collinear)


Residuals:
    Min      1Q  Median      3Q     Max
-13.457  -5.829  -2.056   3.863  18.670


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.883870   6.053708  11.709 1.17e-14 ***
altitude    -0.015641   0.001964  -7.963 7.46e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1


Residual standard error: 8.143 on 41 degrees of freedom
Multiple R-squared:  0.6073,    Adjusted R-squared:
0.5977
F-statistic: 63.41 on 1 and 41 DF,  p-value: 7.462e-10
```

Correlated Predictors: Multiple Regression

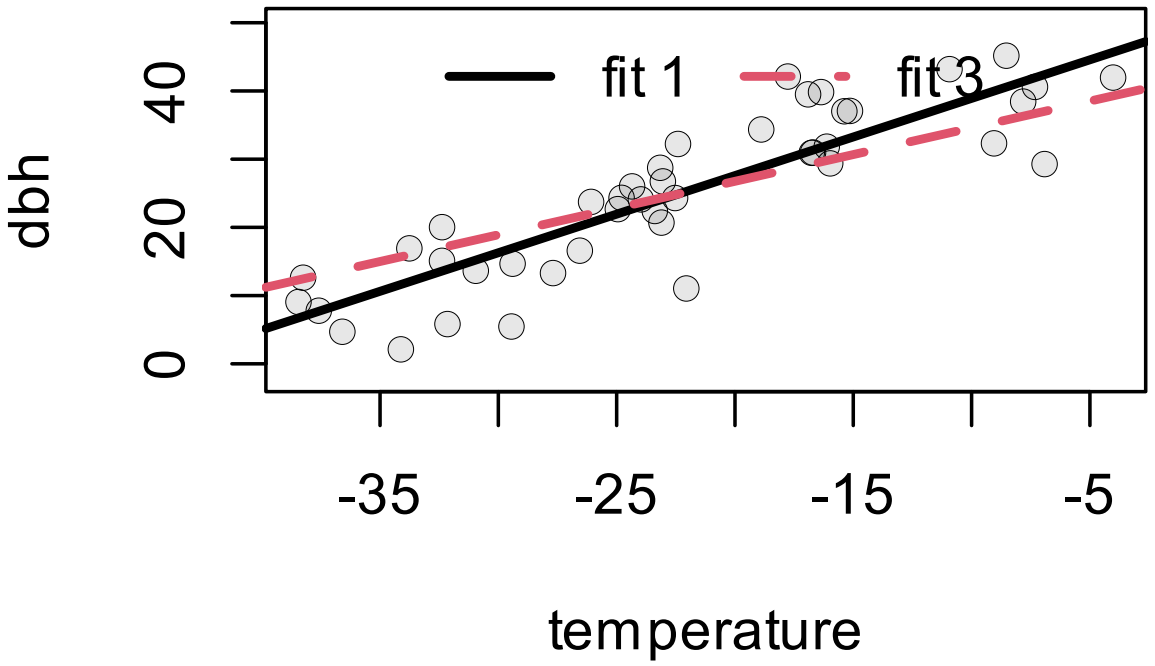
```
Call:
lm(formula = dbh ~ altitude + temperature, data = dat_whitebark_collinear)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1944  -3.6237   0.4966   3.9435   9.1340

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.918987   4.301265  13.698  < 2e-16 ***
altitude    -0.005705   0.001865  -3.059  0.00395 **
temperature  0.790577   0.106972   7.390 5.39e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.36 on 40 degrees of freedom
Multiple R-squared:  0.834,    Adjusted R-squared:  0.8257
F-statistic: 100.5 on 2 and 40 DF,  p-value: 2.527e-16
```

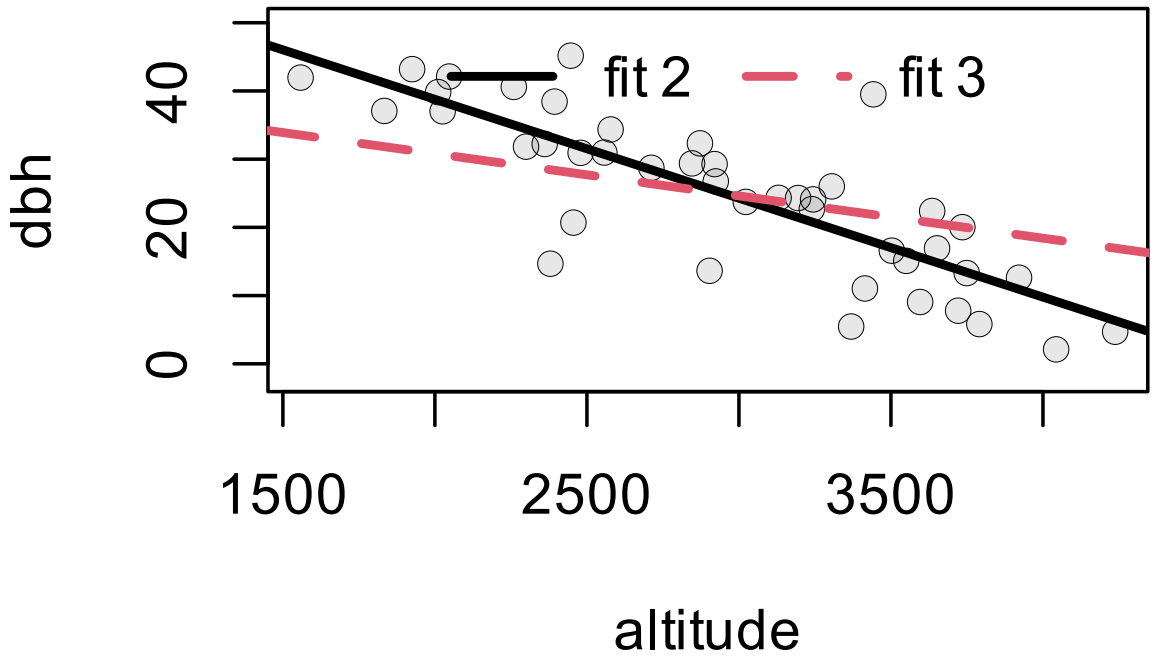
R squared at 80.
When higher up >slope > get smaller trees > -0.005705
Always use adjusted R square > takes into account when we have multiple predictors > takes into account df

Slope altitude > not different if have another predictor > see no relationships
Temperature > Went from single model 1 in double 0.8 > change coefficients > hint that have collinearity

Correlated Predictors: What happened?

What happened?
don't want them to change much > when include both they don't fit the data well
Problems: collinearity > when don't fit the line
Collinearity > can't distinguish which predictor is correlated with outcome



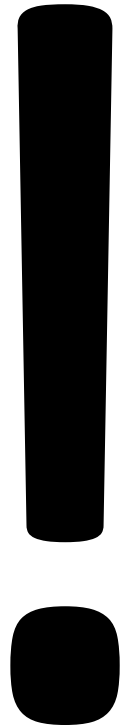


- The coefficients changed. They are now obviously wrong!
- Significance changed: both are still significant, but p-values are higher.

This seems very weird.

Our 2-predictor model performed well for the fish/area/pesticide model.

Why did we have a problem?



Correlated Predictors: Collinearity

Collinearity: when two or more predictors are highly correlated with each other

- Highly correlated predictors contain the **same information**.
- Since they contain the same info, the model can't determine which variable to attribute the effect to!
- The mathematical reasoning is that correlated predictors cause the **design matrix** to be less than **full rank**.
- Don't worry if you don't know what this means, it's not essential for understanding the problem.

What to do?

- Examine a `pair plot`. Base R has the `pairs()` function. Package `psych` has a nice function called "pairs.panels()"
- Remove one of the highly correlated predictors.
- Check for variance inflation using `vif()`

Correlated:
When correlated predictors > reflecting the same information
What to do?
> drop one if highly correlated predictors
> choose predictor that is more highly correlated
> what are the R squared values > temperature > higher R square = higher correlation

Model Coefficients and the ANOVA Table

What's in This Section?

Take-Home Concepts

- Interpreting model coefficient tables for categorical variables
- Interpreting model coefficient tables for continuous variables
- Interpreting the ANOVA table
- Intro to dummy variables

Group 1 model interpretation

Models are linear in parameters
> if increase in predictor we expect increase in response

Group 1 models are *linear in the parameters*

This makes the interpretation of model terms *relatively* easy.

- But note, there is still lots of complexity especially when we mix continuous and categorical terms and interaction terms.

Recall the basic equation:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon$$

- When all of the predictor variables have a value of zero, we expect y to have a value of α, on average.
- For every 1-unit change in x_1 we expect a β_1-unit change in y, on average.

Group 1 model summary presentations

Table of model coefficients model summary.

- This table tells us the strength of effects of predictors, overall model significance test

Model coefficient table > strength of association
If significant we need other table > ANOVA table
> shows which factor strongest influence

ANOVA table.

- This table shows the model variability attributed to each factor, factor-specific significance tests

Group 1 model interpretation

Model coefficient
> get intercept term
Slope > one for each of predictors

Model Coefficient Summary

Intercept: What is the value of the response when the predictor has value zero?

Slope: What is the change in the response with each unit change in the predictor?

Standard Errors: shape of sampling distribution

F-test: overall model significance test

Degree of freedom
> associated with each factor of variable
Sum of Squares > give info on variability
> sense of how much of total variability is explained by each factor
> how much variability explained by factor

ANOVA Table

Degrees of freedom: Reflects the number of samples, number of factor levels, number of individuals per factor level etc.

Sum of squares: Reflects the total squared deviation from the mean explained by a source.

Mean squares: Mean Square due to a source (per DF)

F tests: Test for ratio of variability explained by a particular predictor variable

F-test > compare to normal distributions > f statistic is a ratio of two variances
> tell if distributions is very different from another

ANOVA table vs. model coefficient table

Model coefficient table tells you	ANOVA table tells you
1. Intercept and slope coefficients 2. Overall model significance test, correlation test	1. Variability explained by each factor in the model 2. Significance tests for each factor separately

MCT > strength of relationship
ANOVA > tells if significant or not > but doesn't tell strength of relation

1-way ANOVA

When we have a continuous response and a single categorical predictor with 2 levels we can use a t-test.

What if there are 3 or more levels?

- The t-test is not enough.
- Analysis of Variance is a generalization of the t-test for 3 or more groups.

Single continuous response , single caterogrial predictor > three or more levels

Model Coefficient Tables: Dummy Variables

When you fit a model using a categorical predictor with n levels, the algorithm first detects all of the factor levels present in the data, then creates a set of n - 1 *dummy variables*.

- The dummy variables allow the model-building process to treat each factor level as if it were a separate, numerical predictor that can take on only values of zero or one.

species	speciesGentoo	speciesChinstrap
Adelie	0	0
Gentoo	1	0
Chinstrap	0	1

Create dummy variable
Three species
> Adele is default > set the others variables to zero
If have three levels > two dummy variables
Only one can be 1 the others have to be set to zero!!

Model Coefficient Tables: Interpretation for Categorical Predictors

Since each factor level is treated as a predictor variable, there will be slope parameters for each.

When R builds a model, it selects one of the factor levels to serve as the *base case*.

- When the model contains only categorical variables, the base case is analogous to the *intercept* term in a model, i.e. the α.

It'll be easier to understand with an example.

1-way ANOVA: Palmer Penguins

The procedure for conducting an ANOVA in R is:

- Create a linear model fit with lm().
- Use anova() to perform the Analysis of Variance and print the ANOVA table.

Recall that ANOVA is really a just a different way of looking at a linear model.

- To better understand the relationship, we'll focus on the model coefficient table first:

```
lm(
  formula = body_mass_g ~ species,
  data = penguins)
```

```
Call:
lm(formula = body_mass_g ~ species,
 data = penguins)

Coefficients:
      (Intercept)
          3700.66
speciesChinstrap
            32.43
   speciesGentoo
          1375.35
```

When continuous ignore intercept
> when categorical intercept important > represent base case > here Adelie
Intercept term > base case > Adelie has 3700 grams

Chintrap > that's the amount of additional weight a chinstrap would have
Adding to intercept value is how dummy variables work
Accomodate categorical predictor into linear models

Factor Base Cases

Base case >> is the intercept
Arranged by Alphabet > A chosen

There are slopes for Chinstrap and Gentoo, but where is the Adelie coefficient?

- Recall: the *base case* is the intercept in a 1-way ANOVA.

R assigned "Adelie" to be the base case.

- Notice how R formats the factor-level coefficient names:
  - the variable name prepended to the factor level.

Interpreting the Coefficient Table

```
Call:
lm(formula = body_mass_g ~ species,
 data = penguins)

Coefficients:
     (Intercept)
         3700.66
speciesChinstrap
           32.43
   speciesGentoo
         1375.35
```

- Mean Adelie penguin mass is 3700 grams
- Mean Chinstrap penguin mass is 3700 + 32 grams
- Mean Gentoo penguin mass is 3700 + 1375 grams

Everything is relative to the base case!

Interpreting the Coefficient Table

```
Call:
lm(formula = body_mass_g ~ species,
 data = penguins)

Coefficients:
      (Intercept)
          3700.66
speciesChinstrap
            32.43
   speciesGentoo
          1375.35
```

- The intercept is 3700 grams: Adelie penguins weigh 3700g, on average
- The regression slope for Chinstrap is 32 grams per unit.
  - Adding one ‘Chinstrap penguin unit’ increases the penguin mass by 32 grams, on average.
- The regression slope for Gentoo slope 1375 grams
  - Adding one ‘Gentoo penguin unit’ increases the penguin mass by 1375 grams, on average.

Everything is relative to the base case!

Interpreting the Coefficient Table

```
Call:
lm(formula = body_mass_g ~ species,
 data = penguins)

Coefficients:
     (Intercept)
         3700.66
speciesChinstrap
           32.43
   speciesGentoo
         1375.35
```

We can obtain the mean masses of each species from the model coefficient table.

- Mean Chinstrap penguin mass
  - $3733 = 3701 + 1 \times 32 + 0 \times 1375$
- Mean Gentoo penguin mass:
  - $5076 = 3701 + 0 \times 32 + 1 \times 1375$

Dummy Variables

If we consider x_{chin} a dummy variable which is equal to 1 if the observation is a Chinstrap penguin and 0 otherwise, and likewise for x_{gentoo} we could write the regression equation symbolically as:

$$y_i = \alpha_{adelie} + \beta_{chin} \times x_{chin} + \beta_{gentoo} \times x_{gentoo}$$

What would the coefficient table and equation look like if Chinstrap penguins were lighter than Adelie penguins?

1-way ANOVA: ANOVA Table

We have examined the model coefficients and calculated the group means.

- The masses seem pretty different, but how could we assess the ANOVA *alternative* hypothesis?
  - "The body masses of penguins for *at least one* species are different from the masses of the other species"

```
## Analysis of Variance Table
##
## Response: body_mass_g
##             Df    Sum Sq  Mean Sq F value    Pr(>F)
## species      2 146864214 73432107  343.63 < 2.2e-16 ***
## Residuals 339  72443483   213698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1-way ANOVA: Model Coefficient Table

ANOVA table can say how good species fit
Single coefficient for species instead gentoo …
Sum of squares > amount of variability explained by the factor
F test > ration of variances
Higher f value indicate that one factor is significant in the model
> doesn't give info on each species > need both tables!!!

What can we learn from the model coefficient table?

The *intercept* and *speciesGentoo* coefficients have low p-values, but that's not exactly what we wanted to know!

- We wanted to know about the penguin species *in general*.

1-way ANOVA: ANOVA Table

The ANOVA table gives us a clue

```
Response: body_mass_g
           Df    Sum Sq  Mean Sq F value    Pr(>F)
species     2 146864214 73432107  343.63 < 2.2e-16 ***
Residuals 339  72443483   213698
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Model Coefficients and ANOVA Provide Complementary Information

We'll cover model coefficient interpretation, and the ANOVA table details in greater depth, but for now you should notice:

- Model slope/intercept coefficients: there is one coefficient for each *factor level* of a *categorical predictor*.
- The intercept coefficient corresponds to the *base case*.
- Model coefficient table characterizes the strength and significance of individual intercept and slope coefficients.
    - It *does not* tell us about the overall significance of the categorical predictor.
- The ANOVA table evaluates the ANOVA null hypothesis.
    - It *does not* tell us *which factor levels* are different
    - The two tables each provide part of the picture.

Neither the model coefficient table nor the ANOVA table tell us if a particular pair of factor levels are *significantly* different form one another!

Model Coefficients and ANOVA Provide Complementary Information

Neither the model coefficient table nor the ANOVA table tell us whether a particular pair of factor levels are *significantly* different form one another!

- This is the realm of post-hoc testing.
  - Post-hoc testing is an analysis you perform after (post) you perform the initial analysis (hoc).
- The Tukey Honest Significant Difference is a common post-hoc method.`

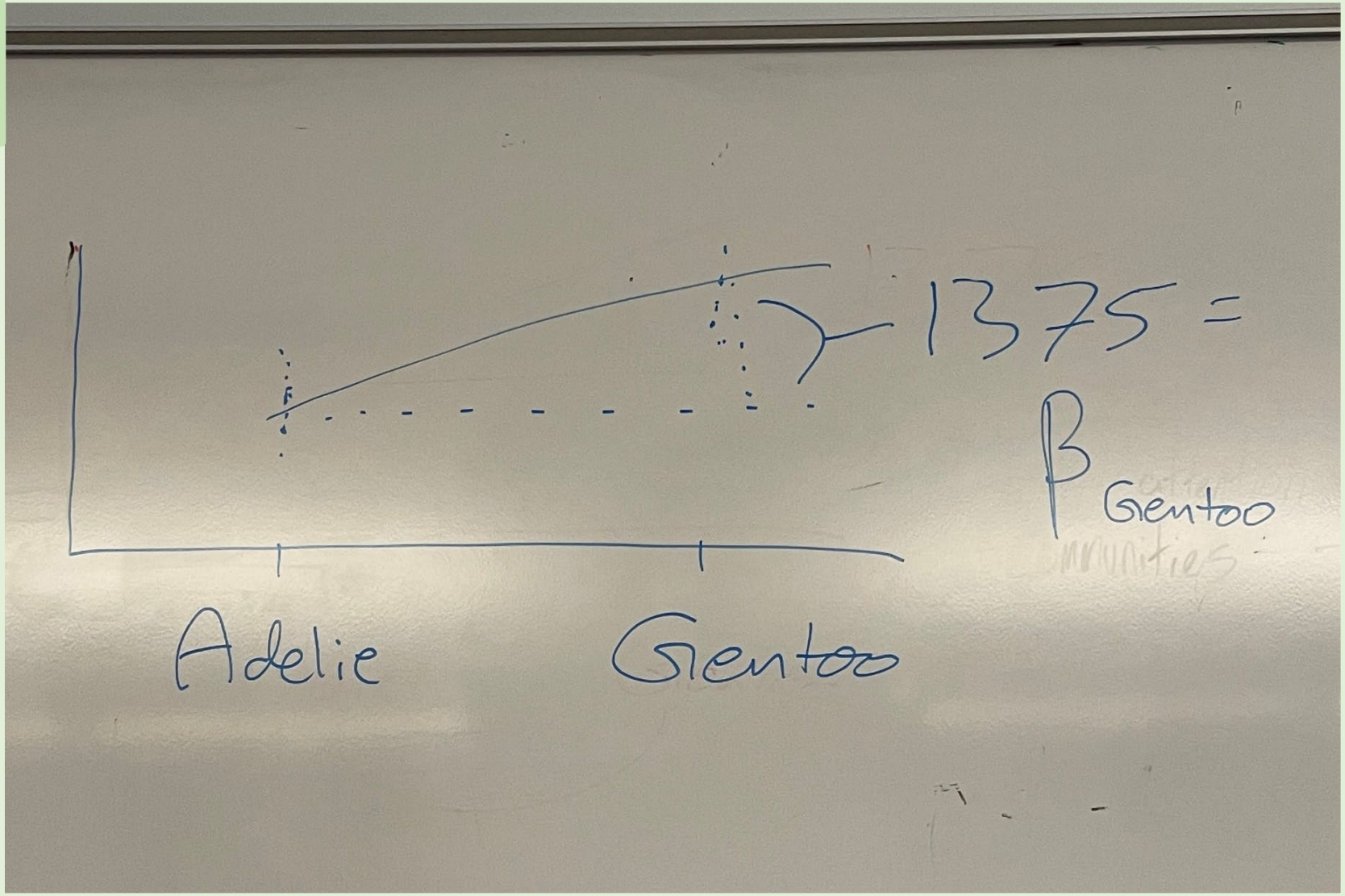
Key Concepts

- Interpreting model coefficient tables for categorical variables
- Interpreting model coefficient tables for continuous variables
- Interpreting the ANOVA table
- Intro to dummy variables

Board Model Art

Dummy Variable Interpretation

- Predictor variable adds one unit of Gentoo
- The coefficient is 1375
- One-unit increase in Gentoo corresponds to a 1375-unit increase in body mass



In-Class t-tests