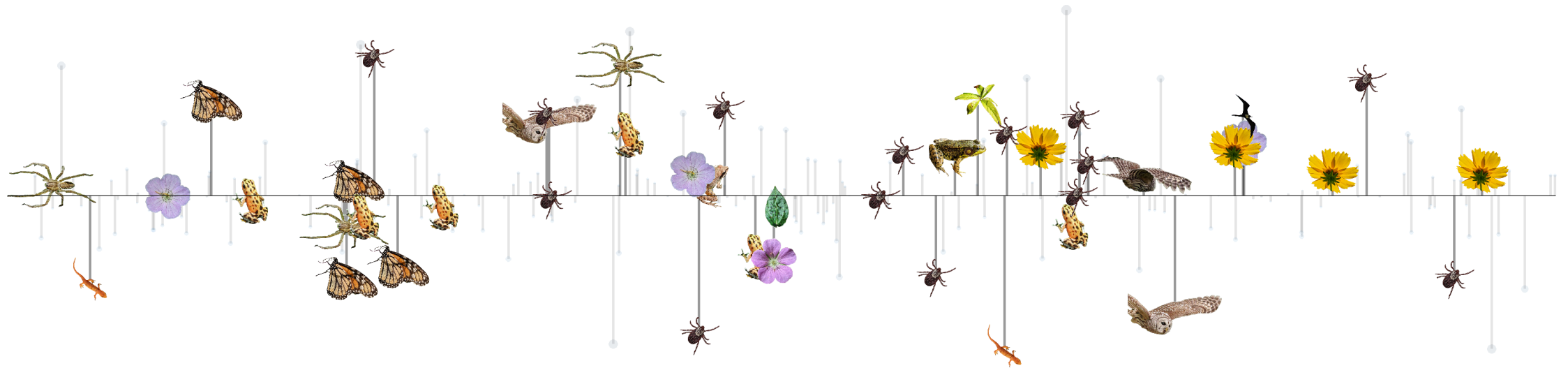


# Intro to Quantitative Ecology

## UMass Amherst – Michael France Nelson

Deck 12B – ANCOVA, Logistic Regression, and Regression Equations



# Analysis of Covariance

Mixing Continuous and Categorical Predictors!

# It's not as complicated as it sounds

We've seen all the components already:

- Slopes
- Intercepts
- Base cases
- Interactions
- Dummy variables
  
- One of the biggest challenges is understanding the base case in an ANCOVA

# Base Case Interpretation

In an ANOVA-style model, the base case was just a group mean.

- With ANCOVA, it's not so simple.
- The base case is no longer a simple group mean: it's the mean of base case when all continuous predictors are zero.

“What is the mean bill length of an Adelie penguin with zero body mass?”

- Base cases in ANCOVA are usually nonsensical (just like intercepts usually are).

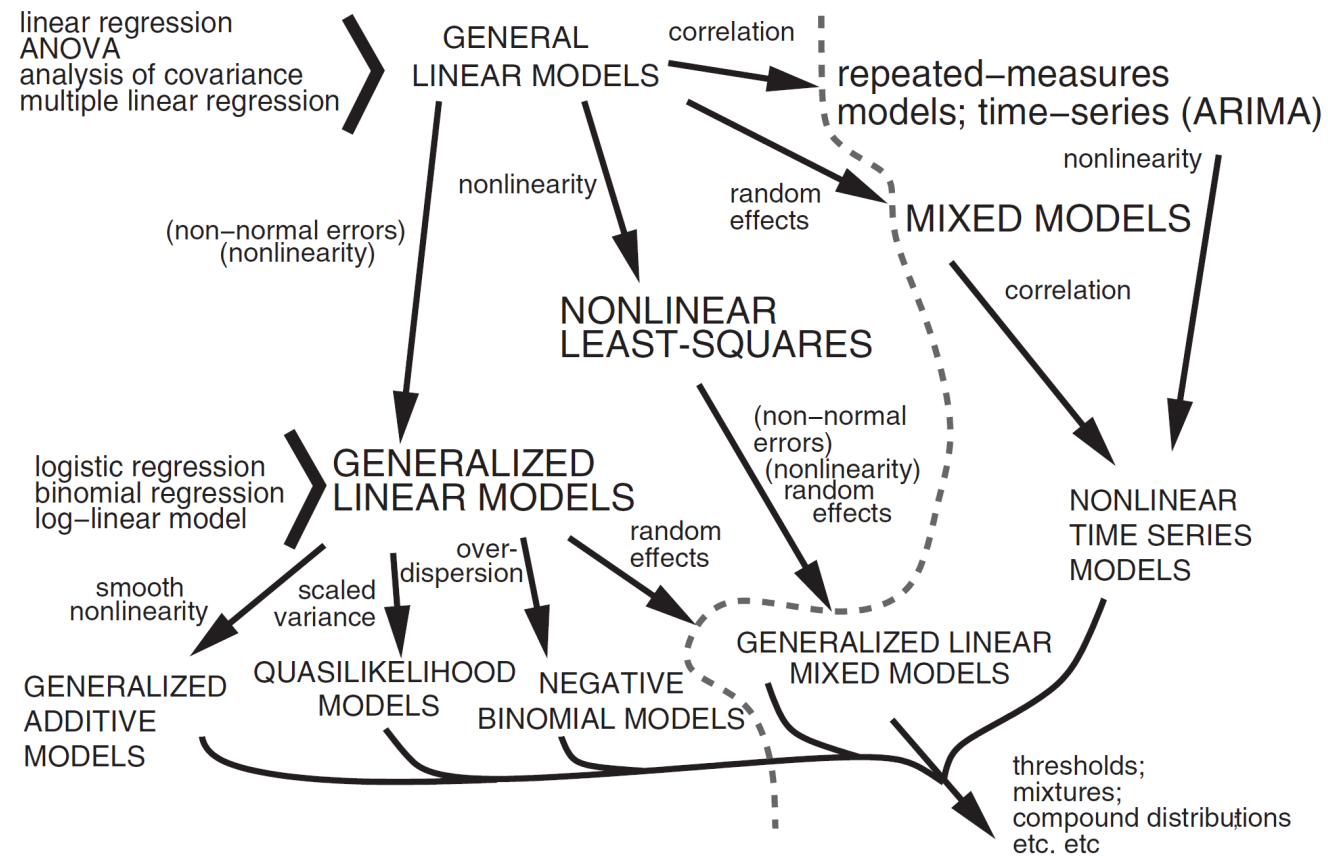
“What is the body mass of a female Adelie penguin that is zero years old?”

“What is the expected stopping distance for a vehicle with six cylinders, that weighs 0 kilograms, and is travelling at 0 miles per hour?”

# Logistic Regression

Welcome to the exciting world of Generalized Linear Models!

# The Constellation of Models



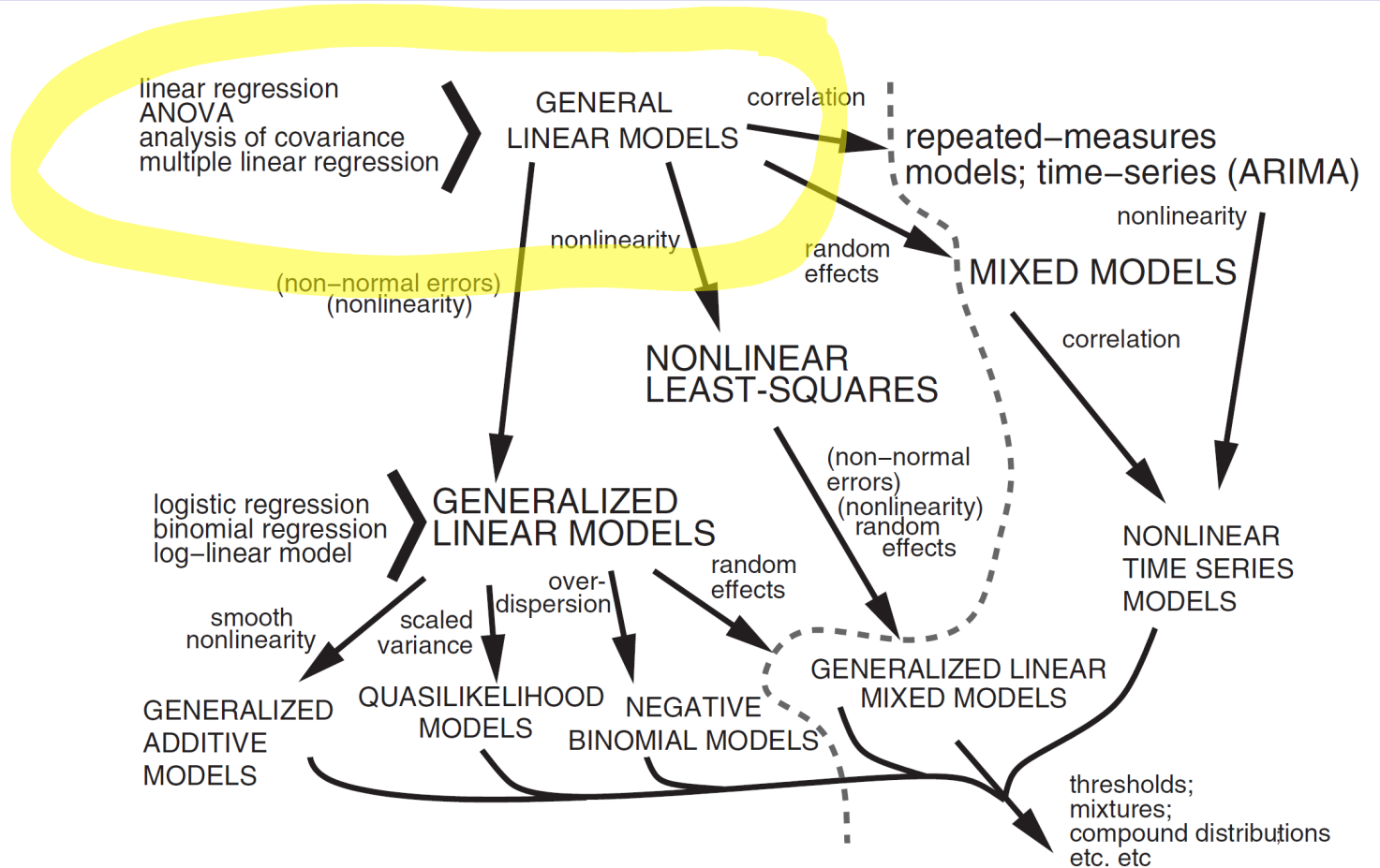
Bolker: Ecological Models and Data in R, Figure 9.2

# The Constellation

We've been exploring the world of General Linear Models.

It's a great place!

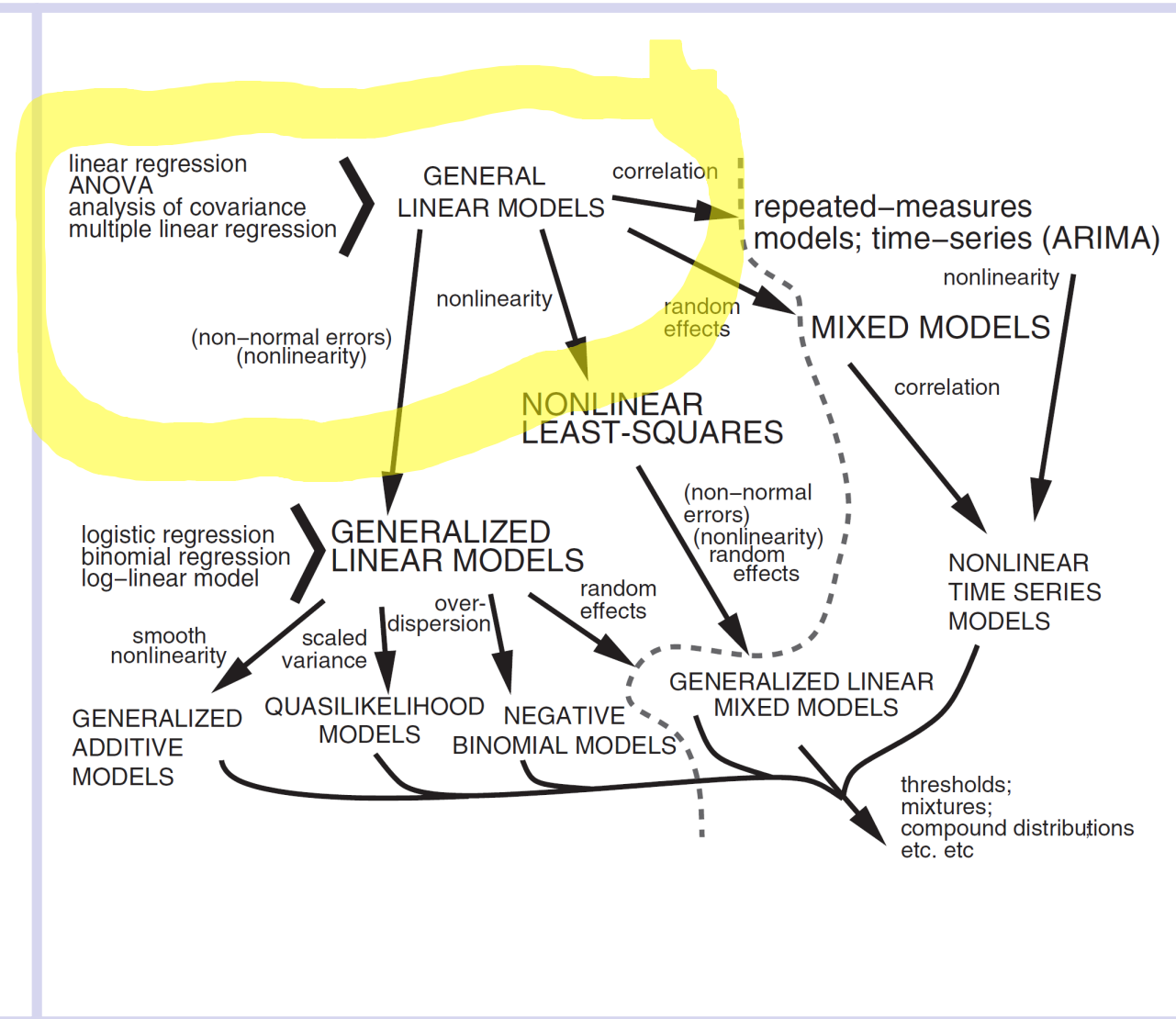
We've learned lots of powerful inference techniques. I call these the **Group 1 Models**.



Bolker: Ecological Models and Data in R, Figure 9.2

# Group 1: General Linear Methods

- Single continuous response variable
- One or more predictor variables
  - They may be continuous or categorical
- Deterministic model must be *linear in the parameters*.
- Stochastic model is the Normal distribution.





# Four key assumptions

Group 1 imposes 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 1 requires that our models be *linear in the parameters* and have a response on a continuous scale.

What does linear in the parameters mean?

- It restricts the types of mathematical relationships we can include in our regression equation.

The different Group 2 models can deal with different violations of these assumptions and requirements.

# Group 1: terms and coefficients

## **response: Y**

- Also called the dependent variable
- This must be continuous (or at least numerical)

## **predictor(s): X**

- Also called the independent variable(s)
- These can be continuous or categorical

## **intercept(s): $\alpha$ . Sometimes symbolized as $\beta_0$**

- This is the expected value of the response when all of the predictors are equal to zero.
- These are often nonsensical; we consider them *tuning parameters*.

## **slope(s): $\beta_i$ :**

- The regression slope is the rate of change in the response variable for each one-unit increase in the predictor variable.

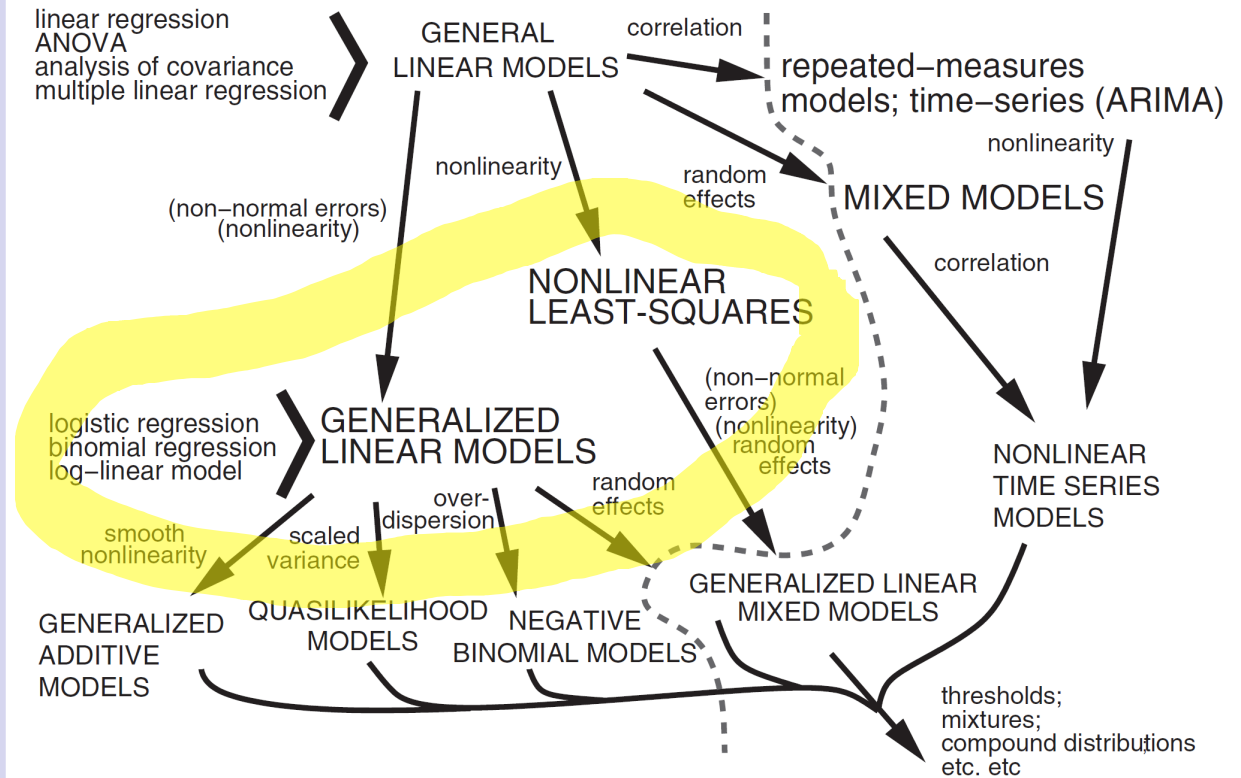
# Group 2: Violations of Assumptions

Recall the key assumptions:

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

Violations of some of these assumptions are more difficult to deal with than others!

Group 2 allows for violations of the normality and homoskedasticity assumptions.





# Group 2: *Generalized* Linear Models

## GLMS and Logistic Regression

GLMs *generalize* general linear models by using a *linearizing link function* that can accommodate certain common types of non-normal errors.

- GLMs work with stochastic models that can be specified by a *exponential family* distribution.
  - Many common distributions belong to this family.

Logistic Regression is a specific type of



# The Binomial and Bernoulli Distributions

## Binomial and Bernoulli: Successes and Failures

The binomial distribution is an important **discrete distribution**.

The binomial distribution describes binary outcomes, often called success and failure (or presence/absence).

The binomial distribution is characterized by 2 parameters:

- $n$ : the number of trials
- $p$ : the probability of success in each trial



# The Simplest Distribution?

One of the easiest distributions to understand is the *Bernoulli Distribution*.

It's a special case of the *binomial distribution*

- Its sample space has only two elements.
- Realizations of a *Bernoulli process* produces a single *binary* outcome.
- It has one parameter: the probability of *success*.

- The Bernoulli distribution is just a binomial distribution when  $n = 1$
- A realization of the *Bernoulli process* is called a **trial**

# Binomial Process: Independent Coin Flips


- Think of each flip as a junction in a tree.
- The first flip has two branches:





# Binomial Process: Independent Coin Flips

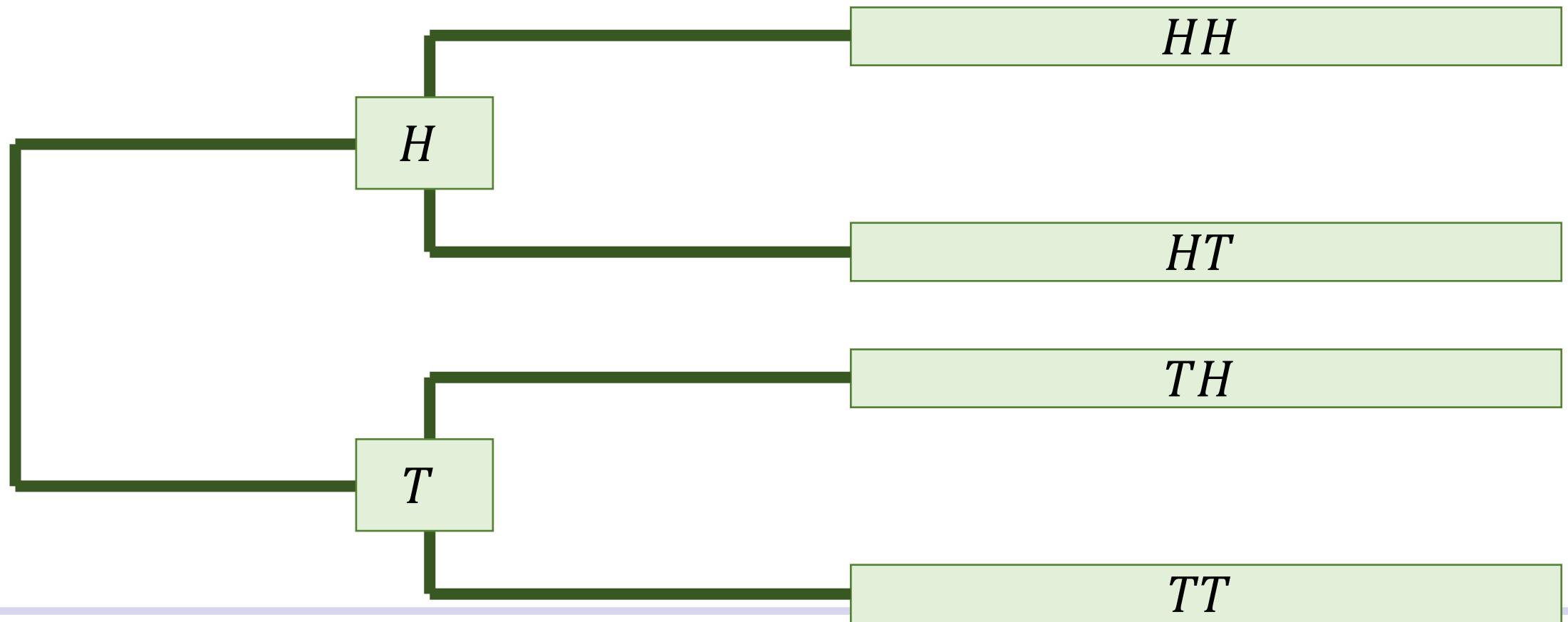
- If the probabilities are equal:


$$\Pr(H) = 0.5$$

$$\Pr(T) = 0.5$$

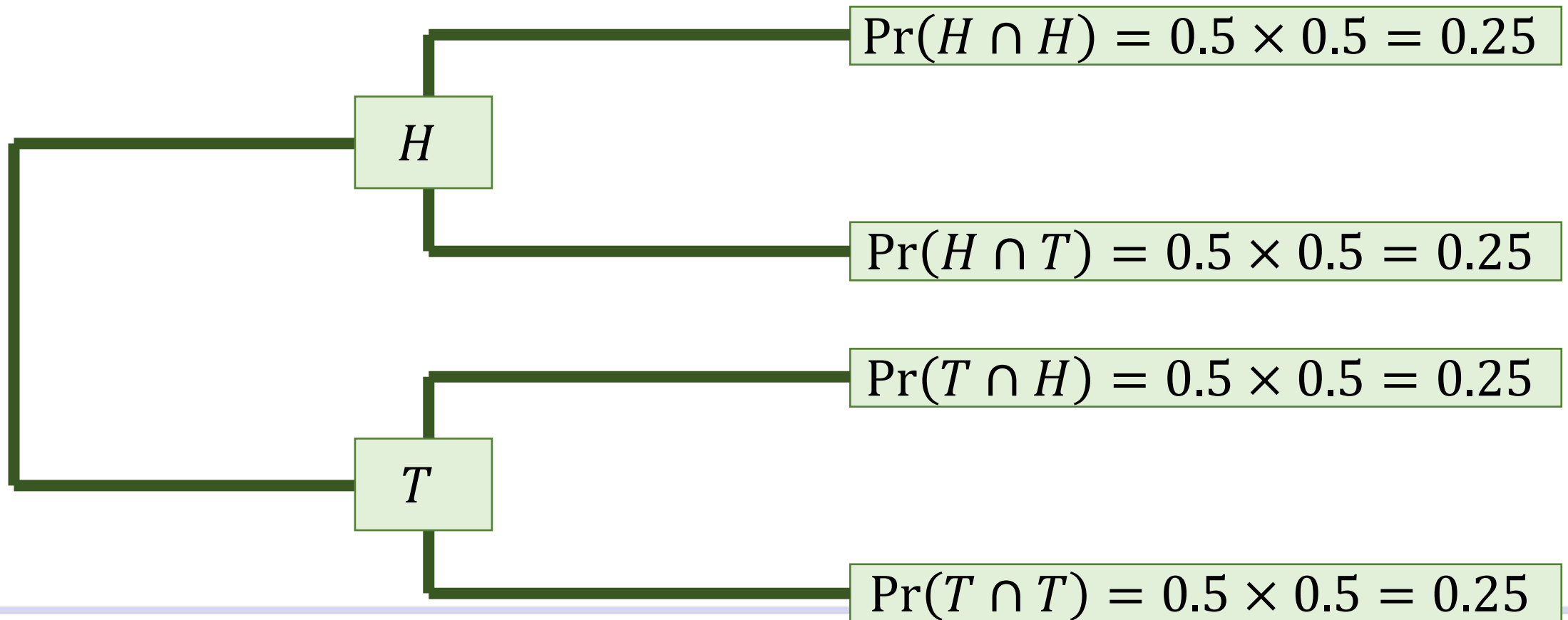
# Binomial Process: Independent Coin Flips

- Each of those branches has two branches:



# Binomial Process: Independent Coin Flips

- Probabilities of independent events are multiplied:



# A Binomial Example: Presence and Absence

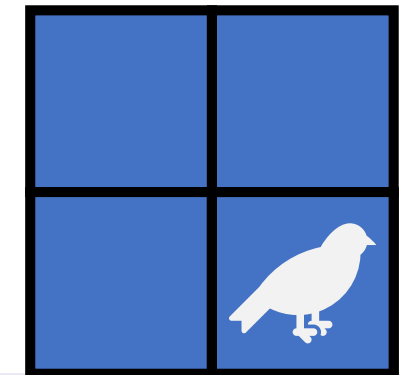
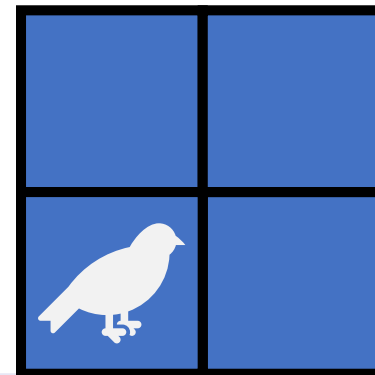
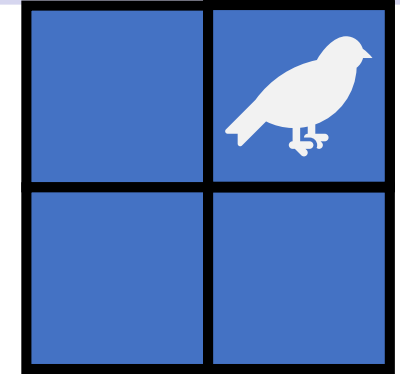
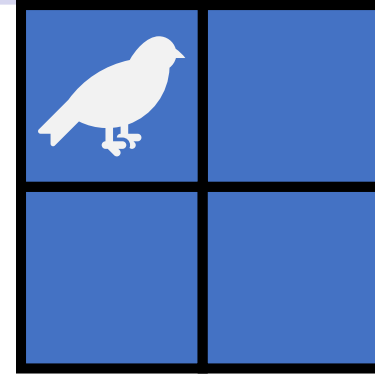
## The Scenario

You have a set of four experimental plots.

You visit each plot and record the presence or absence of a particular bird: the brown creeper.

- This is a binary outcome!
- Sounds like a binomial distribution problem!
- Observation of each plot is a realization of a Bernoulli process

Four ways in which you can observe one presence and 3 absences:



# Modeling Presence/Absence: Logistic Regression

## How can we model the scenario?

**What are some potential predictors?**

**What is the response variable?**

# Modeling Presence/Absence: Logistic Regression

## How can we model the scenario?

**What are some potential predictors?**

- Percent coniferous forest cover?
- Food abundance?
- Presence/absence of water source?

**What is the response variable?**

- Presence/Absence of brown creeper –  
Is there a problem here?

# What would the regression equation look like?

- A potential equation:

$$\mathit{bird} = \alpha + \beta_1 \mathit{food} + \beta_2 \mathit{water}$$

- Any issues?

# What would the regression equation look like?

- A potential equation:

$$bird = \alpha + \beta_1 food + \beta_2 water$$

- Any issues?
  - 'Bird' is not a continuous quantity
  - 'Bird' is either present or absent – maybe we could code it as 0 and 1?
  - This is a linear equation... it can produce outputs less than 0 and greater than 1!
- Let's try another equation...



# What would the regression equation look like?

- A second potential equation:

$$\mathit{prob}(\mathit{bird}) = \alpha + \beta_1 \mathit{food} + \beta_2 \mathit{water}$$

- Any issues?
  - Prob(bird) is now a continuous quantity, but probabilities are bounded by 0 and 1.
  - This is a linear equation... it can produce outputs less than 0 and greater than 1!
- This equation was a little better, but it still doesn't quite work.
- What can we do?

# Let's Consider the Odds

Instead of modeling the probability of observation, we could consider the odds.

But what is the definition of the odds?

We could define the odds as the probability of success divided by the probability of failure. In symbols:

$$odds = \frac{\Pr(bird)}{\Pr(no\ bird)}$$

- If there is a 50% chance of observing the bird, I have a 1:1 odds:  $\frac{50\%}{50\%} = \frac{1}{1}$
- If there is a 25% chance of observing the bird, I have 1:3 odds:  $\frac{25\%}{75\%} = \frac{1}{3}$
- If there is a 75% chance of observing the bird, I have 3:1 odds:  $\frac{75\%}{25\%} = \frac{3}{1}$

# The Log of the Odds

The odds are just a ratio and ratios can be any nonnegative number, they're not bounded by 1.

This gets us even closer to the type of response we want

If we take the logarithm, we can then take on negative and positive numbers!

Logarithms can be mind-bending, but a simple intuition is:

1. Logarithms of large numbers are positive
2. Logarithms of small numbers are positive

# Finally, we've got an equation that works

- We can model the logarithm of the odds (the log-odds) as a linear function:

$$\text{Ln} \left( \frac{\text{prob}(\text{bird})}{\text{prob}(\text{no bird})} \right) = \alpha + \beta_1 \text{food} + \beta_2 \text{water}$$

- Or equivalently

$$\text{Ln}(\text{odds}) = \alpha + \beta_1 \text{food} + \beta_2 \text{water}$$

- The log-of-the-odds function is also known as a **logit** function.

# Logit Function

How does the logit function help us? It seems like we've strayed far away from our original goal of modeling presence and absence. Think about it this way:

1. Instead of directly modeling presence/absence, we can model the probability of presence.
2. Probability was bounded by 0 and 1, so we needed a way to increase the range: the odds ratio.
3. Odds ratios can't be negative, but if we apply a log-function we can get the range we need for a linear function!
4. Finally, we have a workable equation: model the log-odds with a linear function!

# Logistic Regression Wrap-Up: What did we learn?

Logistic regression allows us to use a linear equation to [indirectly] model presence/absence.

Generalized linear models (GLMs), of which logistic regression is one type, use a link function.

Logistic regression uses a logit function as a link.

- Logit is just the log-of-the-odds

But wait...what about the stochastic model?

- We've ignored the stochastic part of the model...we don't have time to go into the details here, but GLMs allow non-normally distributed errors.
- The errors for a logistic regression follow a Bernoulli distribution.

# Model Equations

Using Model Coefficients for Prediction

# Model Equations: Understanding and Prediction

## Recall two of our main modeling goals:

### Understanding

- Does distance to road affect vole population?
- Are male penguins heavier than females?
- Are there inter-species differences in sexual dimorphism?

### Prediction

- How many voles should I expect to observe 100 meters away from a road?
- How much does a male Adelie penguin weigh, on average?
- On average, what's the difference in body mass between male and female Chinstrap penguins?



# Model Coefficients: the Model Equation

- We can use model coefficients to reconstruct a model equation that we can use for prediction. For example, let's examine the coefficient table of a simple model of penguin bill length explained by body mass:

Coefficients:

	Estimate
(Intercept)	2.690e+01
body_mass_g	4.051e-03

$$\text{bill length} = 27 + 4e^{-3} \times \text{body mass}$$

# Prediction with the Model Equation

How can we use the model equation to predict bill length???

- Easy, just plug in the values of our predictor variable!
- Let's predict the bill length of a 1000g penguin:

$$\textit{bill length} = 27 + 4e^{-3} \times \textit{body mass}$$

$$\textit{bill length} = 27 + 4e^{-3} \times 1000$$

$$\textit{bill length} = 27 + 4e^{-3} \times 1000$$

$$\textit{bill length} = 27 + 4 = 31$$

# Predicting With R

- That worked, but it was kind of tedious...
- It would be fantastic if we could use R to do the calculation for us.
- Fortunately, we can use the `predict()` function!
  - Unfortunately, the syntax is a little finicky

The procedure:

1. Fit the linear model
2. Build a new data frame with the values you want to predict.
3. Use the `predict()` function on the new data.

# Step 1: Fit a Model

```
require(palmerpenguins)  
fit1 = lm(  
  bill_length_mm ~ body_mass_g,  
  data = penguins)
```

## Step 2: Dataframe of new data

```
new_dat = data.frame(  
  body_mass_g = 1000  
)
```

Things to note:

1. It has to be a data.frame object
2. It must have columns with the **same names** as the predictor variables in the model.

# Step 3: Apply the predict() function

```
> predict(fit1, new_dat)
      1
30.95029
```

**Success!**

Let's try a more complicated  
model!

# Fit the model

```
fit2 = lm(  
  bill_length_mm ~ body_mass_g + sex,  
  data = penguins)
```

Note the mix of continuous and categorical predictors. We're in ANCOVA territory!



# Examine the Model

- How can we interpret this model?
- What is the base case?
- Do male penguins have shorter bills?

Coefficients:

	Estimate
(Intercept)	2.791e+01
body_mass_g	3.674e-03
sexmale	1.247e+00

# Create the New Data

We'll predict the bill length for two 1-kg penguins: a male and a female:

```
new_dat2 = data.frame(  
  body_mass_g = c(1000, 1000),  
  sex = c("female", "male")  
)
```

# Predict!

```
> predict(fit2, new_dat2)
      1      2
31.58187 32.82901
```