Linear regression

Coding 4 Conservation
June 29, 2022
Sophia Horigan
Using resources from: Rajaonarifara Elinambinia, Andres Garchitorena

Outline

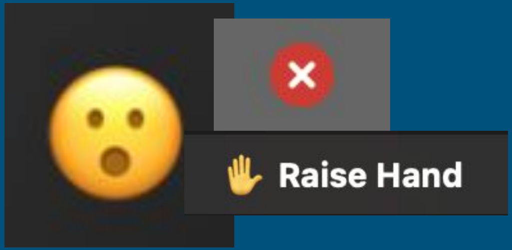
1. Lecture
    a. What is linear regression?
    b. How does linear regression work?
    c. What kinds of linear regression are there?
    d. When do I use linear regression?

    ~ break ~

2. Tutorial
    a. How do I perform linear regression in R?
    b. How do I tell how well my linear regression worked?

Going well?

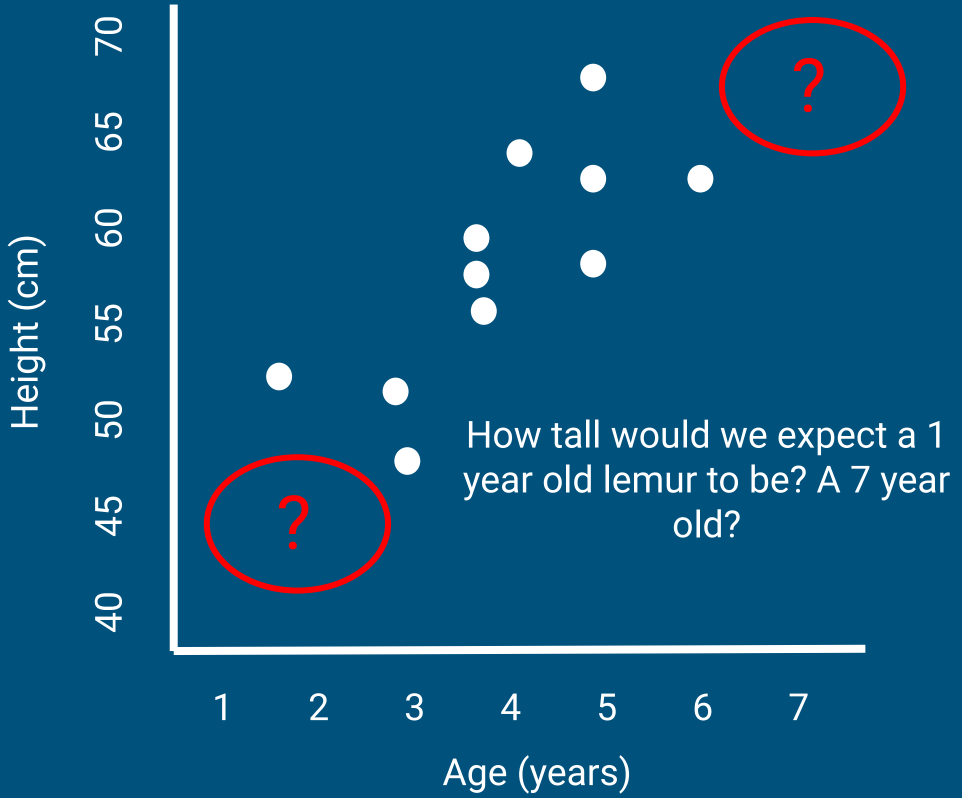
Not going so well?



Starting with a problem

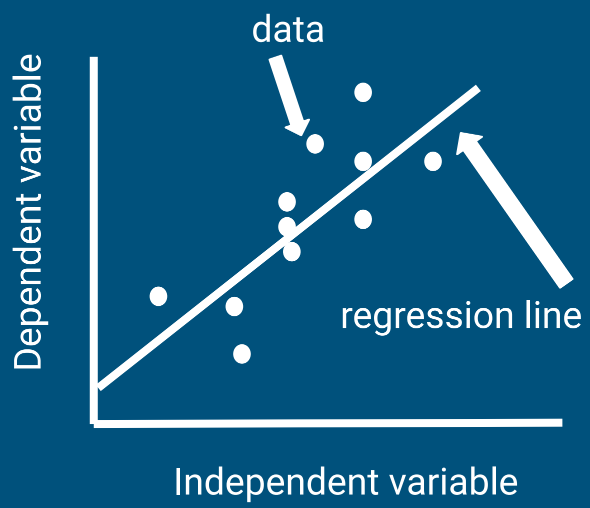
Lemur measurements:

- Age (years)
- Height (cm)



Linear regression basics

Linear regression allows us to implement a model to **predict** the impact of an **independent** variable on a **dependent** variable.



The same as
$Y = mx + b$

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Dependent variable

Y-intercept

slope

Independent variable

error

Linear regression allows us to **predict**

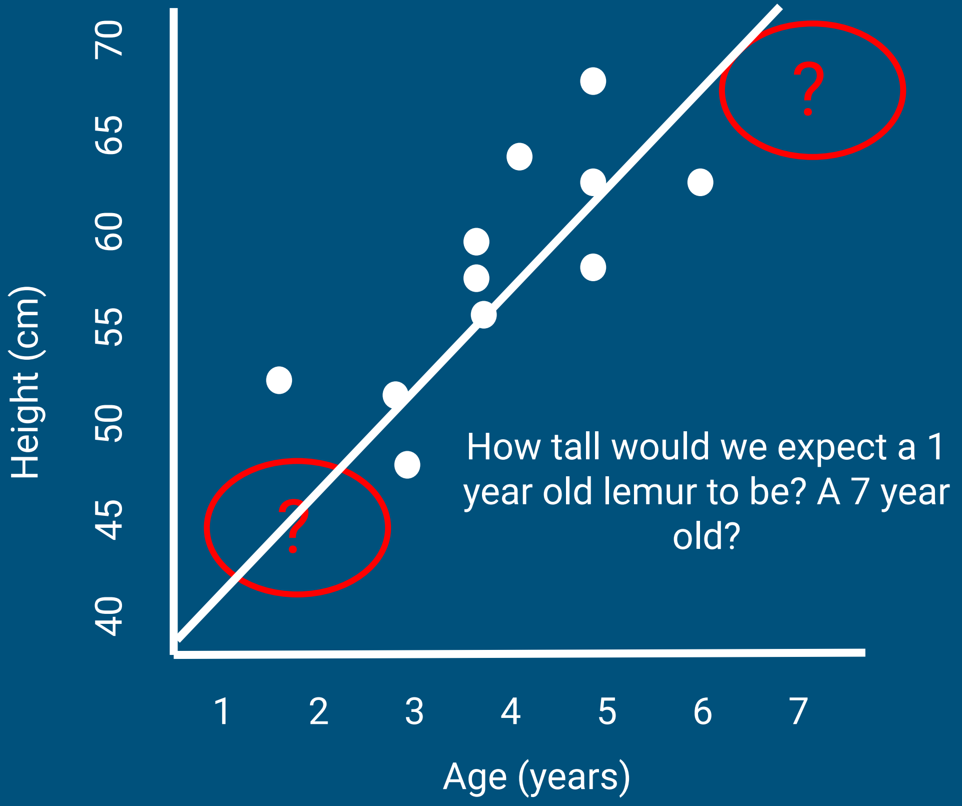
The "truth"

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Our data - a sample of the "truth"

$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

Height = 39 + 2.5*Age + ε

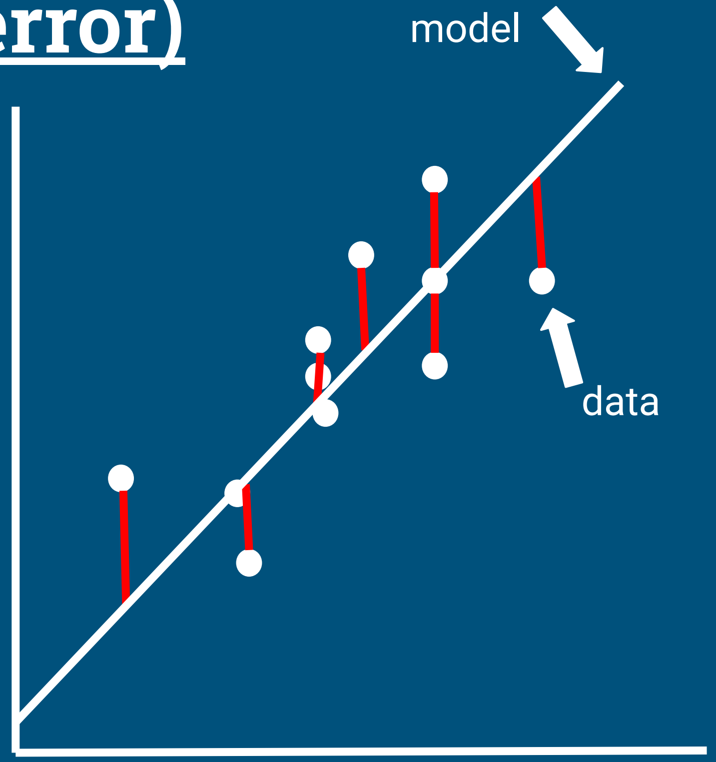


It's all about **residuals (error)**

Residual = difference between **predicted** and **observed**

model data

$$\varepsilon_i = Y_i - \hat{y}_i$$



SSE = sum of squared errors

The goal: **minimize** the squared residuals

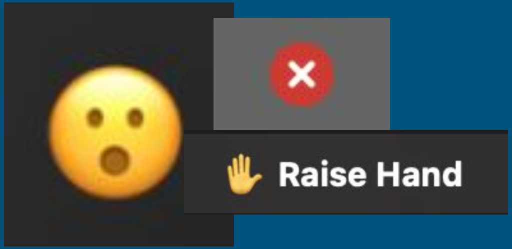
Summary and Check In

Linear regression allows us to implement a model to **predict** the impact of an **independent** variable on a **dependent** variable.

It does so by minimizing the error (residuals).

Going well?

Not going so well?



Types of Linear Regression

Univariate Linear Regression

↓

Multivariate Linear Regression

↓

Generalized Linear Regression

Each type of linear regression has its own **assumptions**.

Univariate Linear Regression

$$X \longrightarrow Y$$

Continuous or near continuous

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent variable

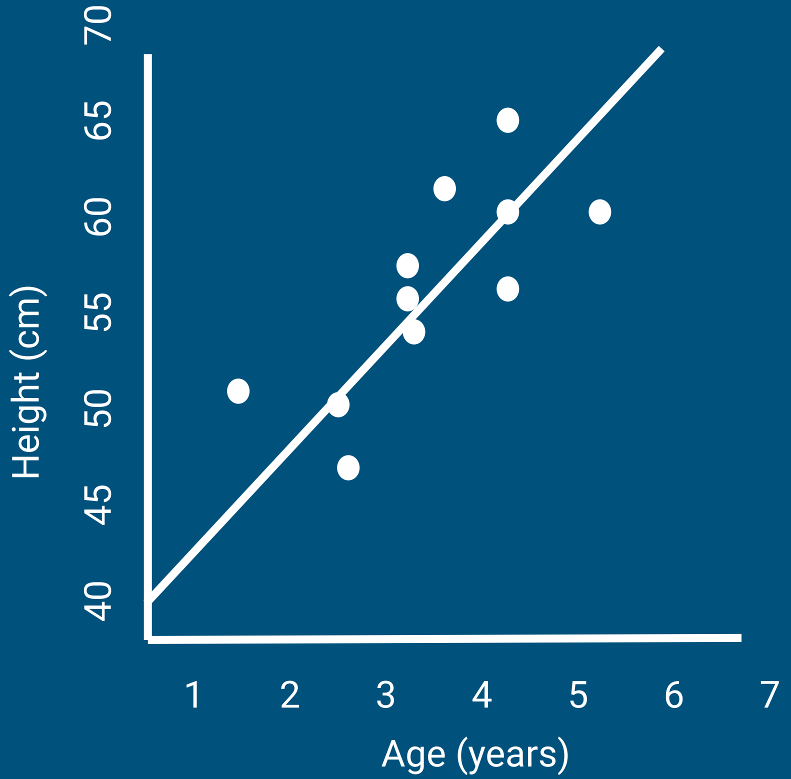
Y-intercept

slope

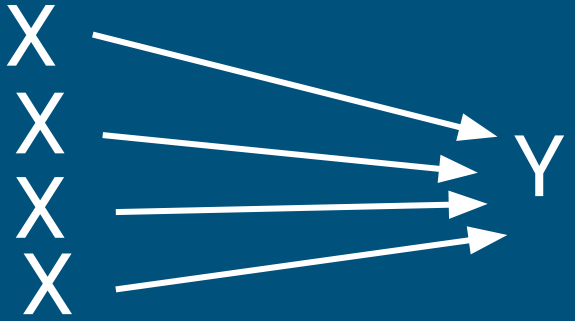
Independent variable

error

Univariate Linear Regression



Multivariate Linear Regression



Continuous or near-continuous

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Dependent variable

Y-intercept

slope 1

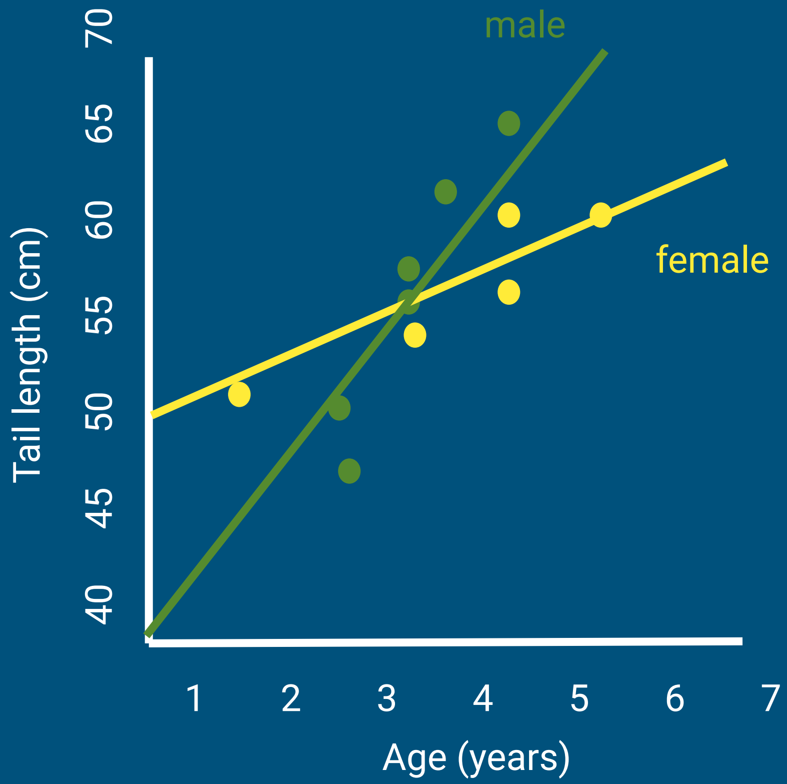
Independent variable 1

slope p

Independent variable p

error

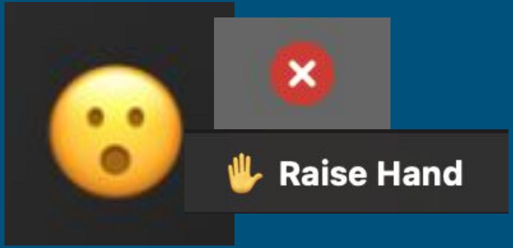
Multivariate Linear Regression



Summary and Check In

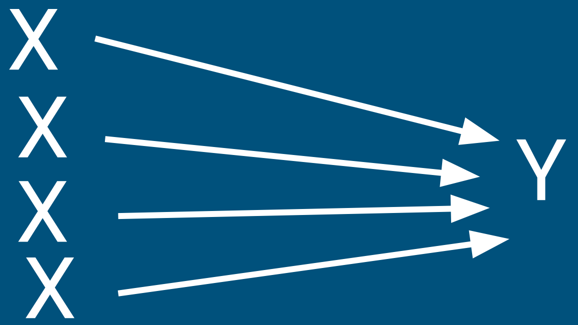
Going well?

Not going so well?



Univariate linear regression

Multivariate linear regression



When can we use a linear model, and how do we make sure we are using it appropriately?

Assumptions → Tests

1. Independence
2. Linearity
3. Homoscedasticity
4. Normality

1. Durbin-Watson Test
2. Plot residuals vs fitted
3. Scale-location plot
4. QQ Plot

All about residuals (error)

ASSUMPTION

1. Independence

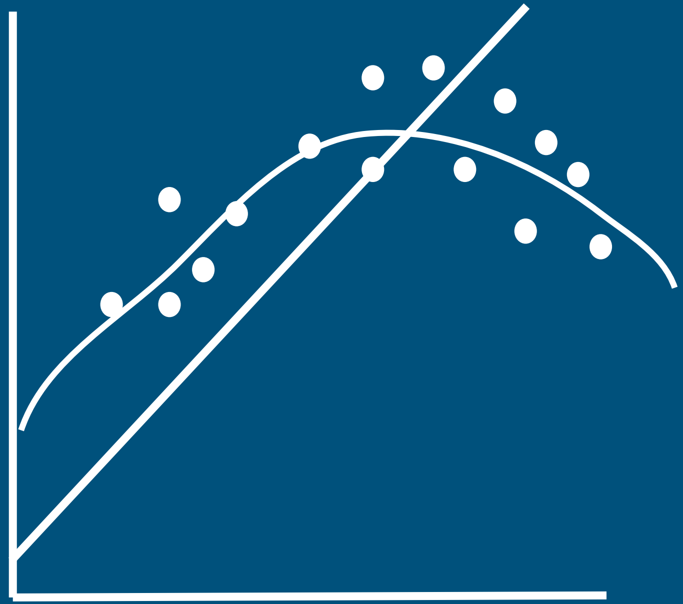
Residual (errors) are assumed to be independent.

TEST

1. Durbin-Watson Test
   a. Examines whether errors are autocorrelated with themselves, returns p-value
2. Plot your data against other variables
   a. i.e. sample date/time

ASSUMPTION

2. Linearity

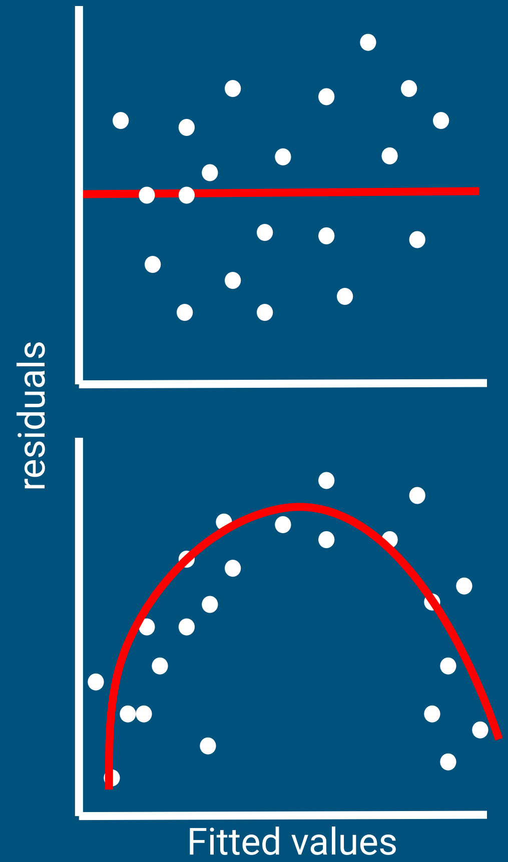


TEST

Plot residuals vs fitted

No pattern

Nonlinear pattern



ASSUMPTION

3. Homoscedasticity

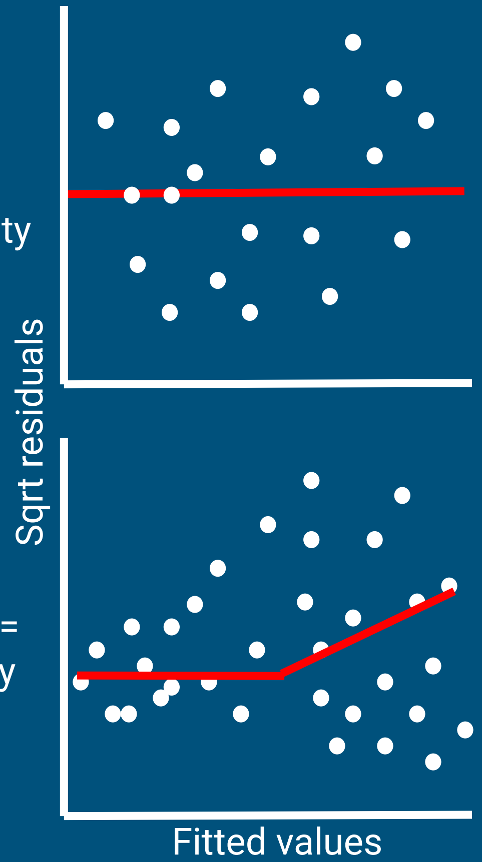
Residual (errors) are assumed to have constant variance (homoscedasticity).

TEST

Scale-Location Plot

No pattern = homoscedasticity

Widening pattern = heteroscedasticity



*can also use residual vs fitted

ASSUMPTION

4. Normality

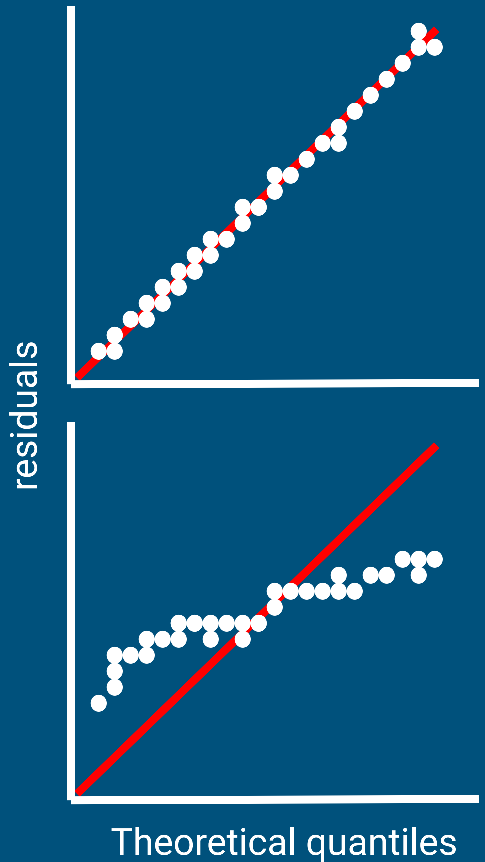
Residual (errors) are assumed to be normally distributed.

TEST

Normal QQ Plot

Residuals follow straight line

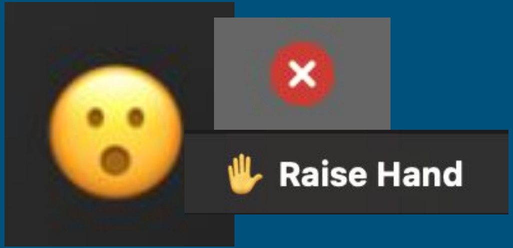
Residuals do not follow straight line



Summary and Check In

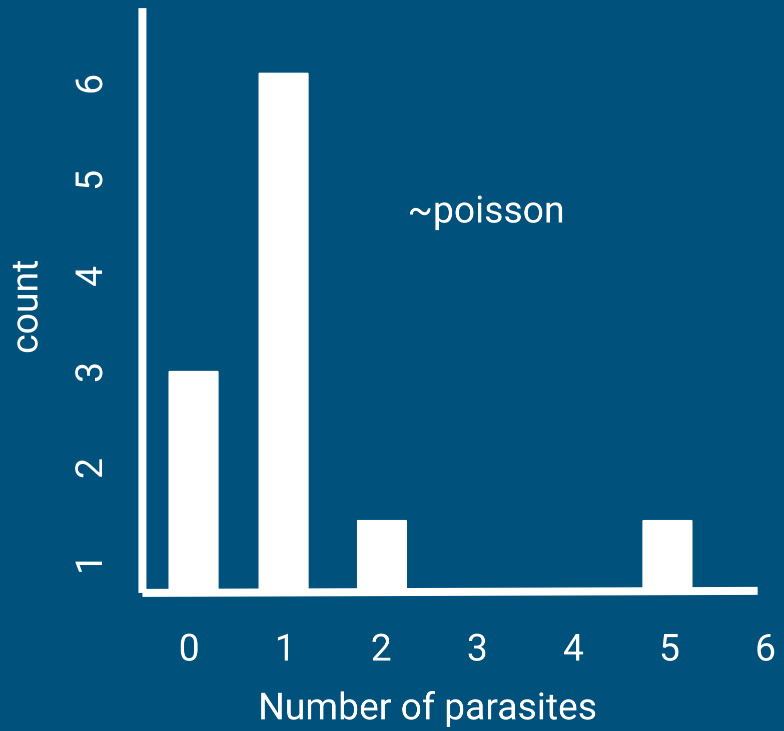
To use linear regression, you have to abide by four main assumptions. There are plots you can generate (based on residuals) that help you determine how well you are meeting the assumptions.

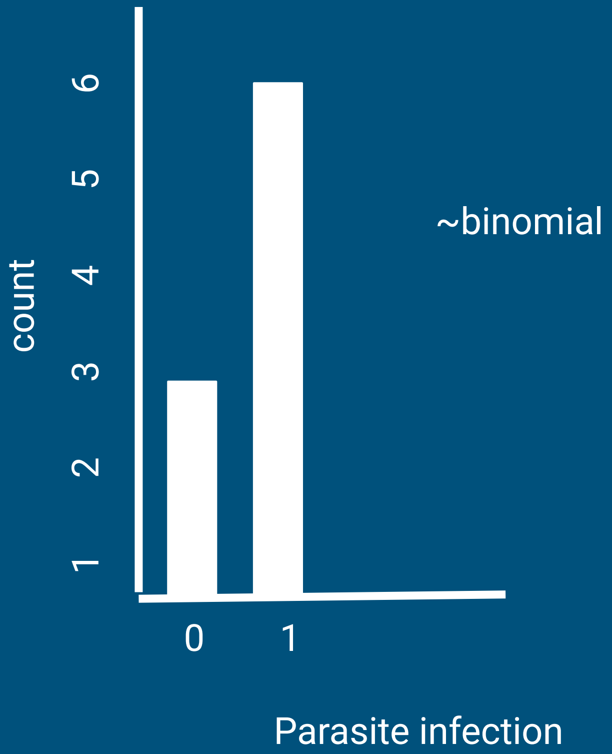
Not going so well?



Going well?

What about when your data isn't normally distributed?





General Linear Model (GLMs)

GLM's extend the linear model framework to allow for non-normal data/residuals by using a **linear predictor** and a **link function.**

Linear predictors

Multivariate regression model

$$Y_i = \beta_0 + \underbrace{\beta_1 X_1 + \ldots + \beta_p X_p}_{\text{Linear predictor}} + \varepsilon$$

$$\nu = \beta_1 X_1 + \ldots + \beta_p X_p$$

Link functions

Describes how the model's mean prediction, μ, depends on the linear predictor, ν

$$\mu = f(\nu)$$

Common Link Functions

Link functions must be within the family of exponential distributions

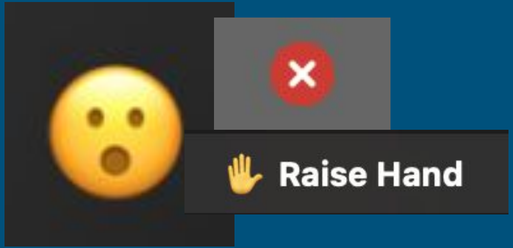
Range of variable	Link function name	Link function formula
Real axis	Identity	μ
Positive real axis	Log	$\log(\mu)$
Positive real axis	Box-Cox	$(\mu^{\lambda}-1)/(\lambda*\gamma^{\lambda})$[2,3]
Positive real axis	Power	μ^{λ}
Reals strictly between 0 and 1	Logit	$\log(\mu/(1-\mu))$
Reals strictly between 0 and 1	Probit	$\Phi^{-1}(\mu)$
Probability vector	Cumulative logit	$\log(\pi/(1-\pi))$[4]
Probability vector	Ordered probit (cumulative probit)	$\Phi^{-1}(\pi)$[4]

Summary and Check In

Generalized linear models (GLMs) allow us to perform linear regression on non-normal data, using a **<u>link function</u>**.

We choose a link function by looking at characteristics of our data.

Not going so well?



Going well?

Model Selection

Stepwise regression : one by one add and remove predictors (X's) in order to find the best-fit model

Nested model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon$$

both

forward

backward

AIC : Akaike Information Criterion

+ how well does a model fit the data
- number of parameters

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \longrightarrow \text{AIC} = 350$$

best fit

$$\boxed{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \longrightarrow \text{AIC} = 200}$$

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon \longrightarrow \text{AIC} = 325$$

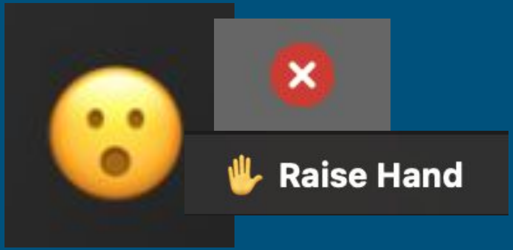
Smaller AIC score is better

Summary and Check In

Stepwise regression allows us to test nested models using an AIC score, in order to select the best-fit model.

There are three types of stepwise regression: forward, backward, and both.

Not going so well?



Going well?

Tutorial Time!