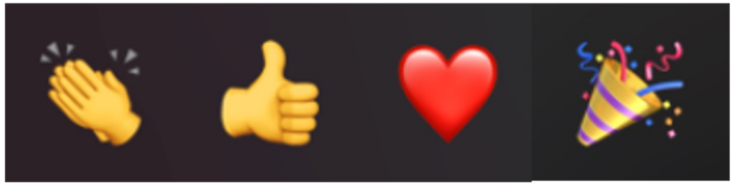
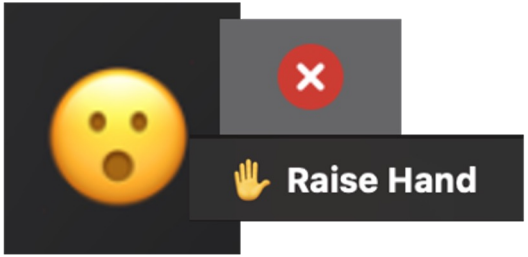
Multivariate Model Formulation

David Klinges, Emily Ruhs

Going well?

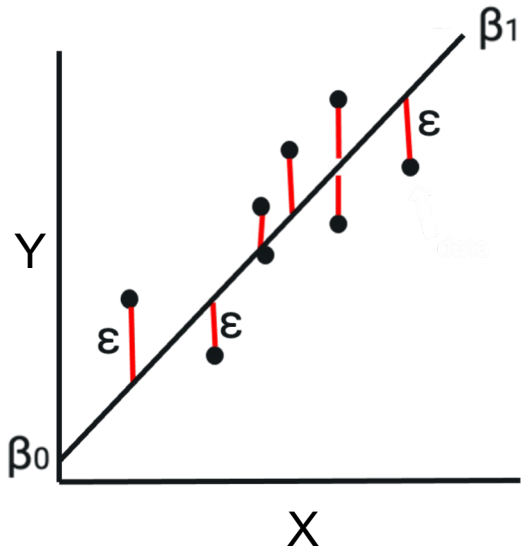


Not going so well?



Review: linear regression

Review: linear regression



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

Dependent variable

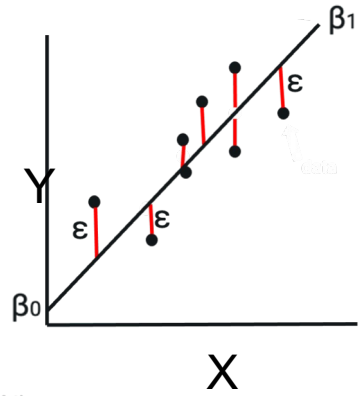
Y-intercept

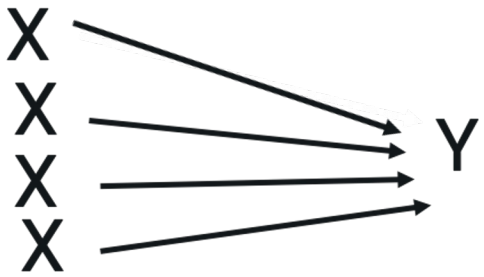
slope

Independent variable

error

Review: multivariate regression





Continuous or near-continuous

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

Dependent variable

Y-intercept

slope 1

Independent variable 1

slope p

Independent variable p

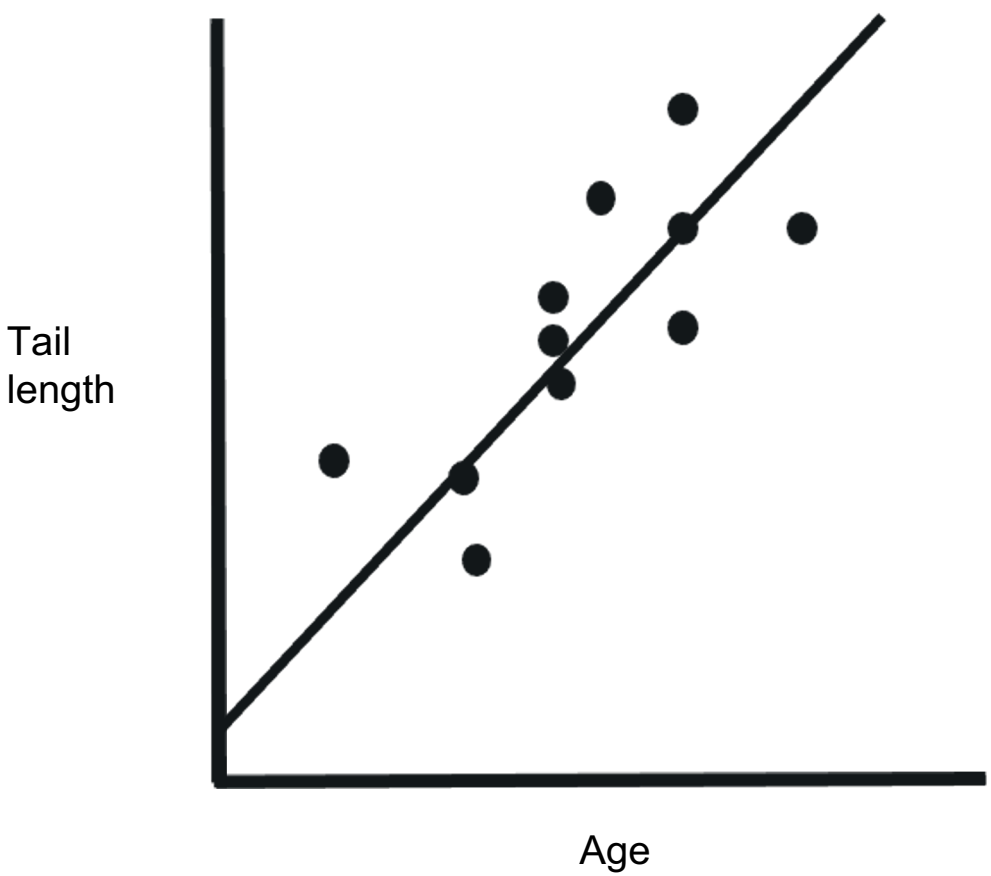
error

Multivariate models

Revisit lemur example:

tail length ~ age

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

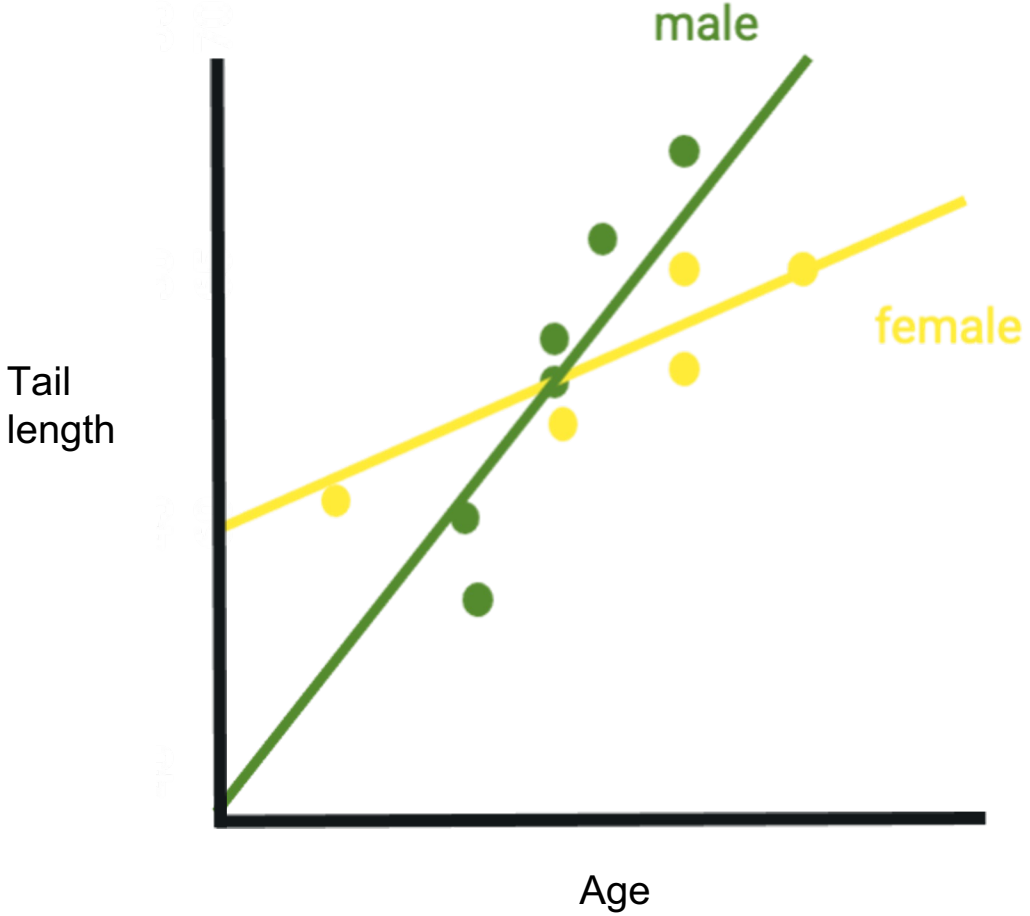


Multivariate models

Revisit lemur example:

tail length ~ age + sex

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



Multivariate models

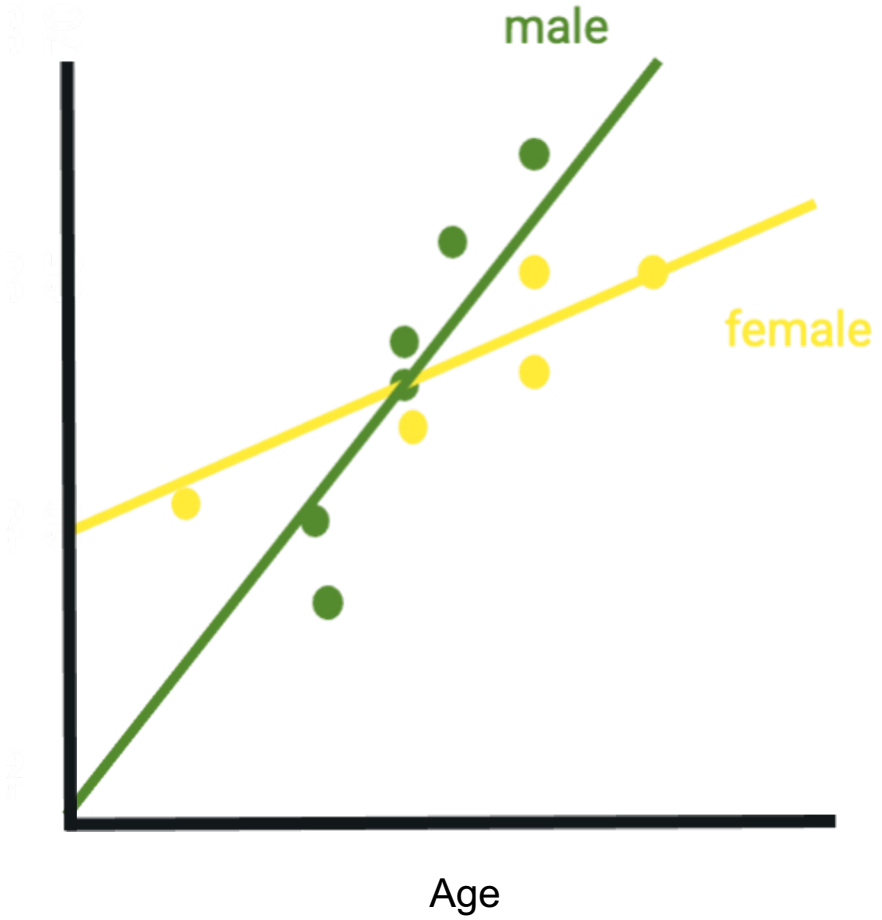
Revisit lemur example:

tail length ~ age + sex

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

```
summary(linear_model)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.30693    0.91506   18.91   <2e-16 ***
age          0.88593    0.04107   21.57   <2e-16 ***
sexMale     11.16501    0.79608   14.03   <2e-16 ***
```

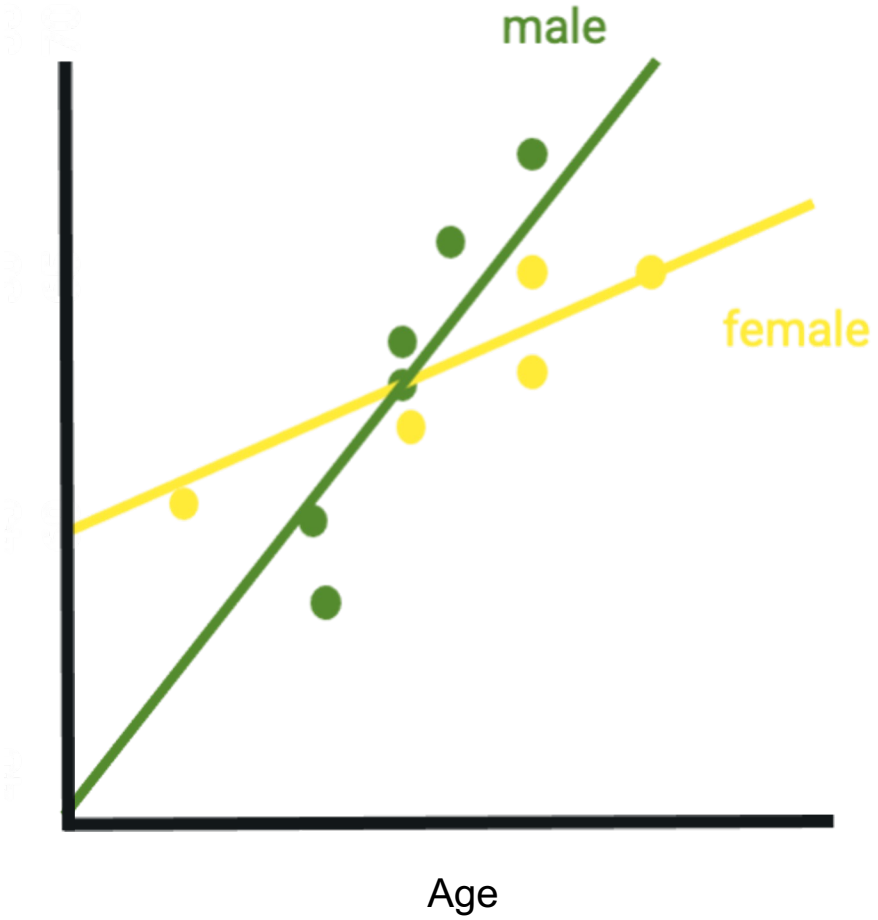
Multivariate models

Revisit lemur example:

tail length ~ age + sex

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

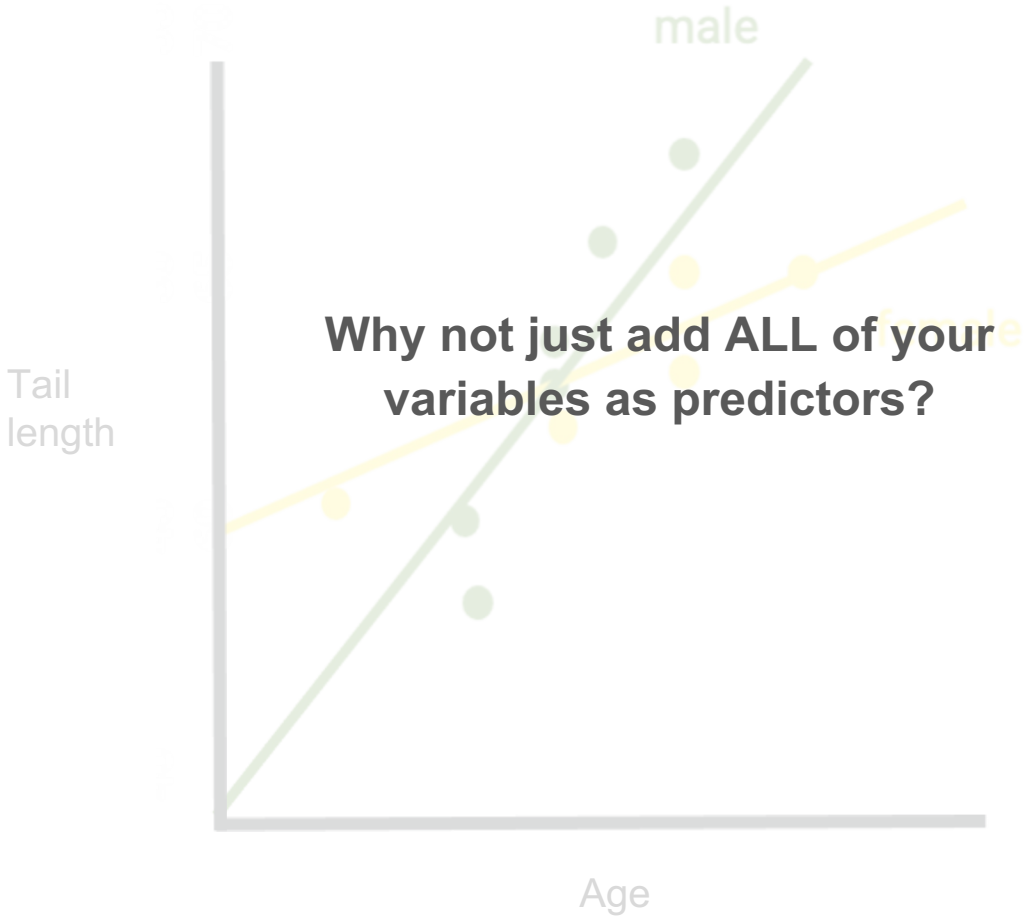
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.30693    0.91506   18.91   <2e-16 ***
age          0.88593    0.04107   21.57   <2e-16 ***
sexMale     11.16501    0.79608   14.03   <2e-16 ***
```

Multivariate models

Revisit lemur example:

tail length ~ age + sex



Why not just add ALL of your variables as predictors?

Multivariate models

Why not just add ALL of your variables as predictors?

Multivariate models

Degrees of freedom

"the number of independent values that can vary in an analysis without breaking any constraints"

$df = N - k - 1$

Low df -> low statistical power -> harder to test effect of any variable

Why not just add ALL of your variables as predictors?

Multivariate models

Degrees of freedom

"the number of independent values that can vary in an analysis without breaking any constraints"

$df = N - k - 1$

Low df -> low statistical power -> harder to test effect of any variable

Multicollinearity

Several of your predictors may be correlated– they vary with each other. This may undermine inference

Why not just add ALL of your variables as predictors?

Multivariate models

Degrees of freedom

"the number of independent values that can vary in an analysis without breaking any constraints"

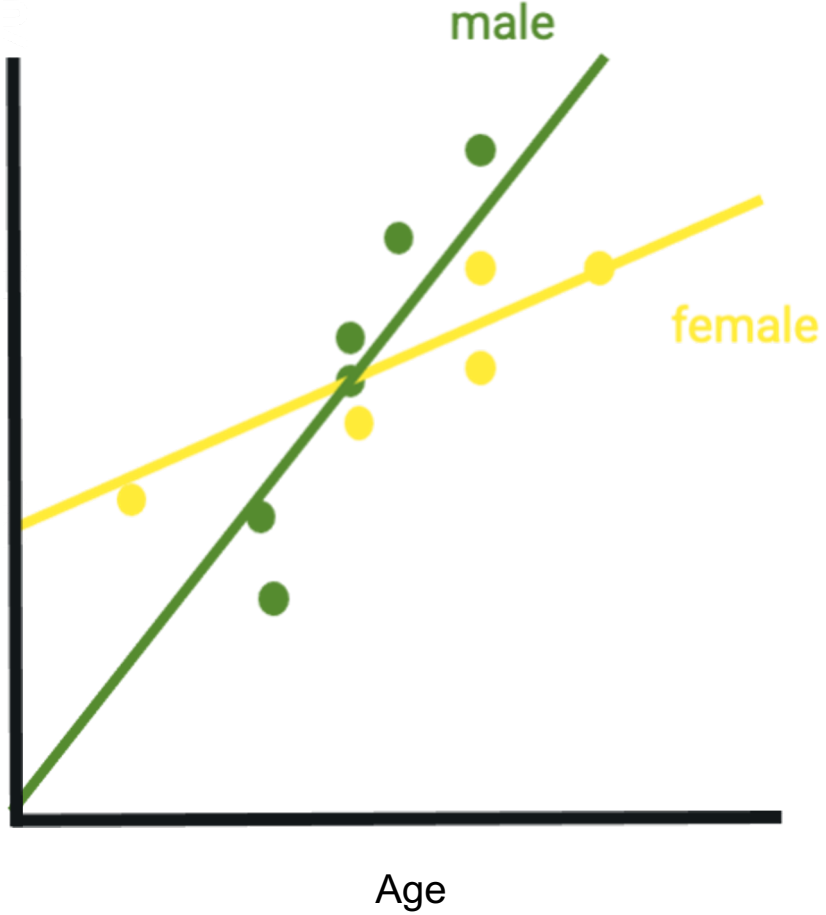
$df = N - k - 1$

Tail length

Low df -> low statistical power -> harder to test effect of any variable

Multicollinearity

Several of your predictors may be correlated– they vary with each other. This may undermine inference



Multivariate models

Degrees of freedom

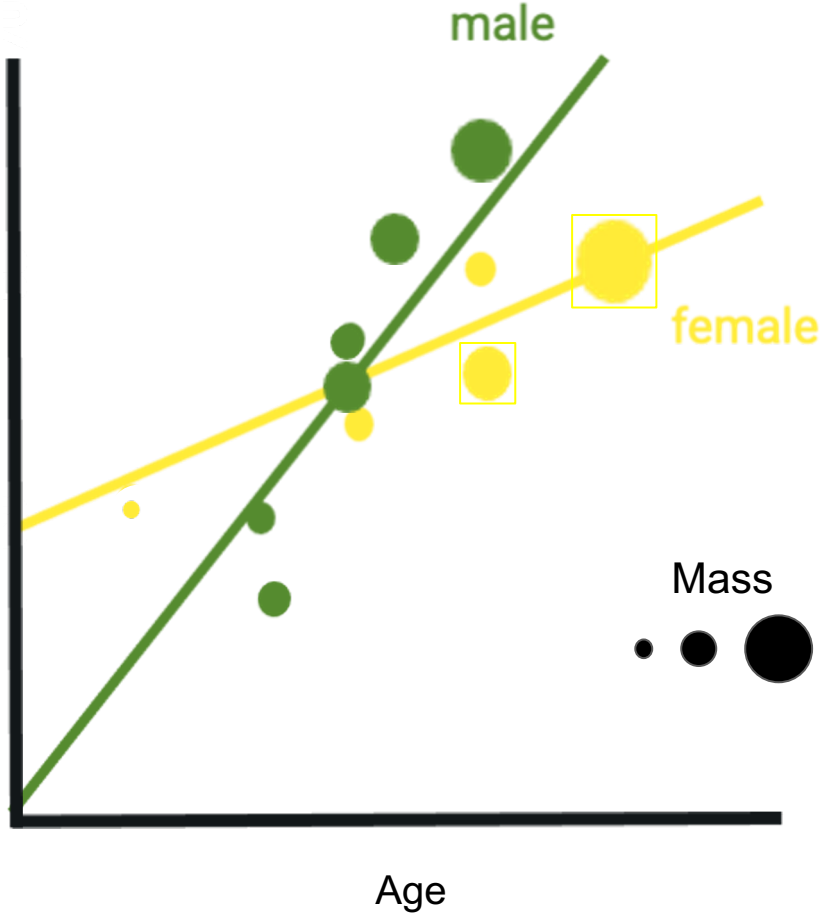
"the number of independent values that can vary in an analysis without breaking any constraints"

$df = N - k - 1$

Low df -> low statistical power -> harder to test effect of any variable

Multicollinearity

Several of your predictors may be correlated– they vary with each other. This may undermine inference



Multivariate models

Degrees of freedom

"the number of independent values that can vary in an analysis without breaking any constraints"

$df = N - k - 1$

Low df -> low statistical power -> harder to test effect of any variable

Multicollinearity

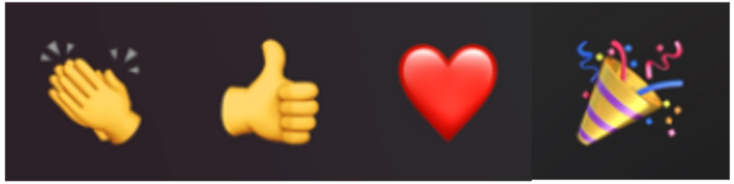
Several of your predictors may be correlated– they vary with each other. This may undermine inference

The structure of your model depends on your research questions, hypotheses, and assumptions.

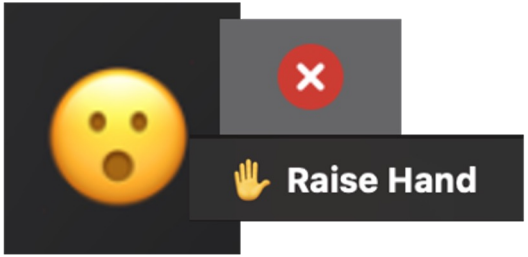
Summary and Check in

- Multivariate models enable us to include multiple predictor variables
- However, the more predictors that we include in a model, the lower our degrees of freedom drops, and we may end up with correlations between predictors (multicollinearity)
- Structuring a model depends upon your research questions, hypothesis, and assumptions

Going well?



Not going so well?



Multivariate model structure

But what if there are things I would like to control for, without losing degrees of freedom?

Multivariate model structure:

But what if there are things I would like to control for, without losing degrees of freedom?

Multivariate model structure:

But what if there are things I would like to control for, without losing degrees of freedom?

Observations must be independent from one another

tail length ~ age + sex + id

How to include parameters in a model?

tail length ~ age + sex + id

According to your research questions and hypotheses, does a categorical variable represent a **driver of primary interest**, or a **cohort** of collected data **representing a broader population or distribution**?

How to include parameters in a model?

tail length ~ age + sex + id

According to your research questions and hypotheses, does a categorical variable represent a **driver of primary interest**, or a **cohort** of collected data **representing a broader population or distribution**?

If a **driver,** consider including it in your model as a **fixed effect.**

How to include parameters in a model?

tail length ~ age + sex + id

According to your research questions and hypotheses, does a categorical variable represent a **driver of primary interest**, or a **cohort** of collected data **representing a broader population or distribution**?

If a **driver,** consider including it in your model as a **fixed effect.**

If a **cohort drawn from a population,** or a **repeated measure** of the same state, consider including it in your model as a **random effect.**

How to include parameters in a model?

tail length ~ age + sex + id

According to your research questions and hypotheses, does a categorical variable represent a **driver of primary interest**, or a **cohort** of collected data **representing a broader population or distribution**?

If a **driver,** consider including it in your model as a **fixed effect.**

If a **cohort drawn from a population,** or a **repeated measure** of the same state, consider including it in your model as a **random effect.**

These are ***guidelines,*** *not* ***definitions****!*

How to include parameters in a model?

Some variables could be best represented as fixed or random, depending on your questions:

How to include parameters in a model?

Some variables could be best represented as fixed or random, depending on your questions:

Site: repeated measures of forest habitat

Site 1

Site 2

Site 3

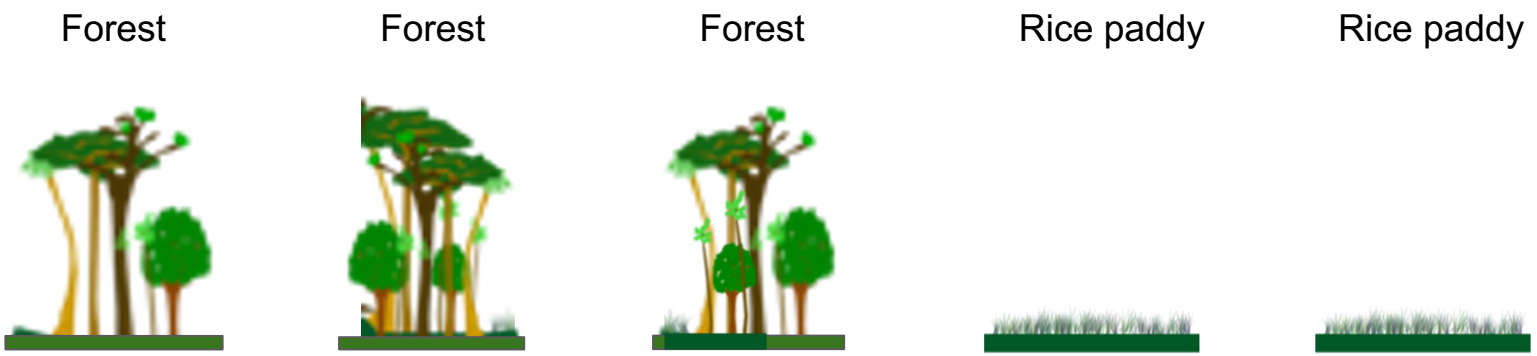
Site 4

Site 5

How to include parameters in a model?

Some variables could be best represented as fixed or random, depending on your questions:

Site: forest and rice paddy habitats



How to include parameters in a model?

Some variables could be best represented as fixed or random, depending on your questions:

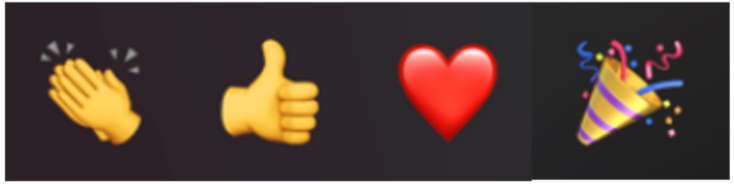
Although there are better and lesser model structures for a given dataset, there is no perfectly "correct" way to structure a model. The structure of your model depends on your research questions, hypotheses, and assumptions.

Forest Forest Forest Rice paddy Rice paddy

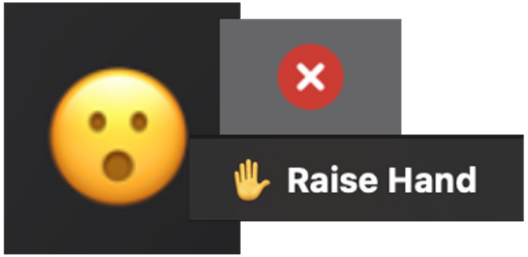
Summary and Check in

- We can represent predictors as either fixed or random effects, in part based upon the role that the predictor plays in our study design and answering our research questions
- Modeling is somewhat of an art and a science: there is no single best recipe to follow!

Going well?



Not going so well?



Model prediction

Model prediction

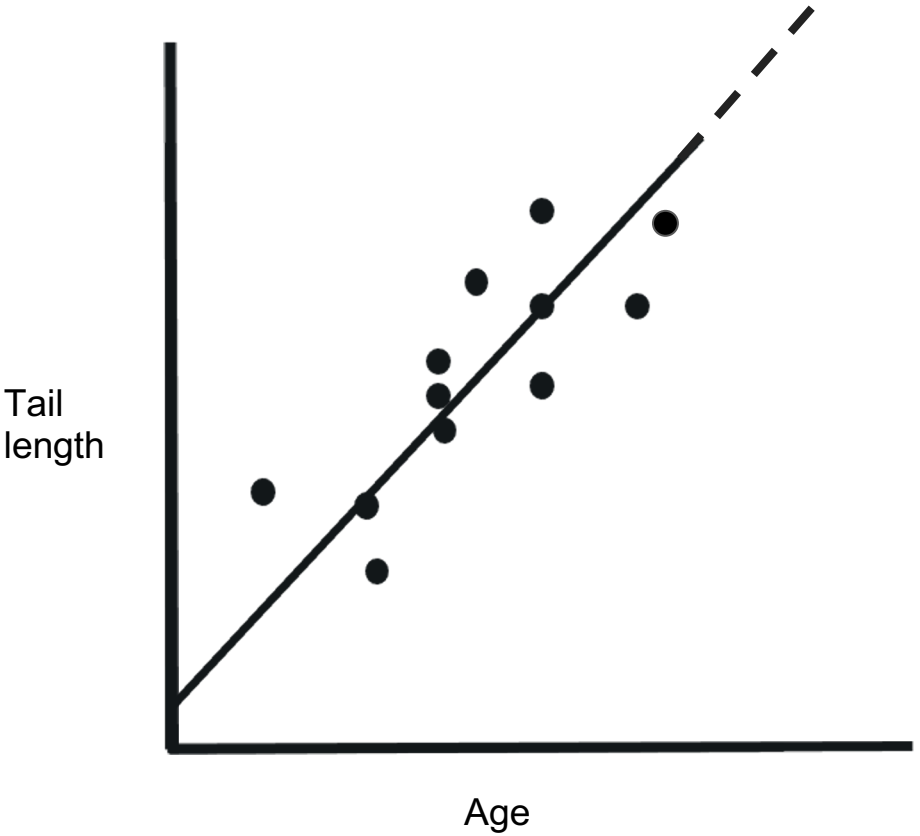
What if we want to know tail lengths for ages that we didn't observe?

Interpolation

Model prediction

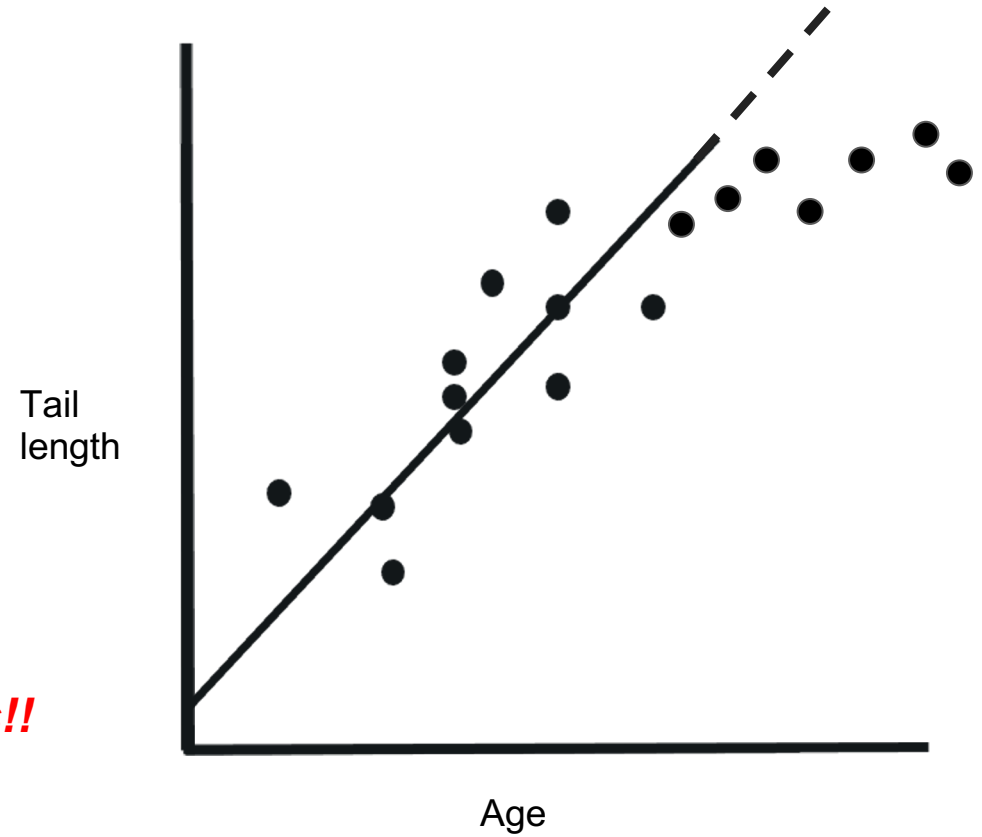
What if we want to know tail lengths for ages that we didn't observe?

Extrapolation



Model prediction

What if we want to know tail lengths for ages that we didn't observe?



*Extrapolation….**dangerous!!***

Model inference and prediction

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

Once we estimate values of Beta coefficients (B0, B1….) from our data, we can ***infer*** what drivers are most important in determining the response.

But, we can also now use this model to ***predict*** new Y, given certain x1, x2, etc.

Models can both be used to draw ***inference*** on relationships between variables, but also ***predict*** to estimate unobserved outcomes

Tutorial time

Move over to R and RStudio….