Model Selection and Comparison

Christian Ranaivoson

E2M2 2025 Andasibe, Madagascar

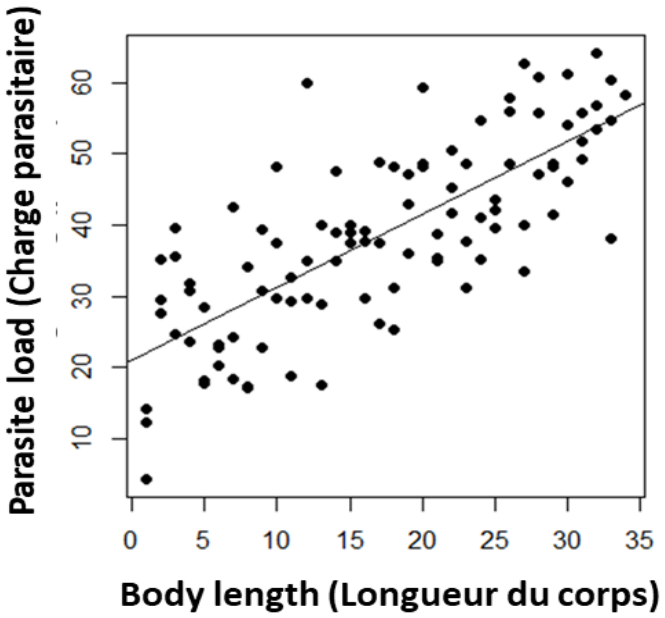
University of Chicago, USA ; Association Ekipa Fanihy; University of Antananarivo, Madagascar

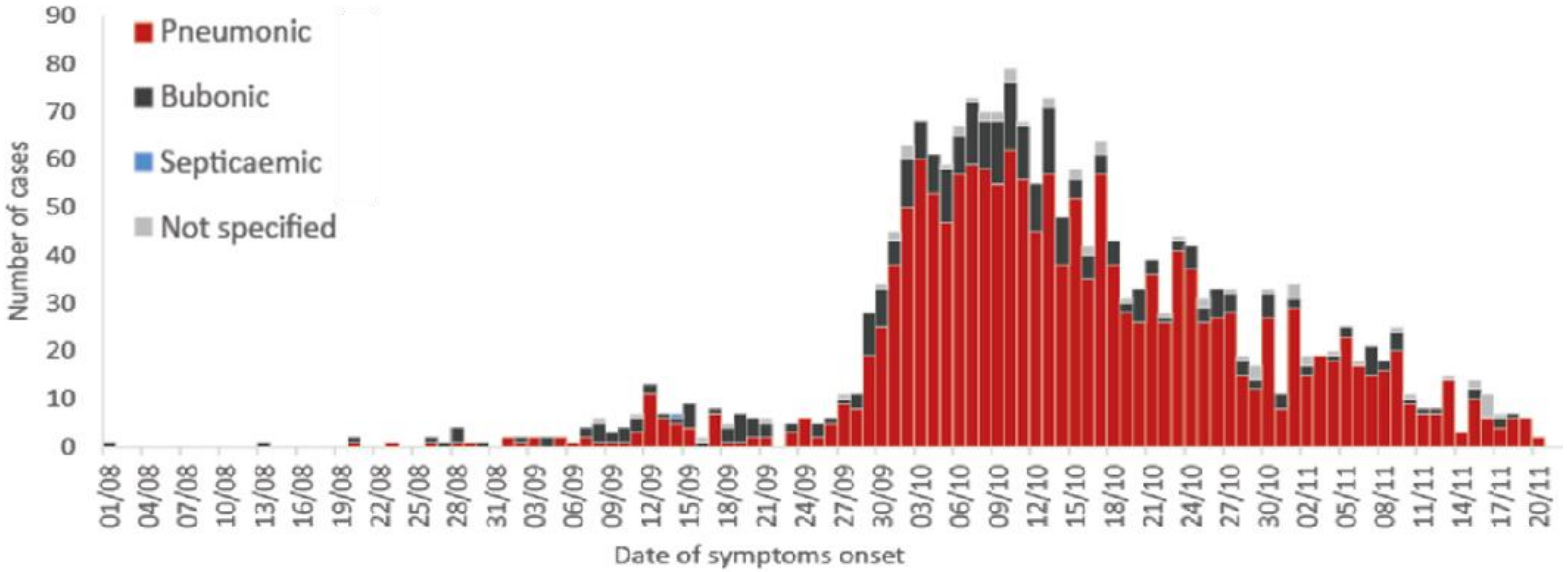
Adapted from Cara Brook, Jessica Metcalf, Christian Ranaivoson E2M2 2022-2024.

Goals of the lecture

- Introduction to model selection
- Be familiar with common measures of model fit
- Tutorial on how to do it in practice (Real example)

Which model is best?





There are many statistical methods used to ‘fit’ models to data and there are many possible scenarios from which mechanical model can be built.

There are many statistical methods used to ‘fit’ models to data and there are many possible scenarios from which mechanical model can be built.

The method best suited for your work will depend on your data and your question.

There are many statistical methods used to ‘fit’ models to data and there are many possible scenarios from which mechanical model can be built.

The method best suited for your work will depend on your data and your question.

What are some measures of model fit that you could use?

What are some measures of model fit that you could use?

Maximum likelihood (Maximum de vraisemblance)

(izay manakaiky indrindra ny tena izy ry reto an)

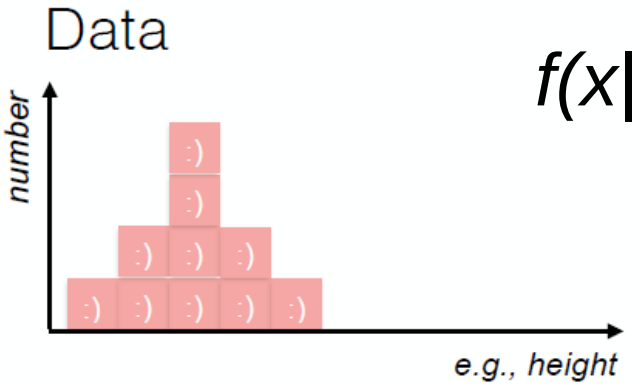
R-squared (R-carré)

Least squares (Moindres carrés)

AIC

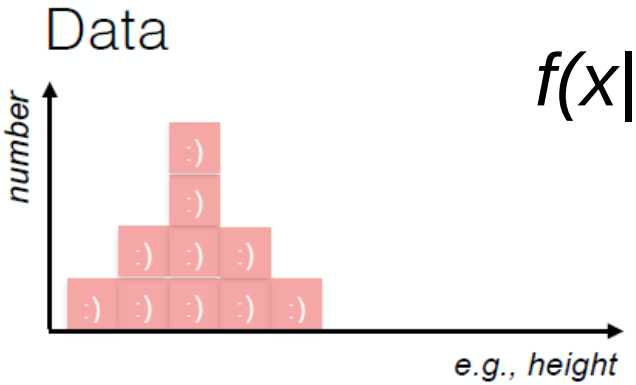
(uses least squares or log-likelihood but penalizes by number of fitted parameters)

Maximum likelihood

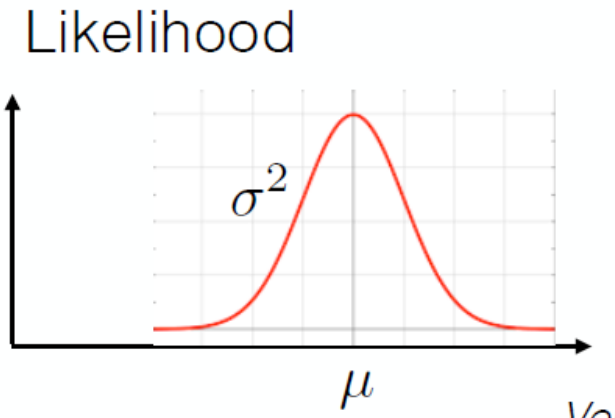


$f(x|\mu, \sigma^2)$

Maximum likelihood



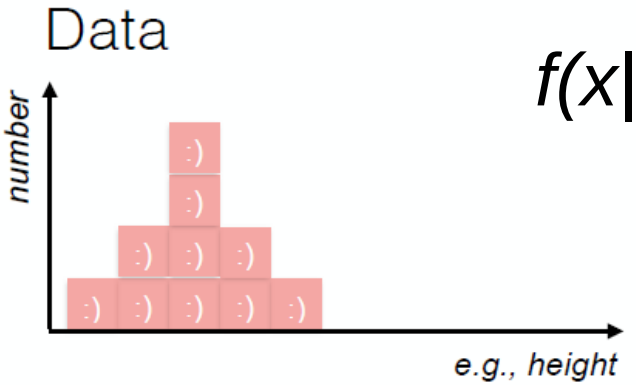
$f(x|\mu, \sigma^2)$



~ the 'probability of seeing the data' given the chosen parameters
…each individual in the data has a corresponding likelihood… multiply them…

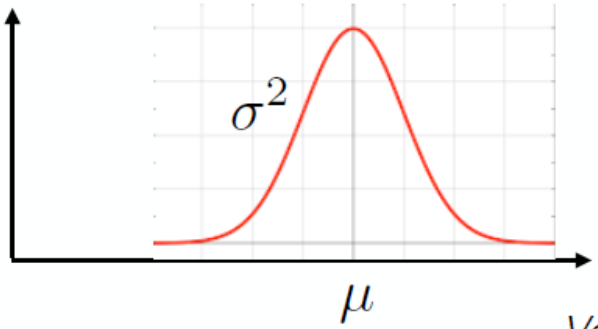
Very low likelihood.

Maximum likelihood



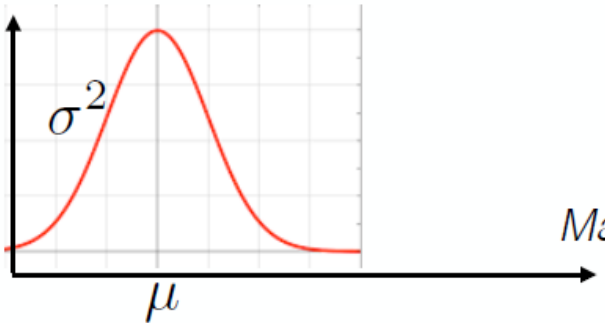
$f(x|\mu, \sigma^2)$

Likelihood



~ the 'probability of seeing the data' given the chosen parameters …each individual in the data has a corresponding likelihood… multiply them…

Very low likelihood.



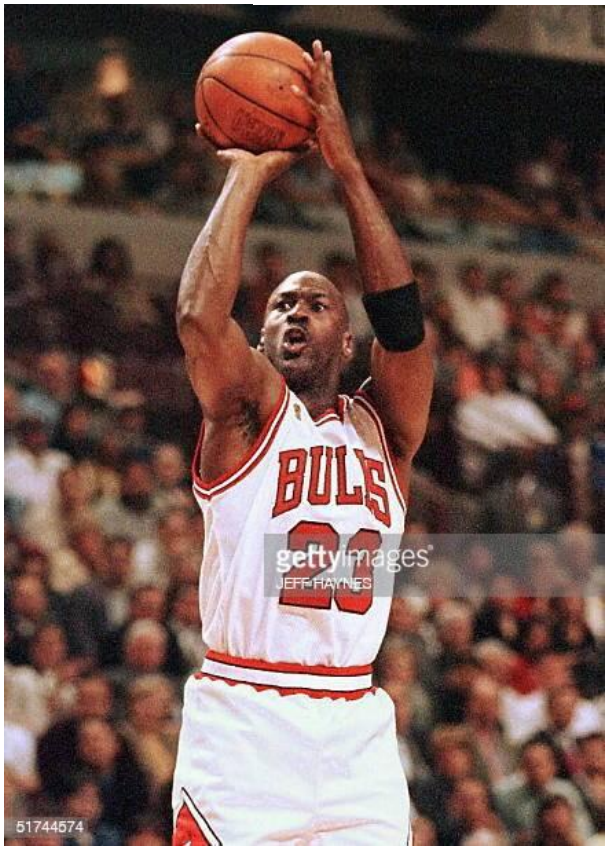
Maximum likelihood

Likelihood Vraisemblance

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$l(\theta|x) = \log L(\theta|x)$$

SUMMARY	FG3%	FT%
Career	32.7	83.5



Examples:

R function : **dbinom(x, size, prob, log=T)**

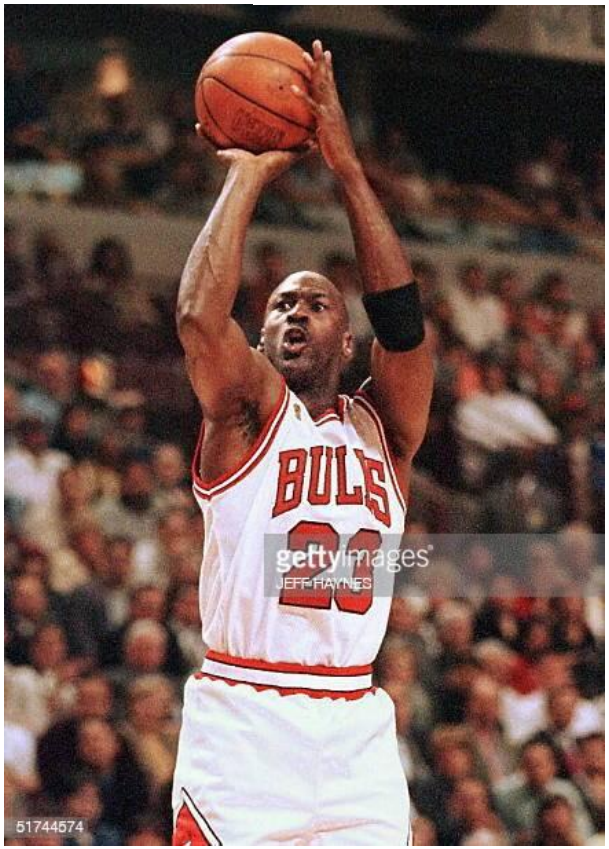
→ R function : **dbinom(8, 10, 0.835) = ?**

Likelihood Vraisemblance

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$l(\theta|x) = \log L(\theta|x)$$

SUMMARY	FG3%	FT%
Career	32.7	83.5



Examples:

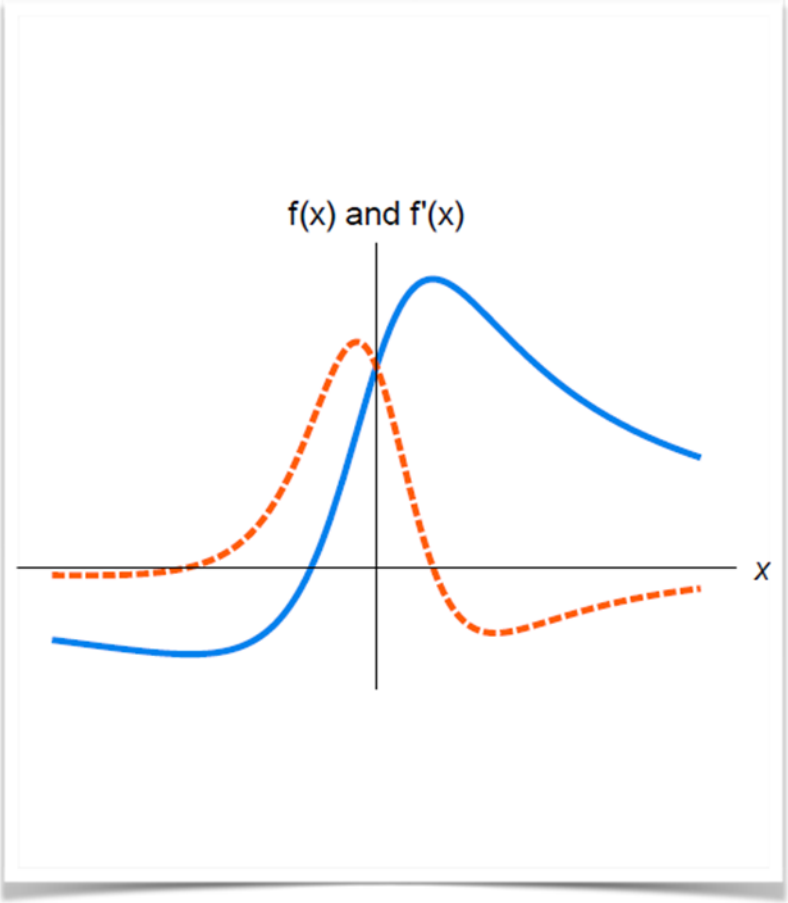
R function : **dbinom(x, size, prob, log=T)**

→ R function : **dbinom(8, 10, 0.835) = 0.289**

Optimization/maximization

A function and its derivative

- What happen when the derivative is:
  - negative?
  - positive?
  - zero?
  - reaching a maximum (finite) value?

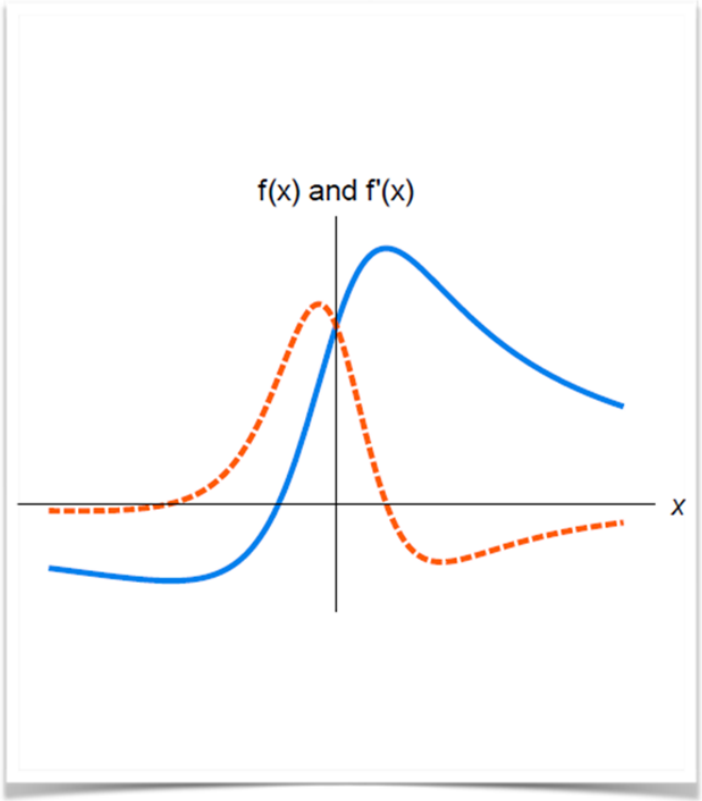


From Tanjona Ramiadantsoa

Optimization/maximization

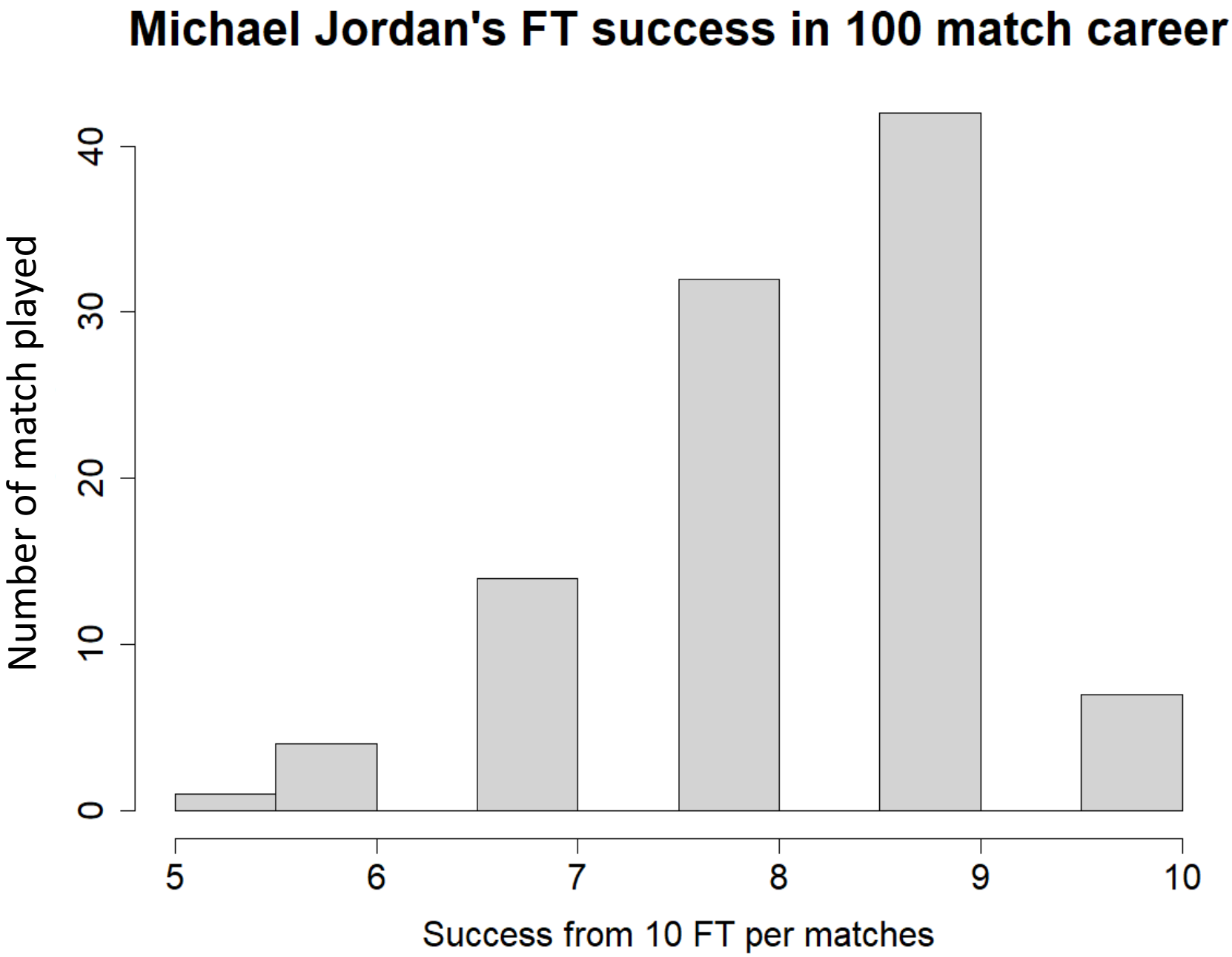
A function and its derivative

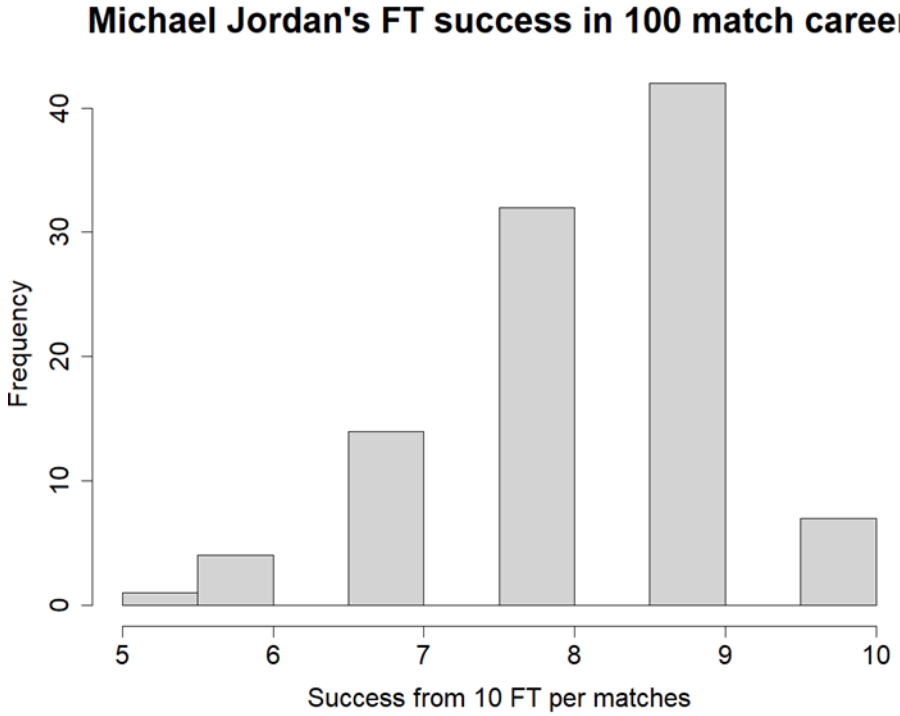
- What happen when the derivative is:
  - negative?
  - positive?
  - zero?
  - reaching a maximum (finite) value?

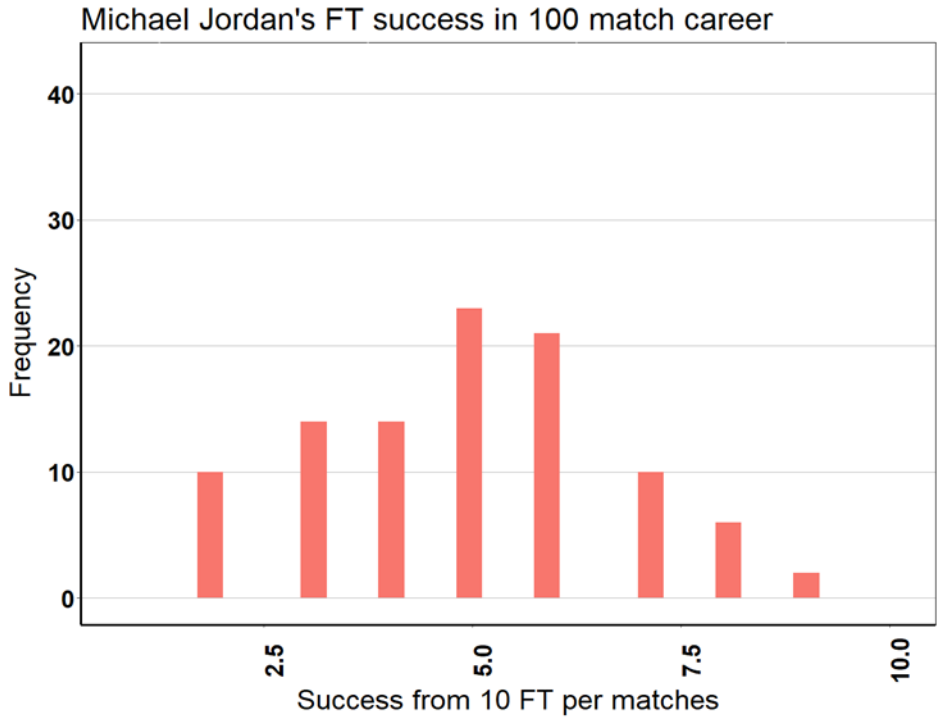


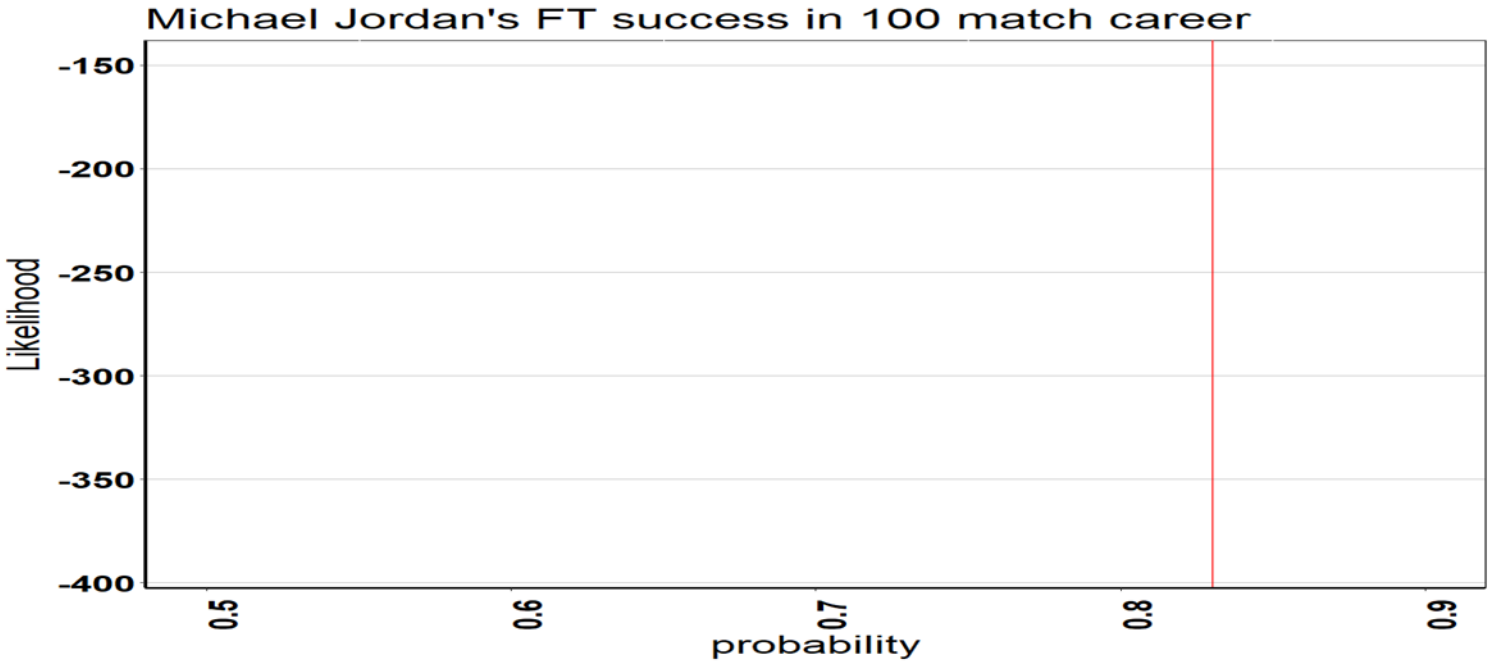
From Tanjona Ramiadantsoa

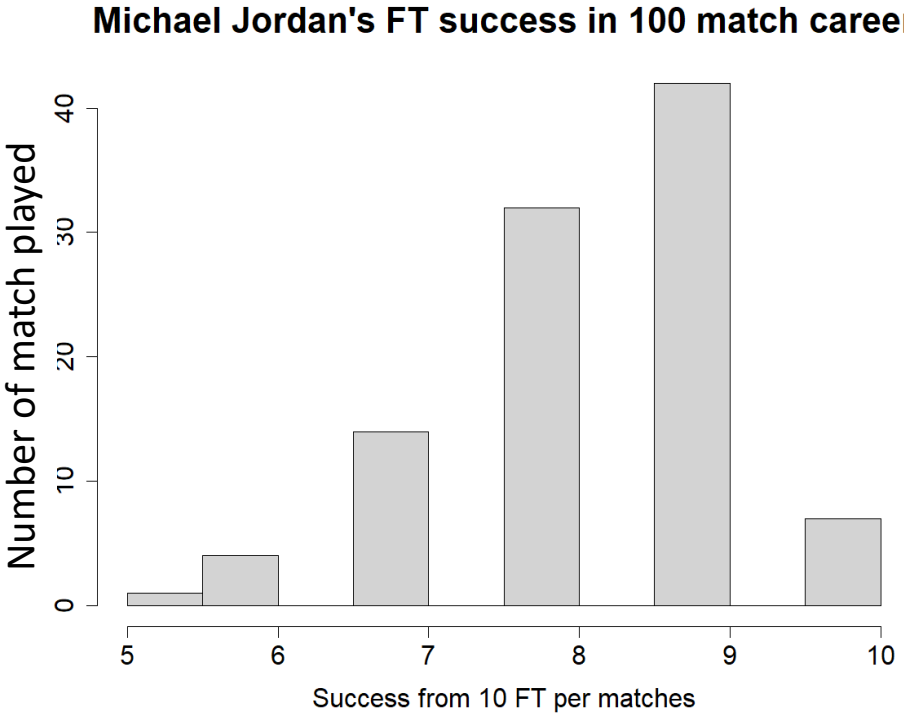
The R function ‘optim’ can be used to find minimum/maximum.

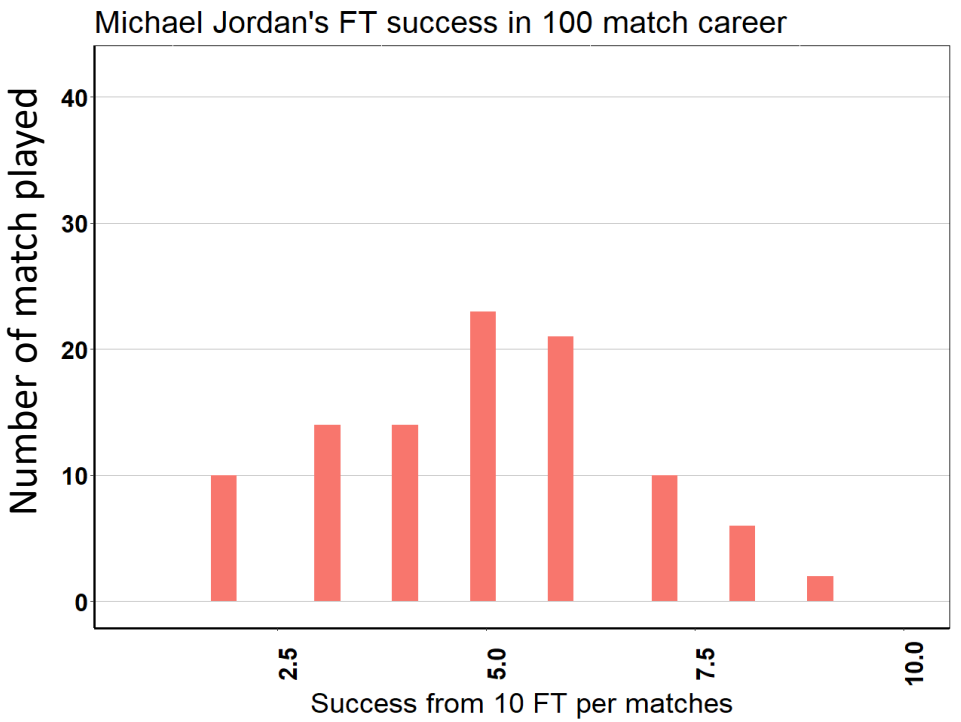

Michael Jordan's FT success in 100 match career
Number of match played
0
10
20
30
40
5
6
7
8
9
10
Success from 10 FT per matches

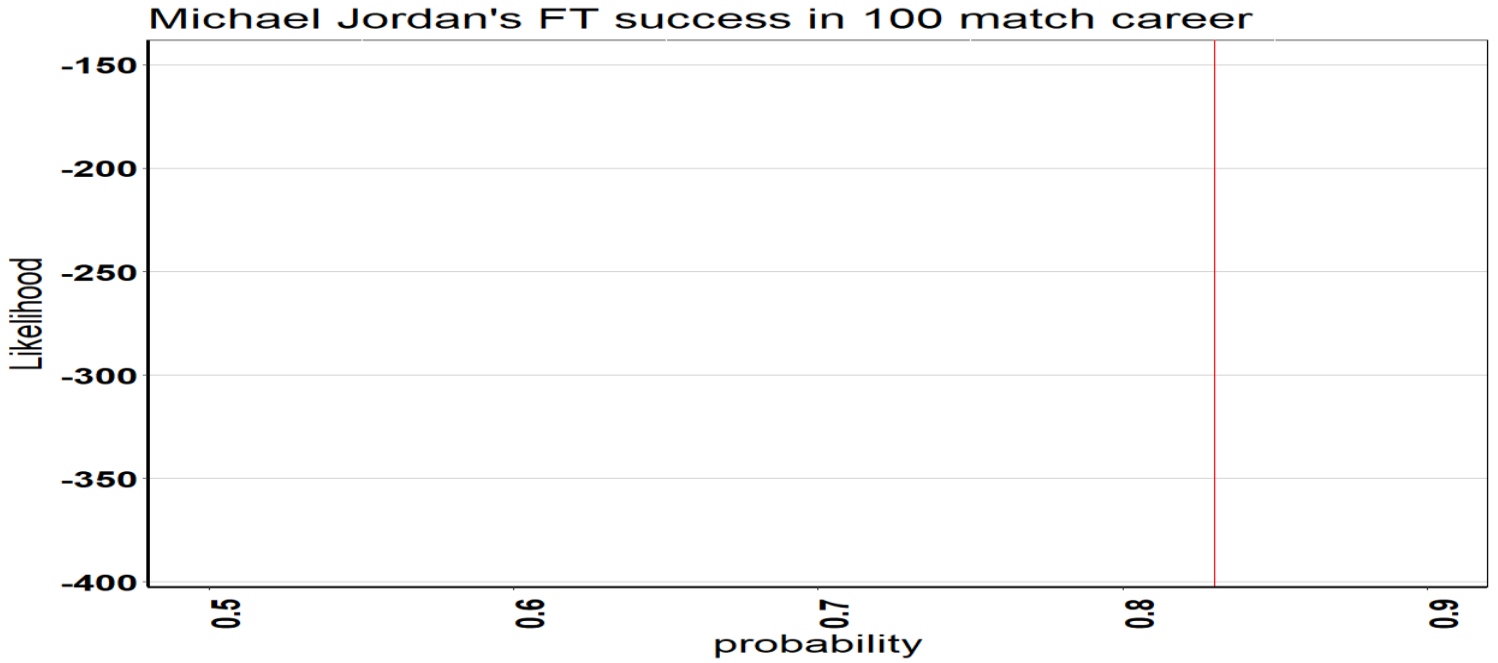

Michael Jordan's FT success in 100 match career
Frequency
0
10
20
30
40
5
6
7
8
9
10
Success from 10 FT per matches


Michael Jordan's FT success in 100 match career
Frequency
0
10
20
30
40
2.5
5.0
7.5
10.0
Success from 10 FT per matches

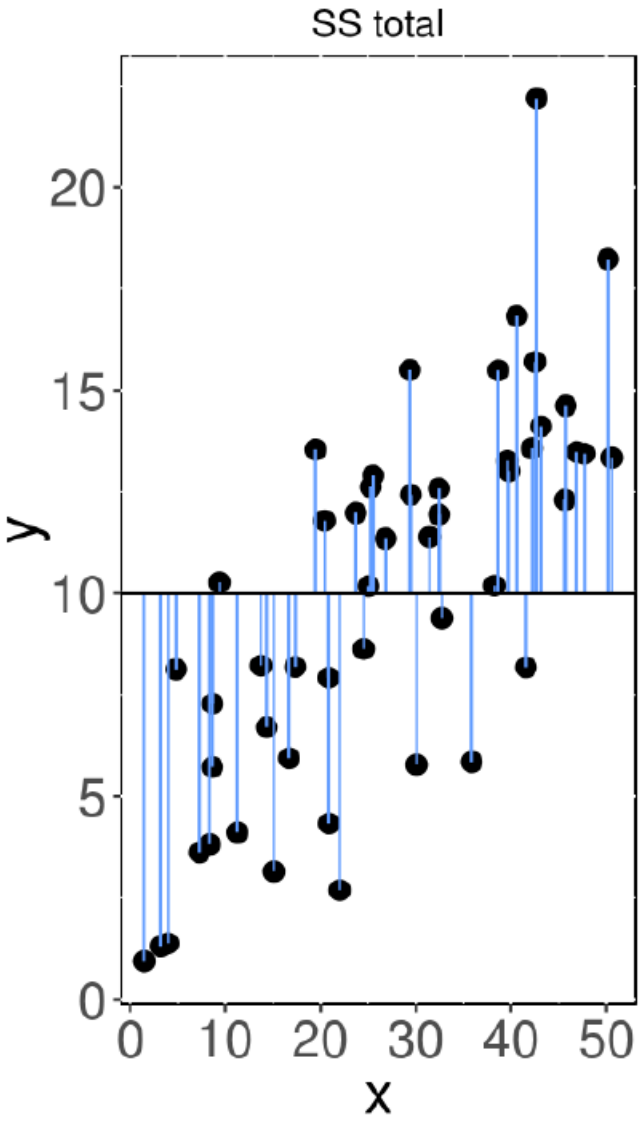

Michael Jordan's FT success in 100 match career
Likelihood
-150
-200
-250
-300
-350
-400
0.5
0.6
0.7
0.8
0.9
probability

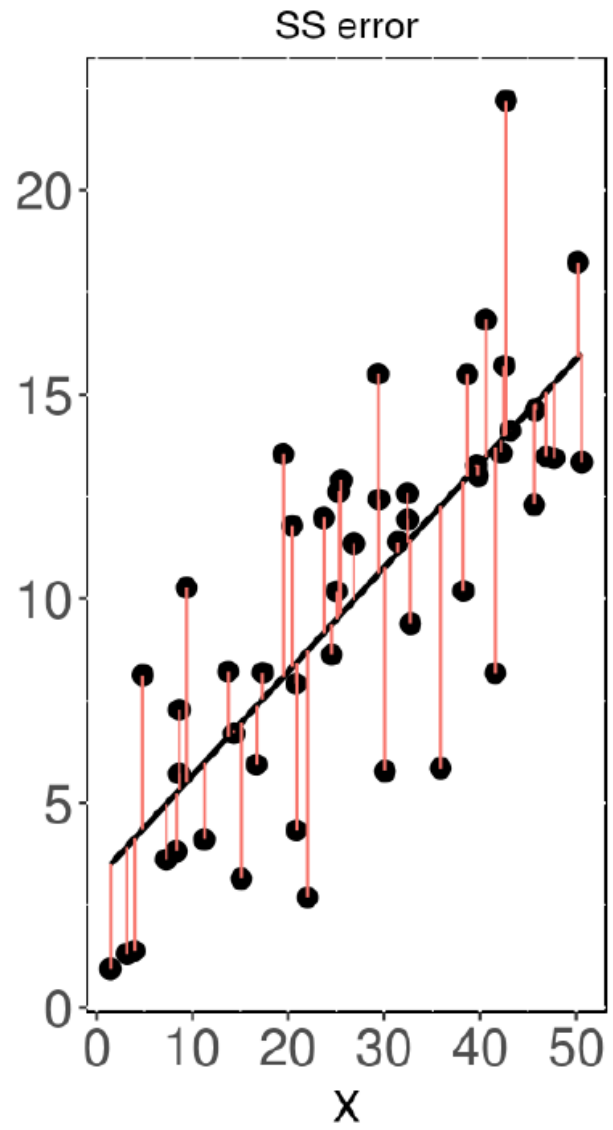

Michael Jordan's FT success in 100 match career
Number of match played
0
10
20
30
40
5
6
7
8
9
10
Success from 10 FT per matches


Michael Jordan's FT success in 100 match career
Number of match played
0
10
20
30
40
2.5
5.0
7.5
10.0
Success from 10 FT per matches


Michael Jordan's FT success in 100 match career
Likelihood
-150
-200
-250
-300
-350
-400
0.5
0.6
0.7
0.8
0.9
probability

Definition r^2

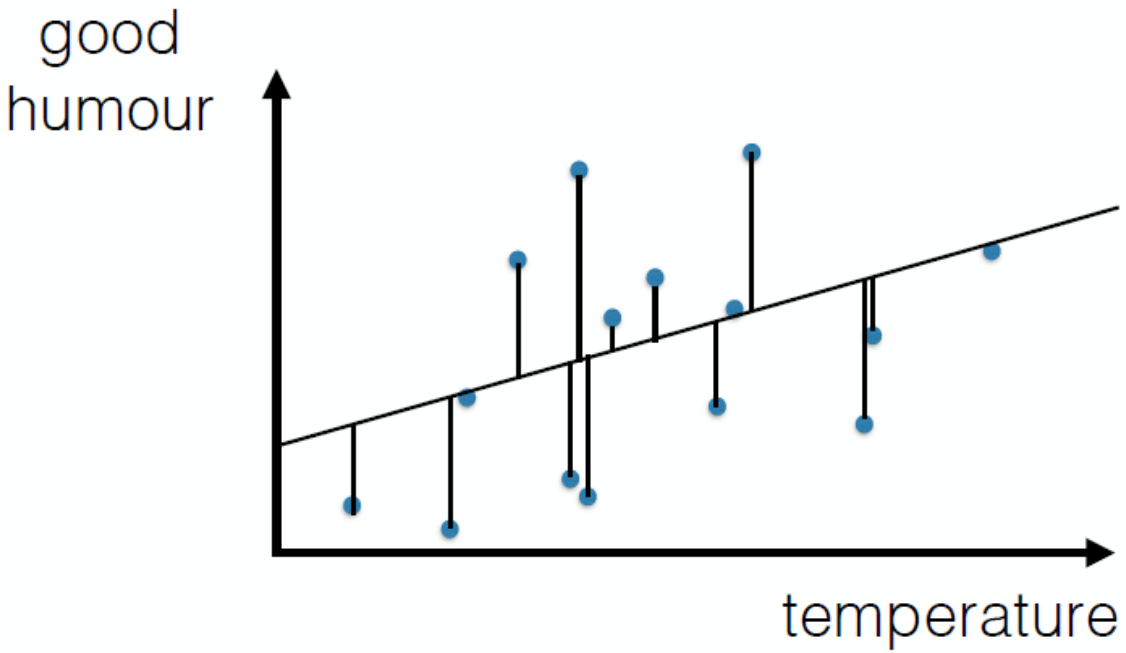




$$R^2 = 1 - \frac{SSE_p}{SST}$$

Least squares

Adding covariates and R^2

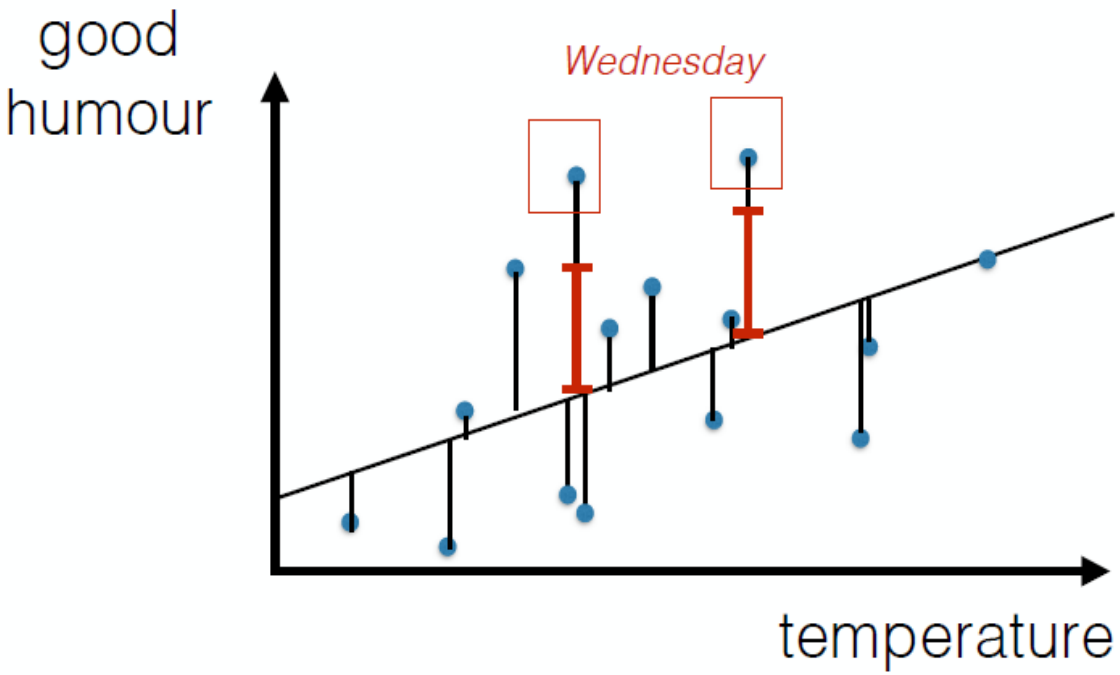


2K

$$\text{humour} = b_0 + b_1\text{temperature} + \text{Error}$$

Least squares

Adding covariates and R^2

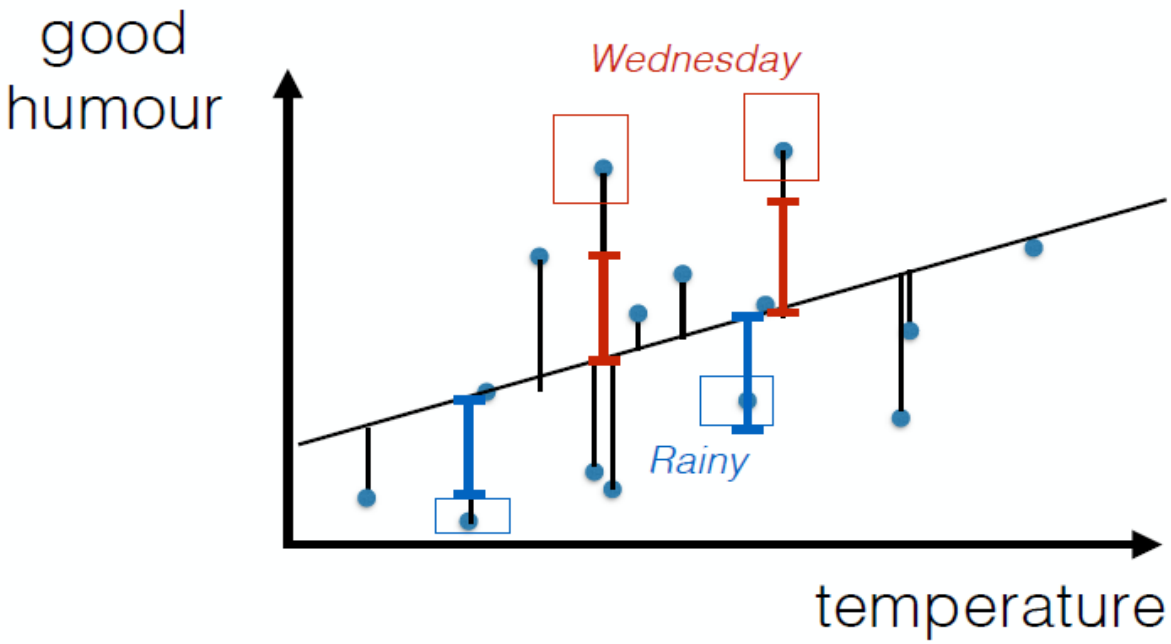


3K

$$\text{humour} = b_0 + b_1\text{temperature} + b_2\text{Wednesday} + \text{Error}$$

Least squares

Adding covariates and R^2

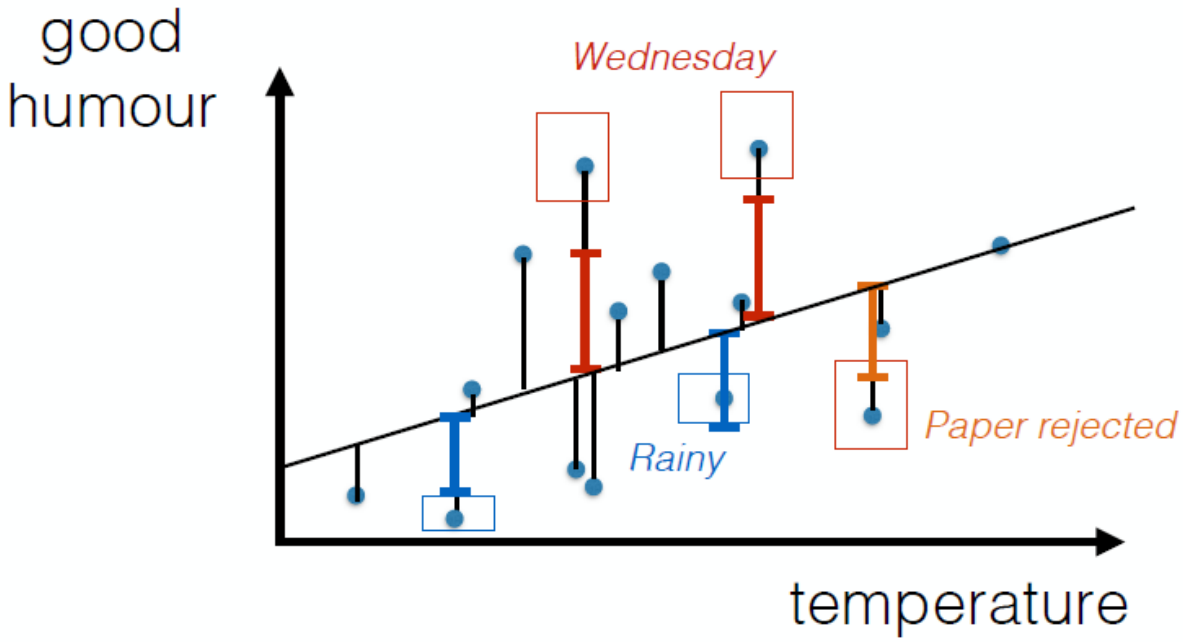


$$humour = b_0 + b_1 temperature + b_2 Wednesday + b_3 rain + Error$$

4K

Least squares

Adding covariates and R^2

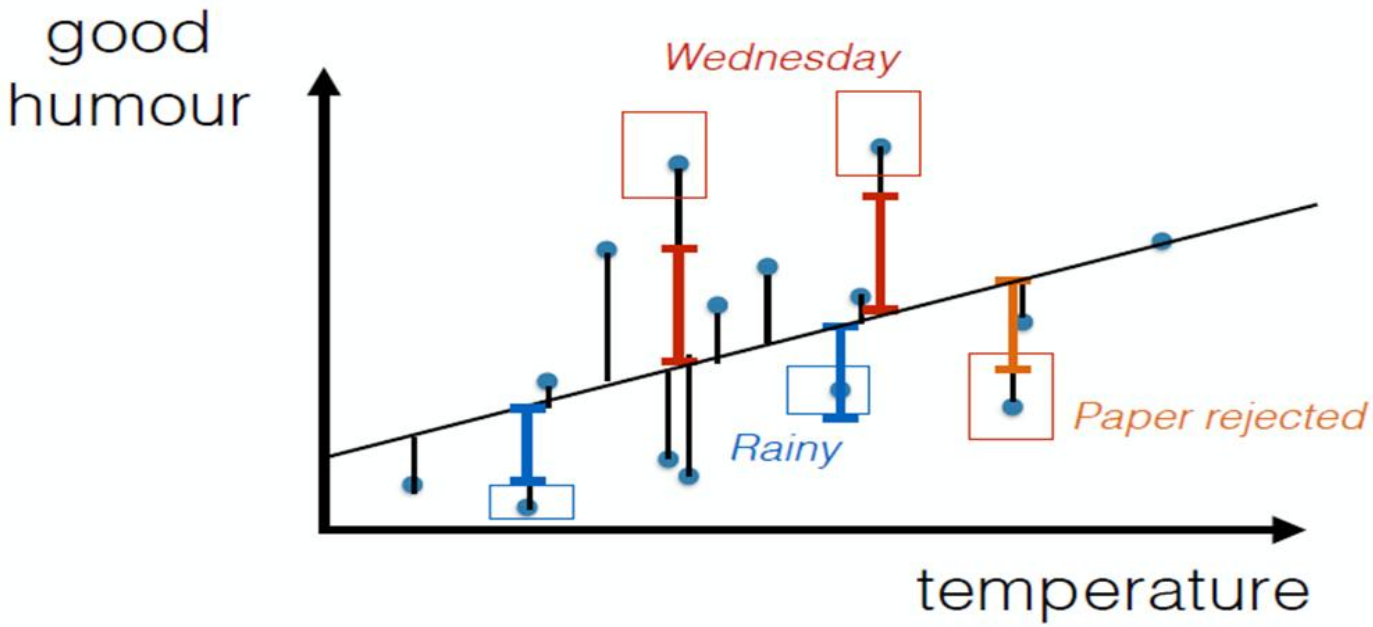


$$\text{humour} = b_0 + b_1\text{temperature} + b_2\text{Wednesday} + b_3\text{rain} + b_4\text{rejection} + \text{Error}$$

5K

Least squares

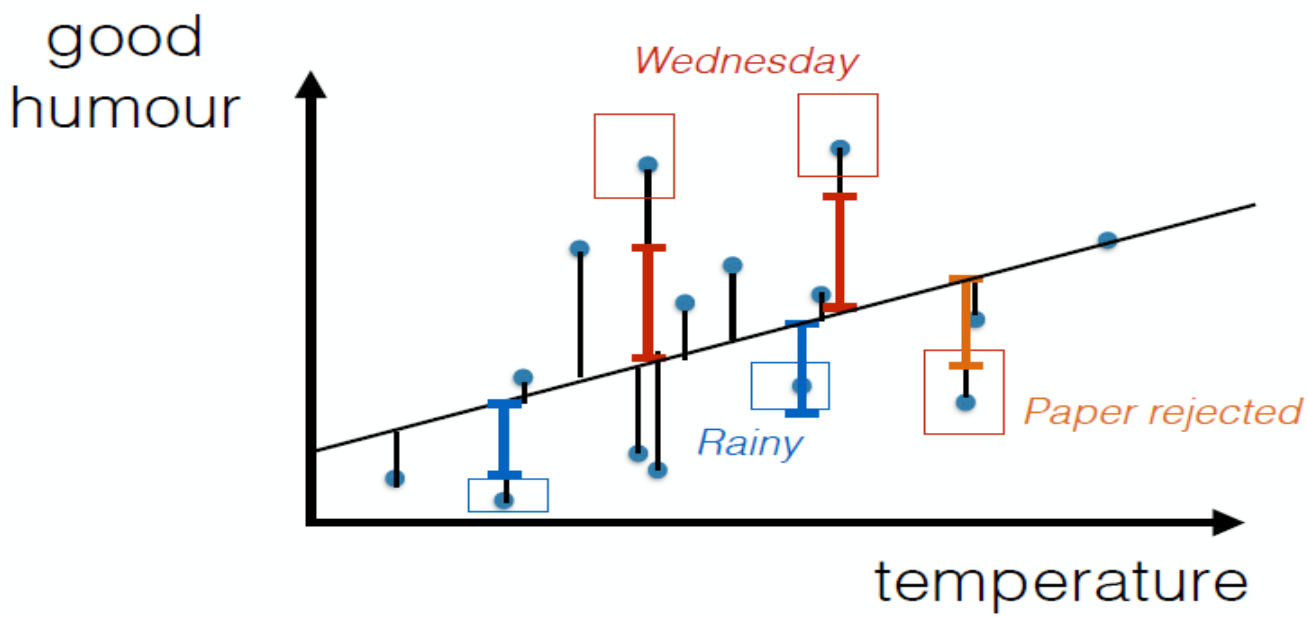
Adding covariates and R^2



Adding covariates almost always increases the R^2 .

Least squares

Adding covariates and R^2



Adding covariates almost always increases the R^2 - so a key question is when to stop.

What to choose?

Least square AIC

$$\mathrm{AIC} = N * ln(\frac{SS_e}{N}) + 2K$$

$\boldsymbol{N}$: *Number of observations*
$\boldsymbol{SS_e}$: *Sum square of errors*
$\boldsymbol{K}$: *Number of parameters*

The smaller the AIC the better

Least square AIC

More parameter is not always good

$$\mathrm{AIC} = N * ln(\frac{SS_e}{N}) + 2K$$

N: *Number of observations*
SS_e: *Sum square of errors*
K: *Number of parameters*

$$(AIC = -2\ln(L) + 2k)$$

The smaller the AIC the better

An example of model selection: *Bartonella spp.* in Madagascar rats

Epidemics 20 (2017) 56–66

Contents lists available at ScienceDirect

Epidemics

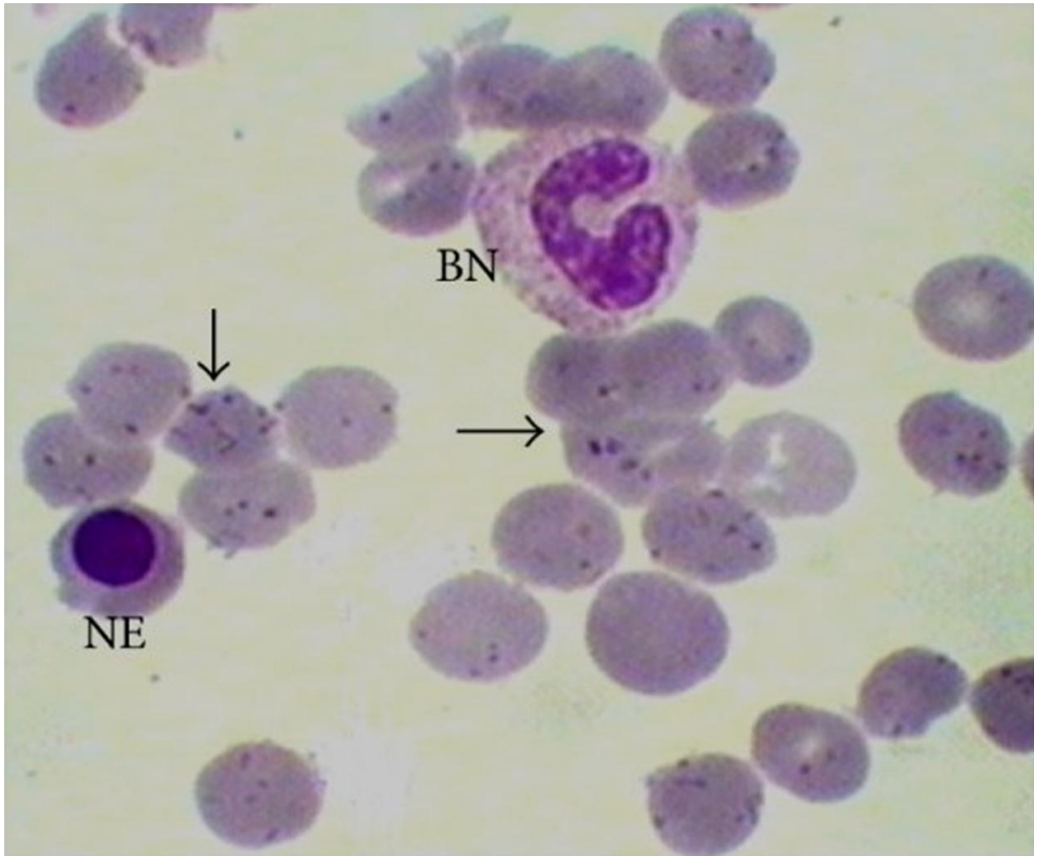
journal homepage: www.elsevier.com/locate/epidemics

Elucidating transmission dynamics and host-parasite-vector relationships for rodent-borne *Bartonella* spp. in Madagascar

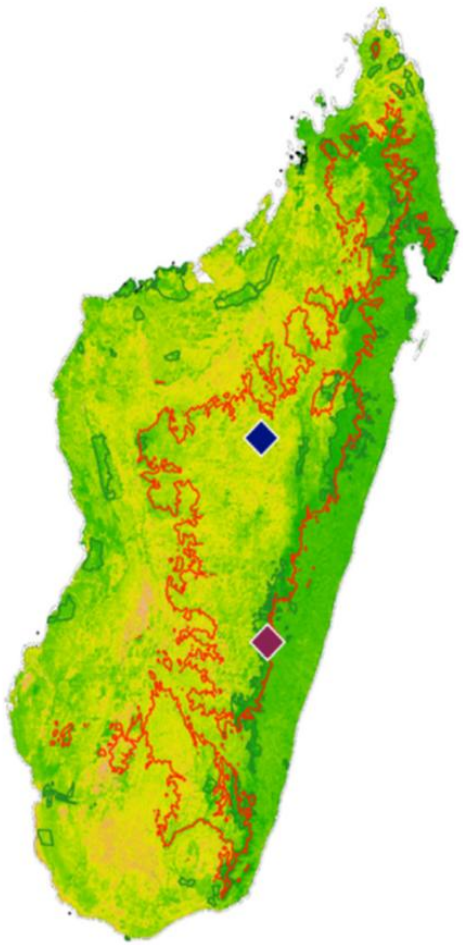
Cara E. Brook[a,*], Ying Bai[b], Emily O. Yu[a], Hafaliana C. Ranaivoson[c,d], Haewon Shin[e], Andrew P. Dobson[a], C. Jessica E. Metcalf[a,1], Michael Y. Kosoy[b,1], Katharina Dittmar[e,1]

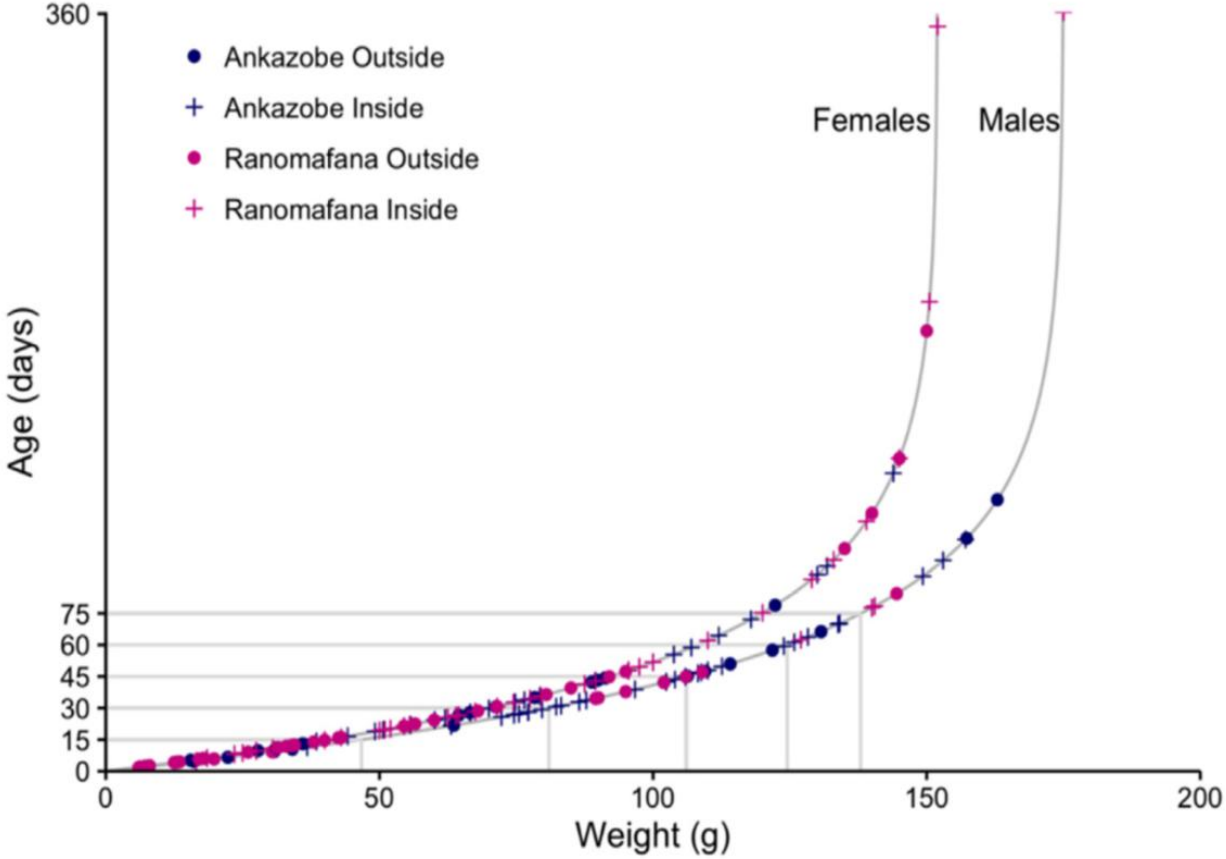
Bartonella spp.

- Persistent erythrocytic bacteria that are sometimes zoonotic
- Vectored by ticks, fleas, sand flies, mosquitoes
- Some species infect humans
  - *Bartonella bacilliformis* = Carrion's disease
  - *Bartonella henselae* = cat scratch fever
  - *Bartonella quintana* = trench fever



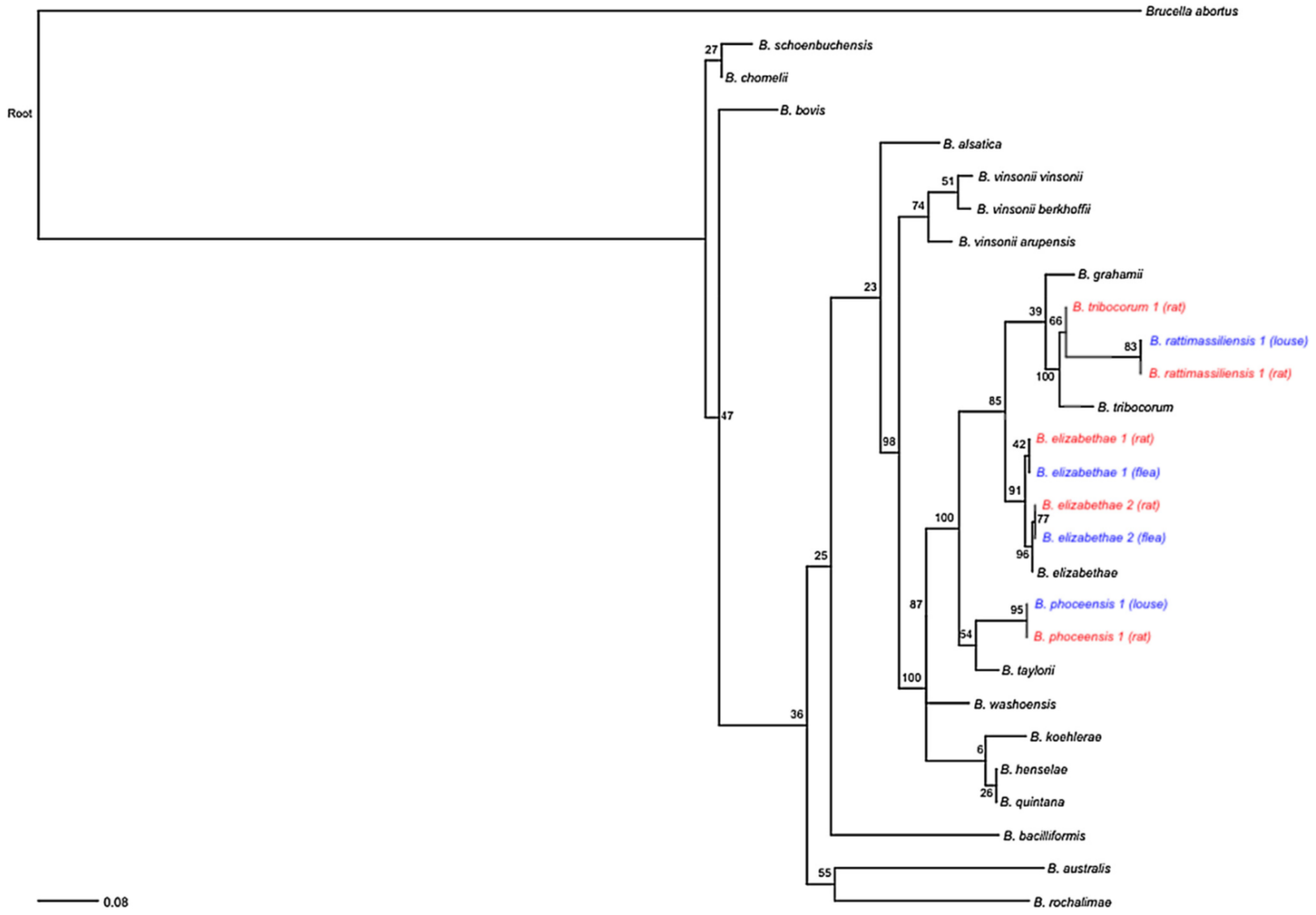
We first collected samples from rats from two sites in Madagascar.





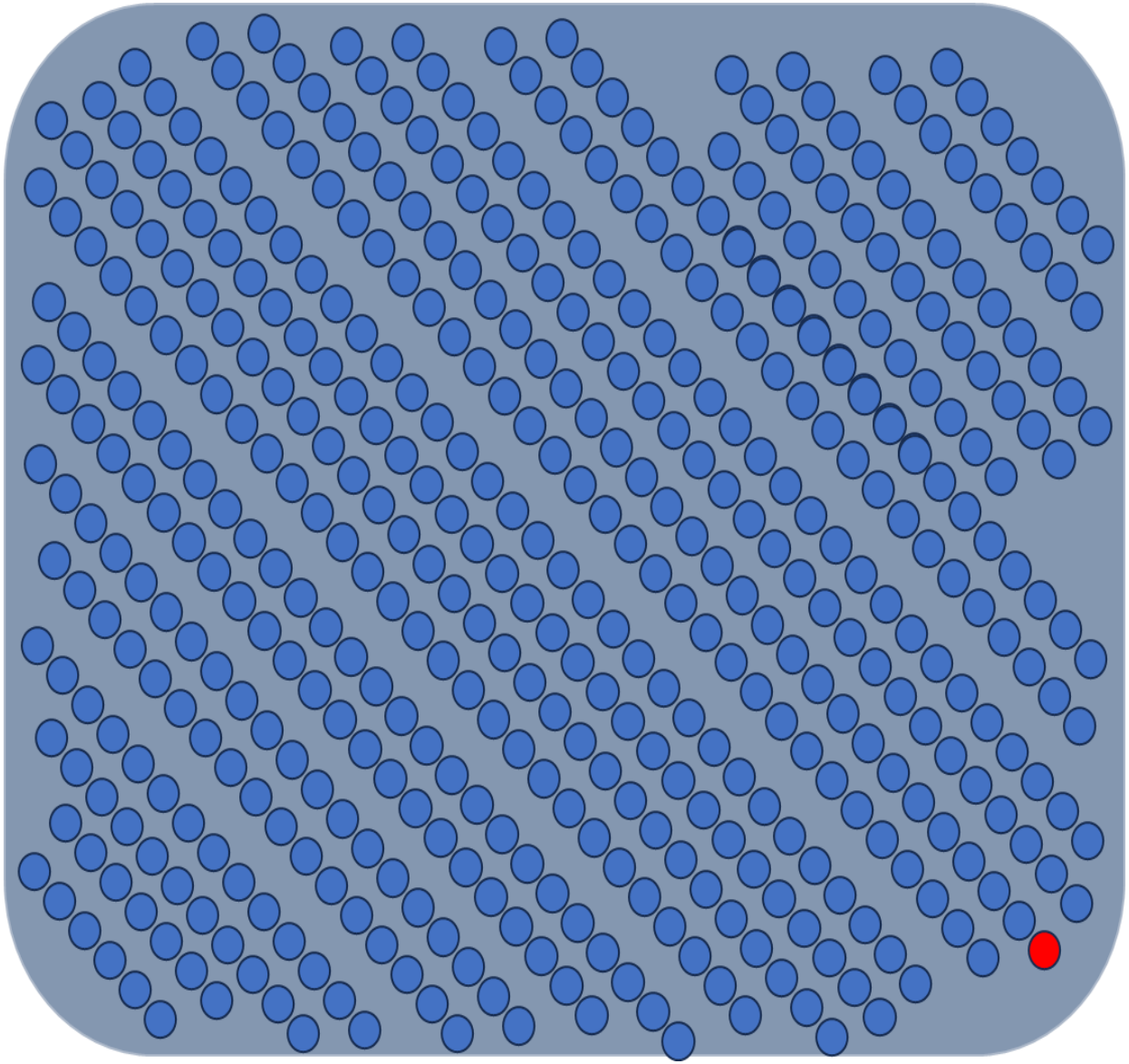
(Ricker 1979)

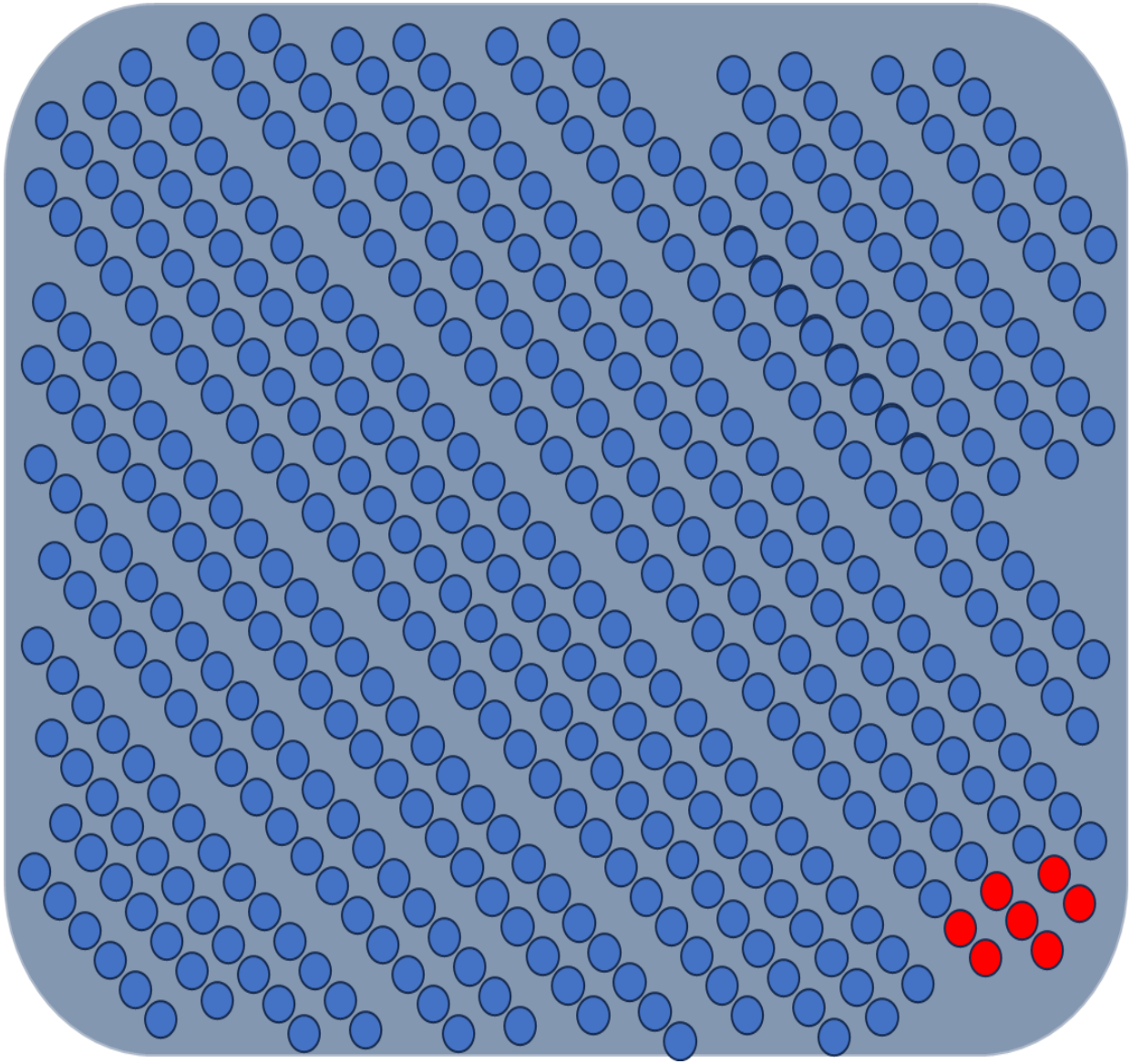
Statistically, we demonstrated an association between genotypes of *Bartonella* spp. found in rats and their ectoparasites.

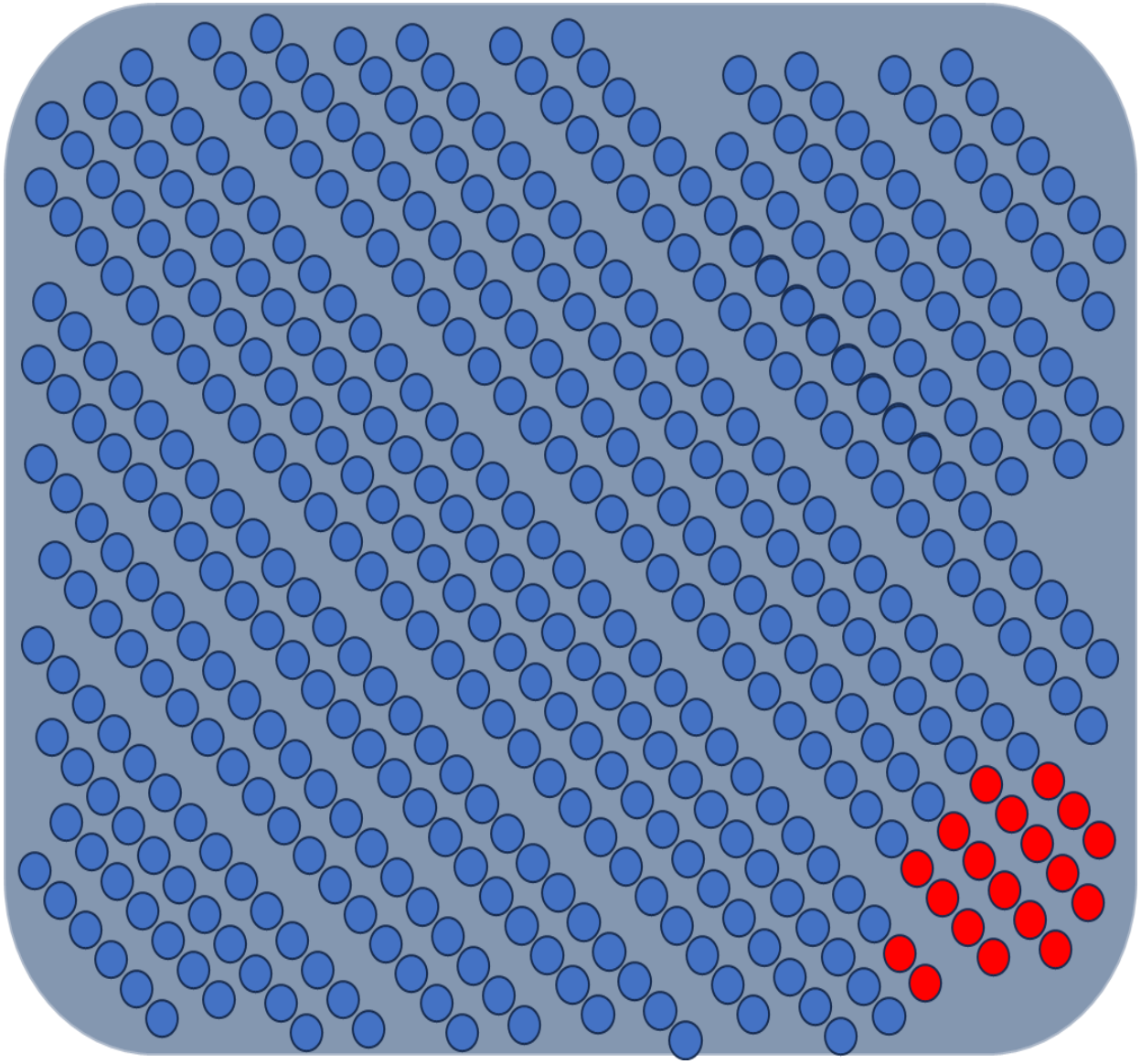


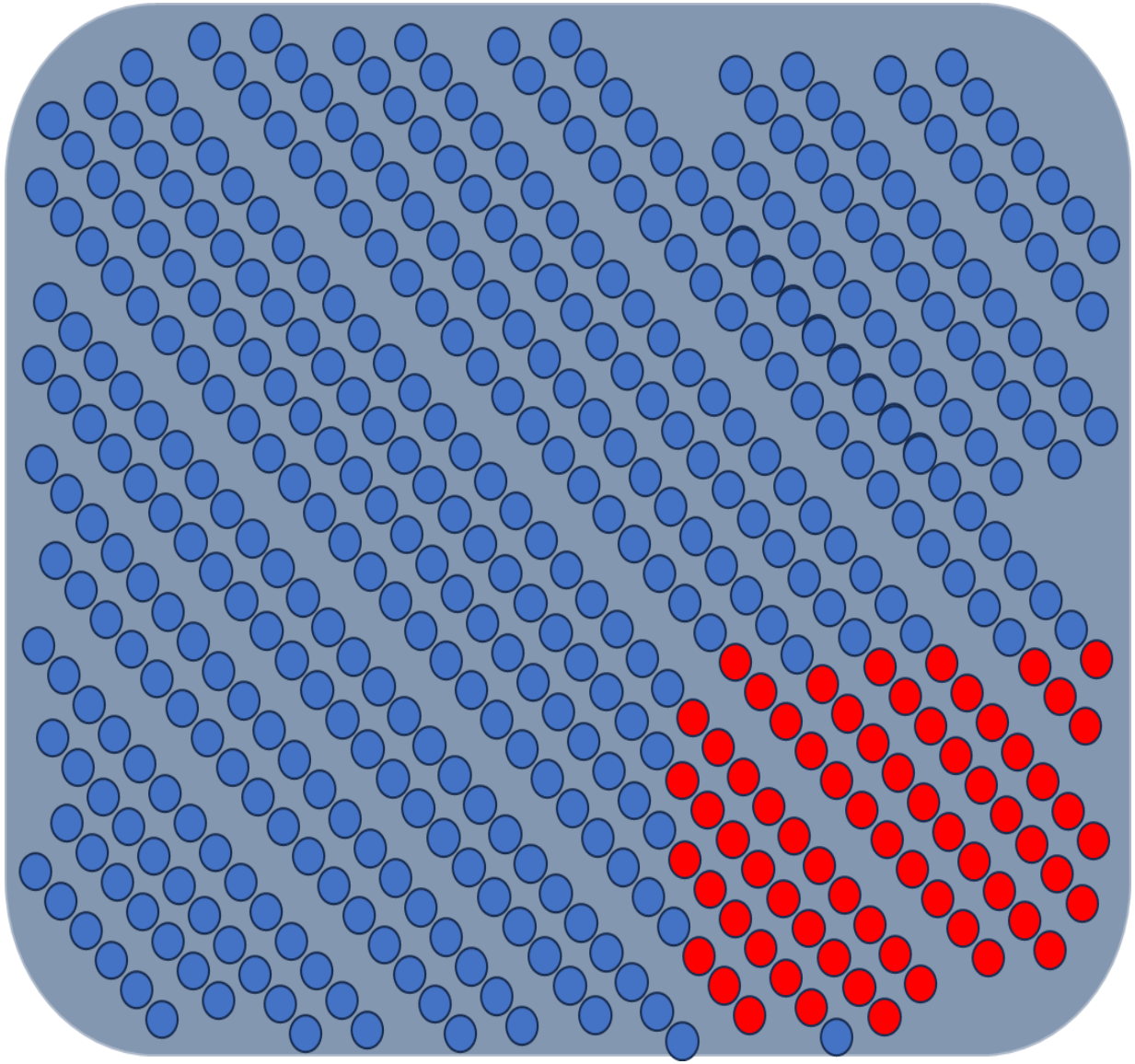
Then, we asked:
How does the rate of becoming infected vary with age?

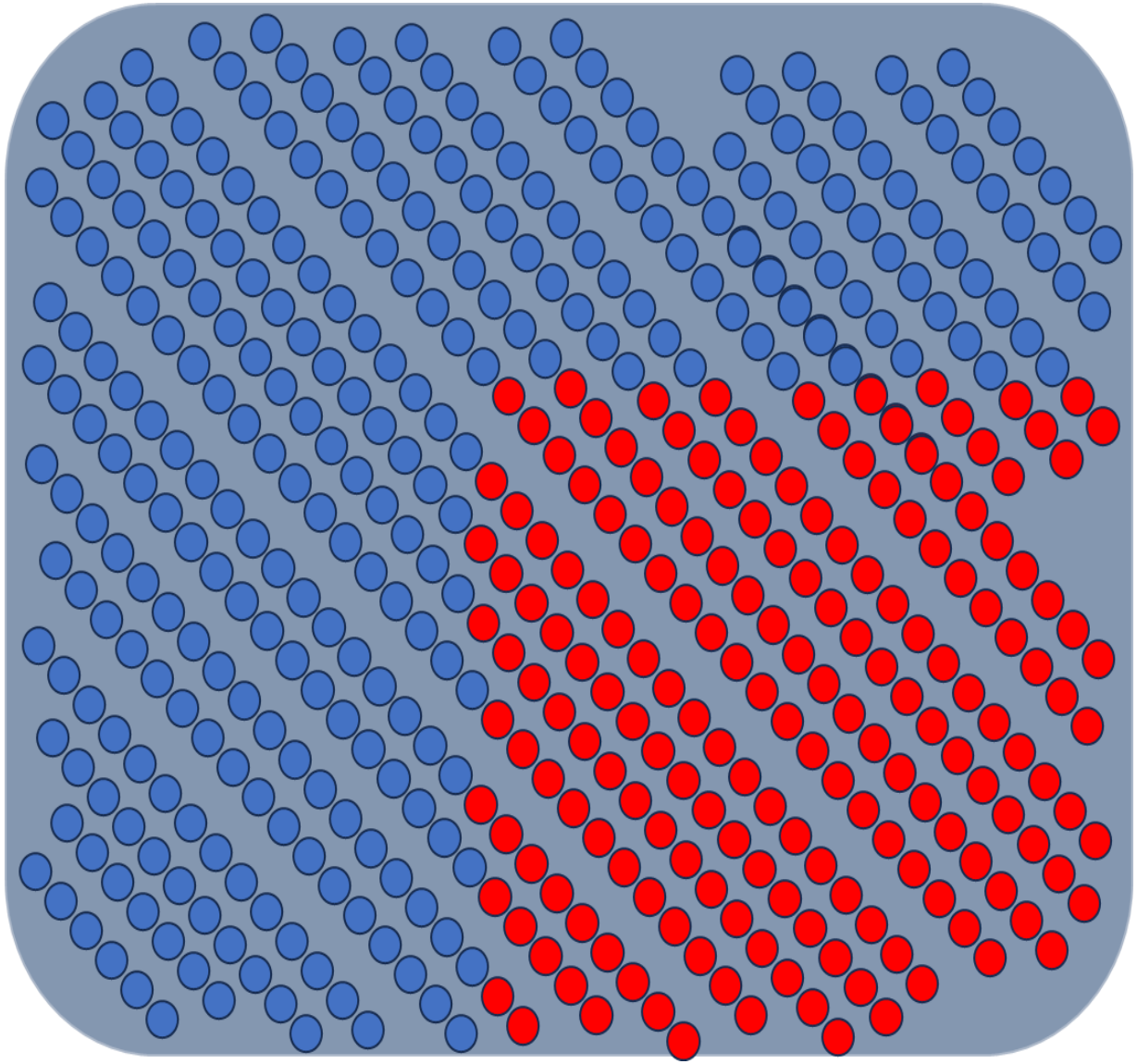
age	sick	age.class	sex
1.685375	0	1	M
2.011821	0	1	F
2.344301	0	1	F
2.611943	0	1	F
4.145697	0	1	F
4.319159	0	1	F
4.319159	0	1	F
4.493246	0	1	F
4.662362	0	1	M
4.784125	0	1	M









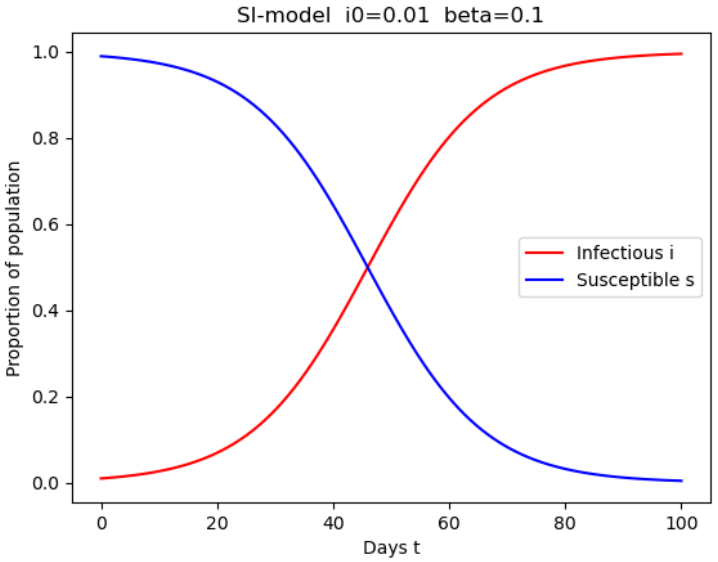


Simple SI model

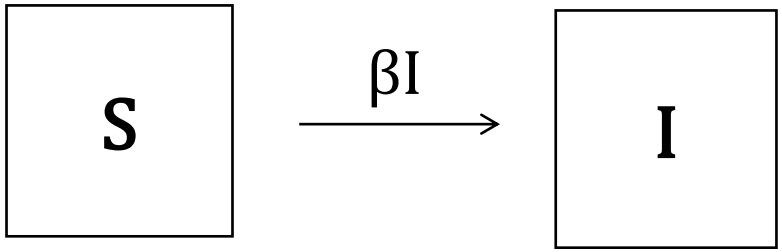
N

S

I

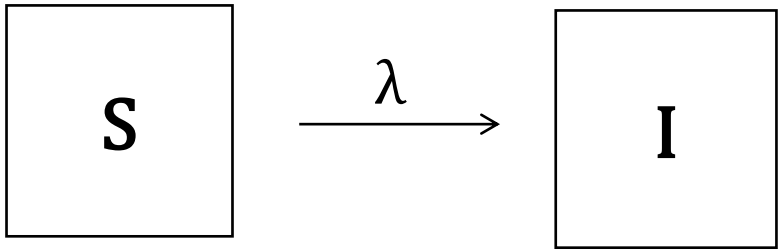


Age-prevalence data allows for powerful inference into the dynamics of pathogen transmission.



for a persistent, non-immunizing infection

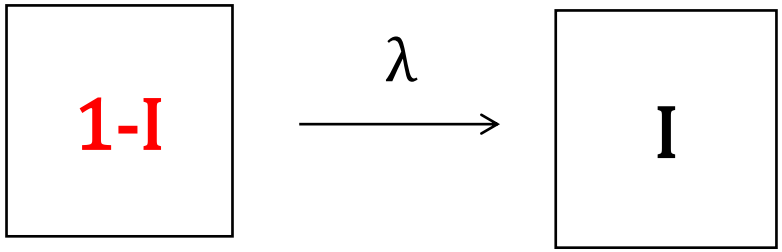
Age-prevalence data allows for powerful inference into the dynamics of pathogen transmission.



where λ, the force of infection, is the per capita rate at which susceptible hosts become infected

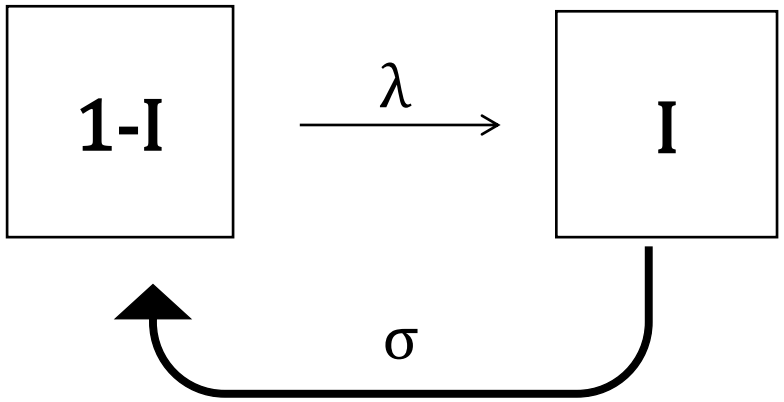
Age-prevalence data allows for powerful inference into the dynamics of pathogen transmission.

with a persistent infection, we can assume that, if not infected, you must be susceptible….



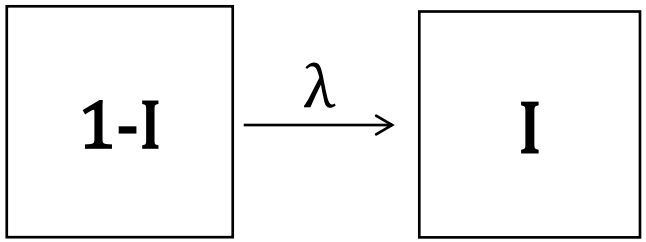
where λ, the force of infection, is the per capita rate at which susceptible hosts become infected

Age-prevalence data allows for powerful inference into the dynamics of pathogen transmission.

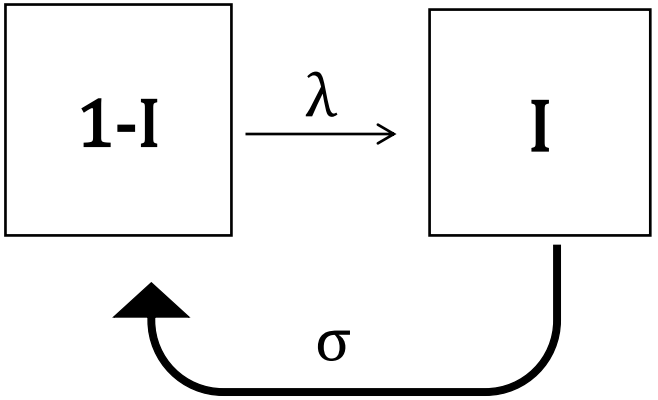


and σ is the rate of recovery from infection

Age-prevalence data allows for powerful inference into the dynamics of pathogen transmission.

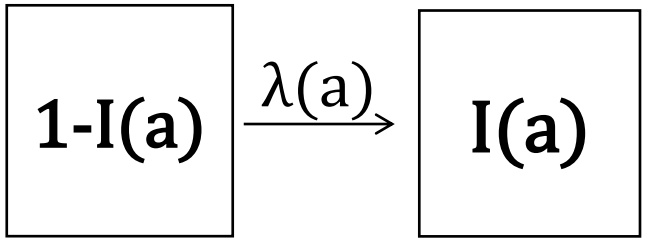


$$\frac{dI}{dt} = \lambda(1-I)$$

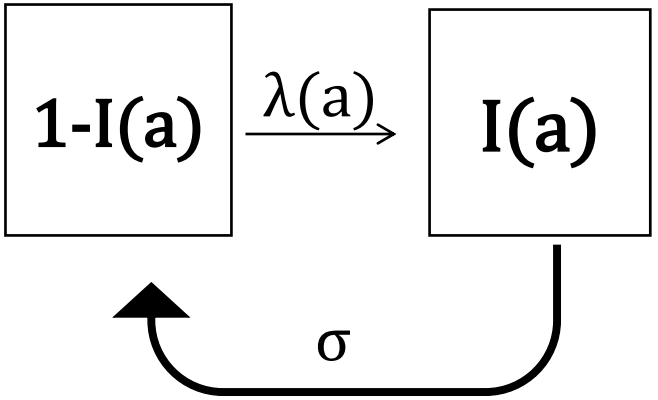


$$\frac{dI}{dt} = \lambda(1-I) - \sigma I$$

Age-prevalence data allows for powerful inference into the dynamics of pathogen transmission.

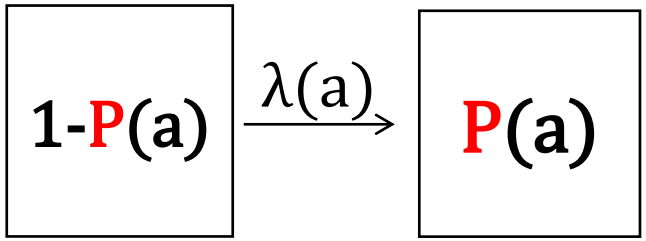


$$\frac{dI(a)}{da} = \lambda(a)\left(1 - I(a)\right)$$

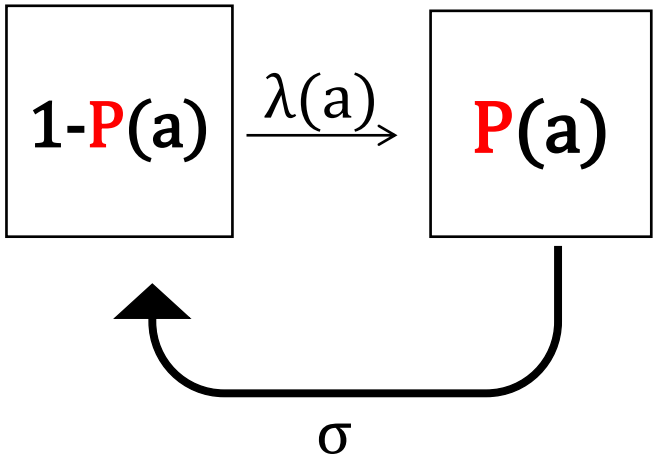


$$\frac{dI(a)}{da} = \lambda(a)\left(1 - I(a)\right) - \sigma I(a)$$

Age-prevalence data allows for powerful inference into the dynamics of pathogen transmission.



$$\frac{d\,P(a)}{da} = \lambda(a)\,(1 - P(a))$$



$$\frac{d\,P(a)}{da} = \lambda(a)\,(1 - P(a)) - \sigma(a)P(a))$$

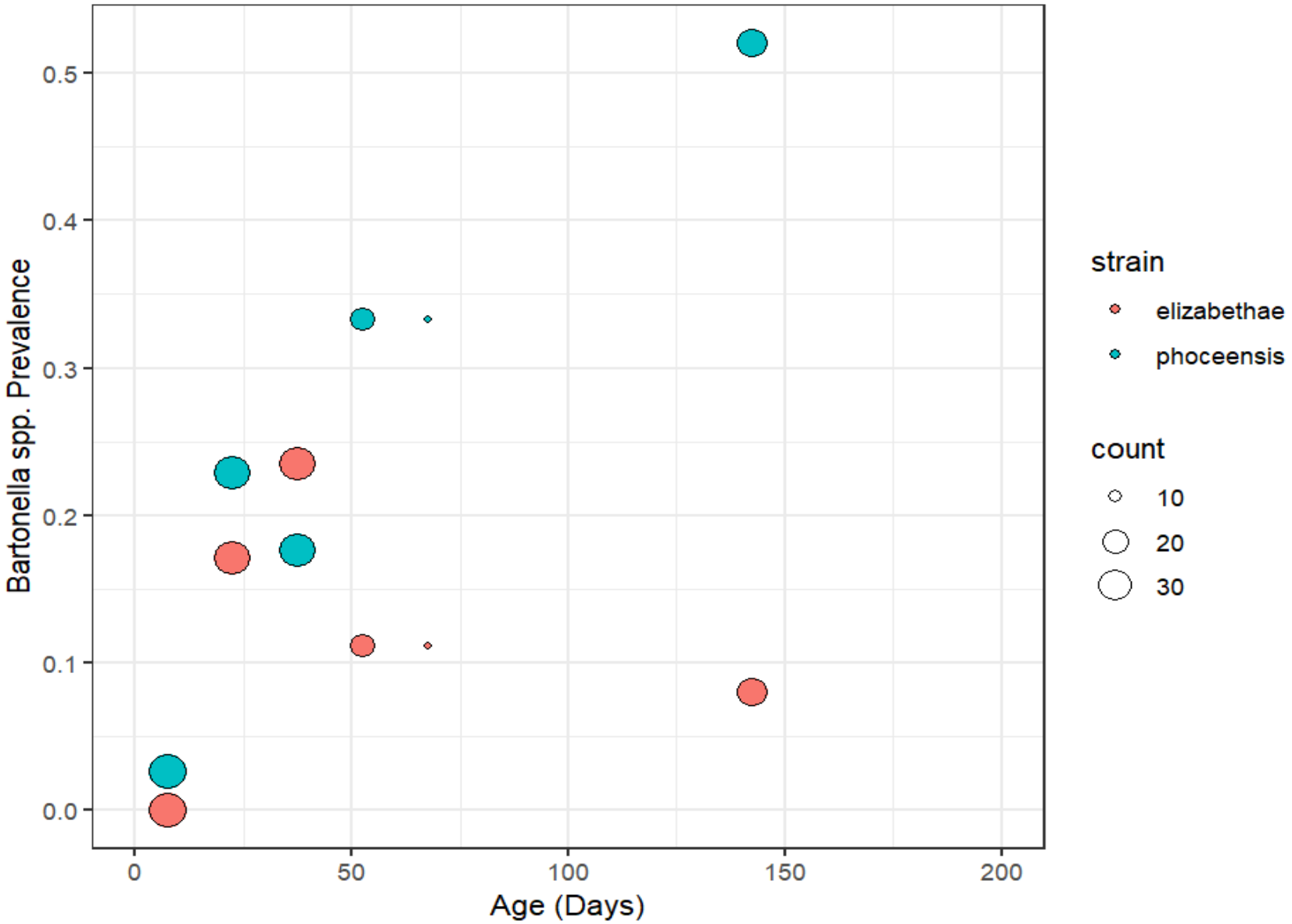
Compare using ***AIC = 2K - 2ln(L)***

similar techniques can also be applied to age-seroprevalence data for immunizing infections

Let’s see which model works best for your data!

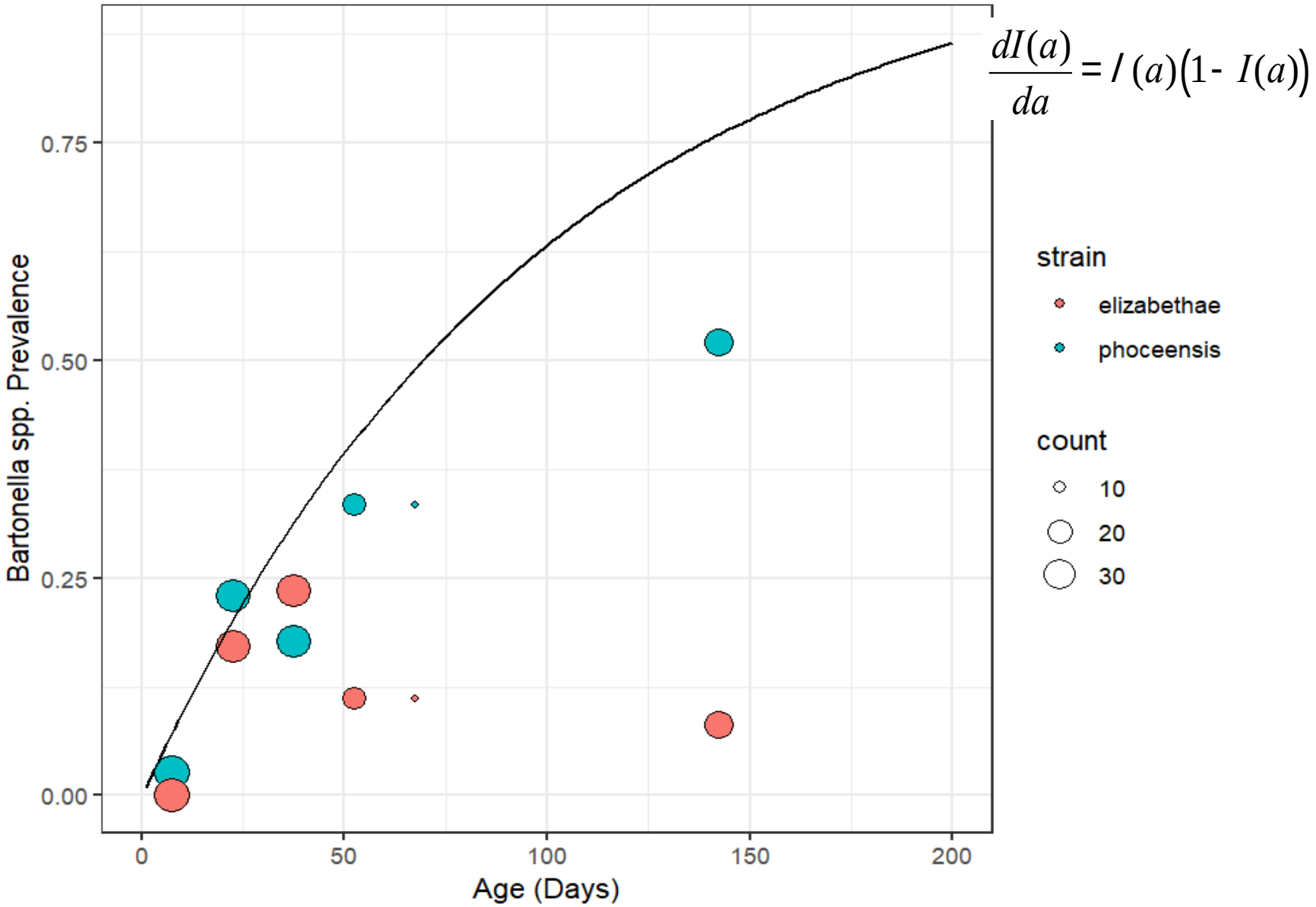
Look at the data !

Jereo aloha hoe manao ahoana



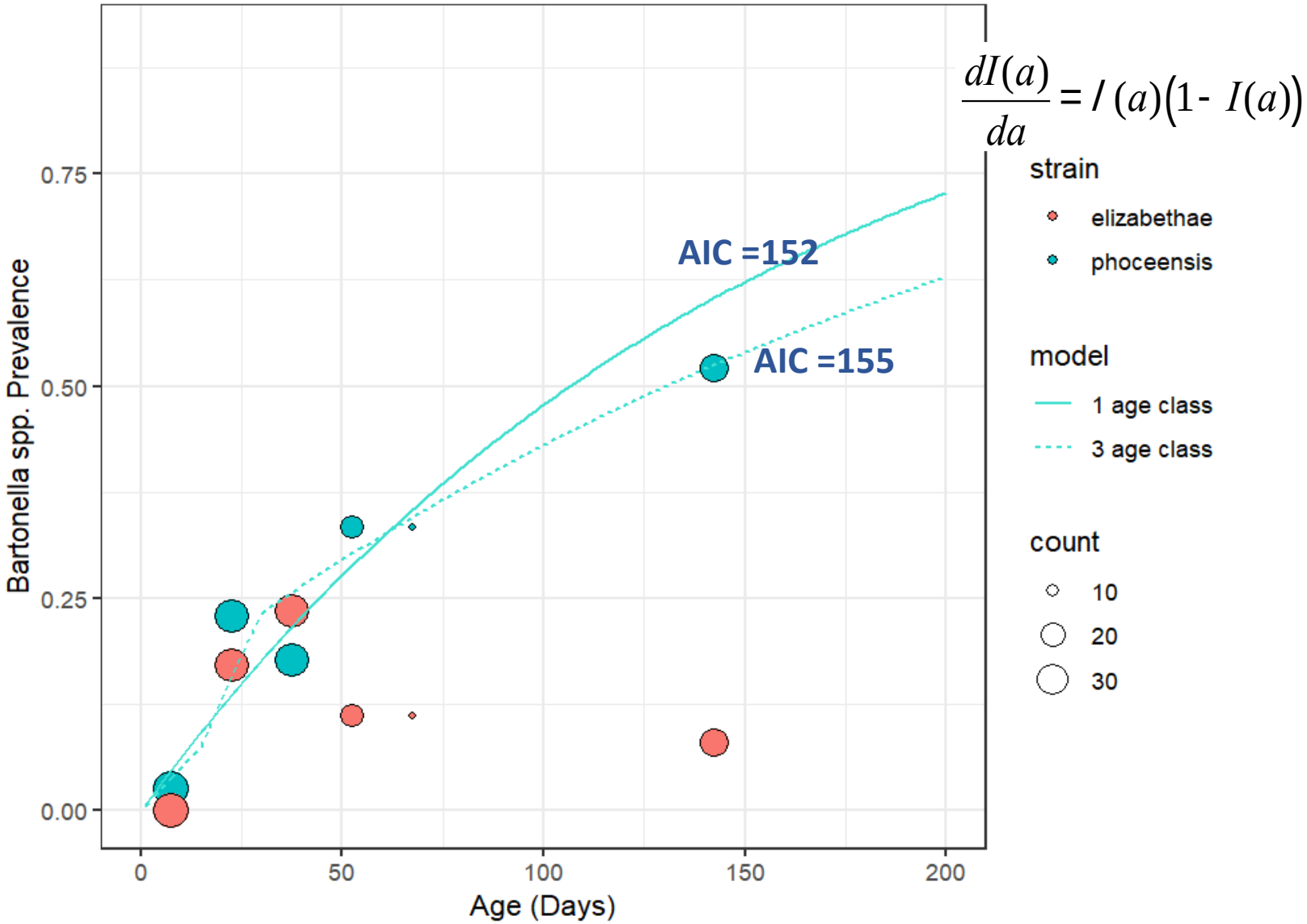
Try the model!

Andramo kely

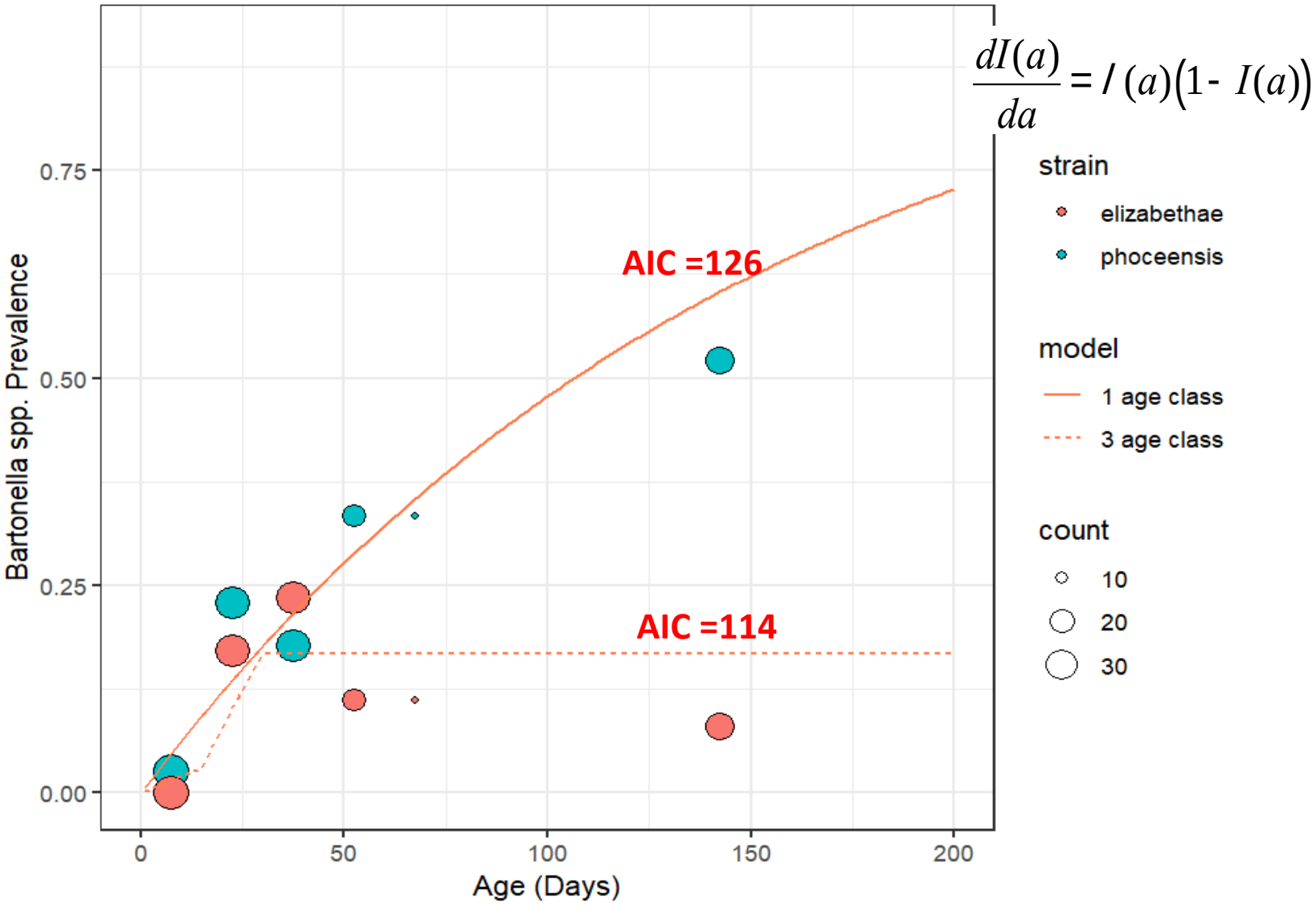


Keep trying!

Alô fô !

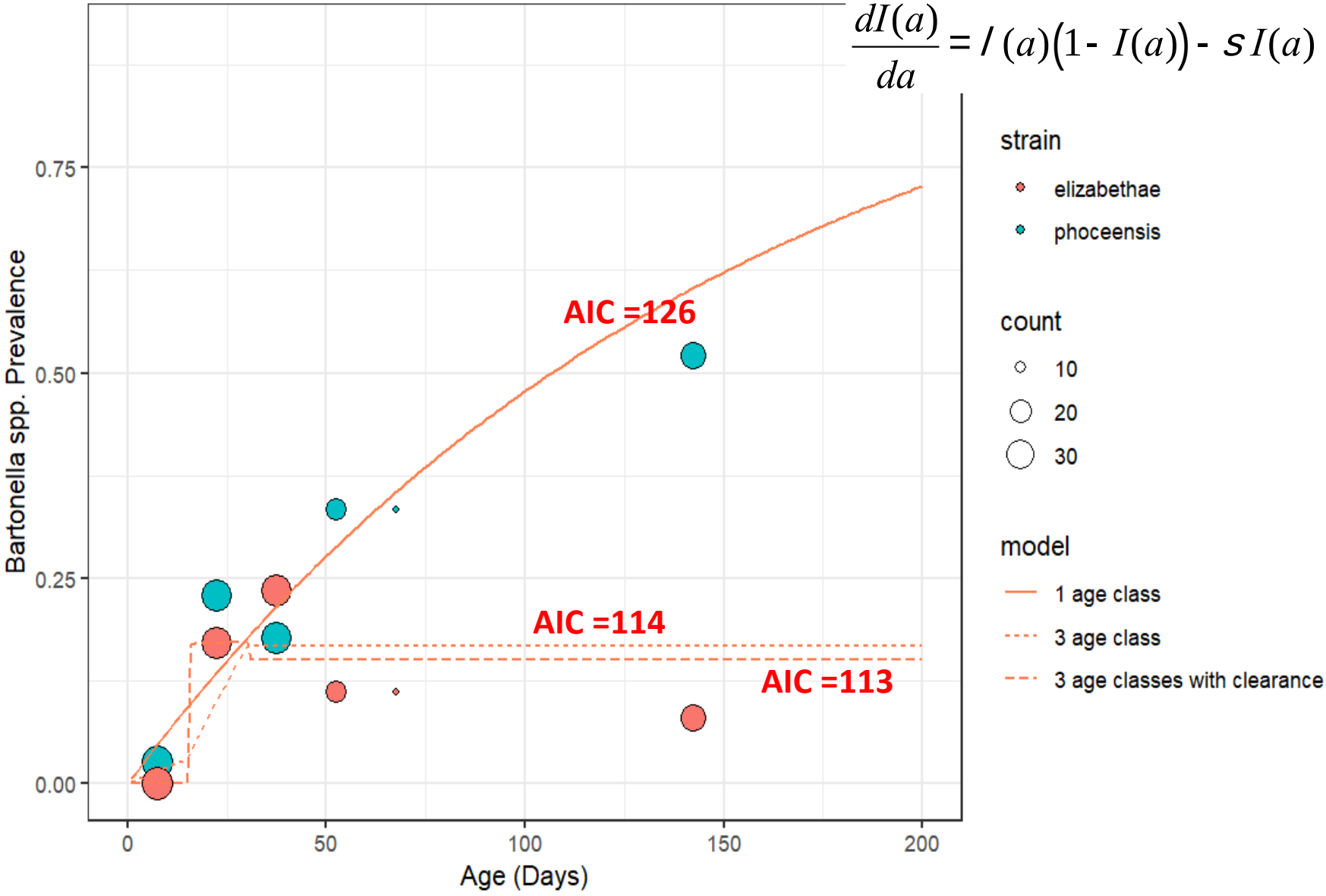


Keep trying!

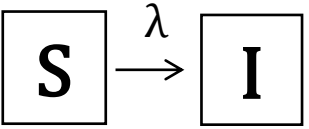

$\frac{dI(a)}{da} = I(a)(1 - I(a))$
AIC =126
AIC =114
Bartonella spp. Prevalence
0.00
0.25
0.50
0.75
0
50
100
150
200
Age (Days)
strain
elizabethae
phoceensis
model
1 age class
3 age class
count
10
20
30

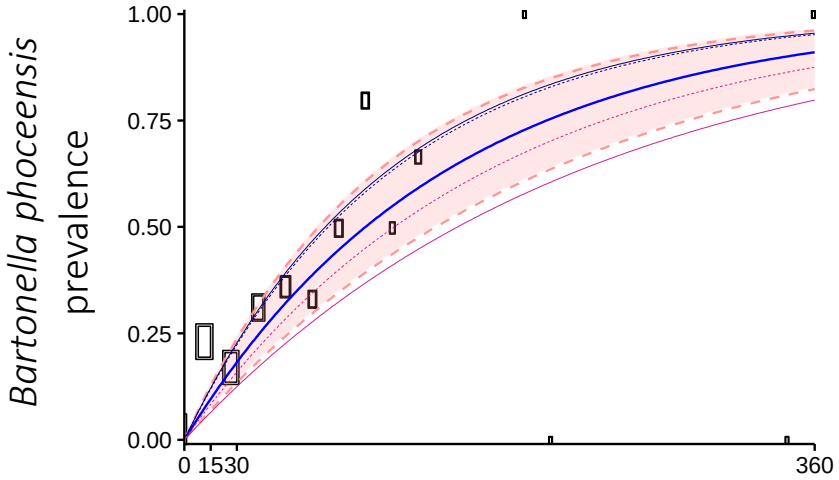
Keep trying!

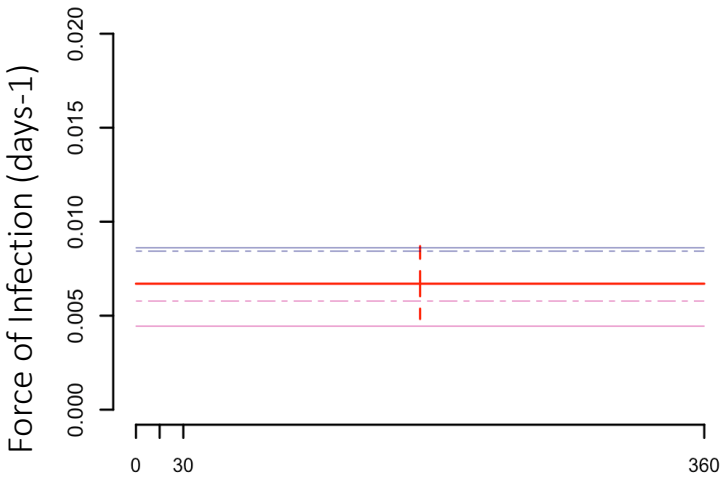
$$\frac{dI(a)}{da} = l(a)\left(1 - I(a)\right) - sI(a)$$

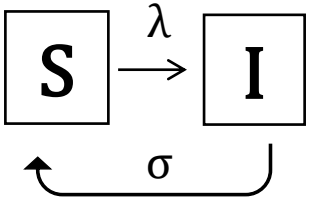


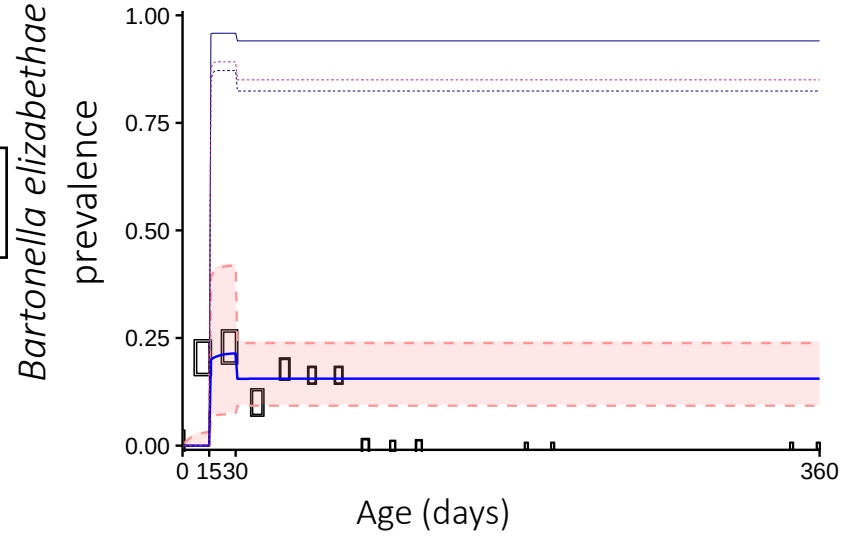
We found that an **SI model** offered the best fit to ***B. phoceensis*** data while the **SIS model** offered the best fit to the ***B. elizabethae*** data.

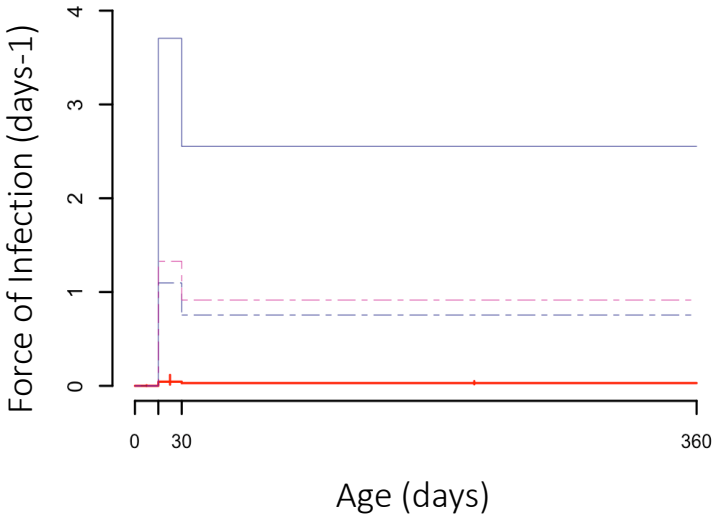




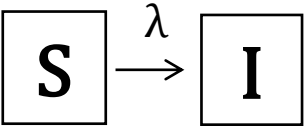


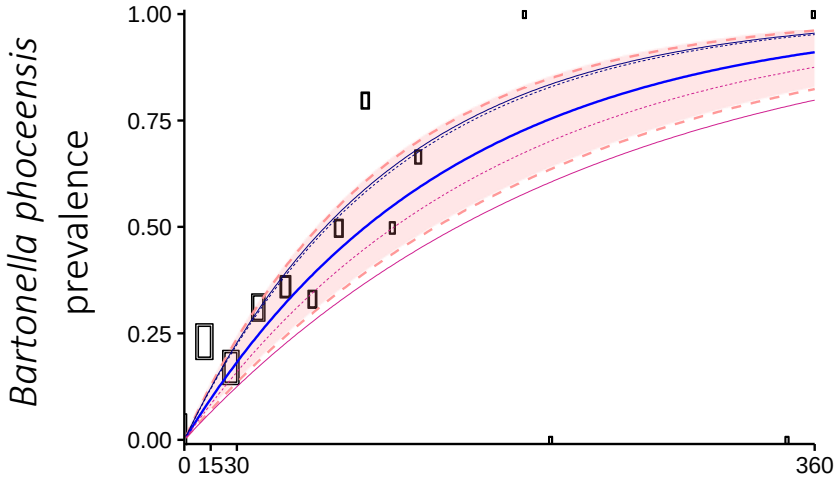


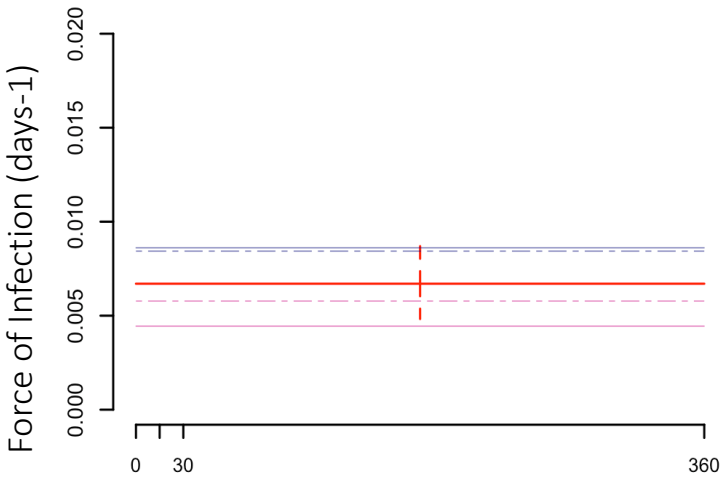


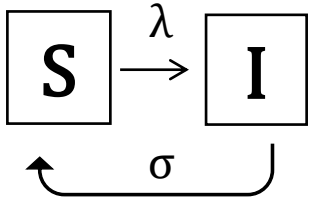


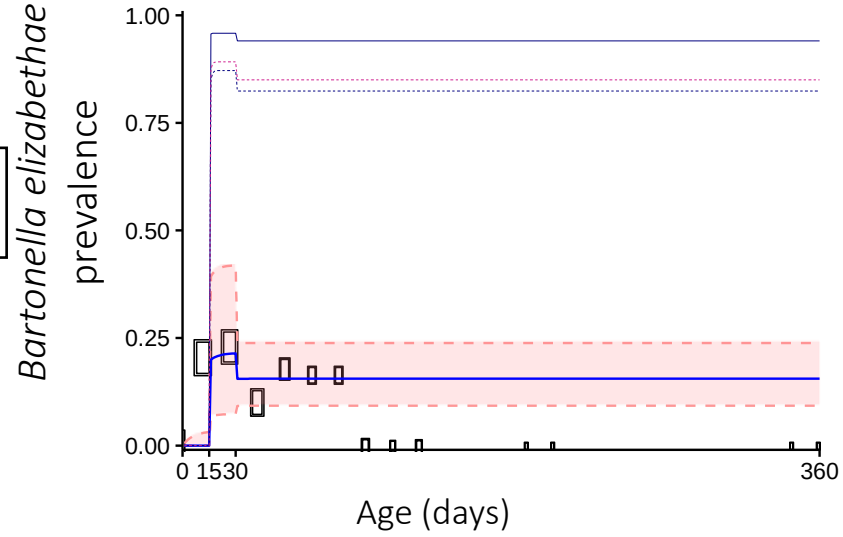
The age-structured FOI identifies age cohorts most influential in an epidemic. Juveniles showed the highest FOI.

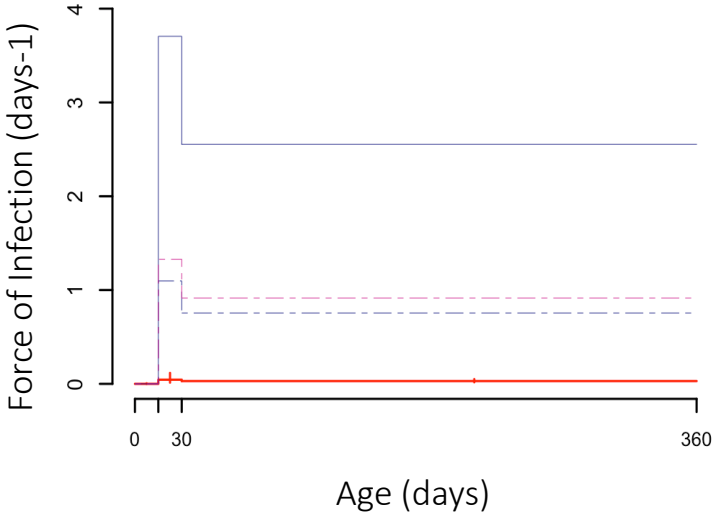


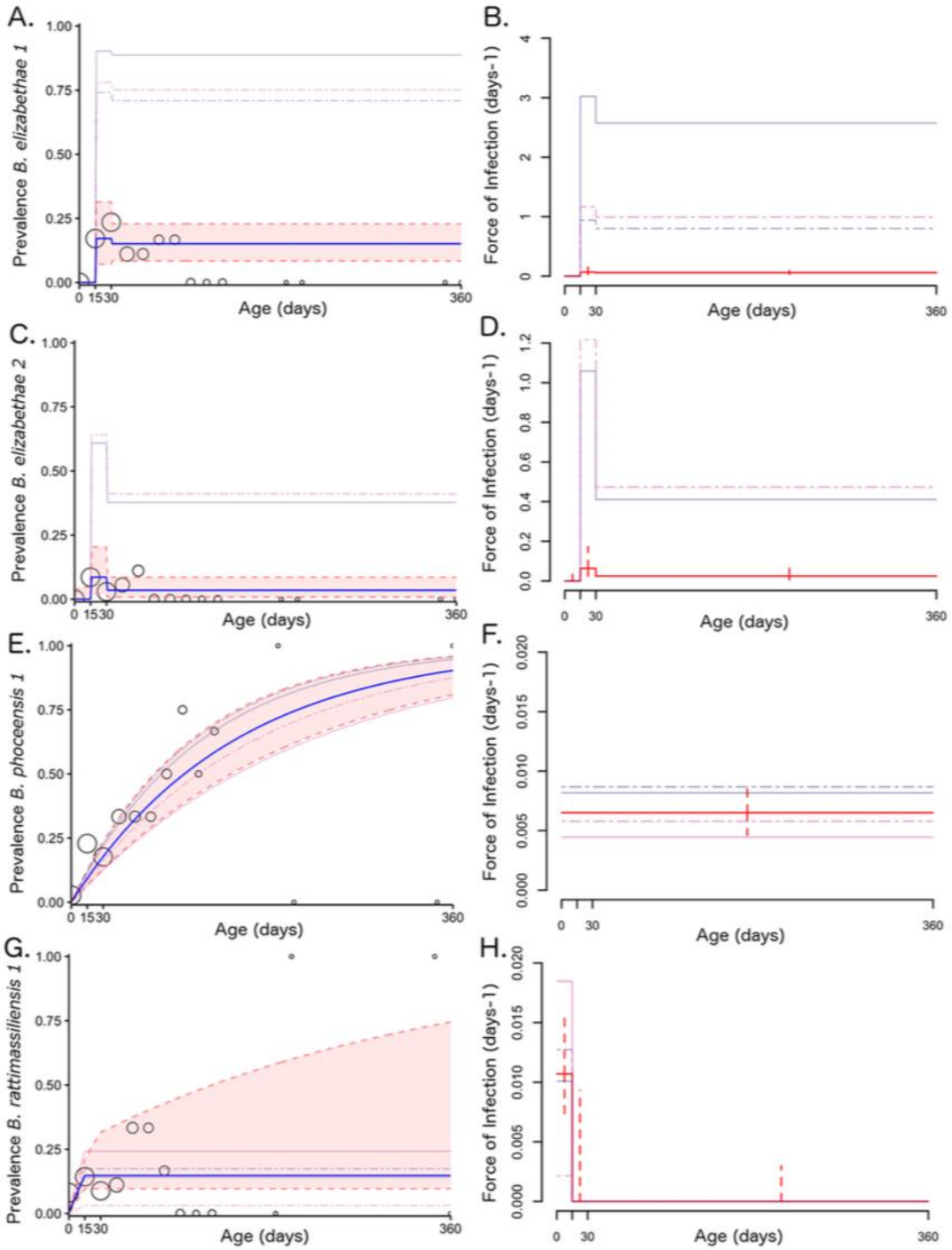











A.
Prevalence B. elizabethae 1
1.00
0.75
0.50
0.25
0.00
0 15 30
360
Age (days)
B.
Force of Infection (days-1)
4
3
2
1
0
0
30
360
Age (days)
C.
Prevalence B. elizabethae 2
1.00
0.75
0.50
0.25
0.00
0 15 30
360
Age (days)
D.
Force of Infection (days-1)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
30
360
Age (days)
E.
Prevalence B. phoceensis 1
1.00
0.75
0.50
0.25
0.00
0 15 30
360
Age (days)
F.
Force of Infection (days-1)
0.020
0.015
0.010
0.005
0.000
0
30
360
Age (days)
G.
Prevalence B. rattimassiliensis 1
1.00
0.75
0.50
0.25
0.00
0 15 30
360
Age (days)
H.
Force of Infection (days-1)
0.020
0.015
0.010
0.005
0.000
0
30
360
Age (days)