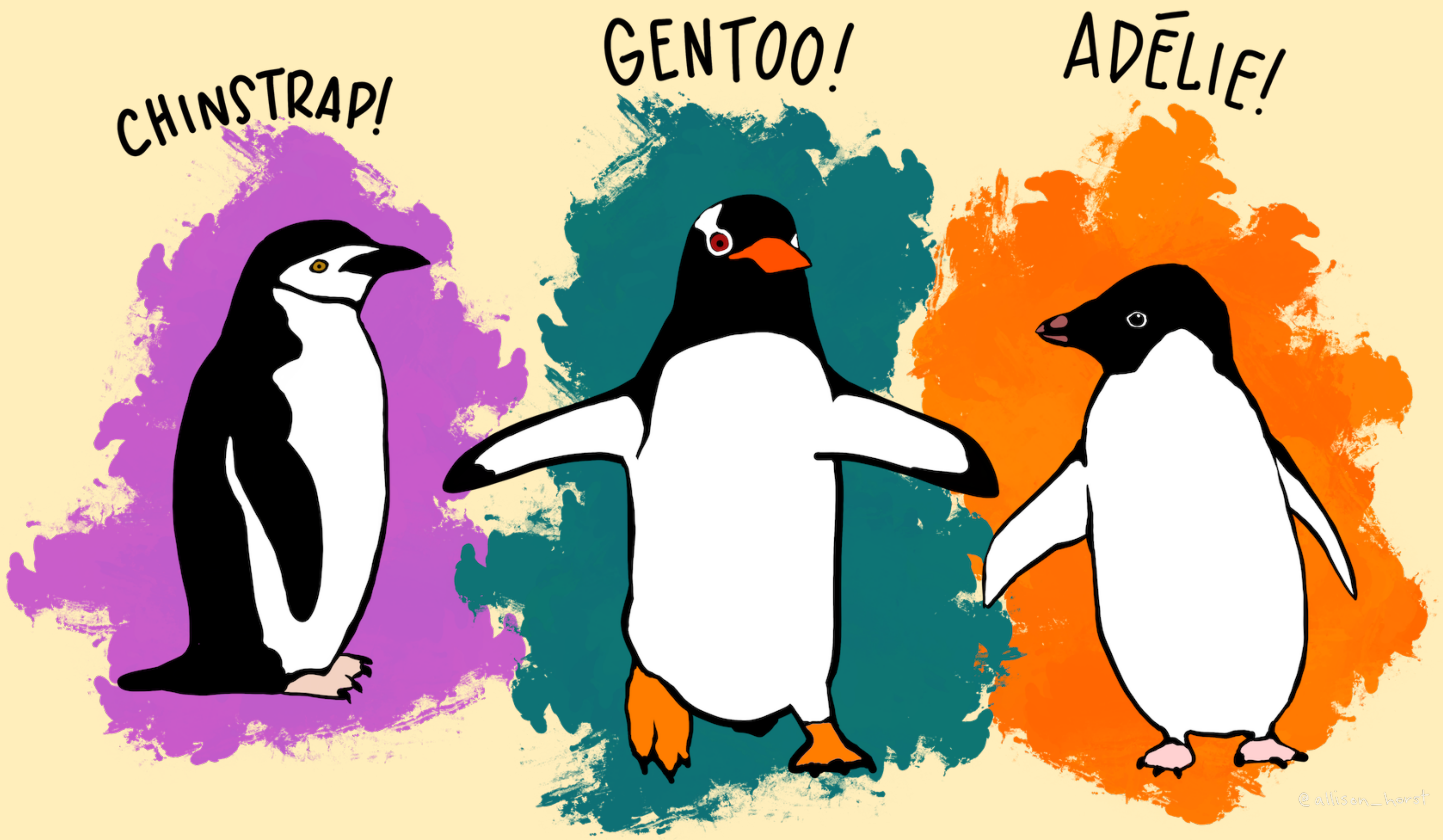
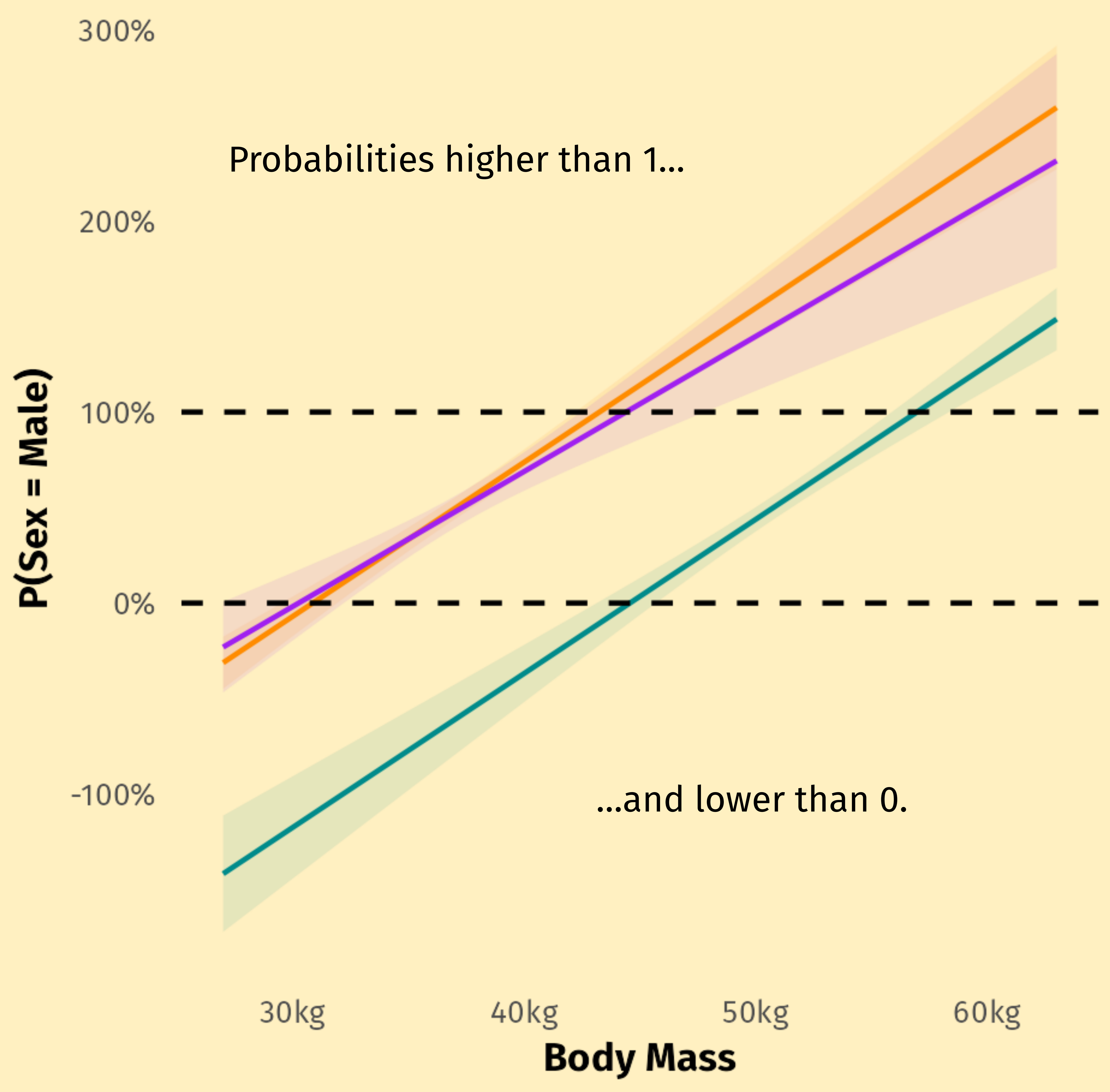
What are Generalised linear models?

Aleš Vomáčka 2023

Linear regression is pretty cool…
… but sometimes it’s not enough

Well, this didn't work



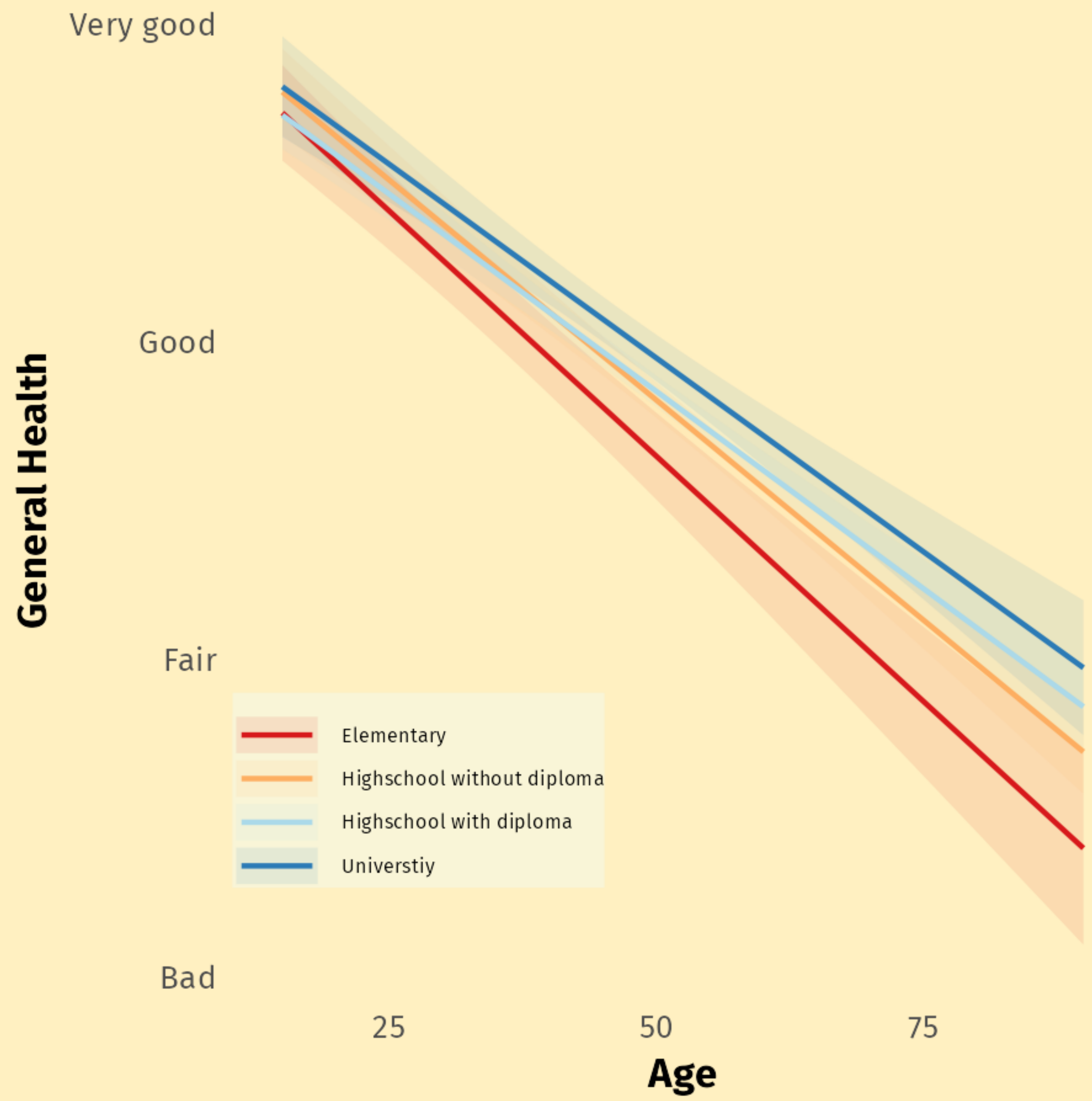


Sometimes it looks good...

- What is your general level of health?

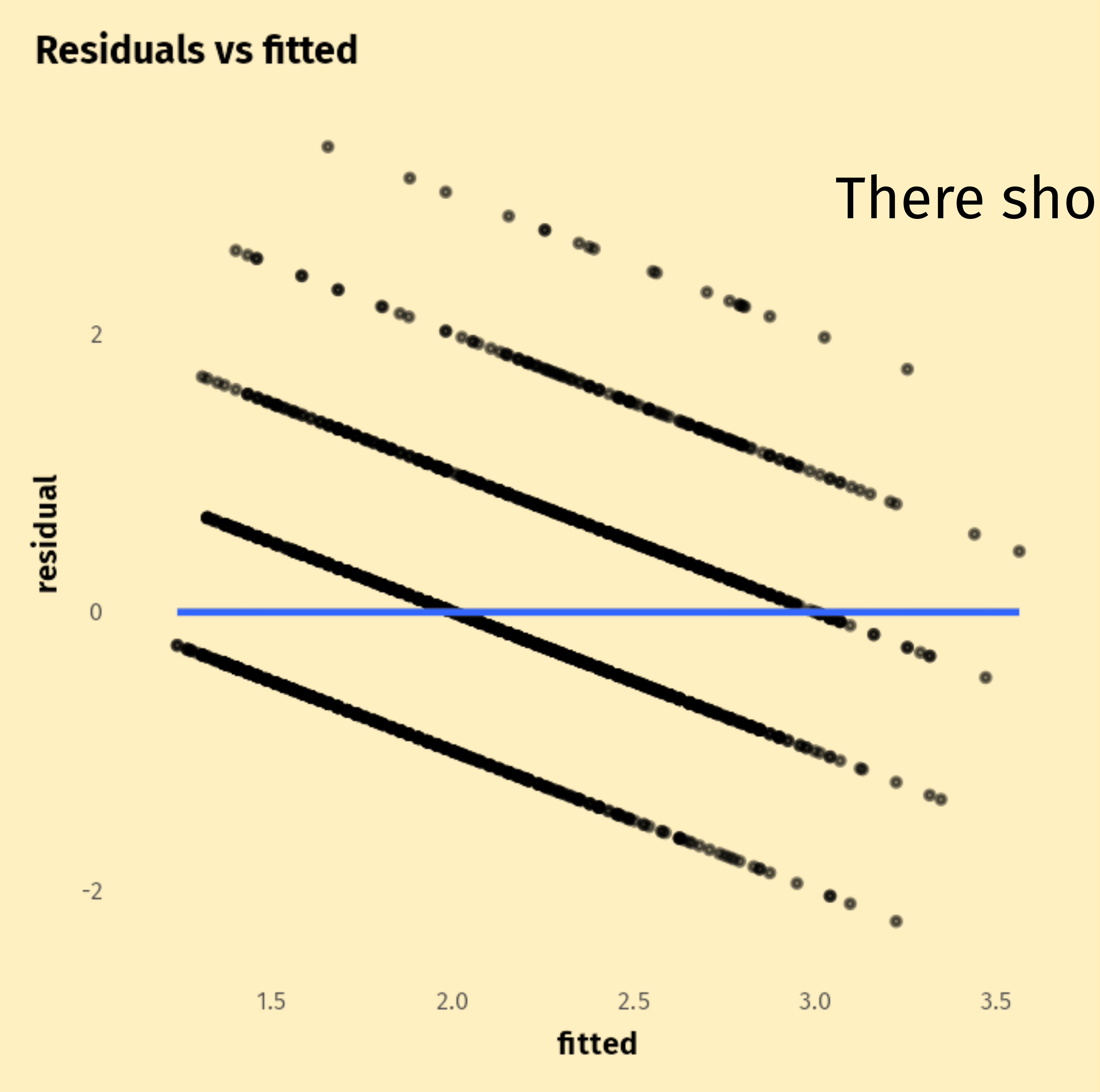
1. Very good
2. Good
3. Fair
4. Bad
5. Very Bad

(European Social Survey 2020)



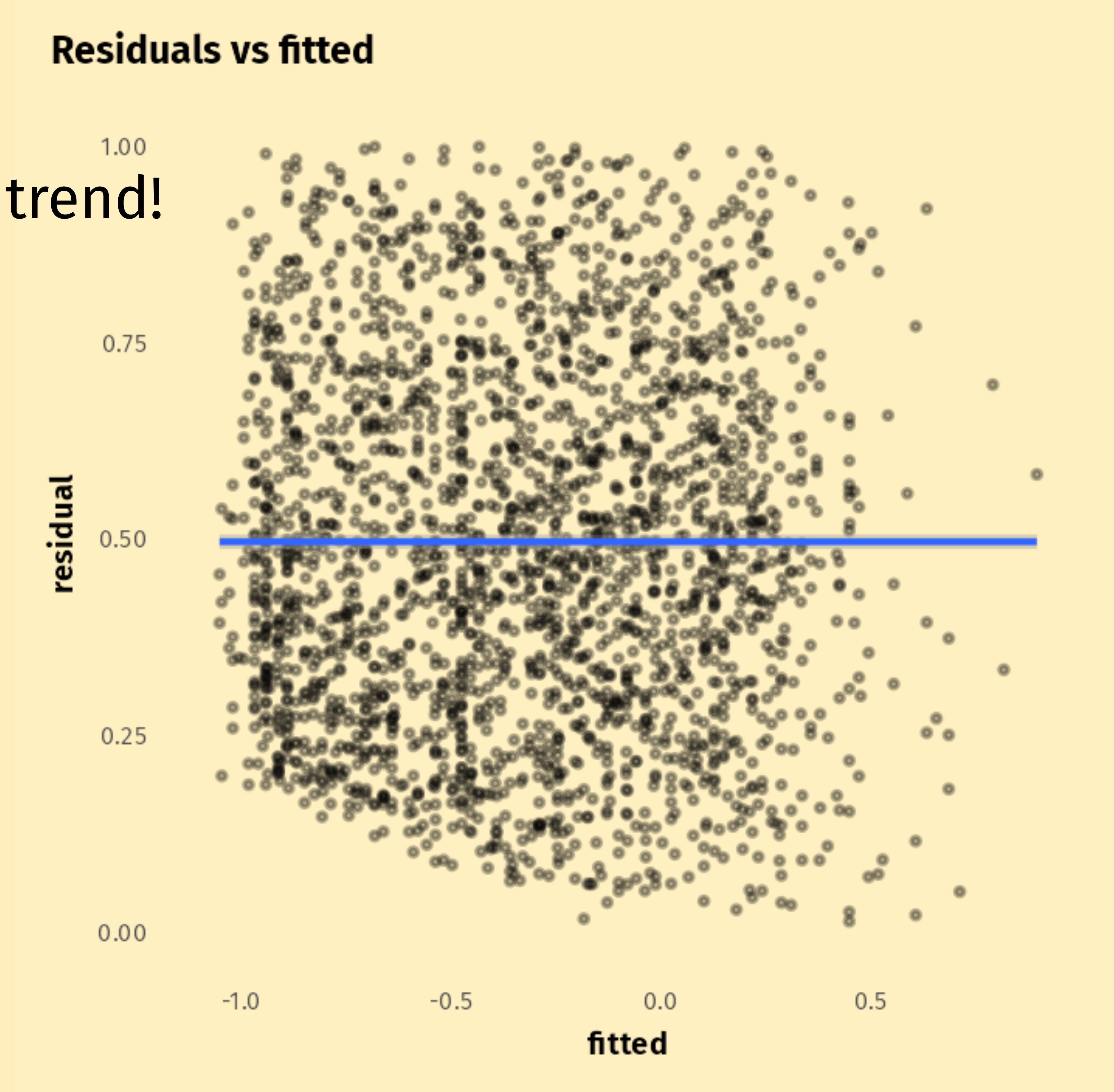
...but then it kinda isn't

Your model



There shouldn't be a trend!

The model she tells you not to worry about



Linear regression is a very robust model, but sometimes the assumptions it makes are not even close to truth.

It’s structure is very rigid:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

We need a more flexible approach. Something more generalised…

Point of view matters

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

The dependent variable Y comes from a normal distribution with the mean of $\beta_0 + \beta_1 + x_i$ and standard deviation of ϵ_i.

In this version, the normal distribution is “baked in”.

Point of view matters

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon$$

The dependent variable Y comes from a normal distribution with the mean of $\beta_0 + \beta_1 + x_i$ and standard deviation of ϵ.

N stands for normal distribution

$$y_i \sim N(\beta_0 + \beta_1 \cdot x_i, \epsilon)$$

"comes from/is distributed as"

The Big Question

Do we *need* to use normal distribution?

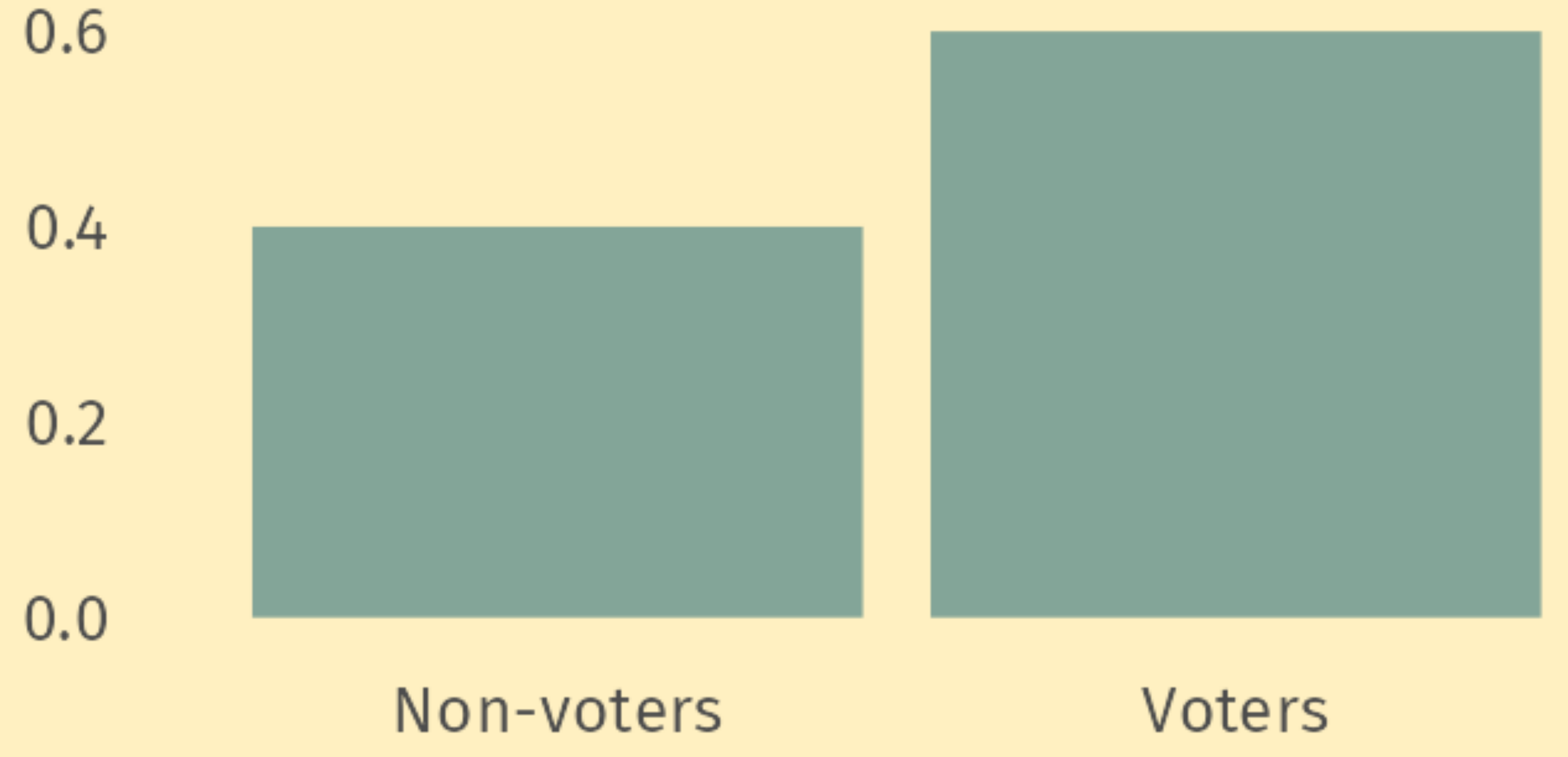
$$y_i \sim N(\beta_0 + \beta_1 \cdot x_i, \epsilon)$$

No

The Generalised Linear Model (GLM) is born

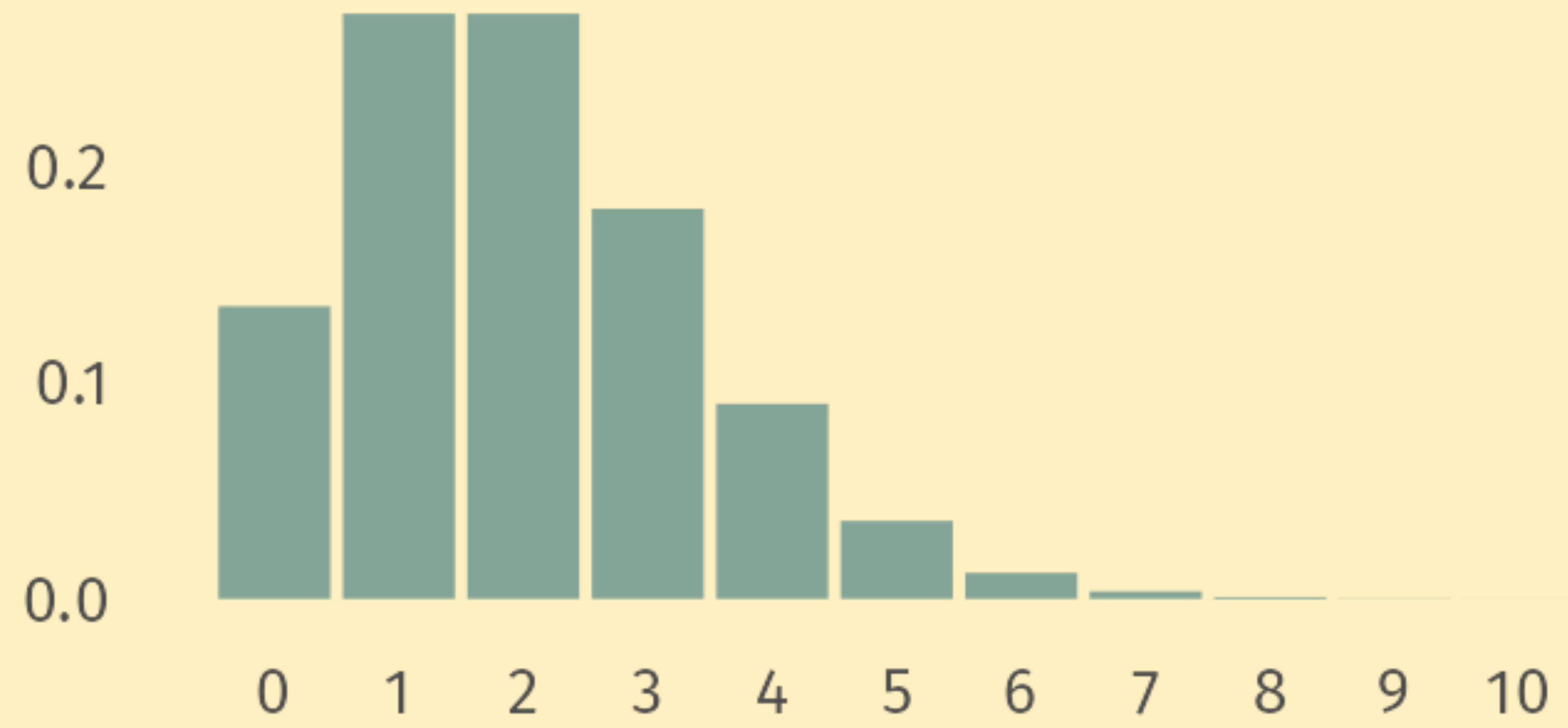
Probability of voting in elections comes from Bernoulli distribution with mean of $\beta_0 + \beta_1 \cdot x$

$$y_i \sim Bernoulli(\beta_0 + \beta_1 \cdot x_i)$$



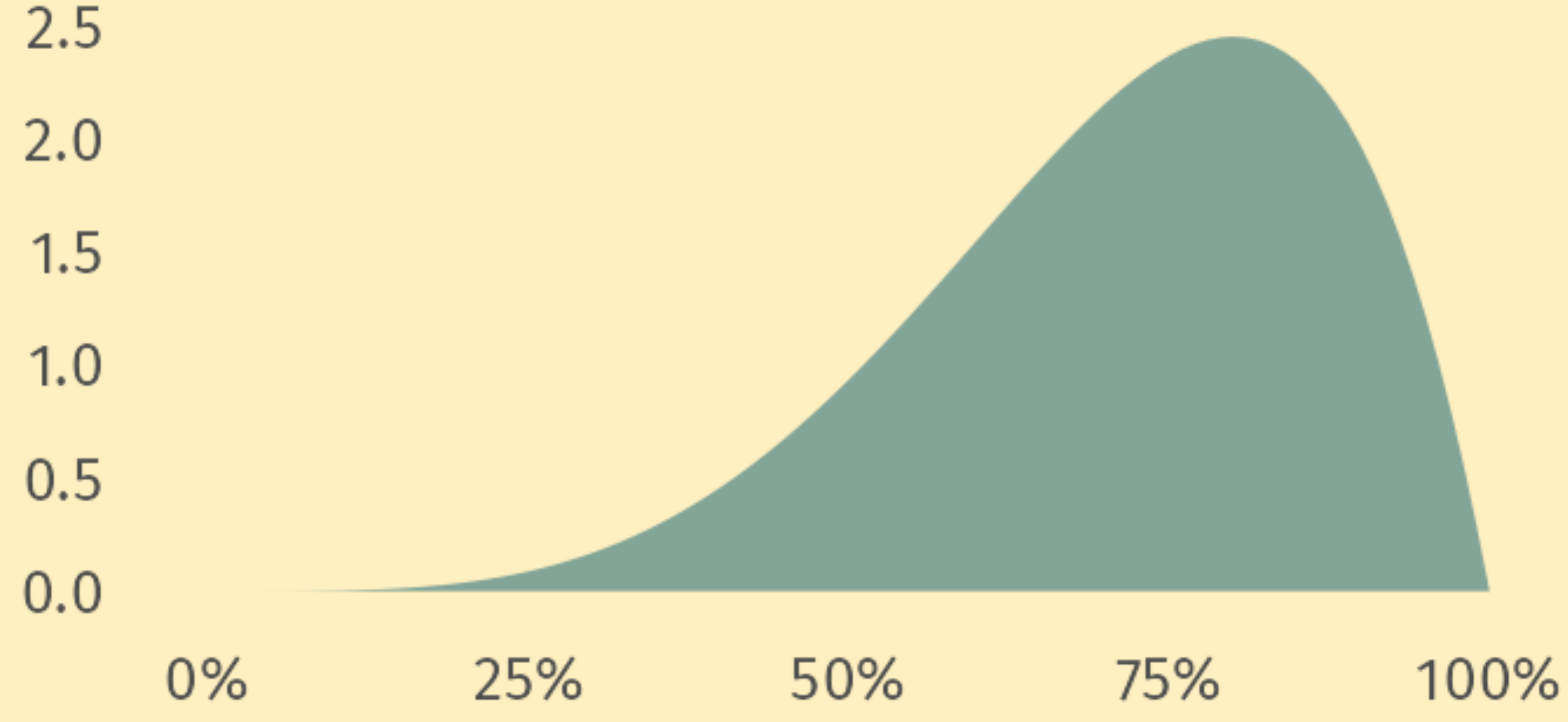
Number of absences in schools comes from Poisson distribution with mean of $\beta_0 + \beta_1 \cdot x$

$$y_i \sim Poisson(\beta_0 + \beta_1 \cdot x_i)$$



Teacher's score in student evaluations comes from Beta distribution with mean of $\beta_0 + \beta_1 \cdot x$

$$y_i \sim Beta(\beta_0 + \beta_1 \cdot x_i)$$



Questions?

Link functions

All estimated parameters have to be properly bounded.

Example: Probability of voting in elections comes from Bernoulli distribution with mean of $\beta_0 + \beta_1 \cdot x$

$$y_i \sim Bernoulli(\beta_0 + \beta_1 \cdot x_i)$$

We need to either make sure $\beta_0 + \beta_1 \cdot x$ is bounded between 0 and 1.

or

Make sure y_i can be take any value.

Link functions

By convention, we transform the dependent variable y_i.

The function that transforms the variable into a proper form is called a link function

(Because it links y_i and $\beta_0 + \beta_1 \cdot x_1$ to make sure they are on the same scale)

Link functions example

Example: Probability of voting in elections comes from Bernoulli distribution with mean of $\beta_0 + \beta_1 \cdot x$

$$y_i \sim Bernoulli(\beta_0 + \beta_1 \cdot x_i)$$

Instead of predicting the probability directly, we predict the logit of y_i

$$logit(y_i = vote) = \frac{P(y_i = vote)}{1 - P(y_1 = vote)}$$

Link functions example

- The full generalised linear model is then

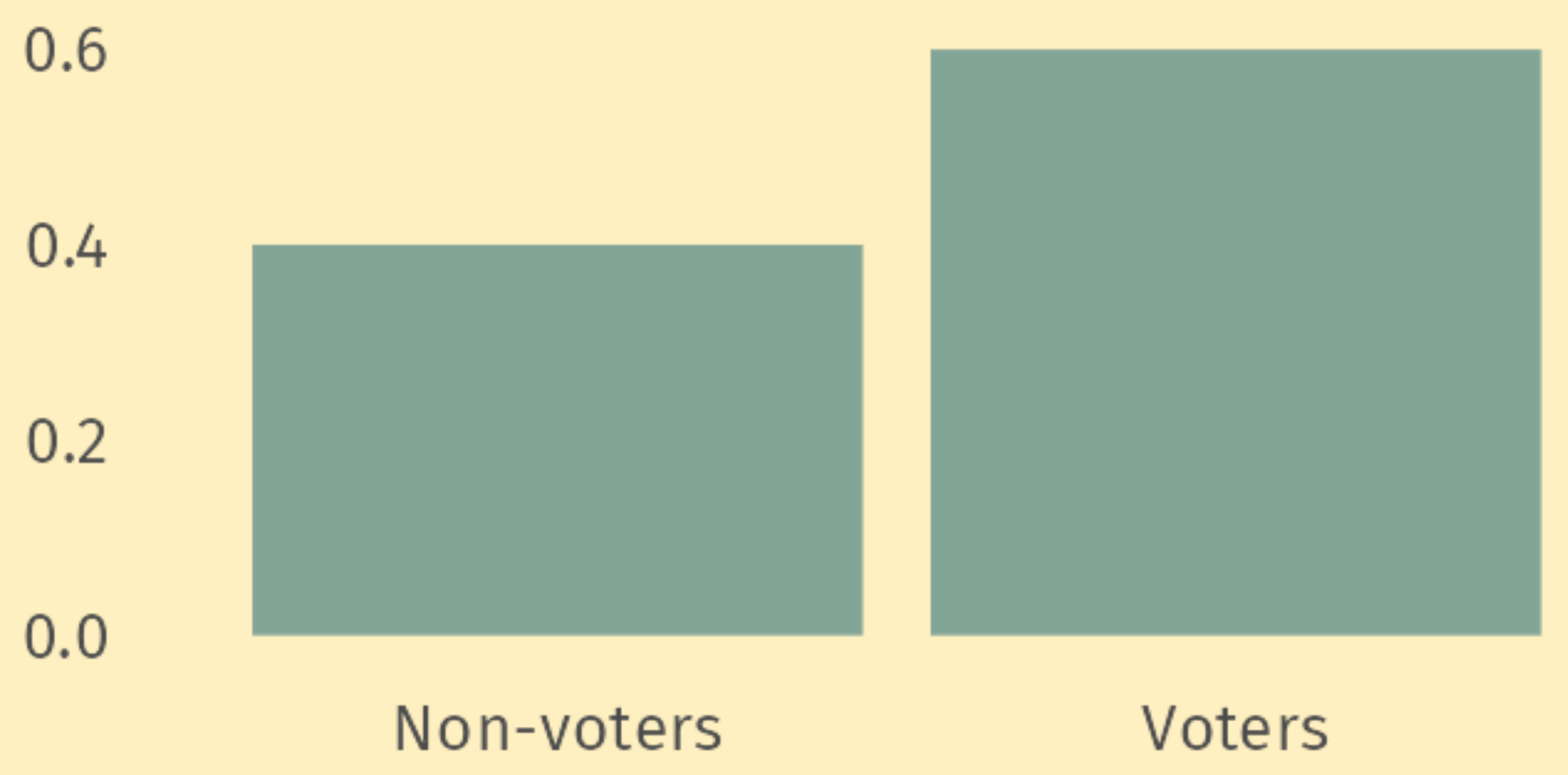
$$logit(y_i) \sim Bernoulli(\beta_0 + \beta_1 \cdot x_i)$$

Where logit is the link function making sure are predicted probabilities are between $(0;1)$.

Link functions make computations easier, we can back transform for interpretation.

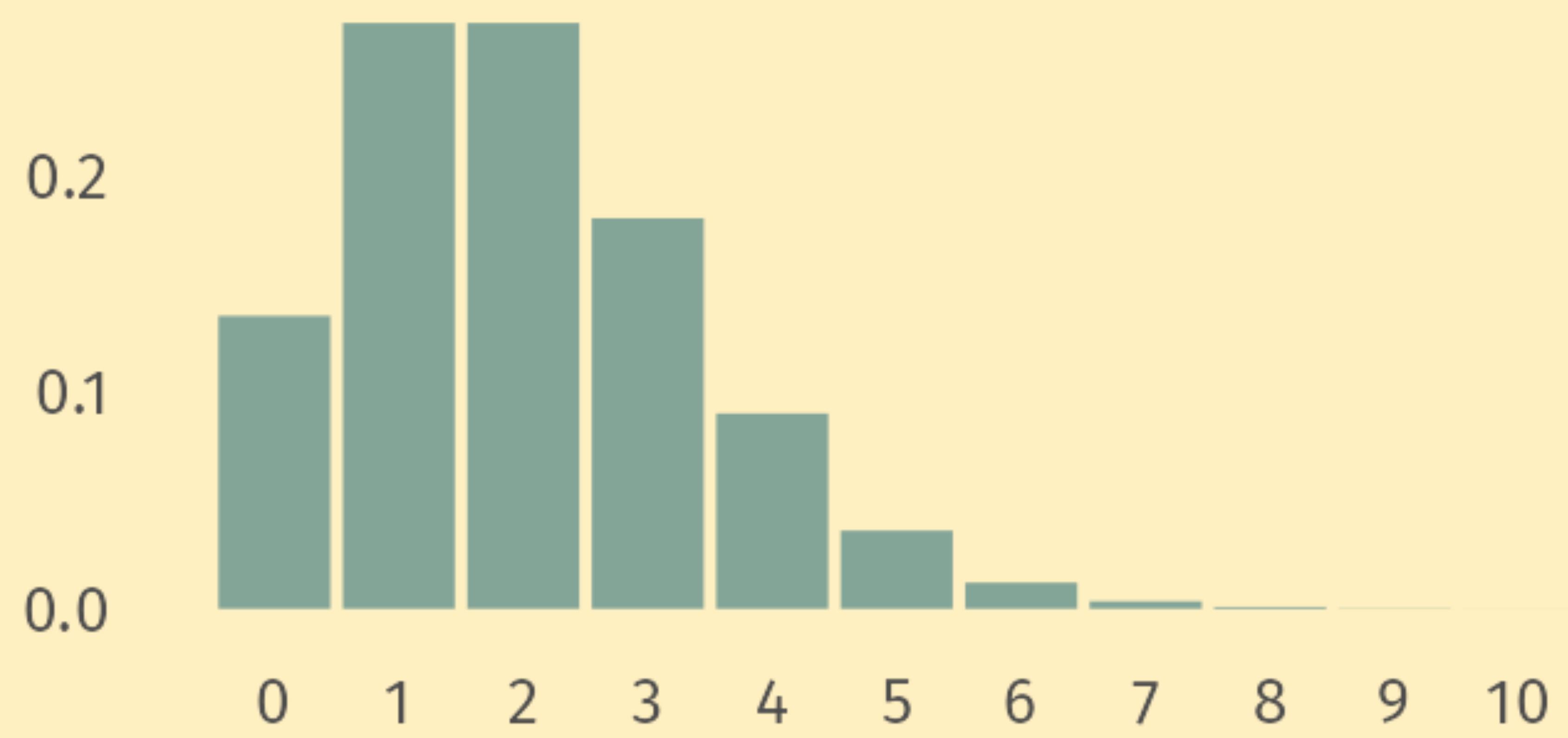
Probability of voting in elections comes from Bernoulli distribution with mean of $\beta_0 + \beta_1 \cdot x$

$$logit(y_i) \sim Bernoulli(\beta_0 + \beta_1 \cdot x_i)$$



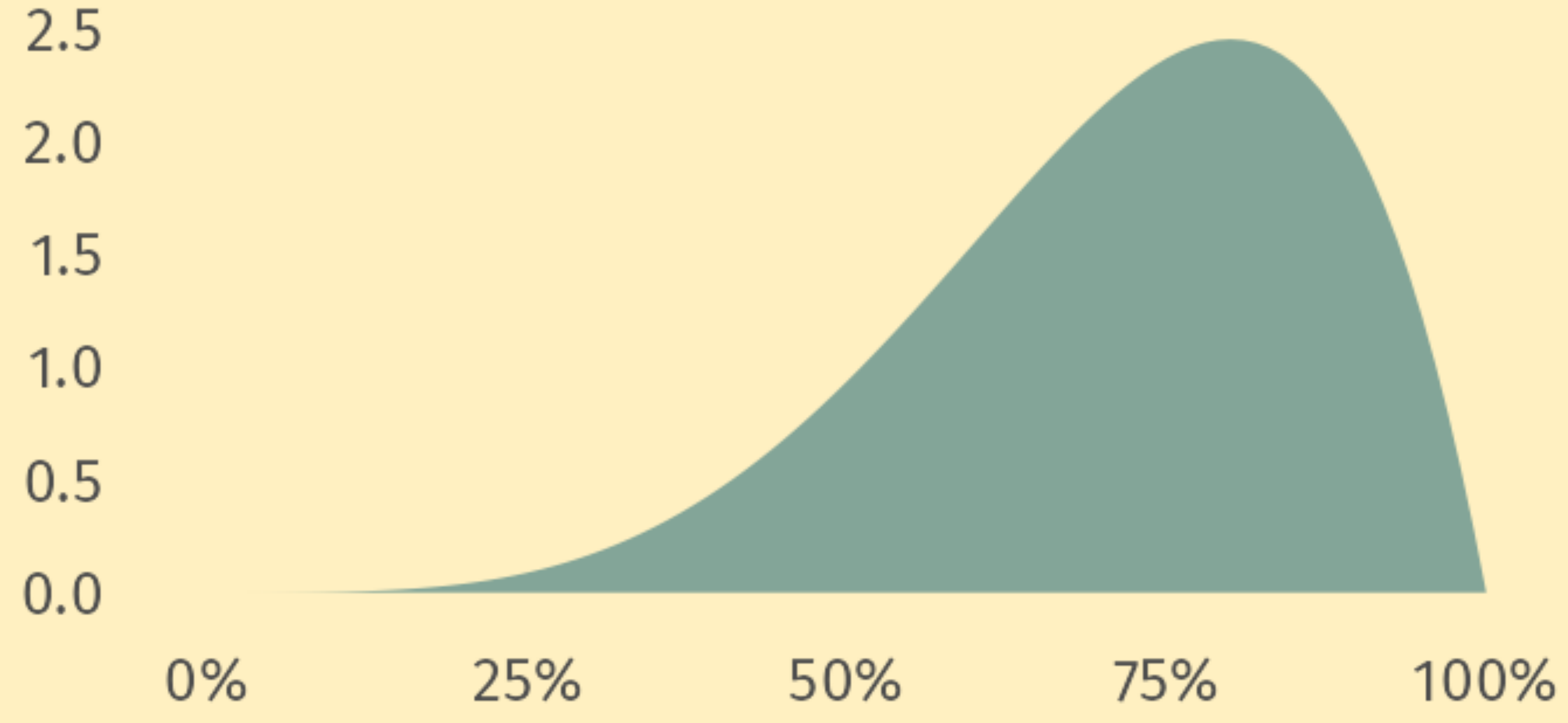
Number of absences in schools comes from Poisson distribution with mean of $\beta_0 + \beta_1 \cdot x$

$$log(y_i) \sim Poisson(\beta_0 + \beta_1 \cdot x_i)$$



Teacher's score in student evaluations comes from Beta distribution with mean of $\beta_0 + \beta_1 \cdot x$

$$logit(y_i) \sim Beta(\beta_0 + \beta_1 \cdot x_i)$$



Link functions - concluding remarks

Even classical linear regression has a link function - identity link function.

$$1 \cdot y_i \sim N(\beta_0 + \beta_1 \cdot x_i, \epsilon)$$

Because linear regression already assume the dependent variable can take any value, we leave it as it is

Link functions - concluding remarks

You don’t have to remember all the link functions.

There is a *canonical* link function for every distribution - always preselected.

But link function play role in interpretation:

- E.g. the regression coefficient in logistic regression are in logit units (log odds)

Putting it all together

Generalised linear models allow us to select distributions better reflecting the nature of our data.

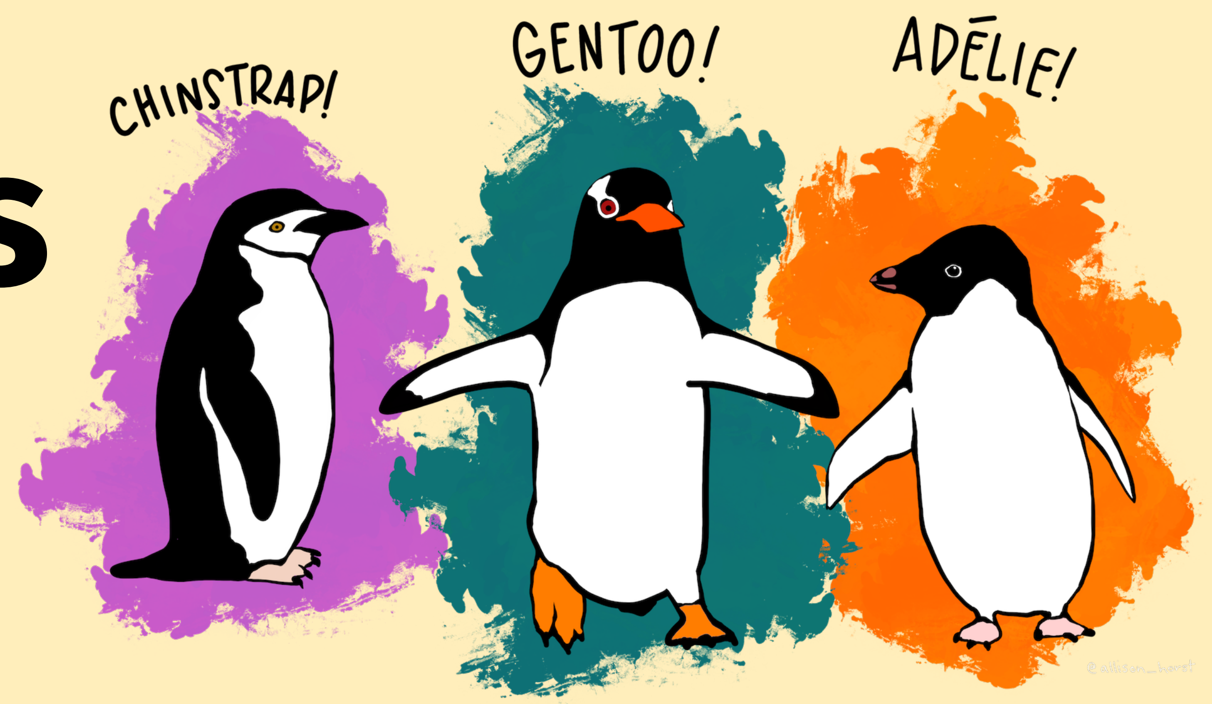
The general form is

$$link(y_i) = Distribution(\beta \cdot X)$$

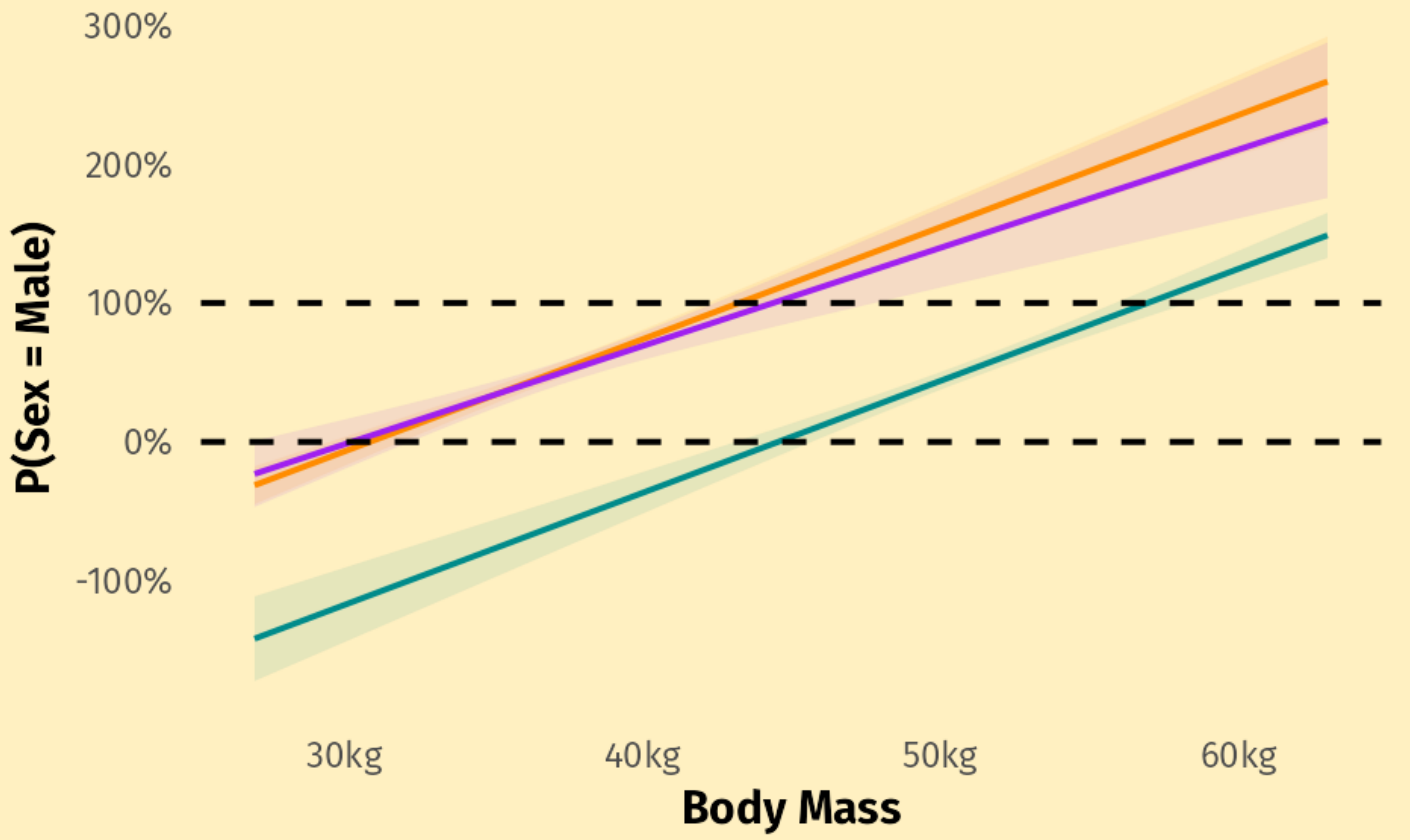
The distribution and link function together make sure that our model respects our data (boundaries, discrete vs continuous etc.)

Respecting your data makes better models

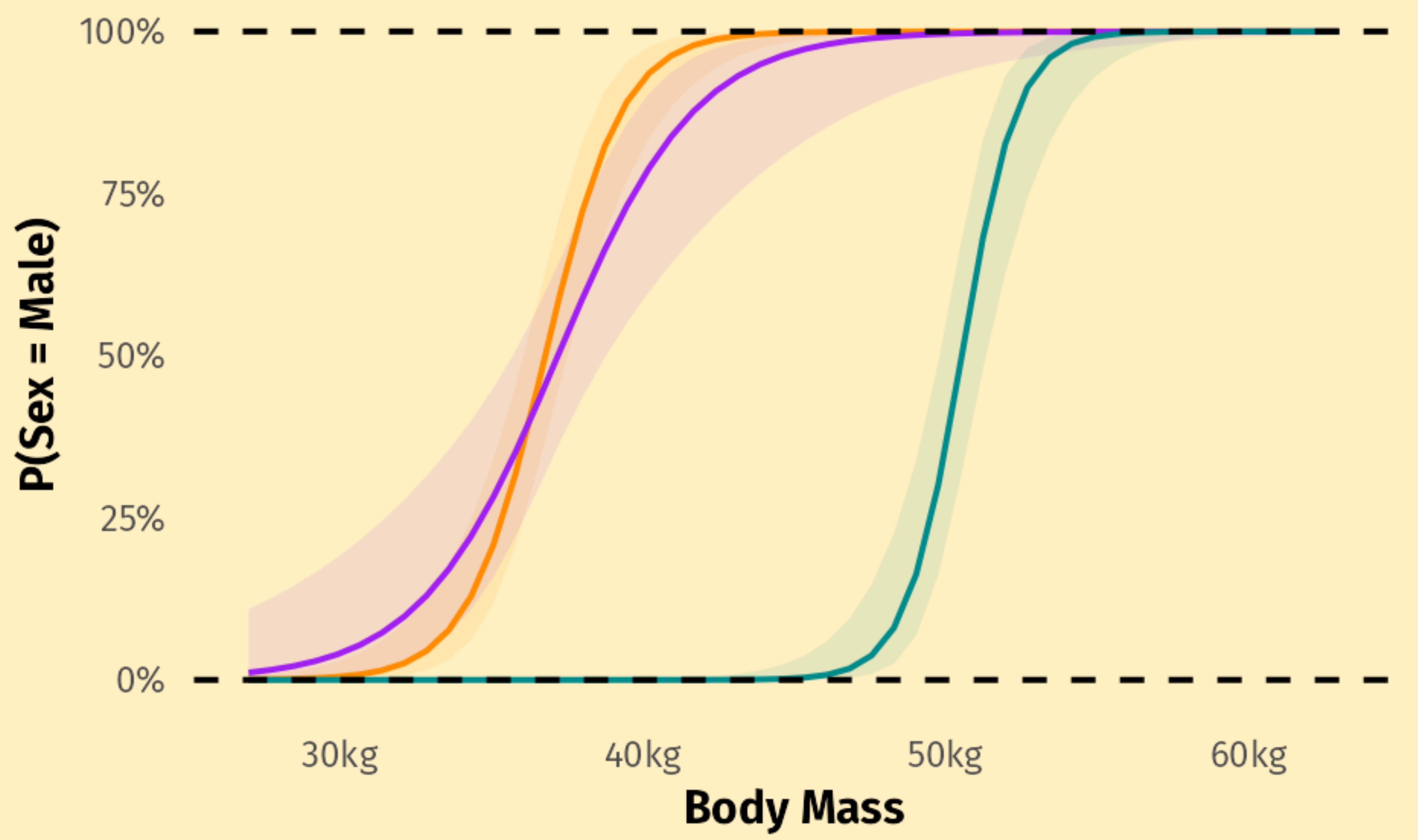
and R makes it easy!



`lm(sex ~ body_mass_g * species)`



`glm(sex ~ body_mass_g * species, family = binomial(“logit”))`



Questions?