Modeling categorical variables 1:

Binary outcomes

Aleš Vomáčka 2023

Lesson Goals

- Compute binary logistic regression
- Interpret binary logistic regression
- Learn how working with GLMs differ from linear regression

Binary variables are everywhere

Smokers/Nonsmokers

Buyer/Nonbuyer

Voter Turnout

Politic. Protest Participation

Election Turnout

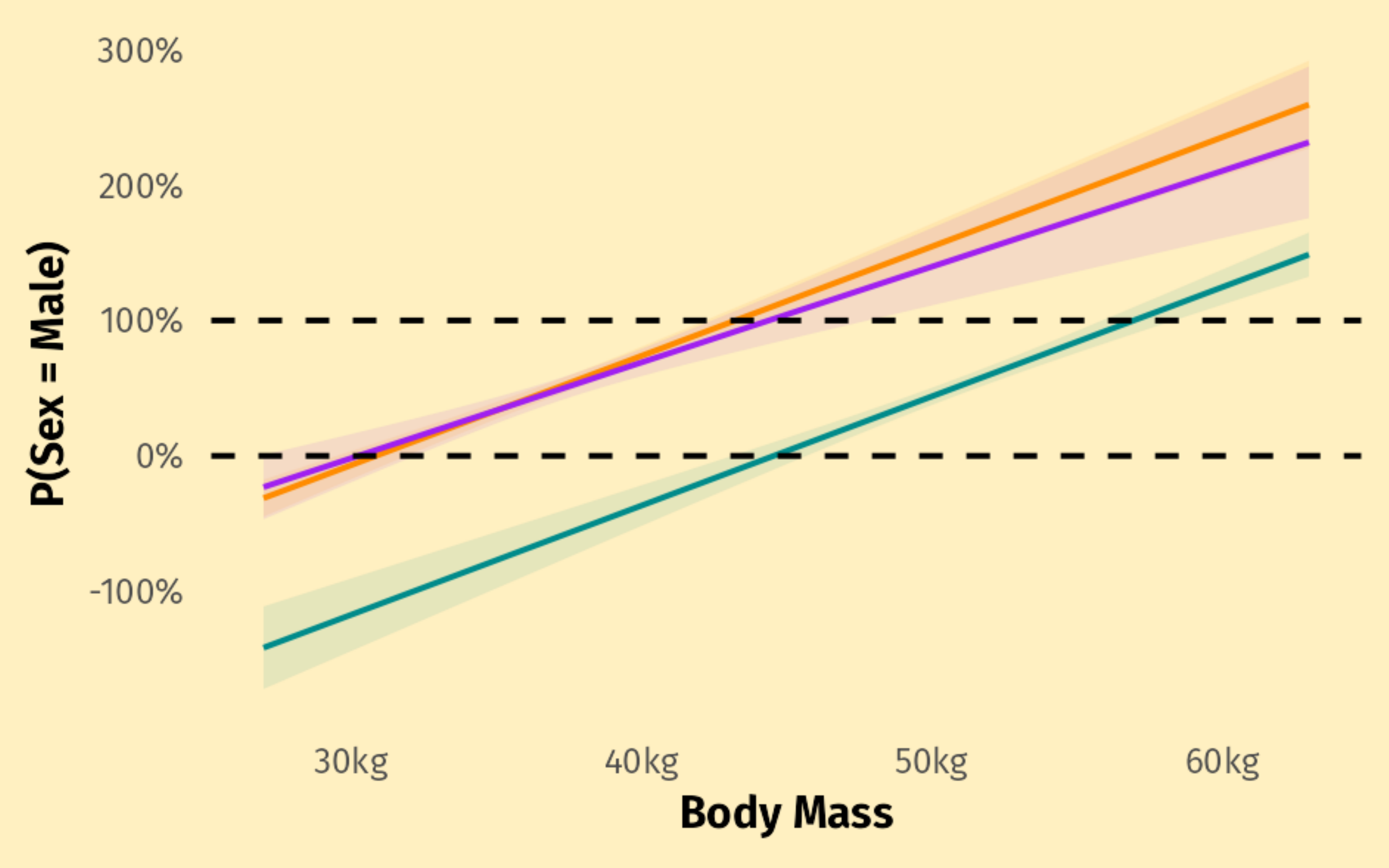
Sex

User/Nonuser

Diploma/Without Diploma

Binary outcomes and linear regression

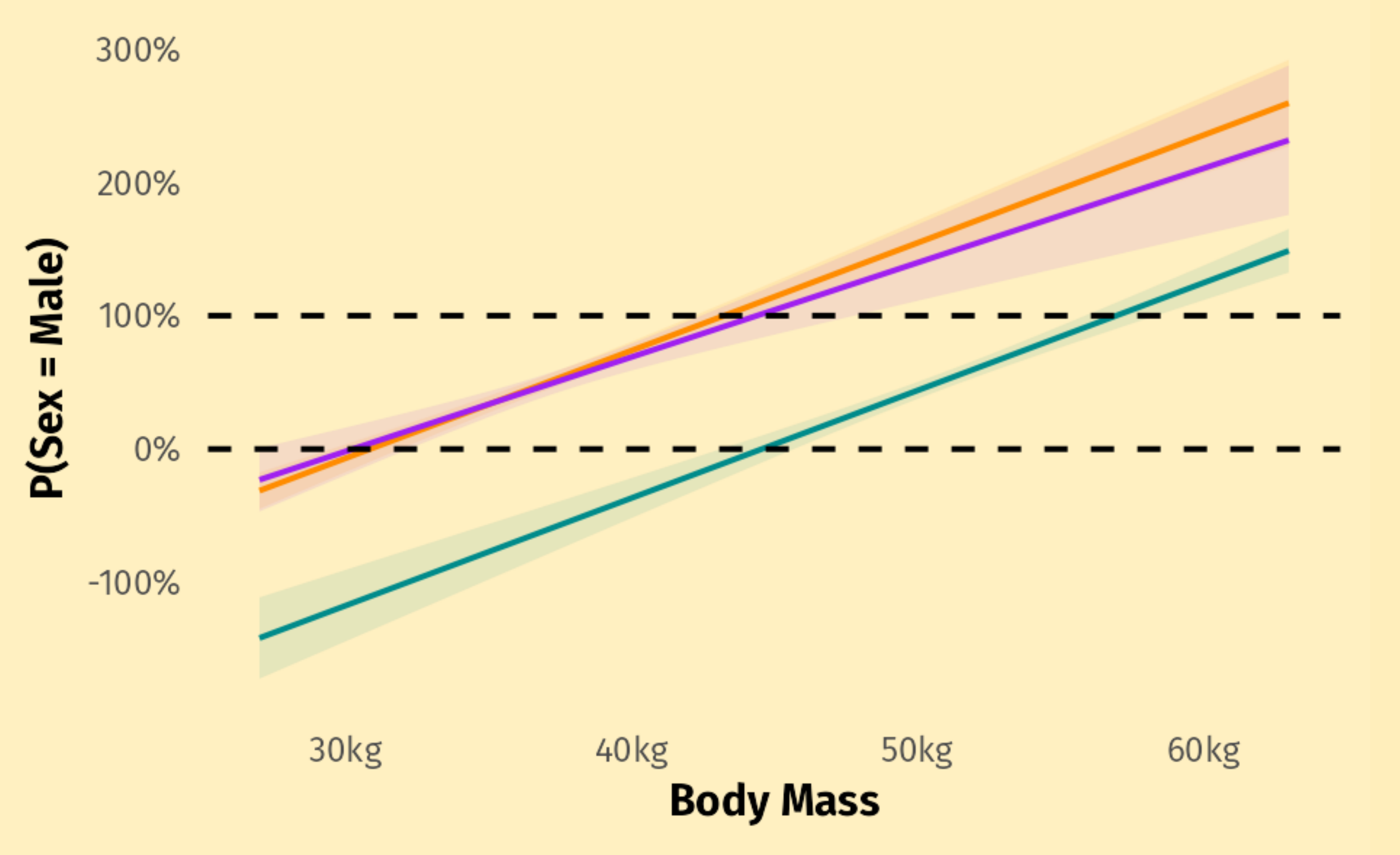
What are the problems?



Binary outcomes and linear regression

<u>What are the problems?</u>

- Linearity - unreliable point estimates
- Homoskedasticity - unreliable std. errors
- Normality - unreliable std. errors



Binary variables are everywhere

Smokers/Nonsmokers

Buyer/Nonbuyer

Voter Turnout

Politic. Protest Participation

Election Turnout

Sex

User/Nonuser

Diploma/Without Diploma

What distribution would fit better?
What do we know about the outcome?

Cooking with GLMs tips:

Pick distribution that matches known properties of the outcome.

Binomial distribution

$$Binomial(p, k)$$

Probability of success → p

Number of trails → k

Predicts probability of “successes” across number of trails.

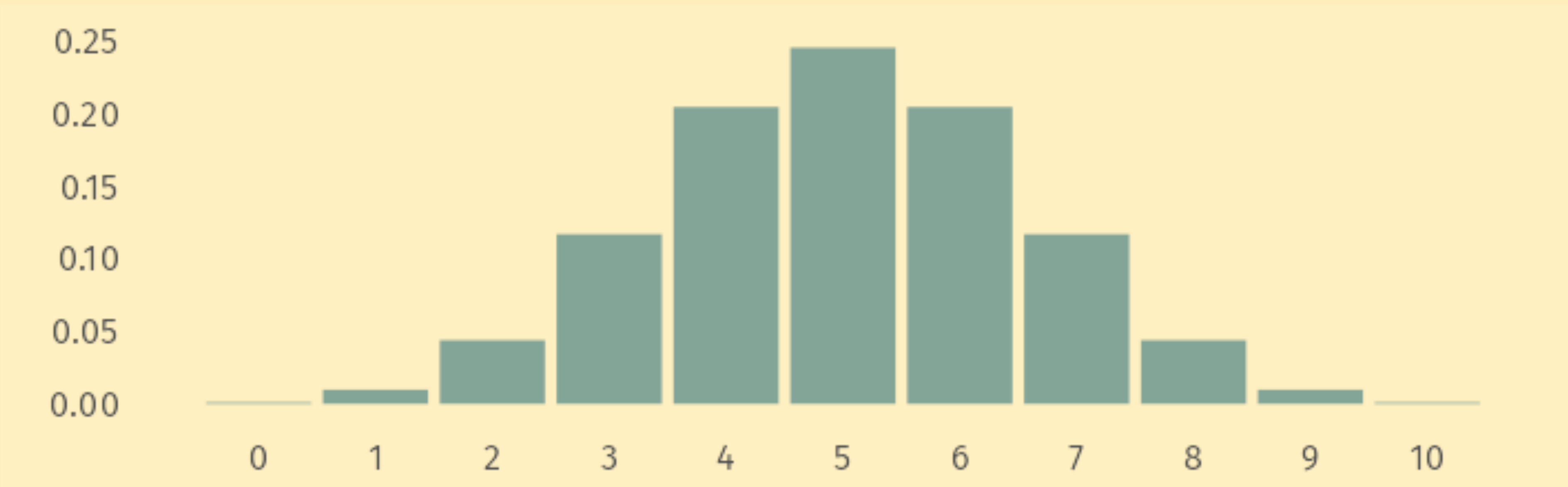
$$Mean(x) = p \cdot k; \quad Var(x) = p \cdot (1-p) \cdot k$$

Number of correct answers on an exam.

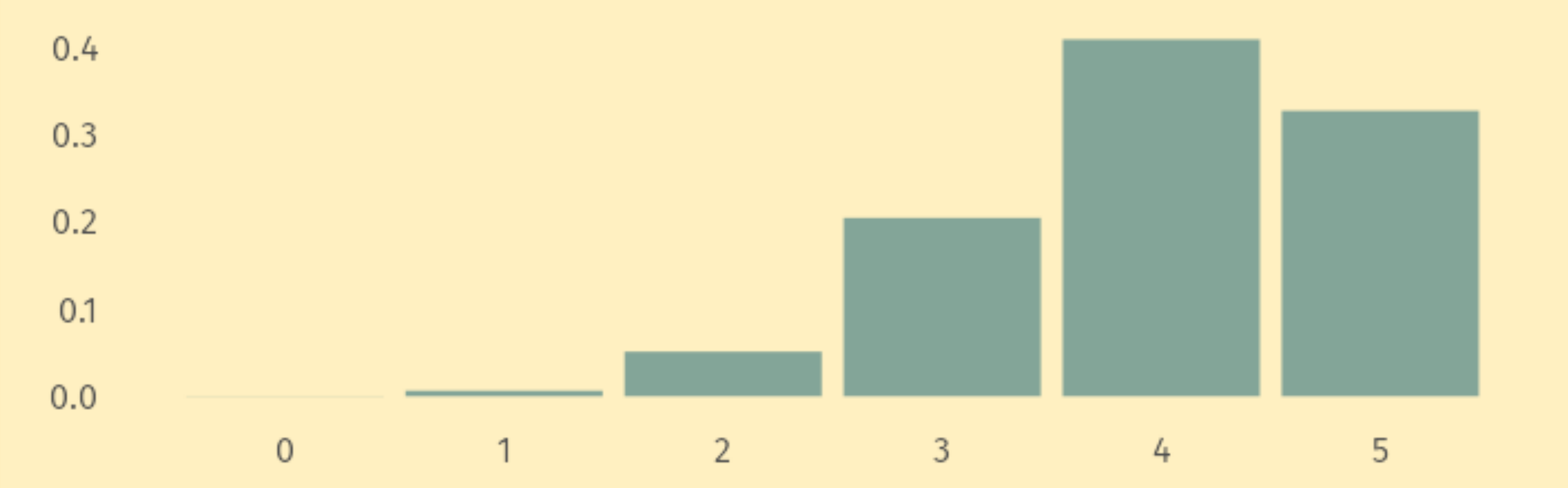
Number of people voting for specific party in population.

Number of times a coin lands on head.

$Binomial(0.5, 10)$



$Binomial(0.8, 5)$

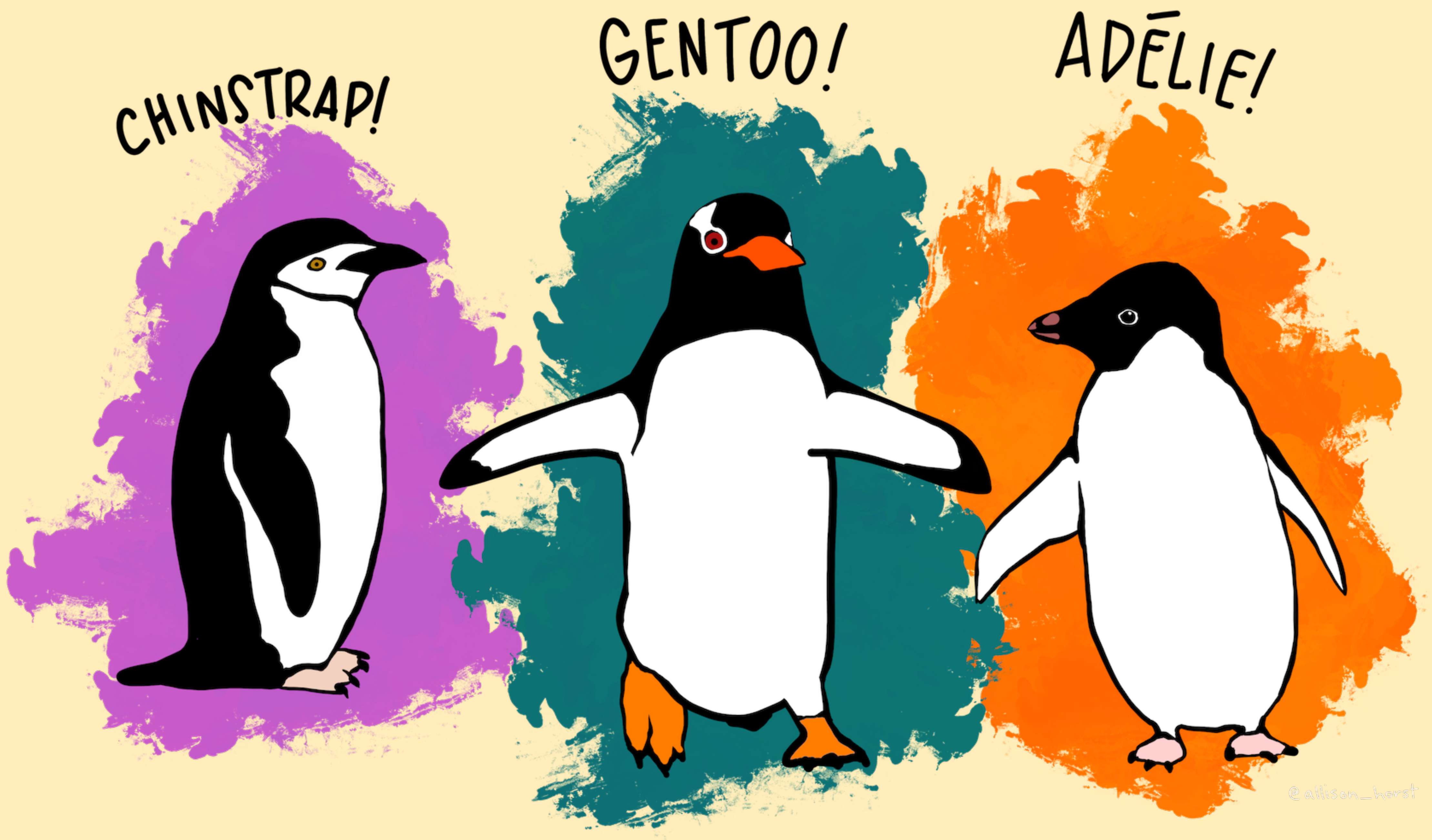


$Binomial(0.2, 1)$

Predicting Palmer Penguins

We want to predict whether a penguin is male or female.

In stats lingo, every penguin has one "attempt" to be male and we are interested in the probability of "success".



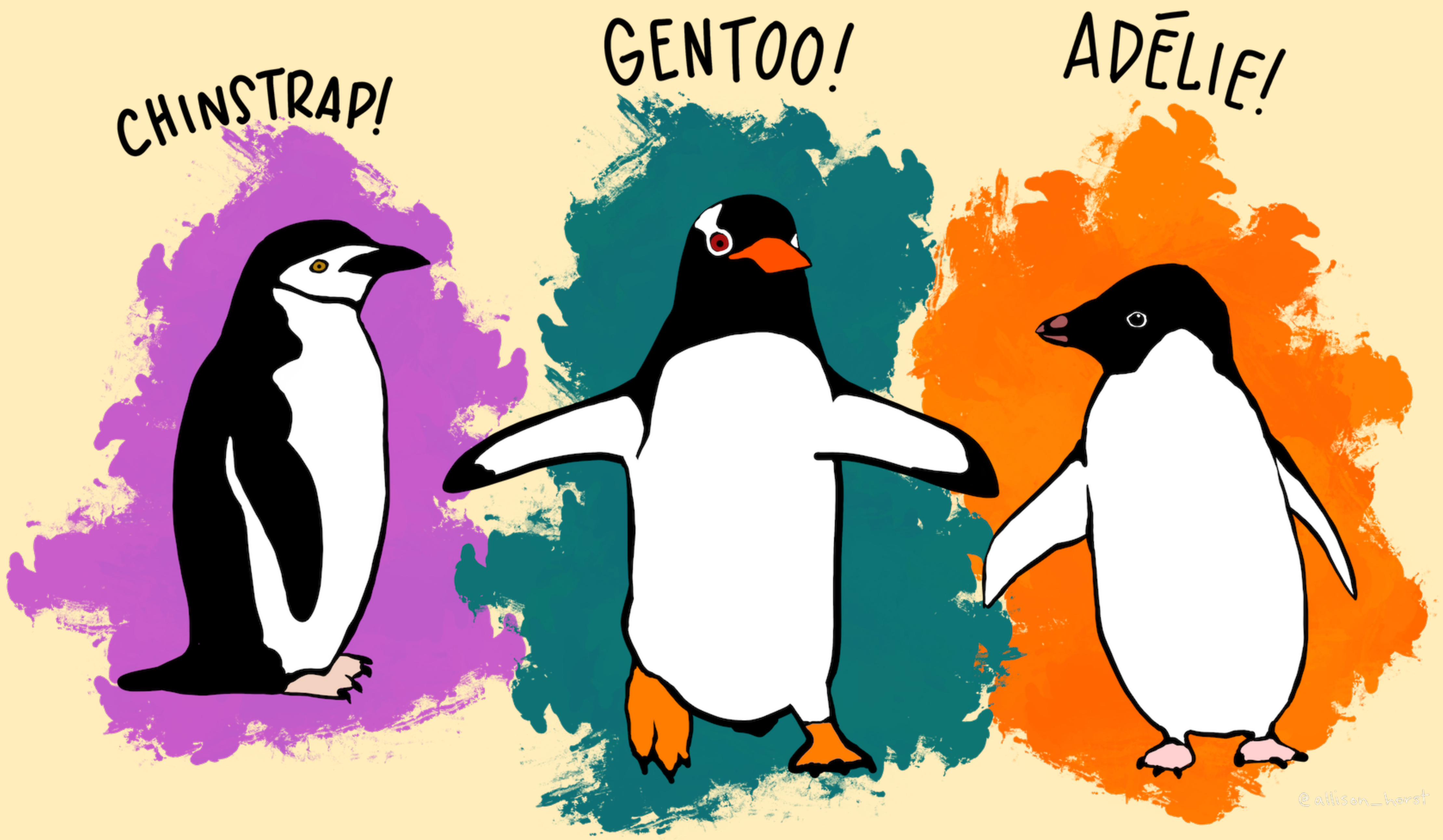
Let's start simple

No predictors, only interested in how many penguins are males.

$$sex \sim Binomial(\beta_0, 1)$$

We know the number of "successes" = 1.

Only have to estimate the probability that a penguin is male.



Problem: Probability is bounded between 0 and 1. How do we make sure our model knows that?

Solution: Link functions to the rescue!

Questions?

Choosing link functions - Common suspects

Link functions are used to make sure outcome y is transformed into an unbounded form

Many possible options, but 3 are the most common

Identity **Log** **Logit**

Identity link function

Literally just $identity(x) = x \cdot 1$

When the outcome is unbounded, no transformation is necessary.

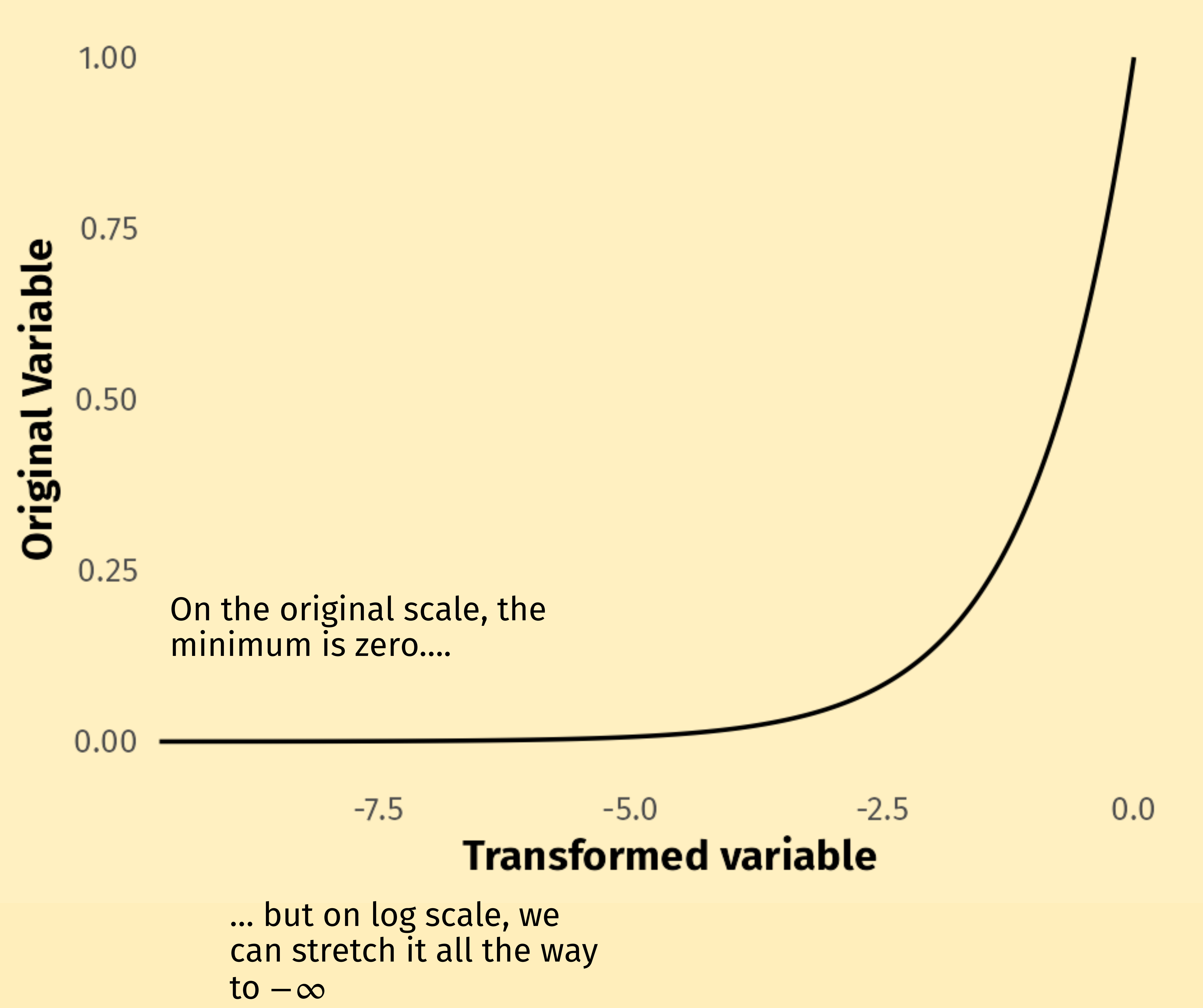
Linear regression is a prominent example.

Log link function

$$link(x) = log(x)$$

When the outcome is outcome is bounded from bottom (e.g. 0), we can “unbound” using logarithms.

Counts (number of children in family, police stops) are bounded by zero.

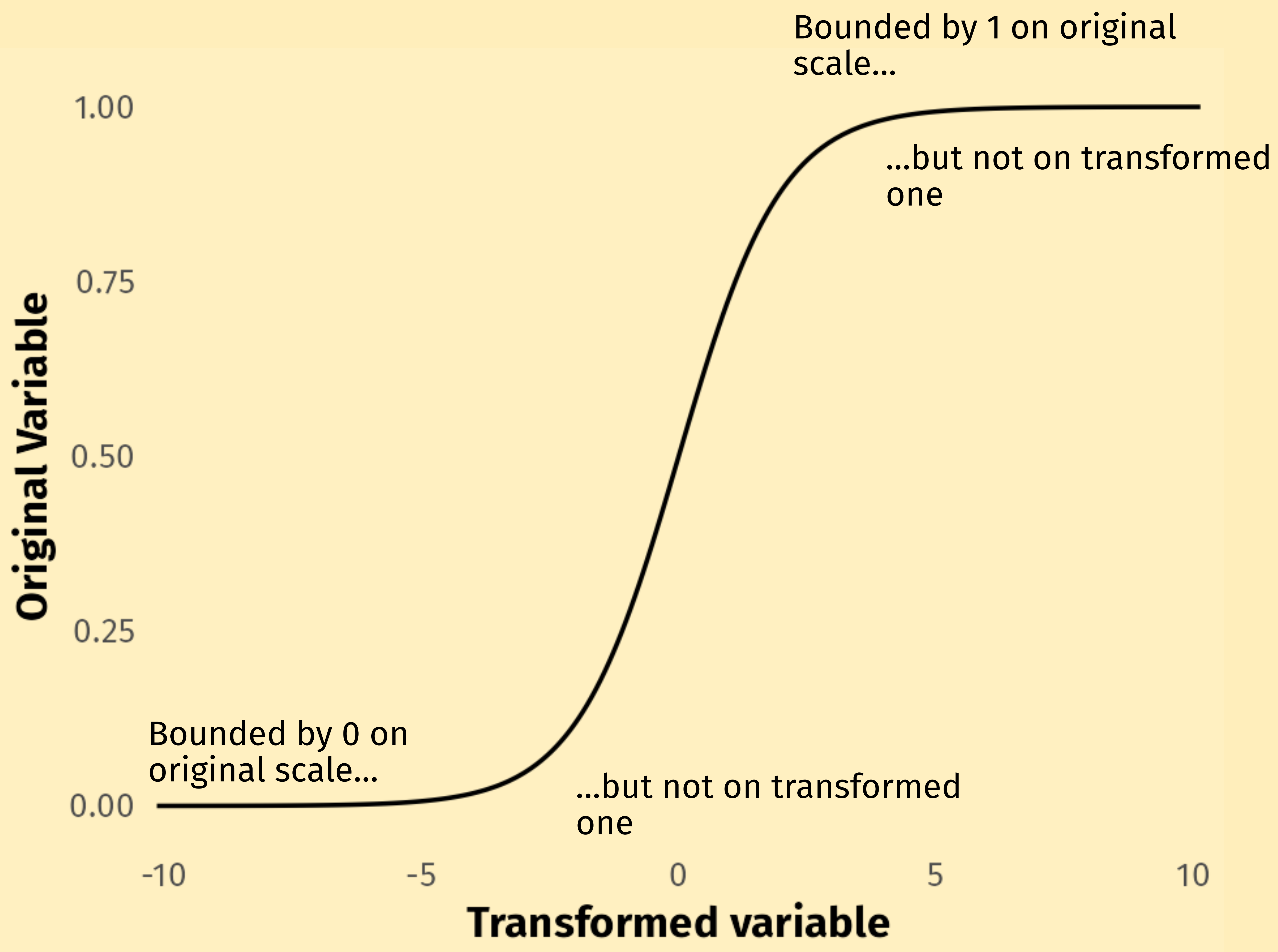


Logit link function

$$link(x) = log\left(\frac{prob_x}{1 - prob_x}\right)$$

When the outcome is outcome is bounded from bottom *and* top (e.g. between 0 and 1), we can "unbound" using logarithms of odds.

Probabilities and concentrations are bounded from both sides.



Choosing link functions - Common suspects

Identity

When outcome is unbounded
(or far from bounds)

Log

When outcome is bounded
from bottom (e.g. 0)

Logit

When outcome is bounded
from both bottom and top
(e.g. 0 and 1)

Attention check: What link function is used to predict probability that a penguin is male?

Questions?

Penguins in their final form

$$logit(sex = male) \sim Binomial(\beta_0, 1)$$

We are predicting penguin's sex using binomial regression with logit link.

(Colloquially, binomial "regression with logit link" = (binary) logistic regression)

The output are log odds of a penguing being male.

Penguins in R

Theoretical model:

$$logit(sex = male) \sim Binomial(\beta_0, 1)$$

Representation in R:

```
glm(sex ~ 1, family = binomial(link = "logit"))
```

or just

```
glm(sex ~ 1, family = binomial()
```

Penguins in R - results

$$logit(sex = male) \sim Binomial(0.018, 1)$$

Where the hell this number came from?

There are 165 female and 168 male penguins in the sample.

The probability of a penguin being male is $\frac{168}{168 + 165} = 0.505 = 50.5\,\%$

The odds of a penguin being male are $\frac{0.505}{1 - 0.505} = \frac{0.505}{0.455} = 1.02.$

In other words, there are 102 males for every 100 females

The log odds (logits) of a penguin being male are $log(1.02) \approx 0.018$

Predict penguin sex using body weight

$$logit(sex = male) \sim Binomial(\beta_0 + \beta_1 \cdot weight_{kg}, 1)$$

$$logit(sex = male) \sim Binomial(-5.16 + 1.24 \cdot weight_{kg}, 1)$$

Interpretation same as for linear regression:

For penguin with 0 weight, the expected log odds of being male are -5.16.

For every 1 kilogram increase, the expected log odds of being male increase by 1.24.

(Remember, this is just a correlation)

Questions?

The big question:

How the f***k we are supposed to interpret this?

There are two ways to interpret logistic regression

1) Exponentiating coefficient

This is what most people do.

It doesn't actually work.

2) Marginal effects on probability scale

This one actually works.

R Intermezzo!

Interpreting logistic regression

Interpreting Logistic regression

Exponentiation of regression coefficients

What most people use, most textbook teach.

Two big problems:

1) Odds ratios are actually not intuitive units.

2) Logistic regression is non-collapsible.

What is non-collapsibility?

!WARNING!

MINDBLOWN INCOMING

(So grab a coffee now, if you need)

The Plan

1. What non-collapsibility causes

2. What actually is non-collapsibility

3. How to fix/avoid non-collapsibility

Linear regression and control variables

Our dependent variable is personal wellbeing (scale -50 to 50)

Independent variables are age, gender and income (in thousands).

All the predictors are independent of each other. Sample size is 1 000 000.

We want to estimate the effect of age on wellbeing. What variables we need to control for?

Linear regression and control variables

This is the result we want

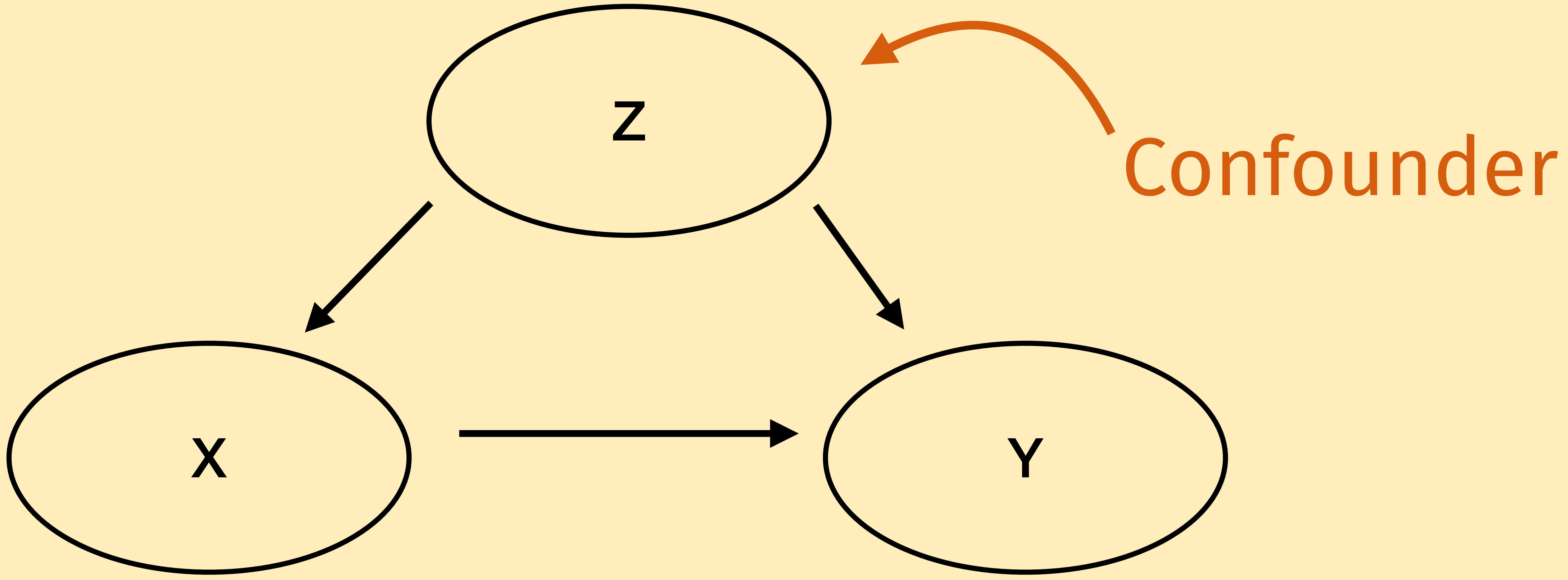
The true model is

$$wellbeing \sim Normal(-150 + 1 \cdot age + 2 \cdot gender + 1 \cdot income, 15)$$

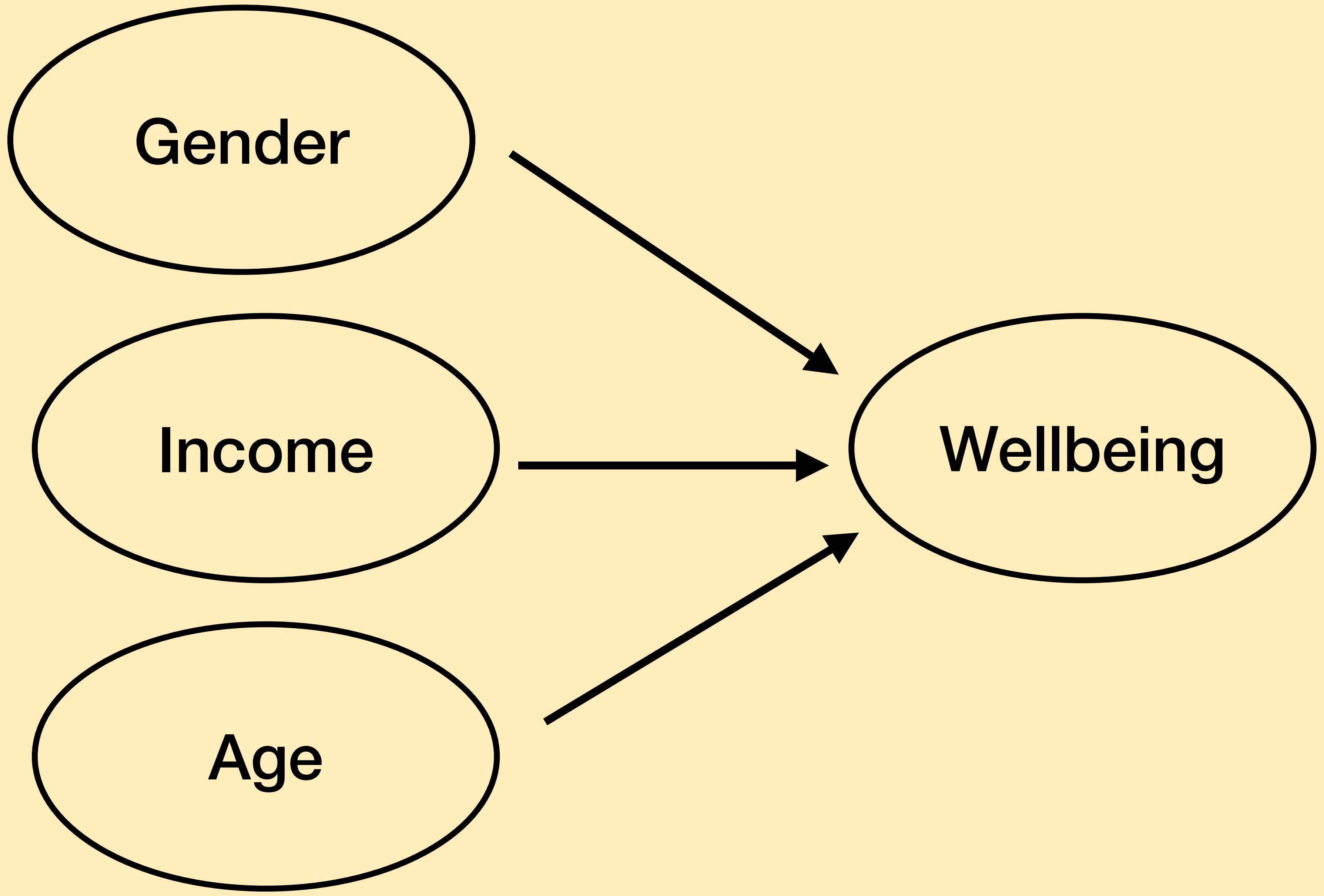
	Model 1	Model 2	Model 3
Age	1	1	1
Gender		2	2
Income			1

Linear regression and control variables

To get correct estimates, we need to control for all confounders.



In our example, the predictors are independent, so no confounders.



Again, but with logistic regression

Our dependent variable is personal voter turnout (binary).

Independent variables are age, gender and income (in thousands).

All the predictors are independent of each other. Sample size is 1 000 000.

We want to estimate the effect of age on voter turnout. What variables we need to control for?

Again, but with logistic regression

This is the result we want

The true model is

$$logit(turnout) \sim Binomial(-80 + 1 \cdot age + 2 \cdot gender + 1 \cdot income, 1)$$

	Model 1	Model 2	Model 3
Age	0.3	0.32	1
Gender		0.64	2
Income			1

In logistic regression model, coefficients are not unbiased estimates of the true relationship, unless you control for all causal determinants.

What you can't do with logistic regression

Interpret regression coefficient as strength of relationship (will underestimate)

Compare coefficients across models

Compare coefficients across datasets

Compare coefficients across subpopulations

What you can't do with logistic regression - Example

We want to compare relationship between turnout and age across Czechia and Germany.

Results:

	Czech model	German model
Coefficient for Age	0.3	0.09

Naive interpretation: In Czechia, age is more important than in Germany.

But actually, these are the true values:

Czech model: $logit(turnout) \sim Binomial(-80 + 1 \cdot age + 2 \cdot gender + 1 \cdot income, 1)$

German model: $logit(turnout) \sim Binomial(-180 + 1 \cdot age + 2 \cdot gender + 4 \cdot income, 1)$

Effect of Age is the same for both countries

It's the effect of (uncontrolled for) income that's different

What you can't do with logistic regression

Interpret regression coefficient as strength of relationship (will underestimate)

Compare coefficients across models

Compare coefficients across datasets

Compare coefficients across subpopulations

Silver lining

- Statistical significance is not affected
- Direction of effect (positive/negative) is not affected

Questions?

Non-collapsibility so far...

1. Logistic regression can't be interpreted using regression coefficients.

So, regression coefficients in logistic regression are uninterpretable.

Why?

Non-collapsibility

Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, *26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

What is non-collapsibility?

Two explanations:

1) Graphical explanation (what is happening)

2) Latent variable explanation (why it's happening)

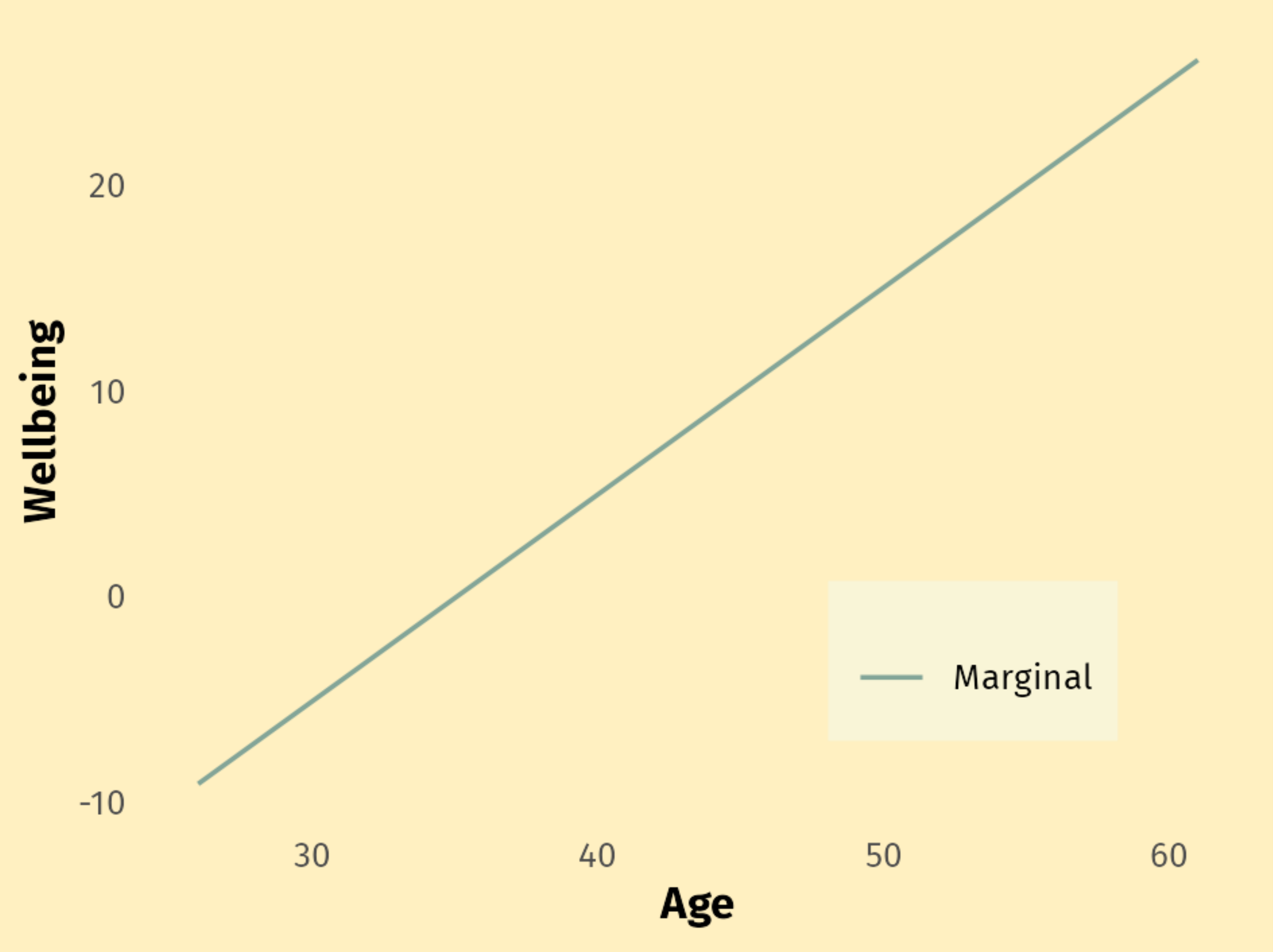
Graphical Explanation

Collapsibility - Linear regression

$$wellbeing \sim N(-150 + 1 \cdot age, 15)$$

Model with only age as predictor.

The slope is 1.



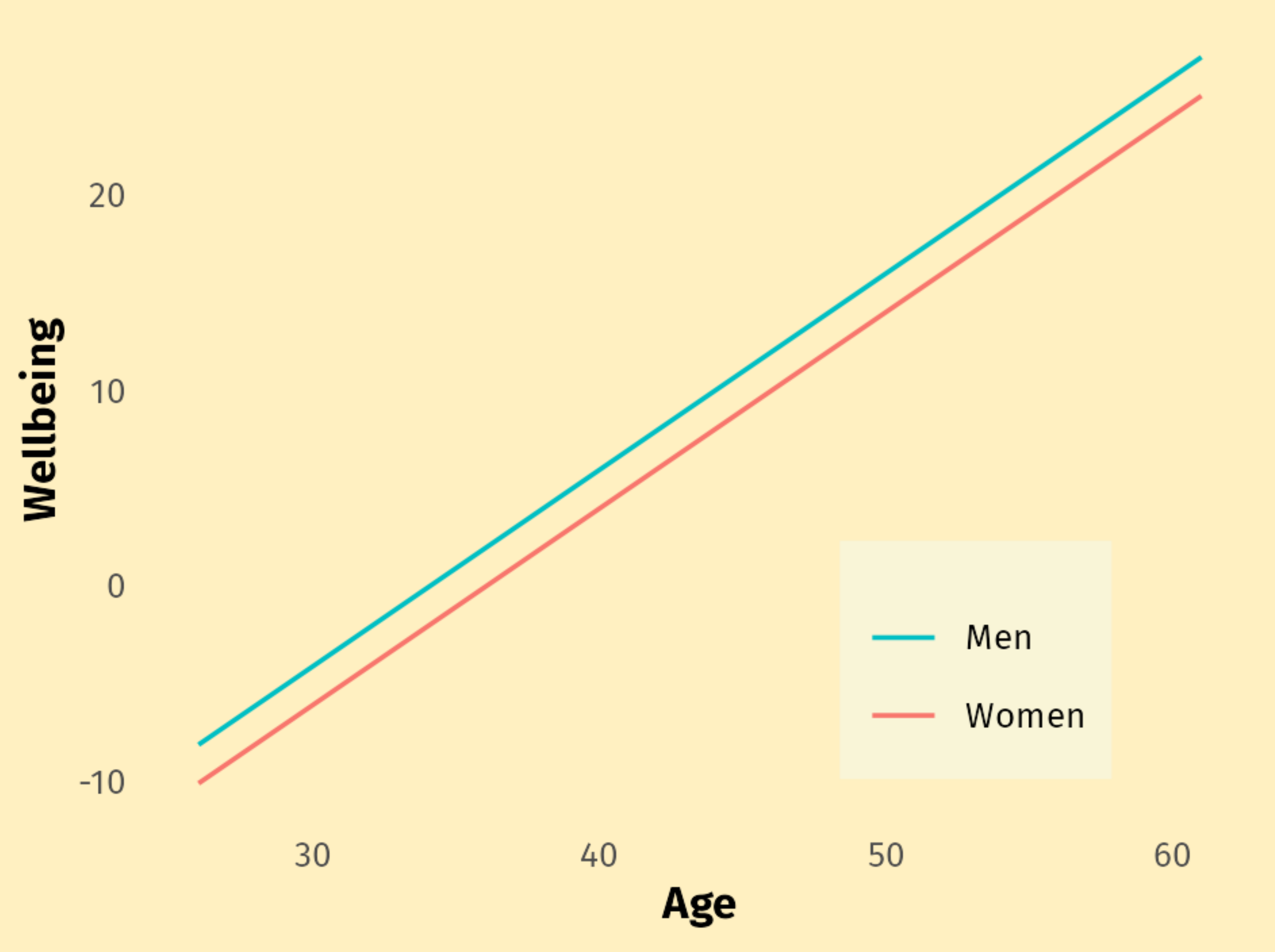
Collapsibility - Linear regression

$$wellbeing \sim N(-150 + 1 \cdot age + 2 \cdot gender, 15)$$

Model with age and gender as predictors.

The slope is for age is still 1

(age and gender are independent).



Collapsibility - Linear regression

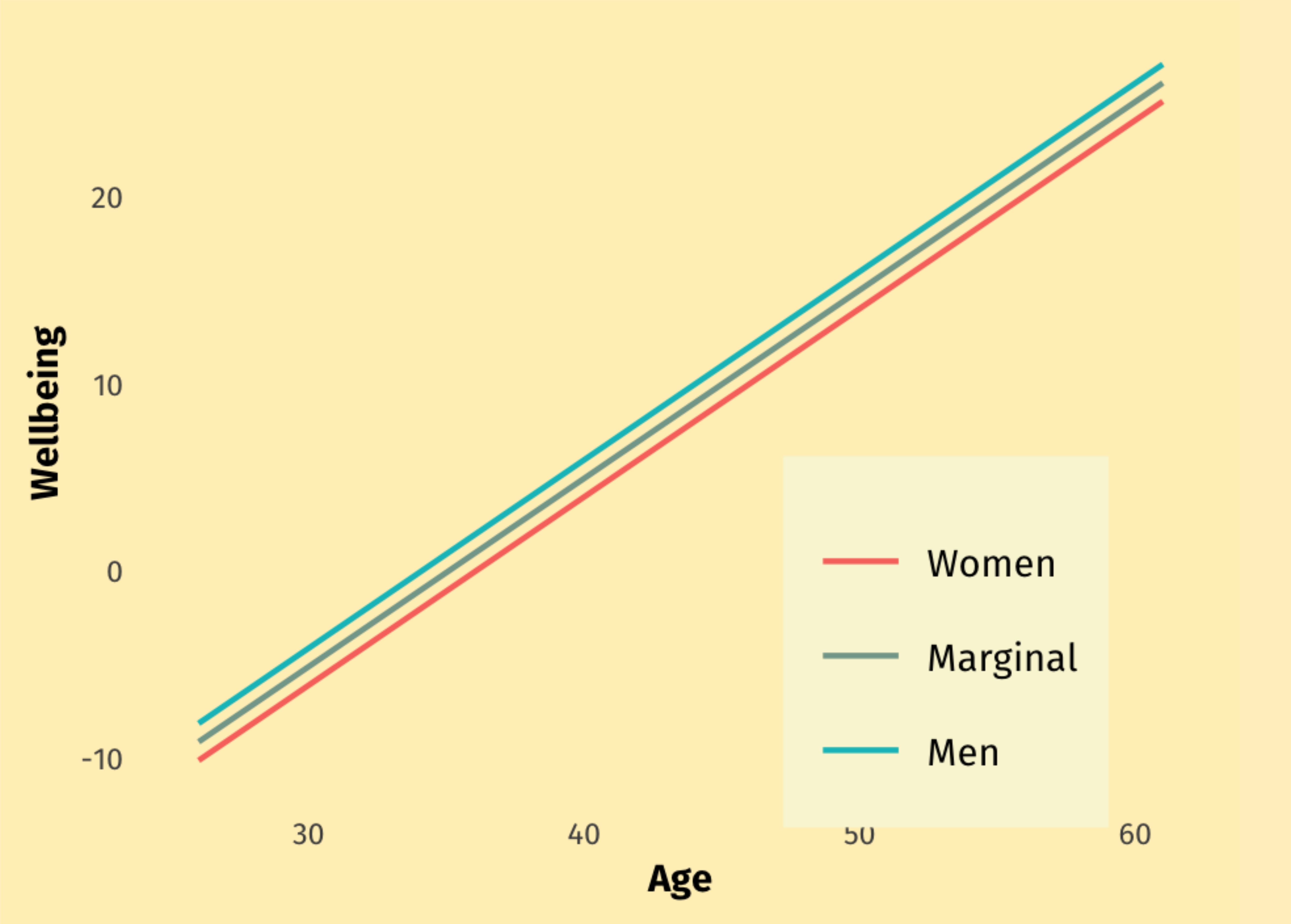
$wellbeing \sim N(-150 + 1 \cdot age + 2 \cdot gender, 15)$

It doesn't matter, if we control gender or not.

Removing gender from model just squishes/*collapses* the lines.

Slope remains the same.

Linear regression is collapsible.



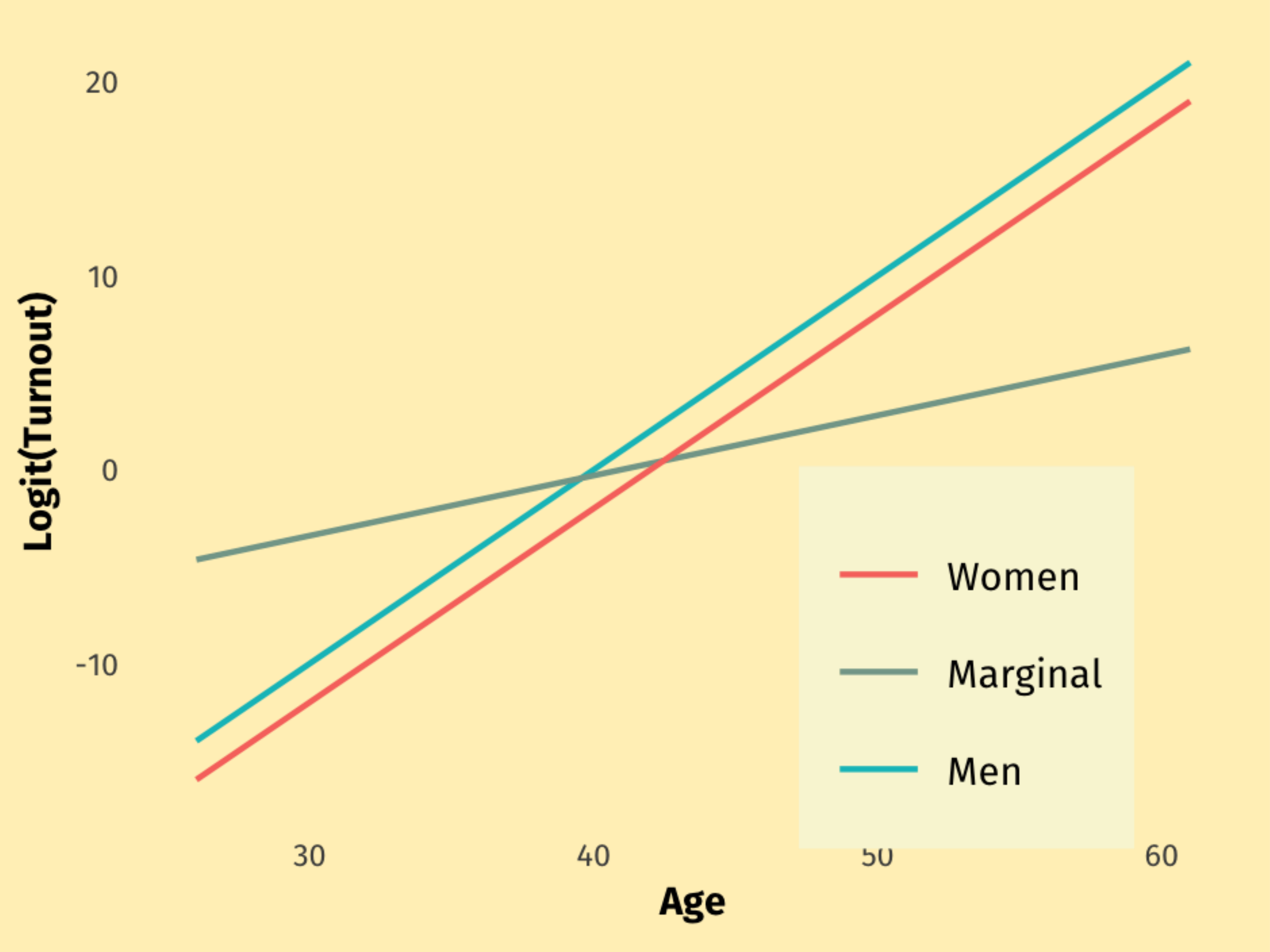
Collapsibility - Logistic regression

$logit(turnout) \sim Binomial(-80 + 1 \cdot age + 2 \cdot gender + 1 \cdot income, 1)$

Slopes across models are not the same!

squishing/*collapsing* the subpopulation lines doesn't give the population effect.

Logistic regression is noncollapsible.



Questions?

Latent variable explanation

(Un)explained variance in linear regression

In linear regression, the dependent variable has fixed variance.

We can predict/"explain" this variance by adding predictors into the model.

(Predicted variance measured by R^2)

In logistic regression, this is not the true!

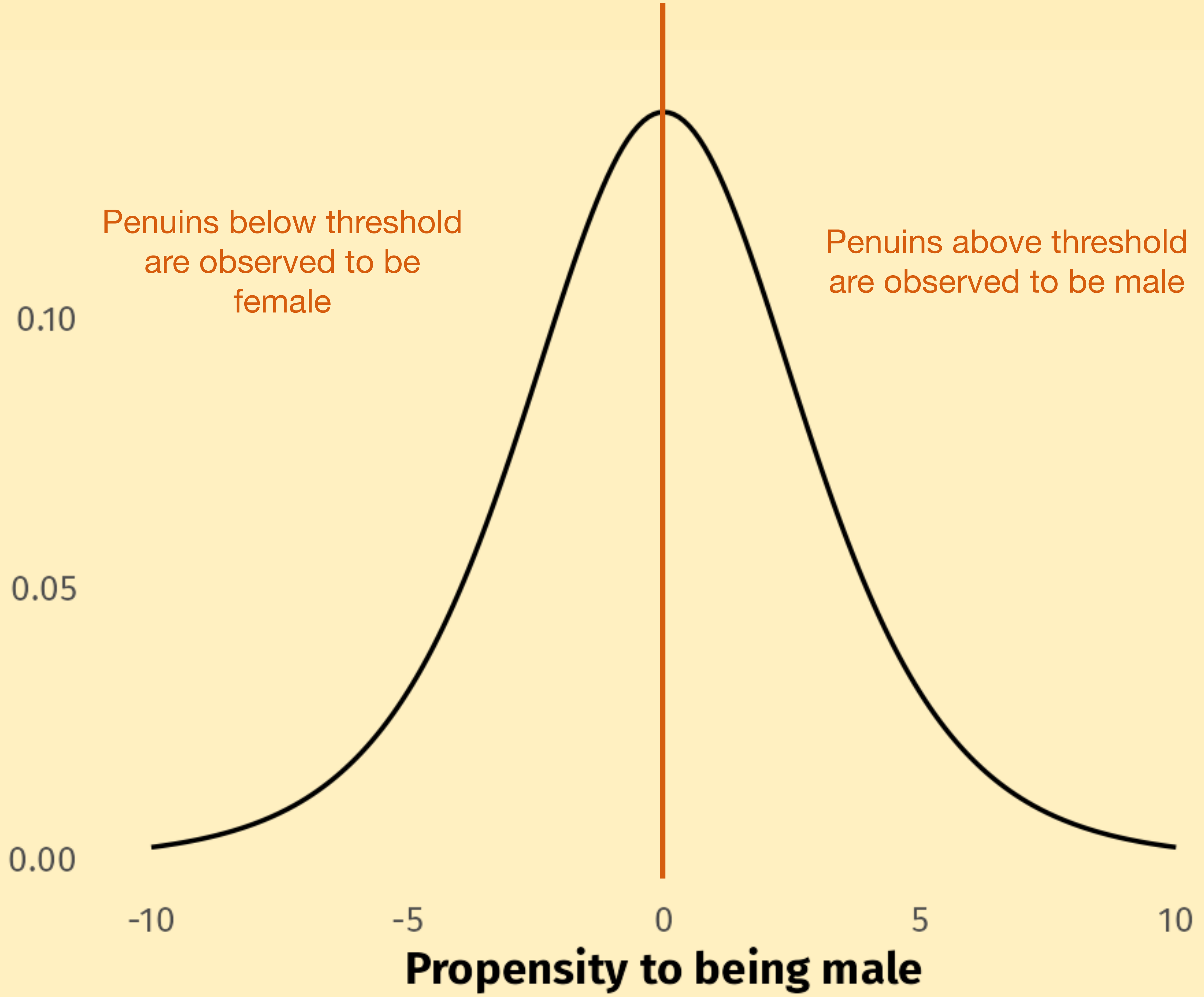
(Un)explained variance in logistic regression

Logistic regression assumes there is a latent (= unobservable) variable.

(e.g. propensity for being male, propensity to go to elections)

The observed binary variable is a manifestation of the latent one.

We are estimating threshold on the latent scale.

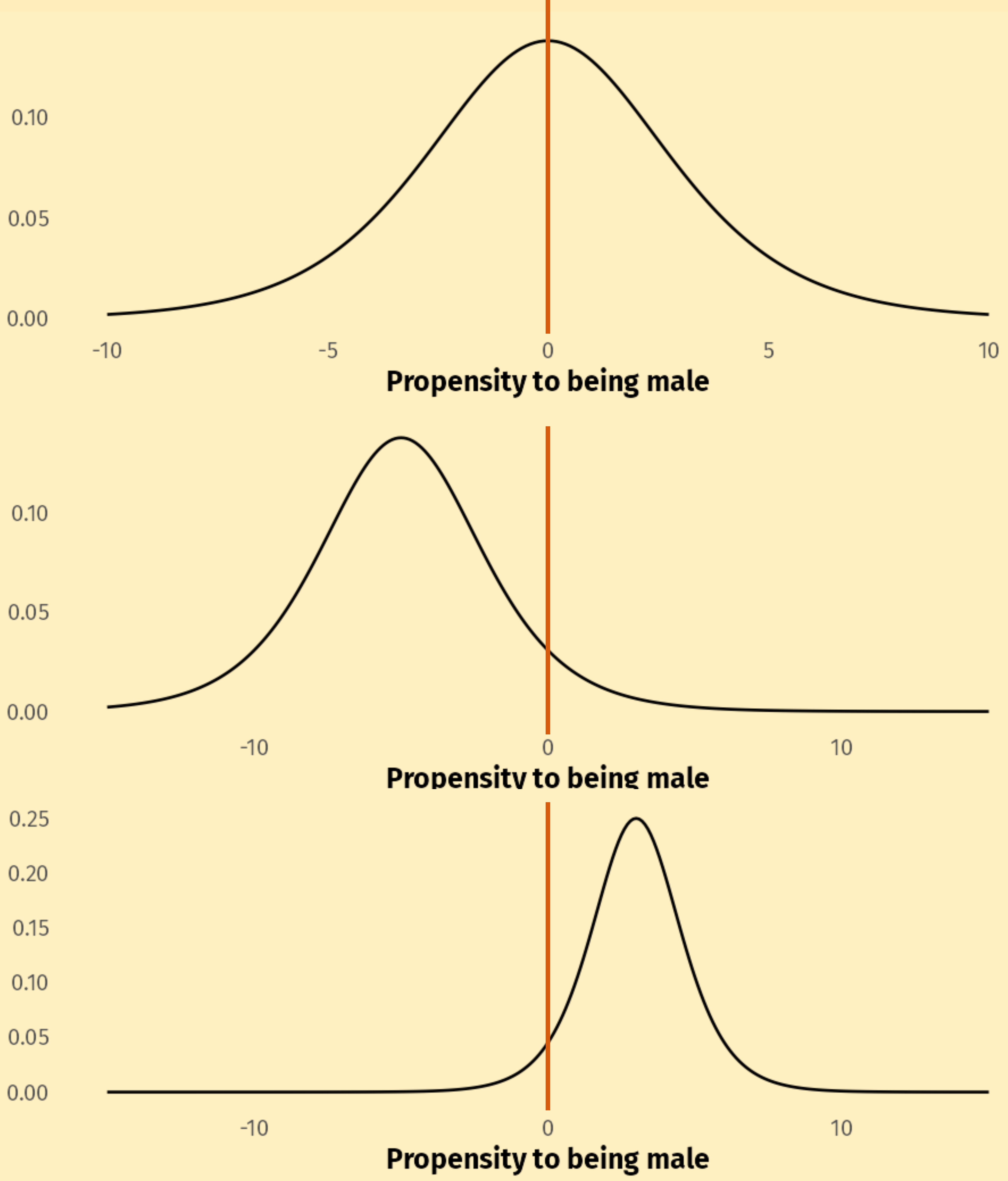


(Un)explained variance in logistic regression

The problem: To estimate the threshold, we need to know what the latent variable looks like.

We don't know what it looks like.

Solution: Just assume a fixed mean and standard deviation (usually 0 and 1.81)



Core of noncollapsibility

Null model: $logit(turnout) \sim Binomial(\beta_0, 1)$

The assumed total variance of the dependent variable is 1.81

Age model: $logit(turnout) \sim Binomial(\beta_0 + \beta_1 \cdot age, 1)$

The assumed total variance of the dependent variable is 1.81 + var. explained by age!

The total variance of the dependent variable changes, based on predictors!

In logistic regression, coefficients are in different units, depending on population and predictors.

Questions?

Non-collapsibility so far...

1. Logistic regression can't be interpreted using regression coefficients.
2. Coefficients are non-collapsible due to being in different units - we can't get the population slopes by squishing subpopulation ones together.

How to solve noncollapsibility

1) The good solution - Marginal effects on probability scale

2) The quick & dirty solution - Linear probability models

Marginal effects

Marginal effects

Problem - the regression coefficients are not interpretable on logistic scale.

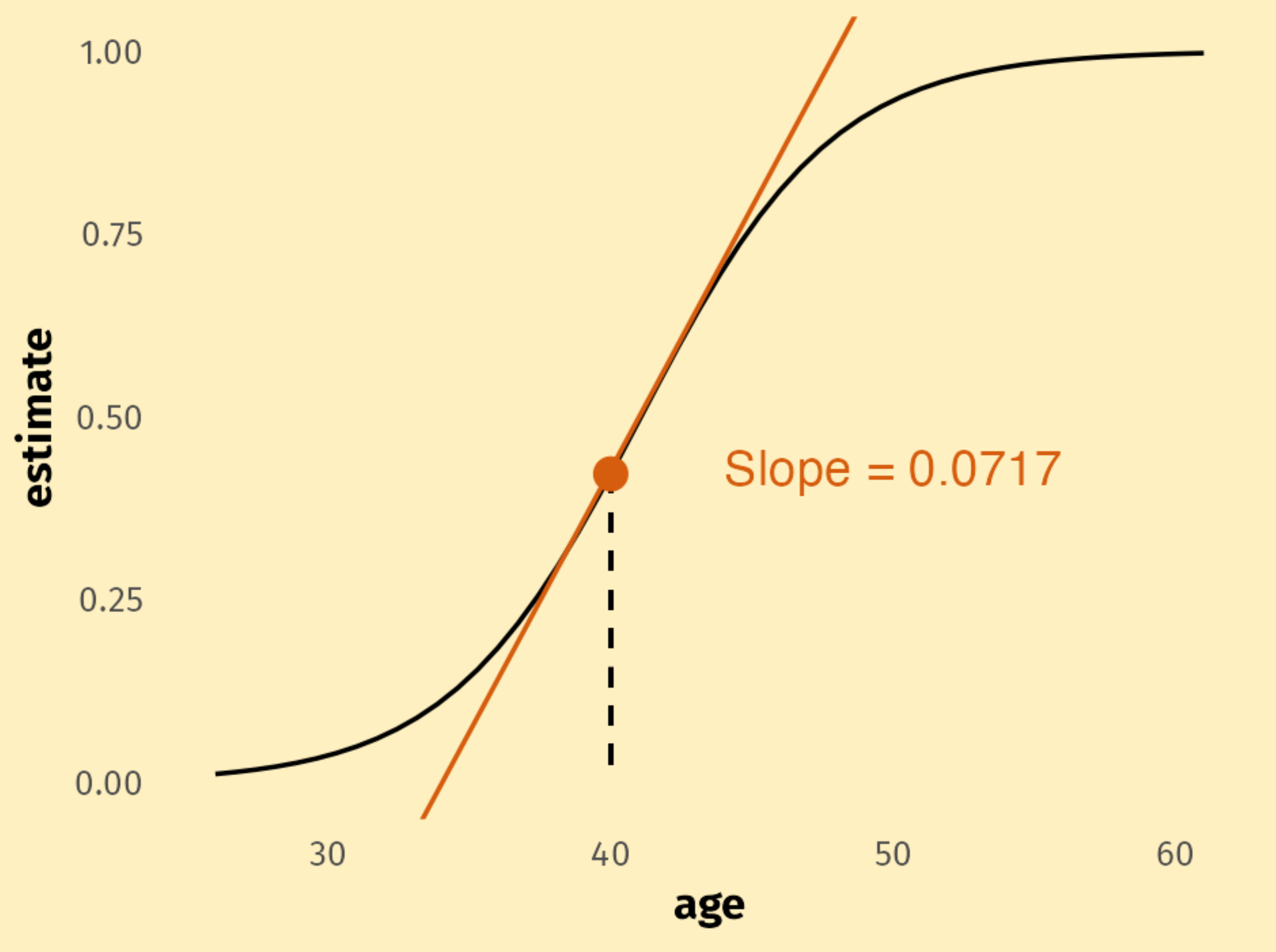
Solution - Work on probability scale! Enter marginal effects.

Marginal effects are slopes

Marginal effect - slope of the regression line at given value of predictor.

Marginal effects at the age of 40 is 0.0717

For people who are 40 years old, one year change in age is associated with 0.0717 increase in the probability of going to elections.



Marginal effects as difference in predictions

Computing slopes analytically is pretty hard. We approximate them instead.

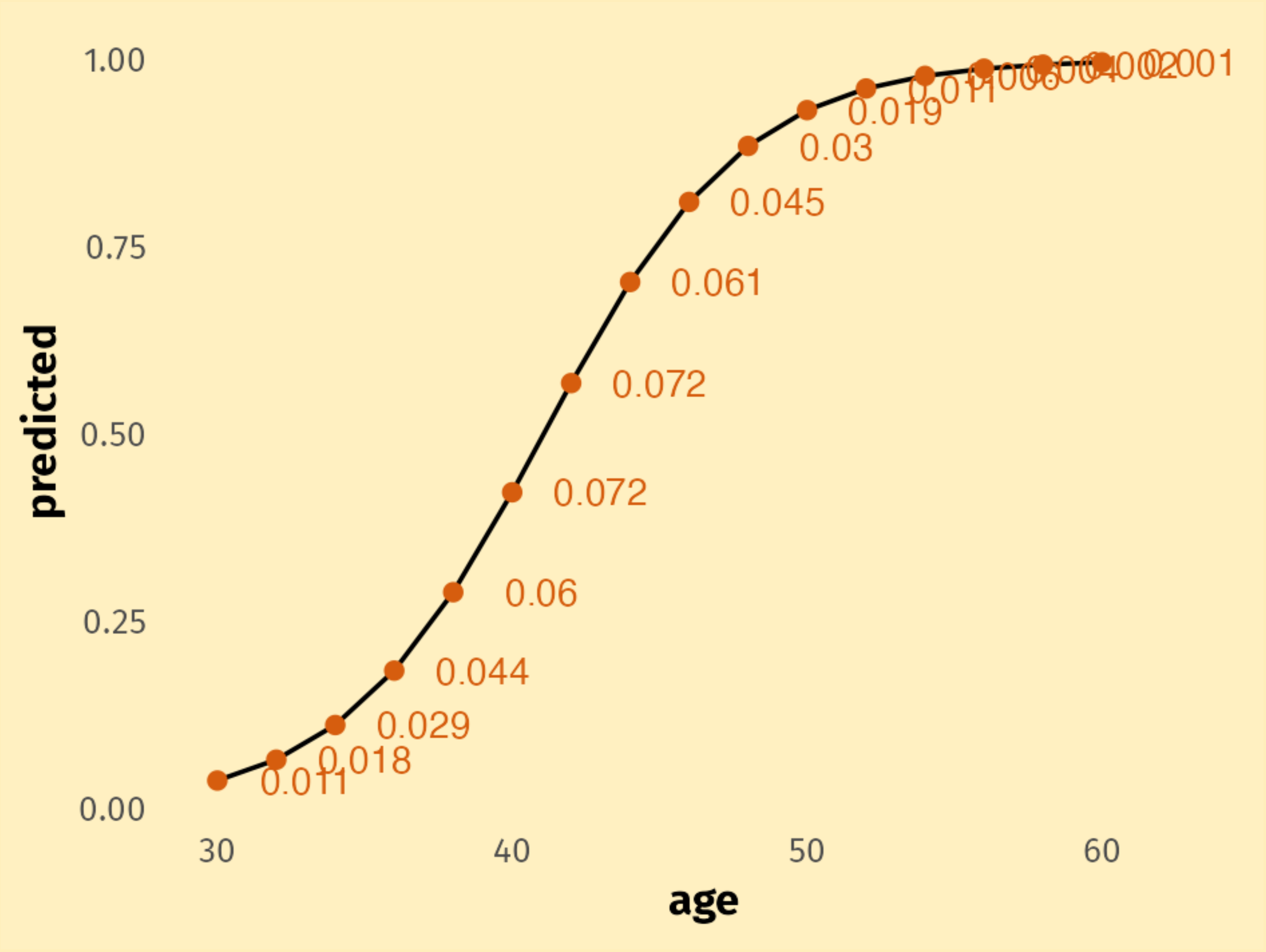
Age	Predicted Turnout for Age	Age + 0.001	Predicted Turnout for Age + 000.1	Difference * 100 (aka marginal effects)
40	0.4221270	40.001	0.4221986	0.00717
41	0.4949359	41.001	0.4950093	0.00734
41	0.4949359	41.001	0.4950093	0.00734
42	0.5679603	42.001	0.5680323	0.00721
43	0.6381457	43.001	0.6382135	0.00679
44	0.7028937	44.001	0.7029551	0.00613
44	0.7028937	44.001	0.7029551	0.00613
45	0.7604060	45.001	0.7604595	0.00535

Average marginal effects

Individual marginal effects (for each observation) are hard to interpret.

Average marginal effects (AME) = average of individual effects (duh)

On average, people who are 1 year older have 0.0563 points higher probability to go to elections.



(Average) marginal effects are collapsible!

On average, people who are 1 year older have 0.0563 points higher probability to go to elections.

They are the same!

	Model 1	Model 2	Model 3
Age	0.056	0.056	0.056
Gender		0.124	0.124
Income			0.061

Marginal effects

Advantages

- Probabilities are easy to interpret.
- Collapsibility is gone.

Disadvantages

- Marginal effects are not linear (so always also plot look at the prediction plot)
- Not available in all software (looking at you, SPSS)

Linear probability models

Statistics has a dirty secret....

You *can* use linear regression for binary data!*

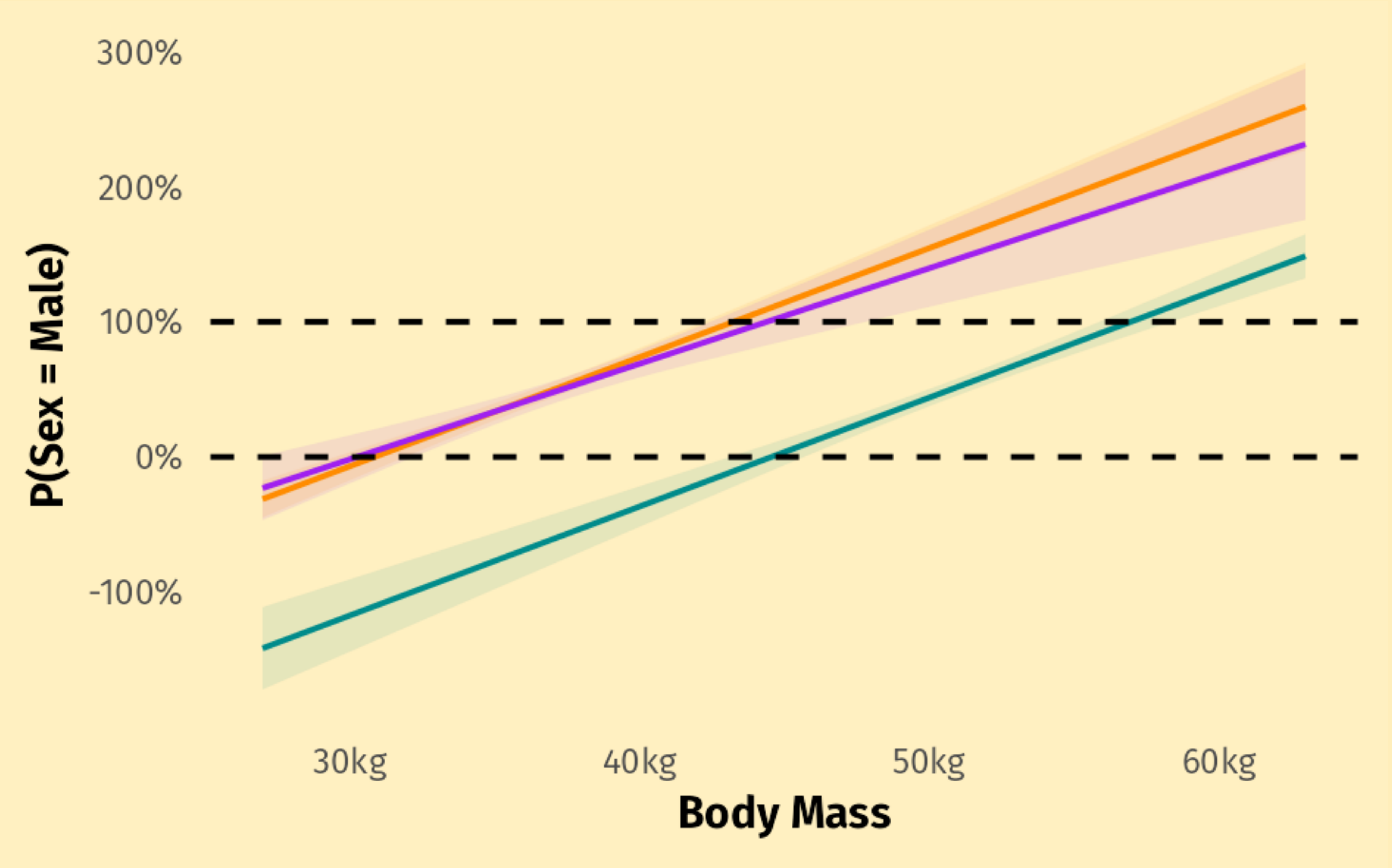
(*Sometimes)

Linear probability models...

... are just linear regression with binary dependent variable.

This shouldn't work.

But with some tweaks, it can actually give you correct average marginal effects estimates.



It just works...

Just use lm() as usual.

Notice that the coefficients are the same as average margin effects from logistic regression!

	Model 1	Model 2	Model 3
Age	0.056	0.056	0.056
Gender		0.124	0.124
Income			0.0612

… but there are caveats

Linear regression assumptions are still important.

Normality assumptions is violated (so std. errors. are wrong), but can be ignored if sample size is big enough.

Homoscedasticity assumption is violated (so std. errors are wrong), but you can salvage it by using robust standard errors.

Linearity assumption is violated, but if you are lucky, you can still get good marginal effects estimate.

Sometimes it doesn't work.

"If the main purpose of estimating a binary response model is to approximate the partial effects of the explanatory variables, ... then the LPM often does a very good job.

But there is no guarantee that the LPM provides good estimates of the partial effects for a wide range of covariate values, especially for extreme values of x."

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data, second edition. MIT Press.

Linear probability models

Advantages:

- Super easy to implement in every software.
- Often provide good approximation of Average marginal effects.

Disadvantages

- Sometime just fails for (seemingly) no reason.

Questions?

Non-collapsibility in nutshell

1. Logistic regression can't be interpreted using regression coefficients.

2. Coefficients are non-collapsible due to being in different units - we can't get the population slopes by squishing subpopulation ones together.

3. Interpret logistic regression using (average) marginal effects on probability scale.

InteRmezzo!

Fit indices

Model fit indices

In linear regression, we can asses how well the model fits the data, we use indices like coefficient of determination (R^2).

In generalised linear models, we can't use R^2.

Two options instead:

1) Information criteria.

2) Pseudo R^2

Information criteria

How much information we loose by reducing data into model.

The number itself is not interpretable.

Information criteria are used to compare models against each other.

Aikake Information Criterium (AIC)	Penalizes for number of predictors, basically R squared
Bayesian Information Criterium (BIC)	Penalizes for number of observations

Information criteria for our models

	Predictors	AIC	BIC
Model 1	Age	1 128 000	1 128 000
Model 2	Age, Gender	1 108 000	1 108 000
Model 3	Age, Gender, Income	254 700	254 700

Lowest number means, using this model leads to smallest loss of information

Questions?

Pseudo Coefficient of Determination

Tries to imitate classical R^2 - values between 0 and 1.

Based on comparing intercept-only and our models (usually).

Actually many different versions (Cohen's, Cox & Snell's, Nagelkerke's, Tjur's).

Caveats:

- Not actually proportion of explained variance.

- Magnitude has to be interpreted differently from LM (e.g. 0.8 is actually almost perfect)

Pseudo R^2 for our models

	Predictors	Pseudo R Sqr. (Tjur's version)
Model 1	Age	0.204
Model 2	Age, Gender	0.222
Model 3	Age, Gender, Income	0.841

Highest number means this models fits best.

Fit indices

- We can't use classical R^2
- Two other options:
  - Information criteria - how much information we lose (AIC, BIC)
  - Pseudo R^2 - kinda like "explained variance" (but not really)

Questions?

Assumptions

Logistic regression assumptions

1) Valid and reliable measurement

2) Representative sample

3) Linearity (between logit and predictors)

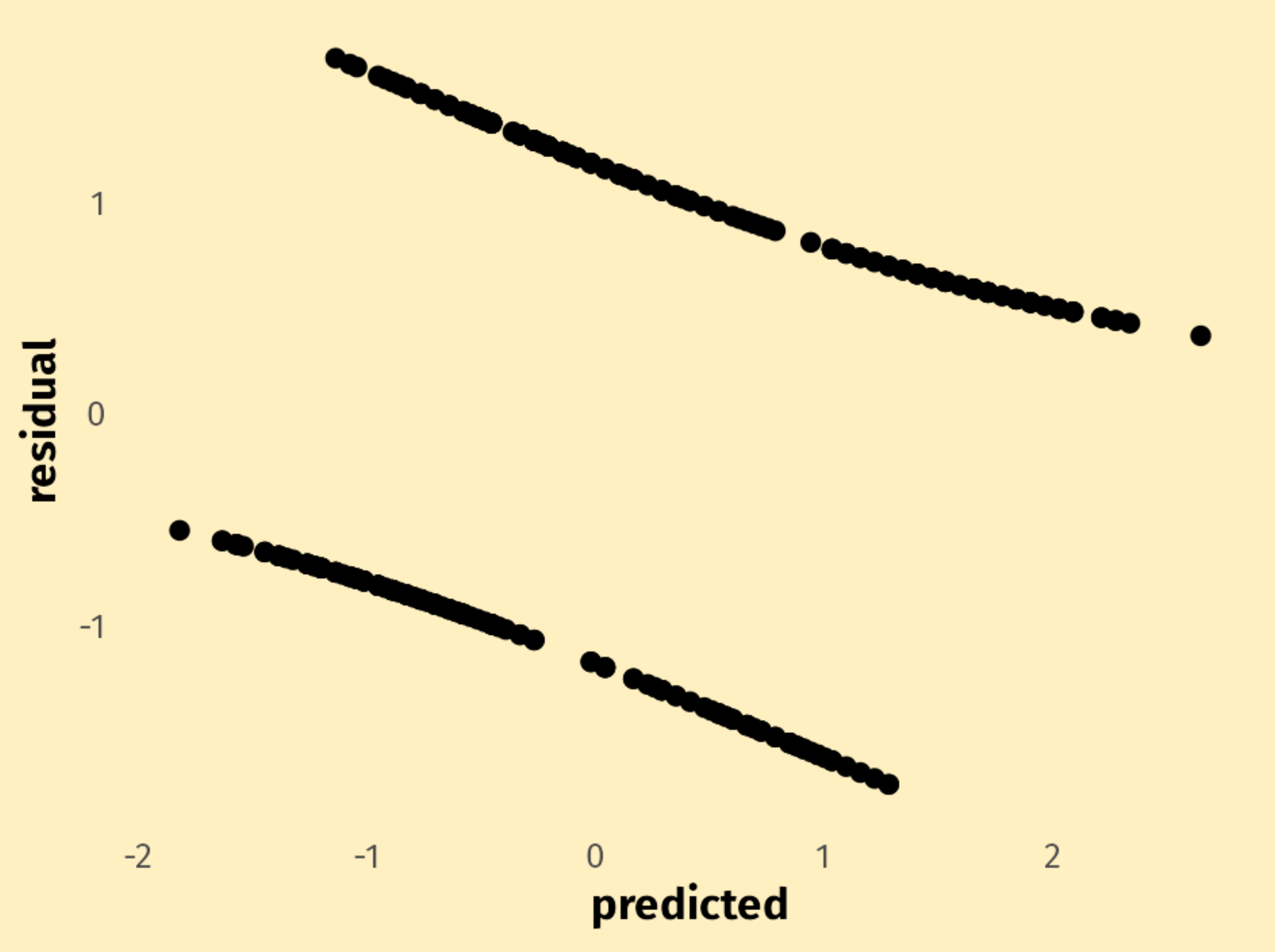
4) Independence of observations

Linearity Assumption

Classical residual plots are not great here.

Observed values are either 0 or 1.

Residual plot will look even if the model is correct.



Linearity Assumption - what to do?

If not classical residuals, then what?

1) Binned residuals (great, but only for logistic regression)

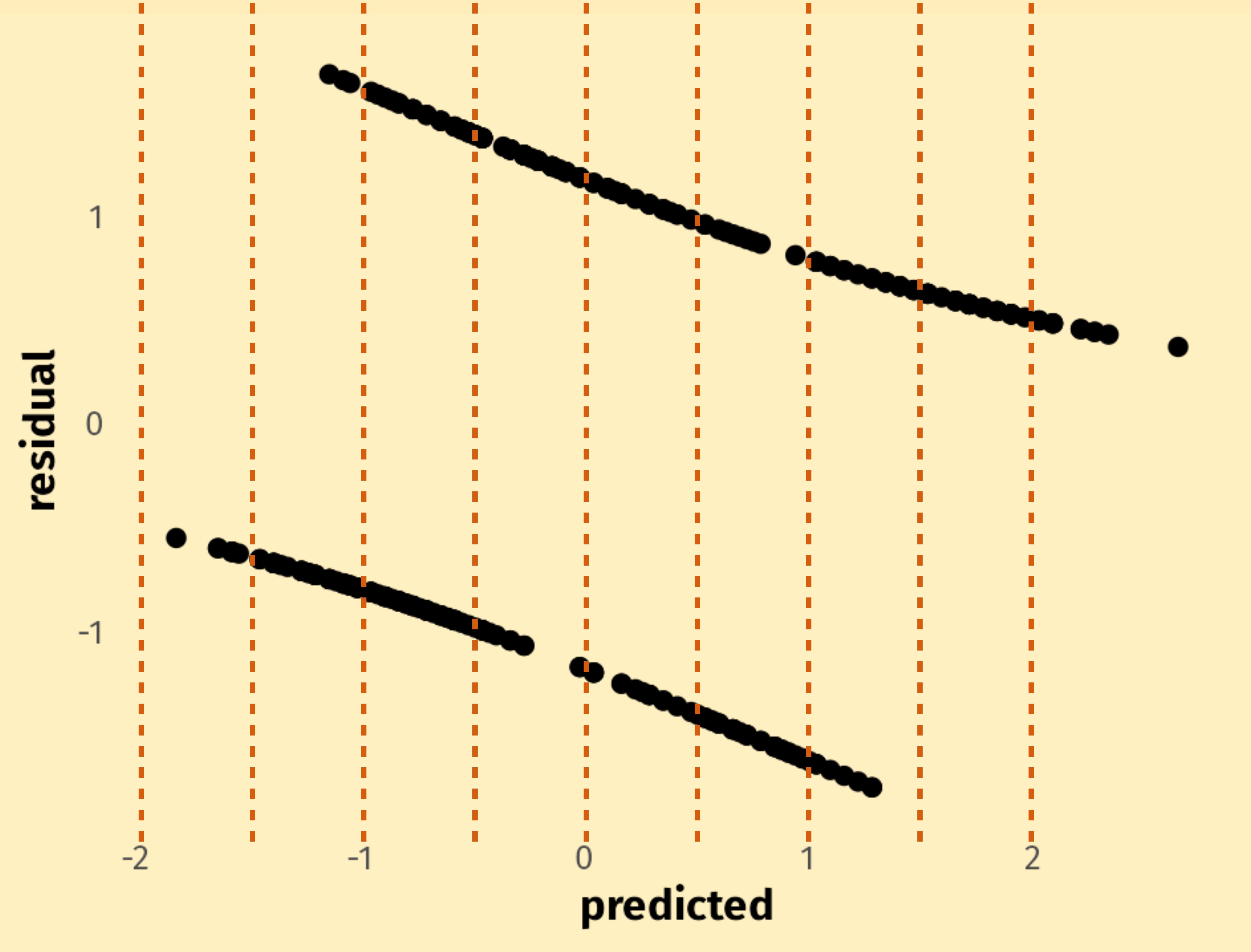
2) Randomized residuals (less common, but works for every model)

Linearity assumption - binned residuals

1) Cut the plots into bins.

2) Compute the mean of residuals for each bin.

3) Interpret as usual.

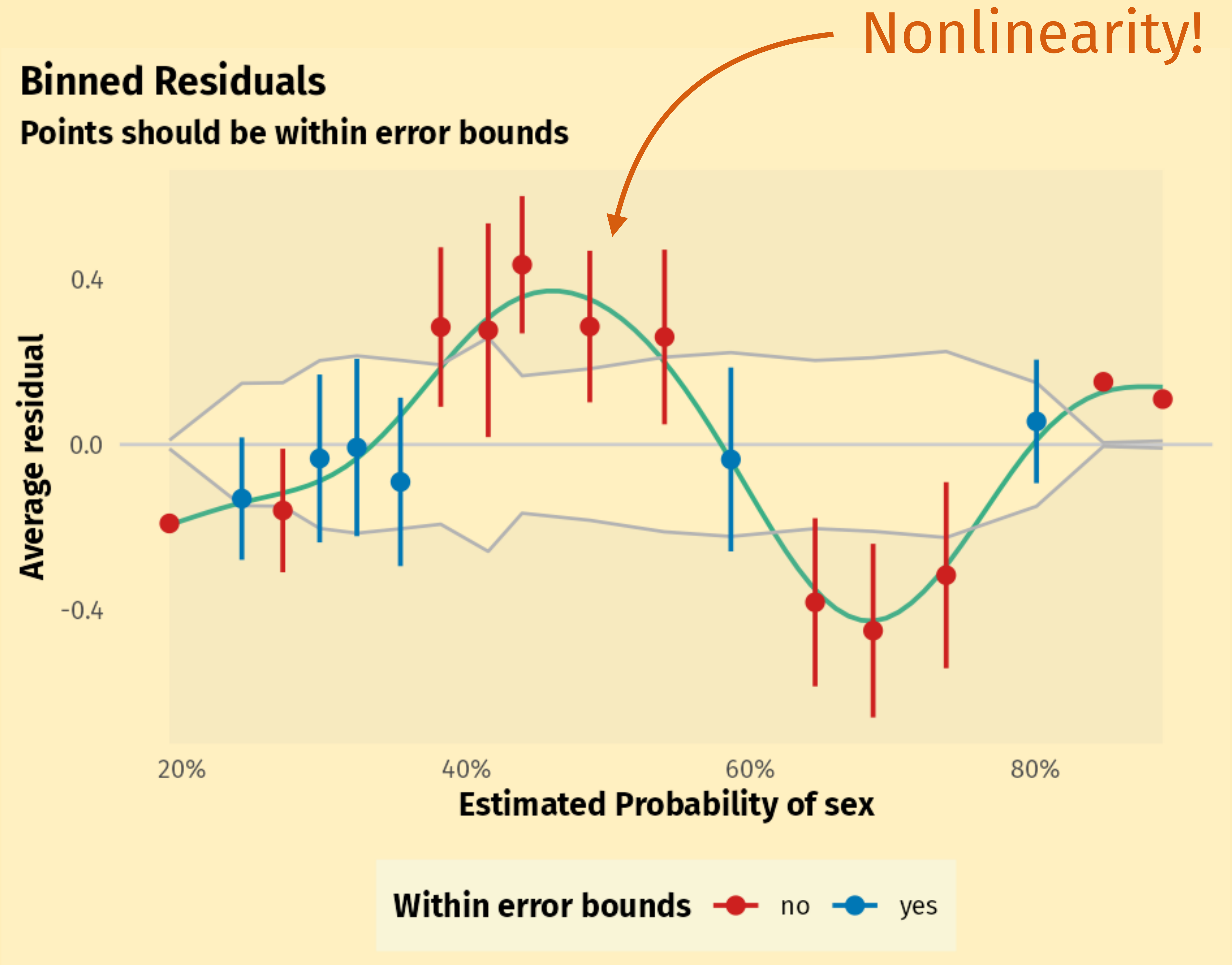


Linearity assumption - binned residuals

1) Cut the plots into bins.

2) Compute the mean of residuals for each bin.

3) Interpret as usual. Points should be spread around zero with no pattern.

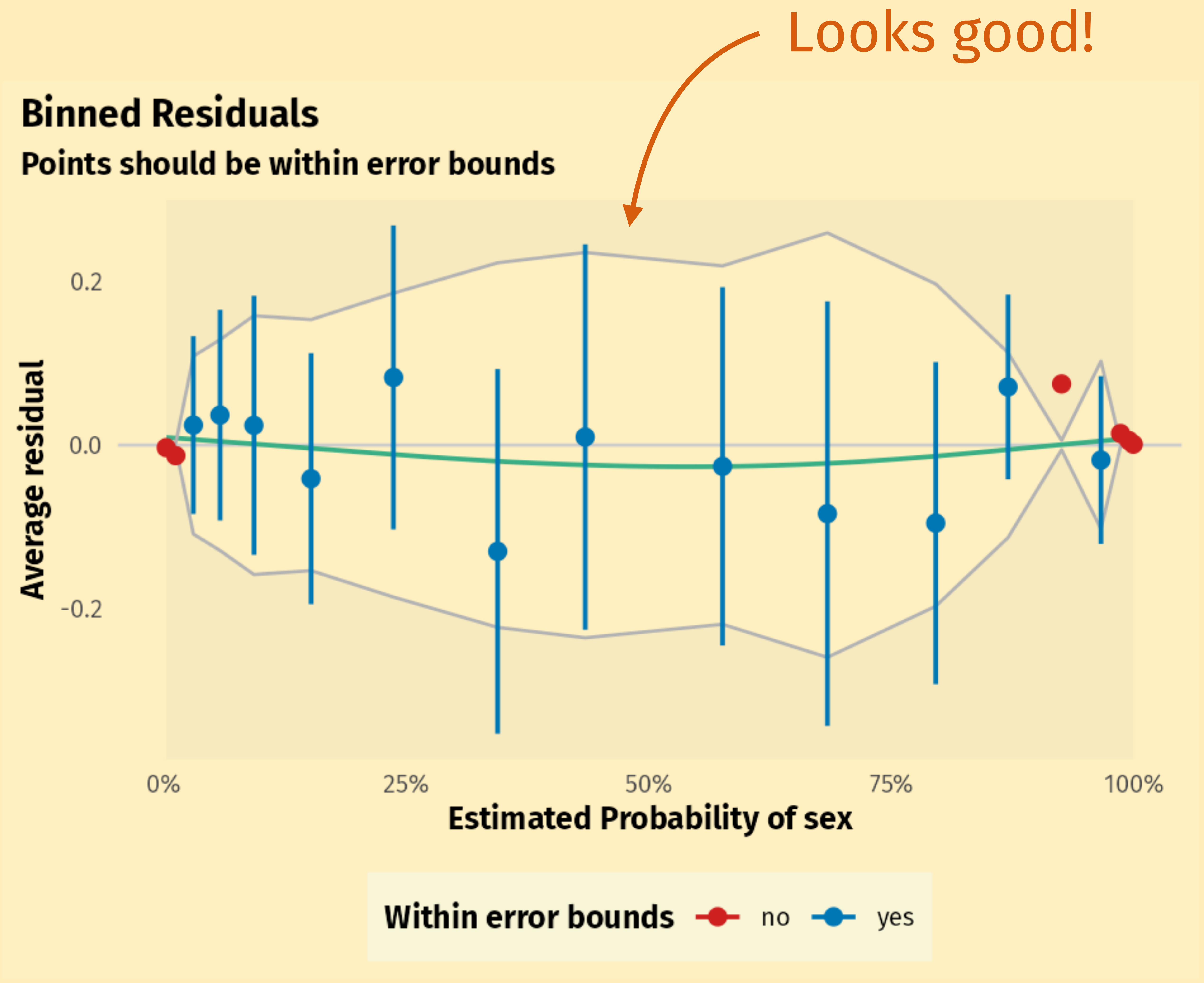


Linearity assumption - binned residuals

1) Cut the plots into bins.

2) Compute the mean of residuals for each bin.

3) Interpret as usual. Points should be spread around zero with no pattern.



Questions?

Linearity assumption - Randomised residuals

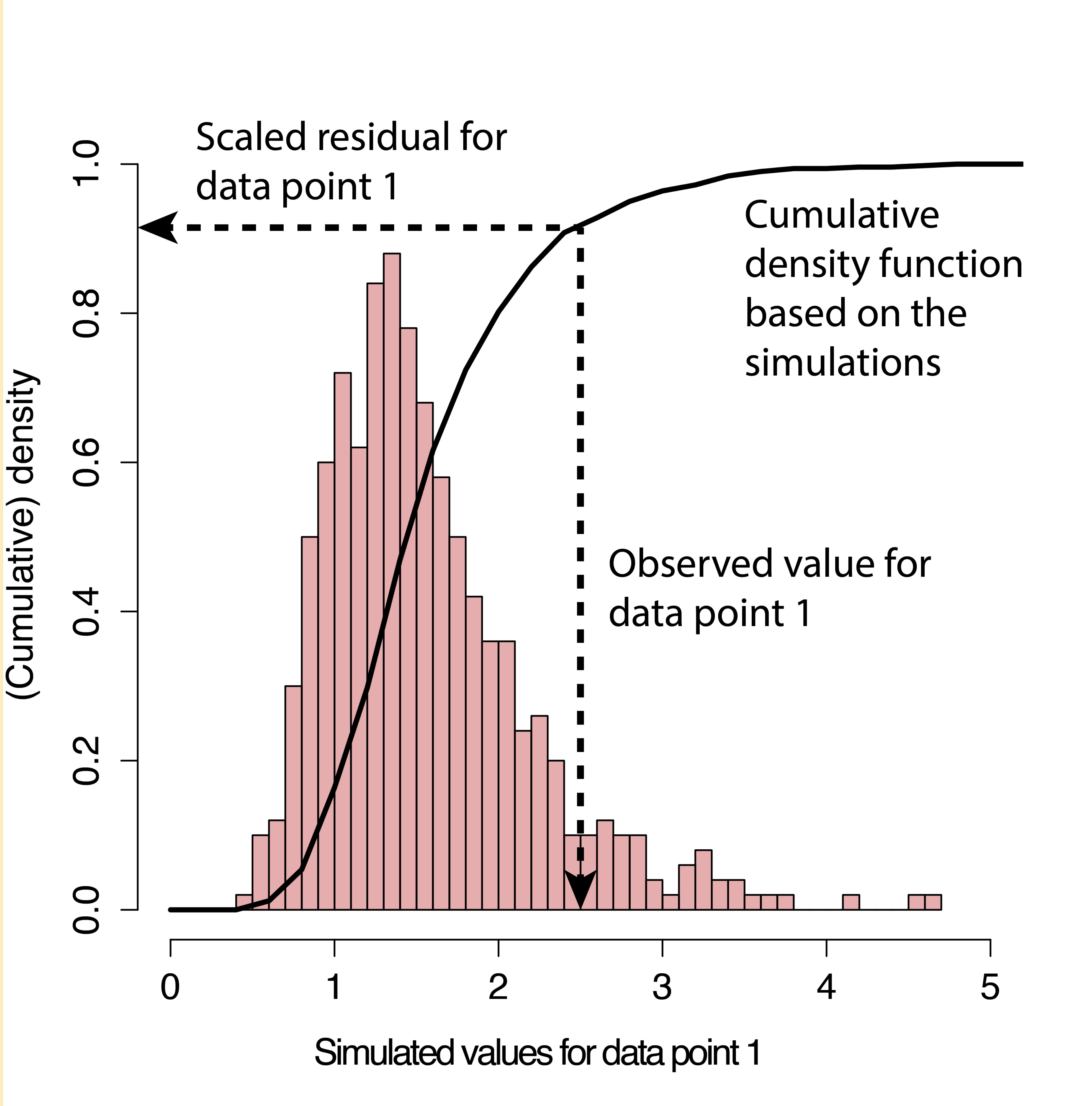
Randomised residuals based on simulating data from our model.

How to make them:

1) Simulate new values for every respondent/ observations based on you model.

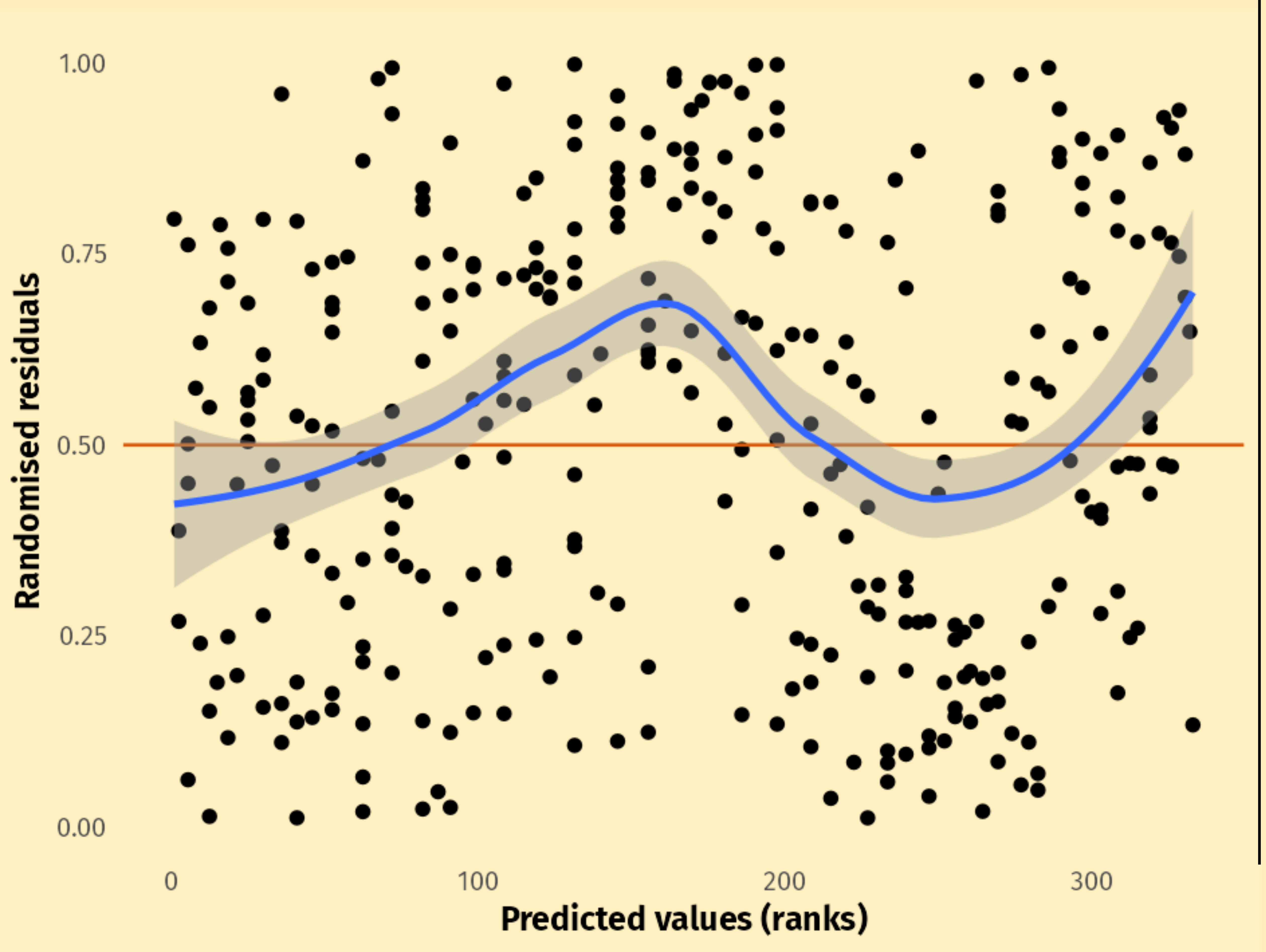
2) Count how many times the simulated values are lower than the observed value.

3) Randomised residual is the proportion of times the simulated values were lower than the observed value.

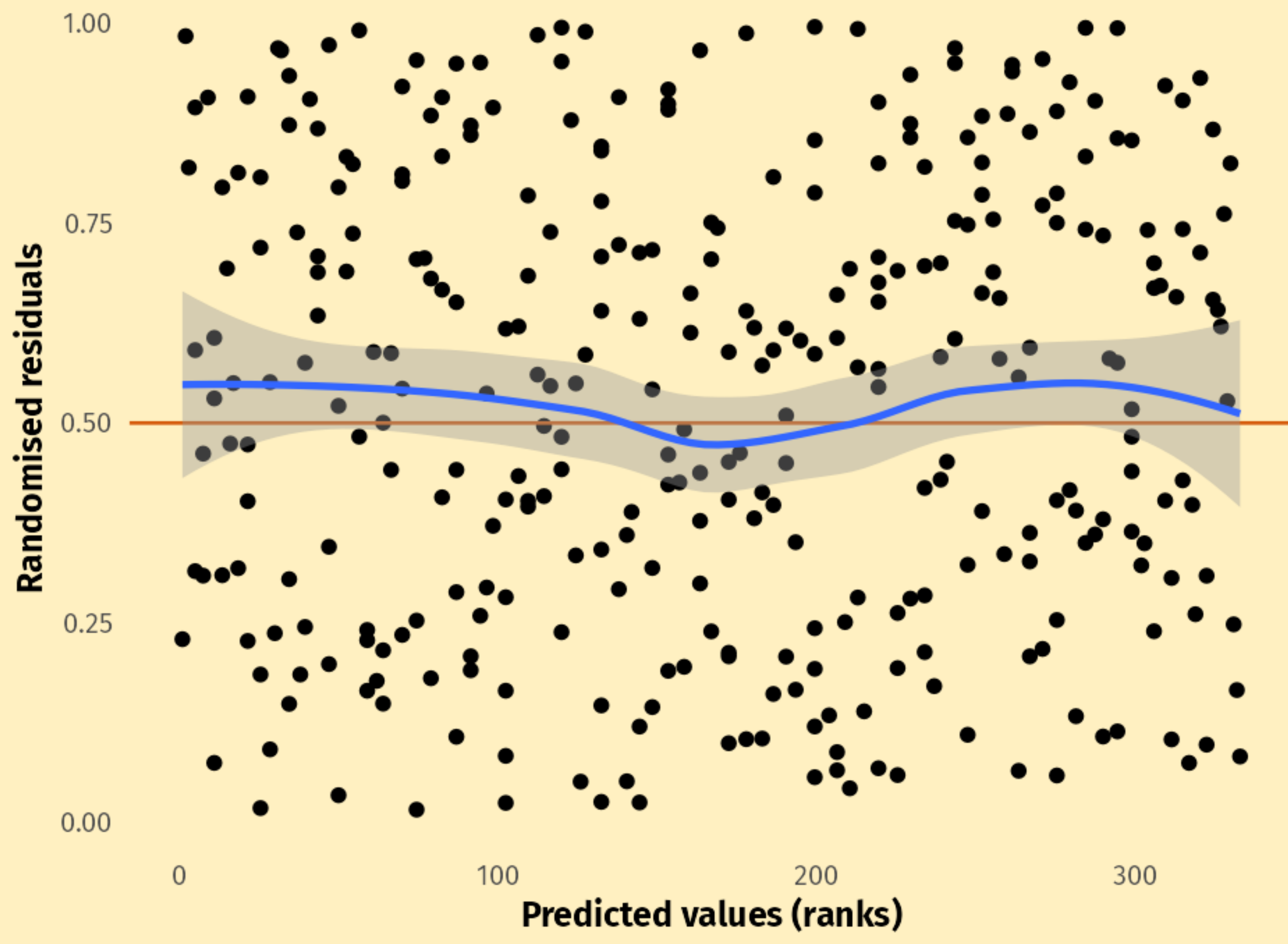


Linearity assumption - Randomised residuals

Linearity violated



Linearity fulfilled



Binned vs Randomised residuals

Binned residuals

- (Perhaps) easier to understand
- Only work with binary logistic regression

Randomised residuals

- Computation more involved
- Work for *every* model

For binary logistic regression, both approaches work!

Questions?

InteRmezzo!