Dealing with (Non)linearity

Applied Regression in R

Aleš Vomáčka

29. 03. 2026

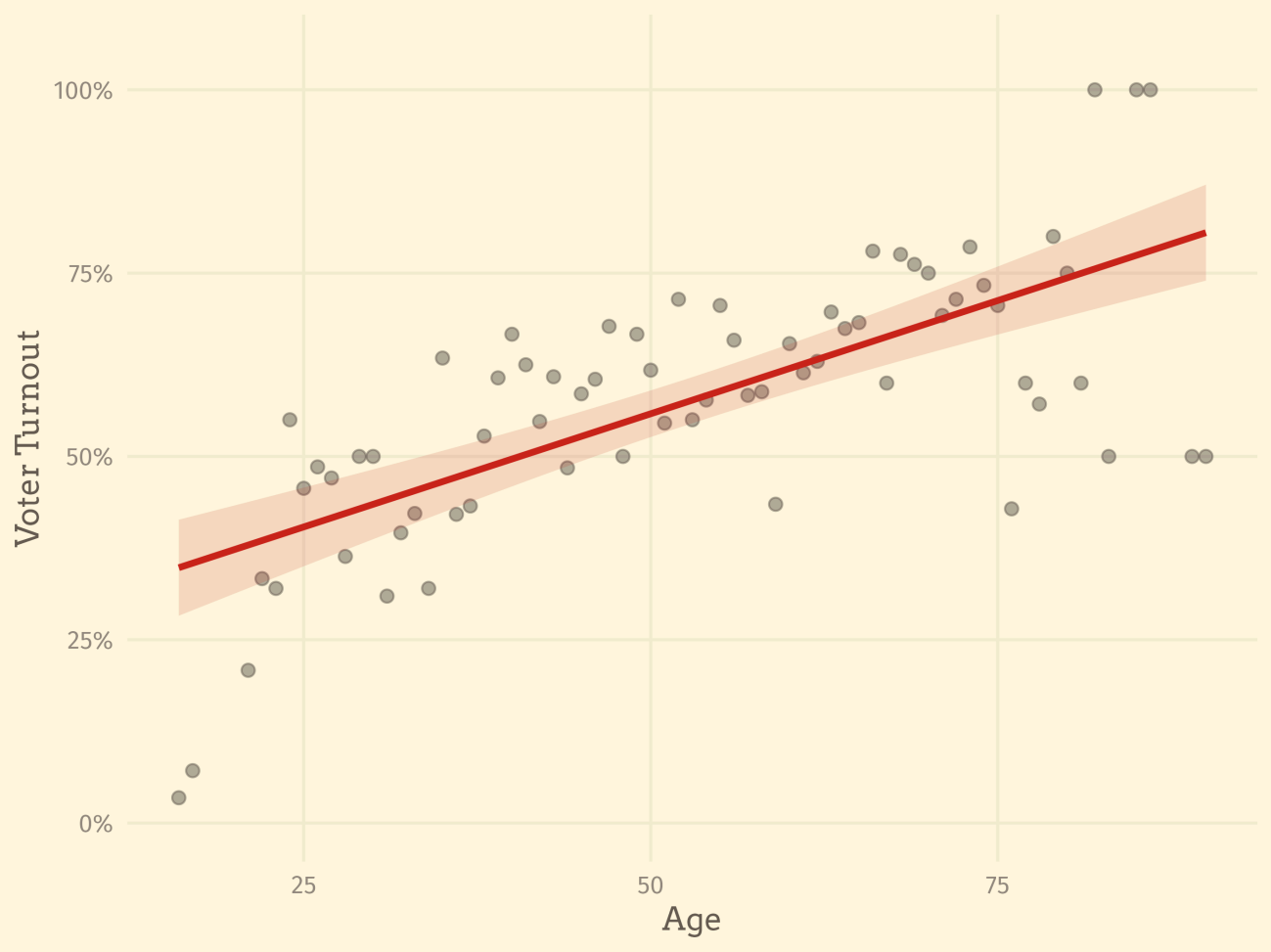
Faculty of Arts, Charles University

Lesson Goals

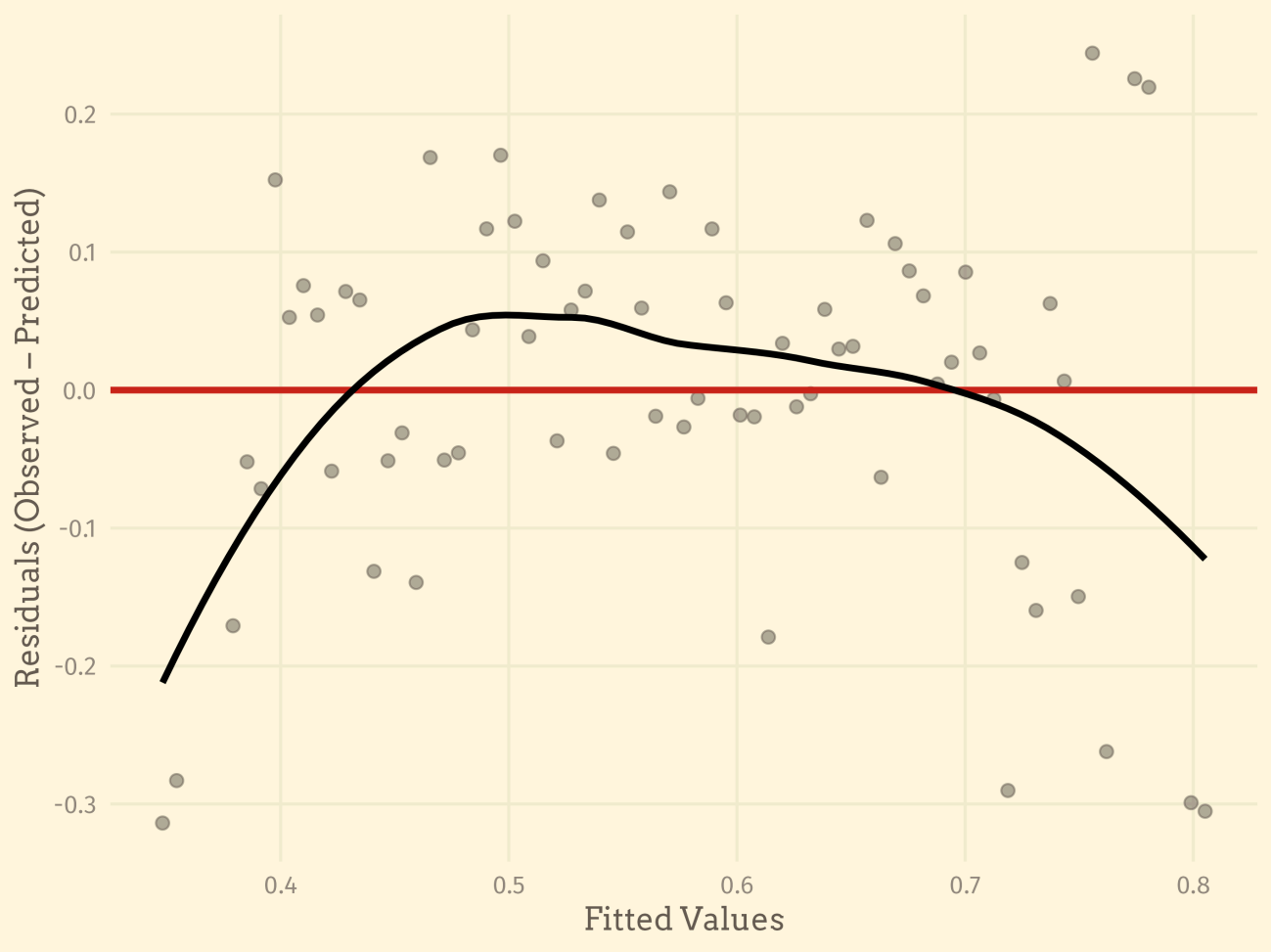
Learn about:

1. Categorisation
2. Simple Polynomials
3. Linear Splines
4. Natural Splines

Turnout versus Age



We Are Missing Some Nonlinearity

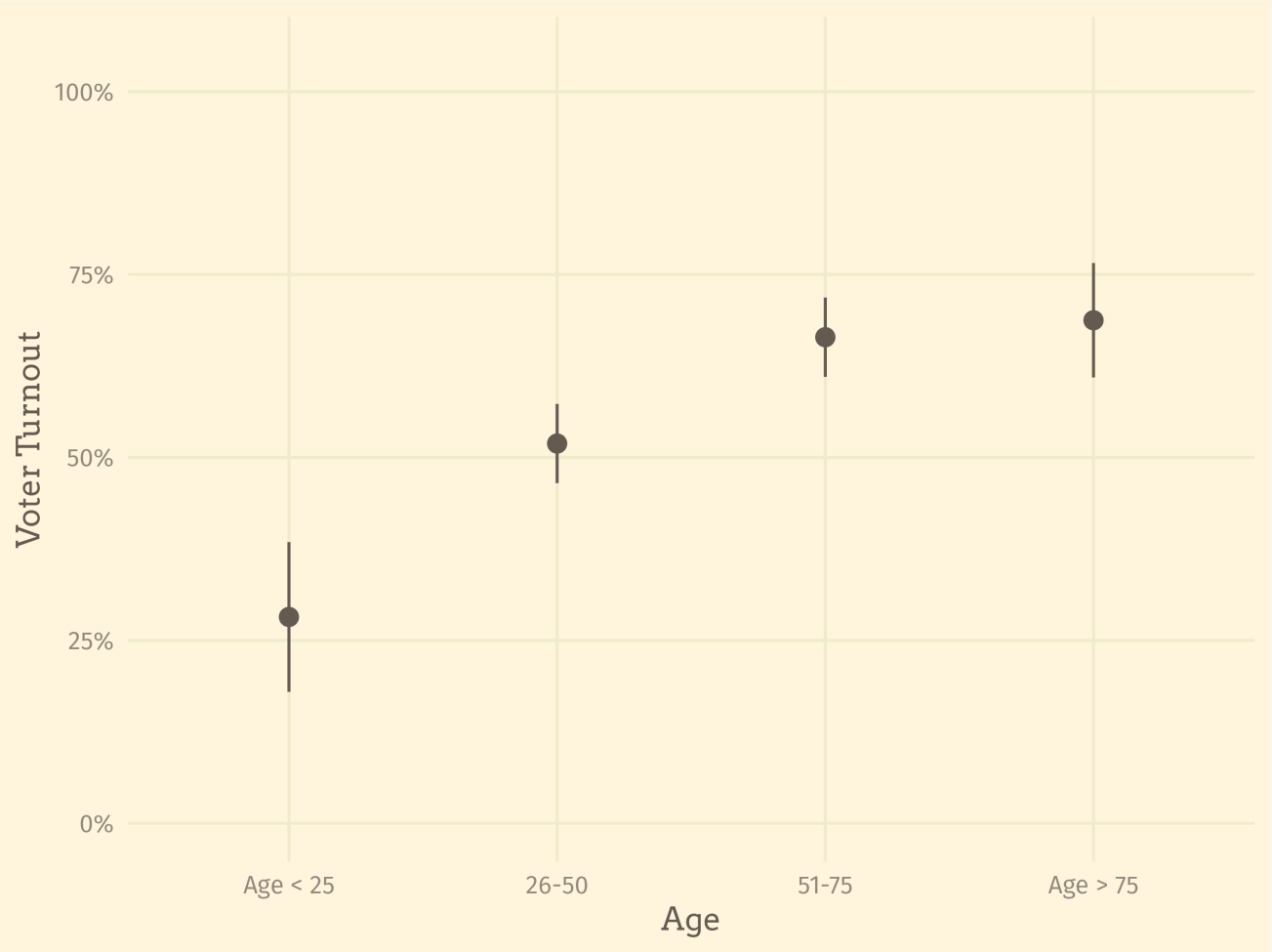


Categorisation

Categorisation

Transforming numeric variable into a **set of categories**.

For example, we can split age into four groups.



Many Ways to Cut the Data

There isn't an optimal way to categorise numerical predictors.

Some options are:

- Based on theory
- Set of groups with the same range (`cut_interval()`)
- Quantiles (`cut_number()`)
- Intervals with the same width (`cut_width()`)

Advantages and Disadvantages of Categorisation

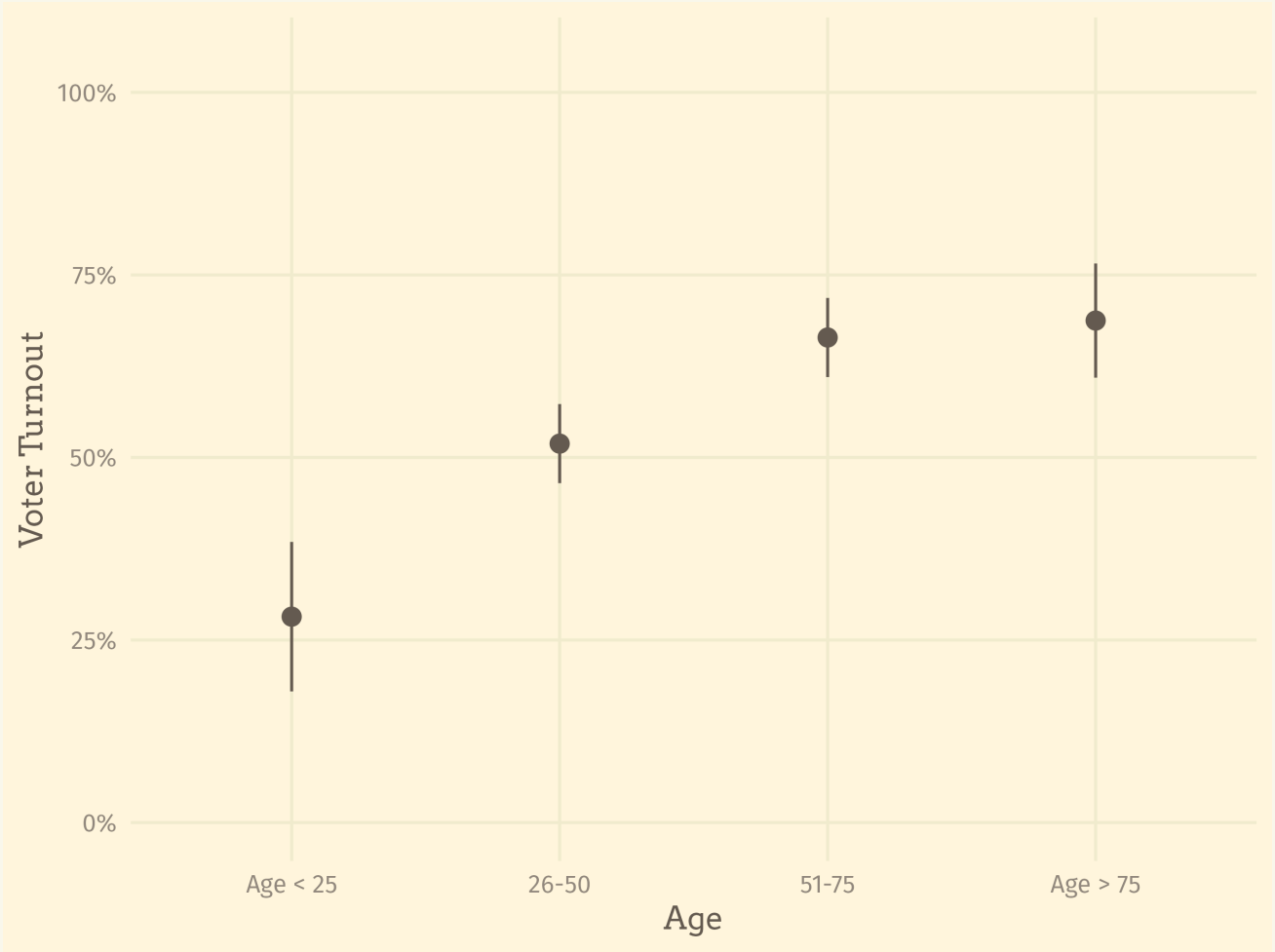
Advantages

- Easy to interpret and communicate, even to nontechnical audience.

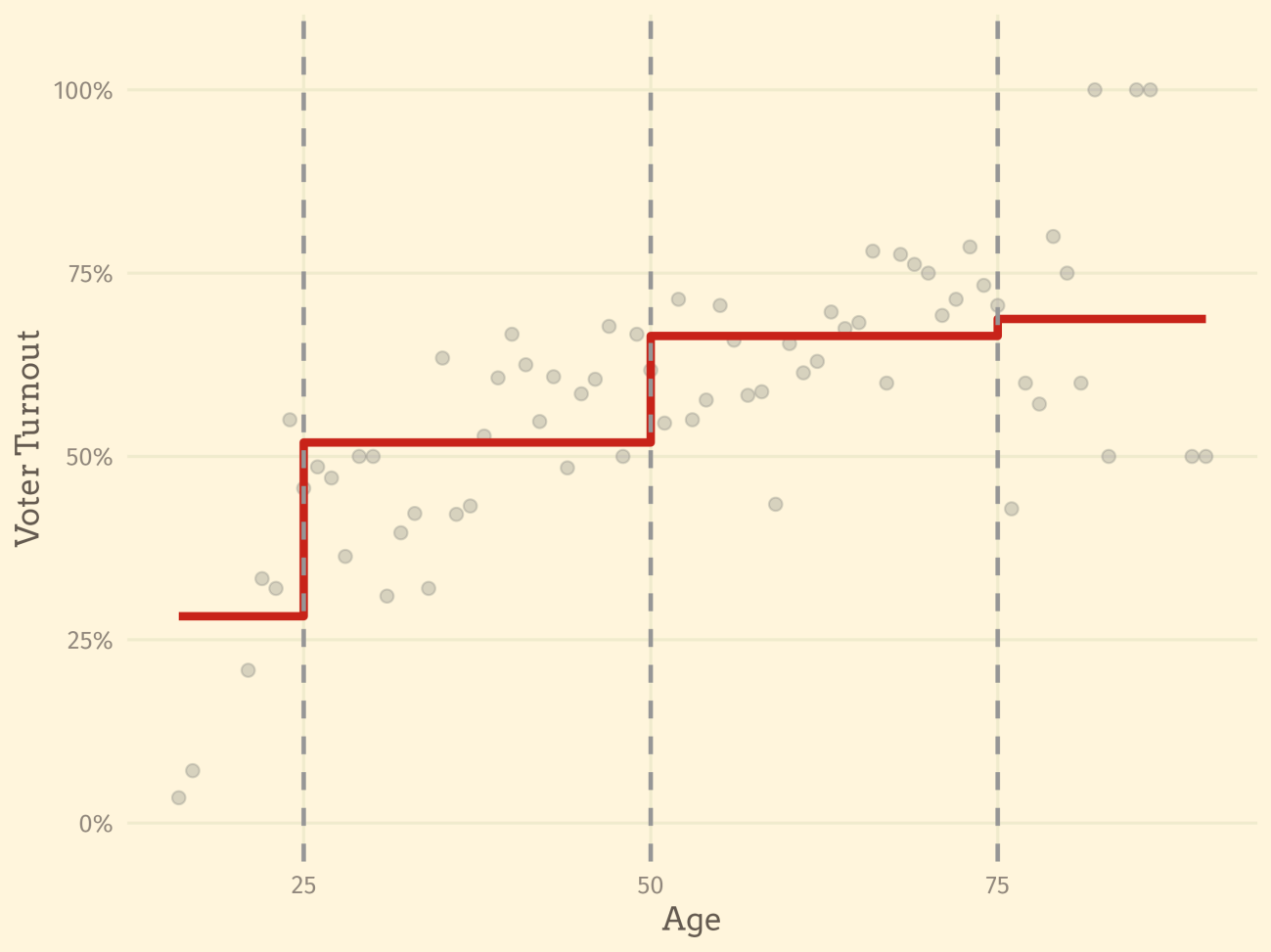
Disadvantages

- Less precision and power.
- Assumes the relationship is flat inside intervals.
- Assumes there is discontinuity between intervals.
- Cutpoints are arbitrary.

Categorised Predictors Showed As This...



… But the Model Actually Says This



Questions?

Simple Polynomials

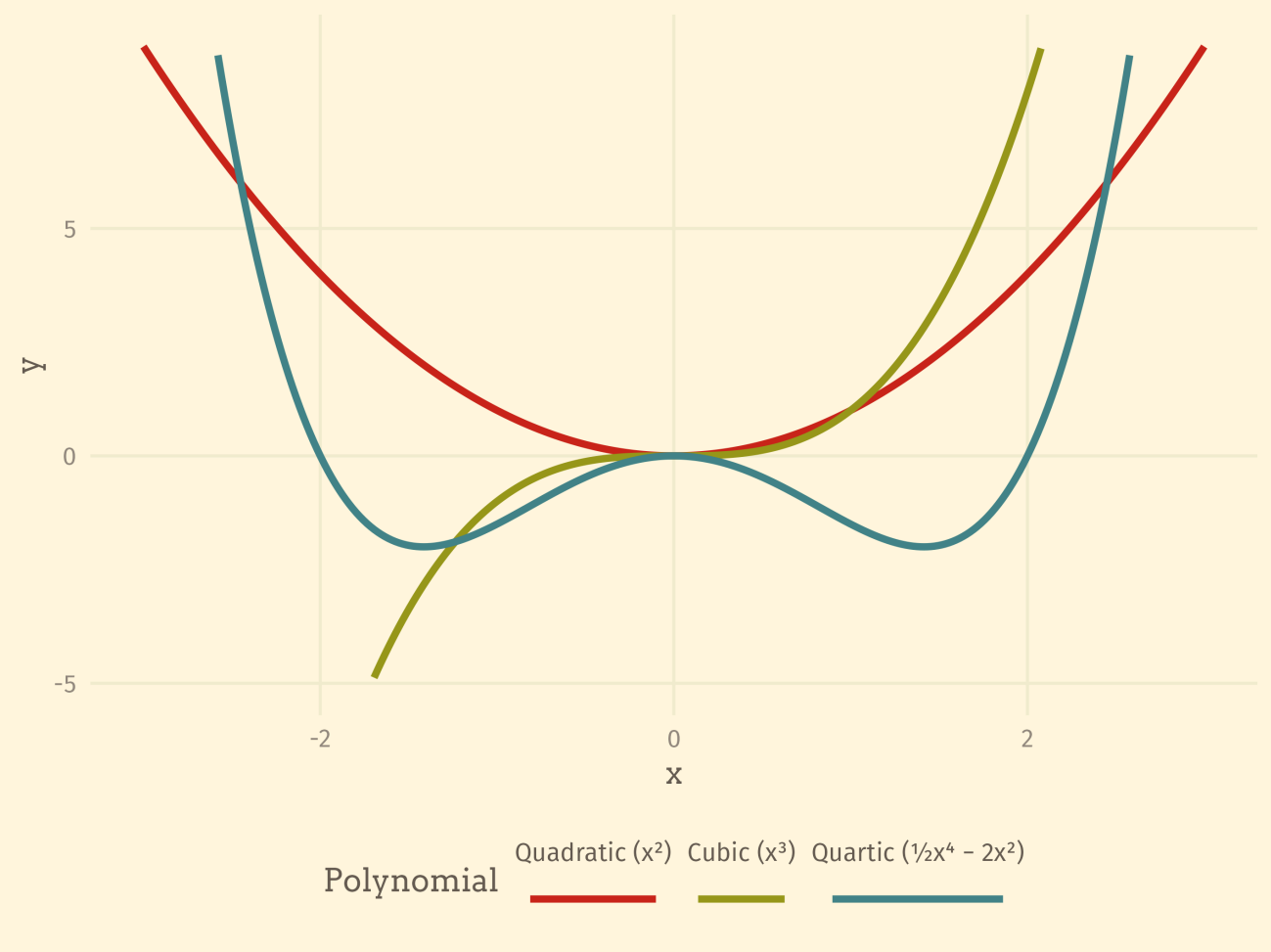
Simple Polynomials

How can we fit a **smooth function**, but still capture nonlinear relationships?

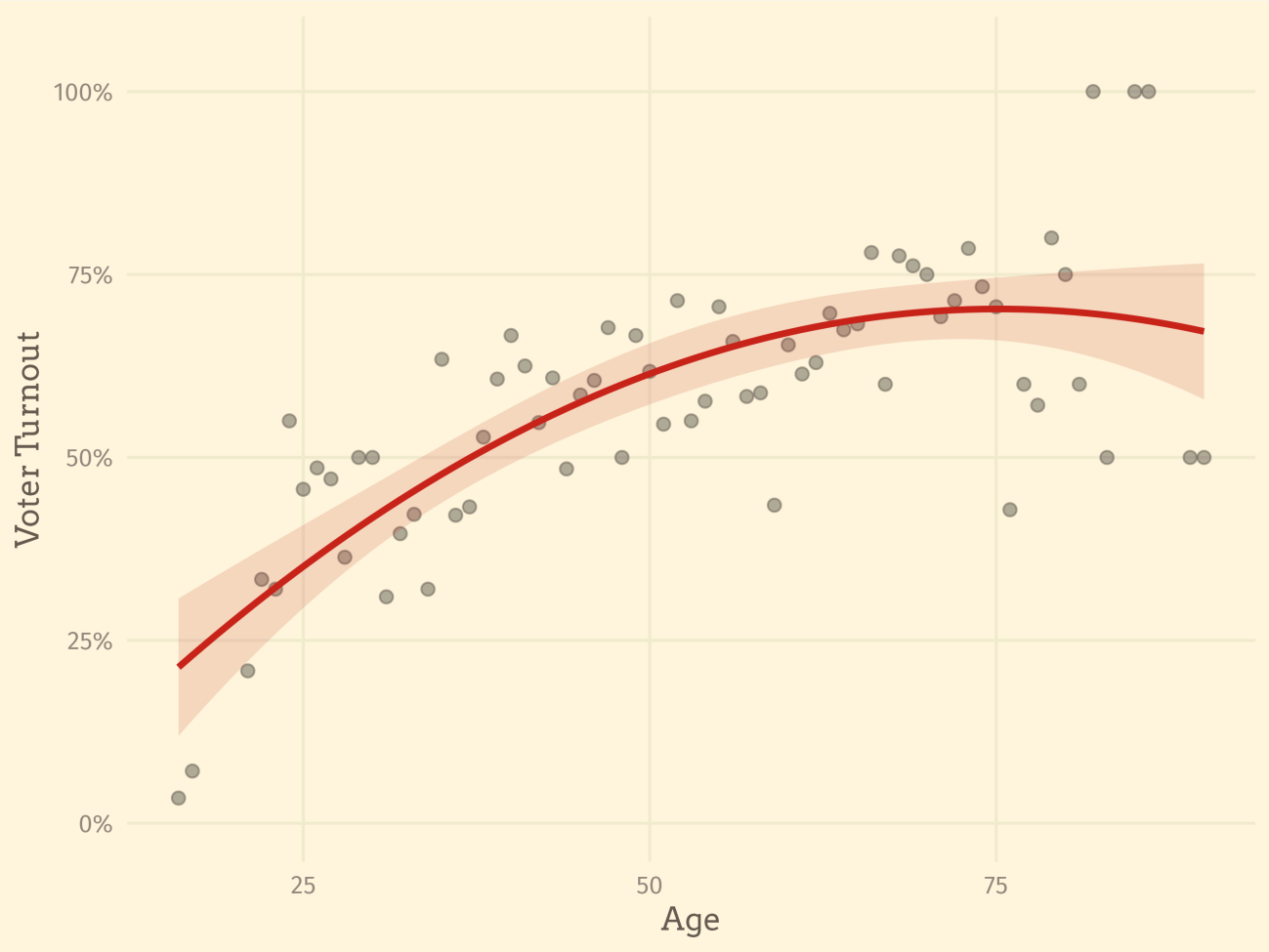
The most common way is by using **simple polynomials** (parabolas).

E.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ...$

Some Simple Polynomials



Simple Polynomials in Action



Questions?

InteRmezzo!

Advantages and Disadvantages of Simple Polynomials

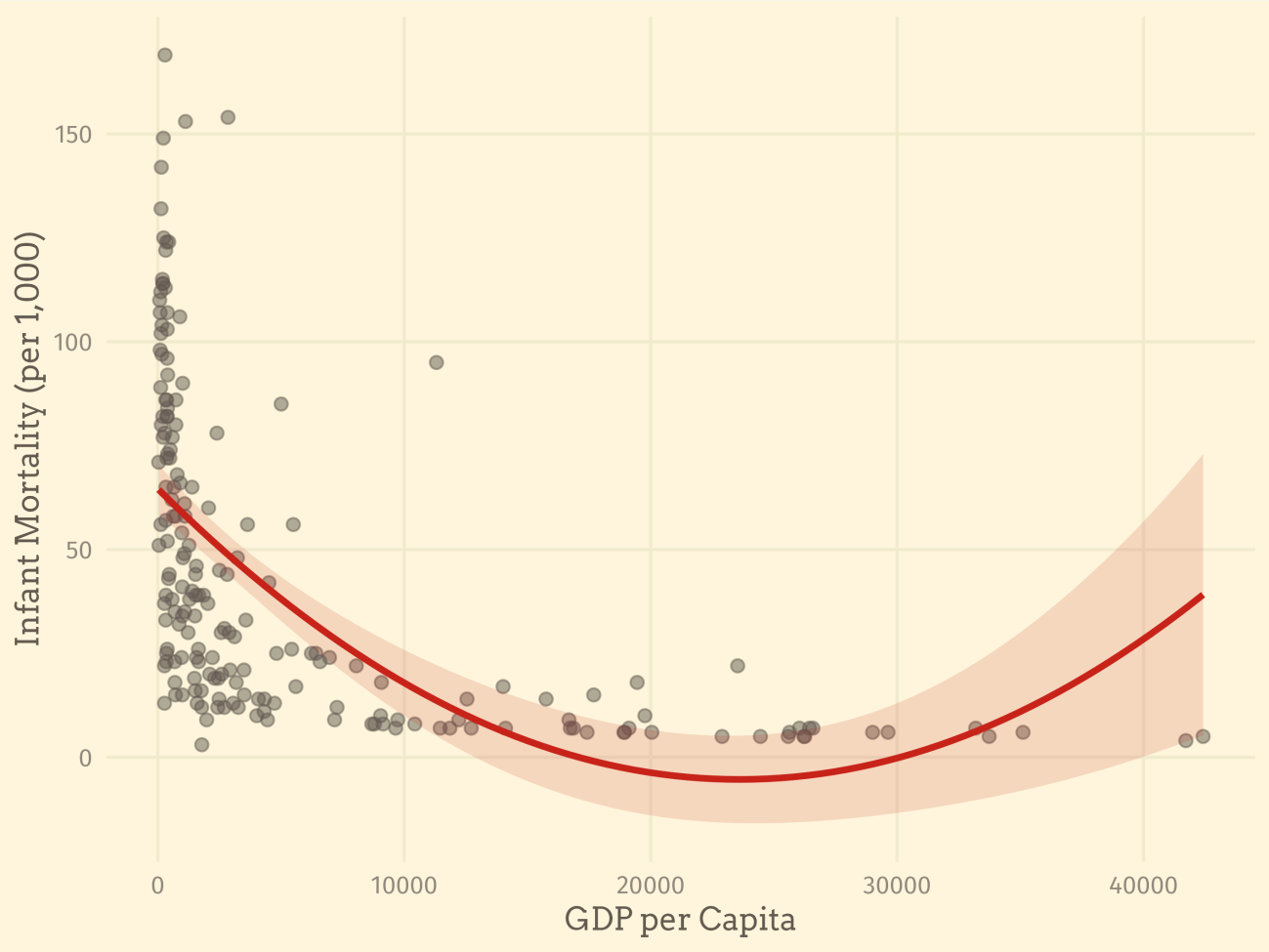
Advantages

- Fits continuous curve, usually better fit than categorisation.

Disadvantages

- Can only fit simple types of nonlinear relationship (parabolas).
- Polynomials tend to get *way* too „curvy“ at the edges.
- Coefficients not interpretable.

Simple Polynomials Fail For Nonsimple Relationships



When Covid Ended in 2020

Polynomials behave well in the middle of the data, but produce **implausible extrapolations** near the boundaries — they tend to curl sharply upward or downward beyond the observed range.

This instability at the edges is one of the key weaknesses of polynomial regression.

Questions?

Linear Splines

Splines

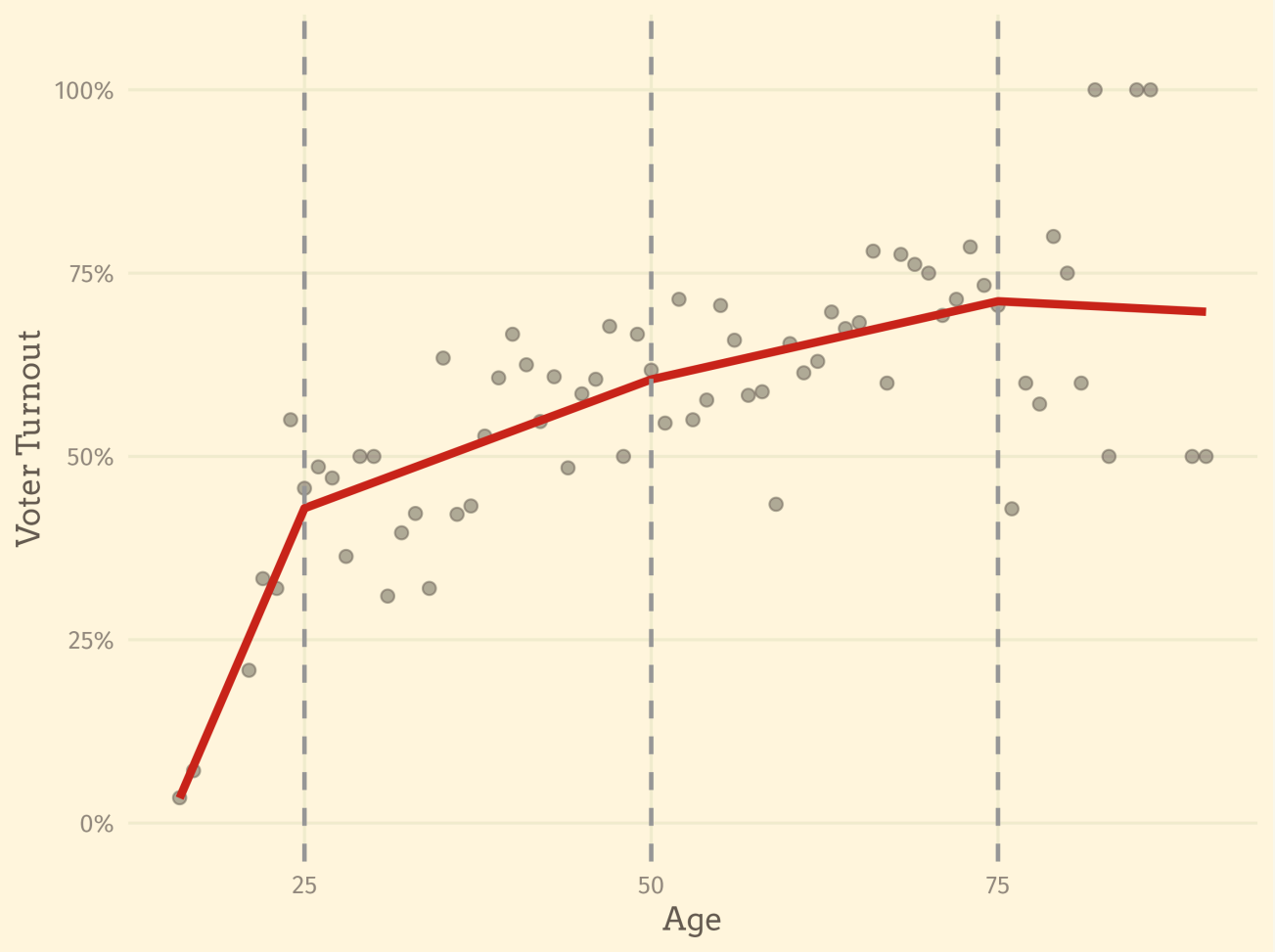
Why try to fit the data using a single line, when we can **split it into multiple smaller ones**?

- Predictor is cut into smaller **segments**. The cutpoints are called **knots**.
- We fit a regression line/curve **inside each segment**.

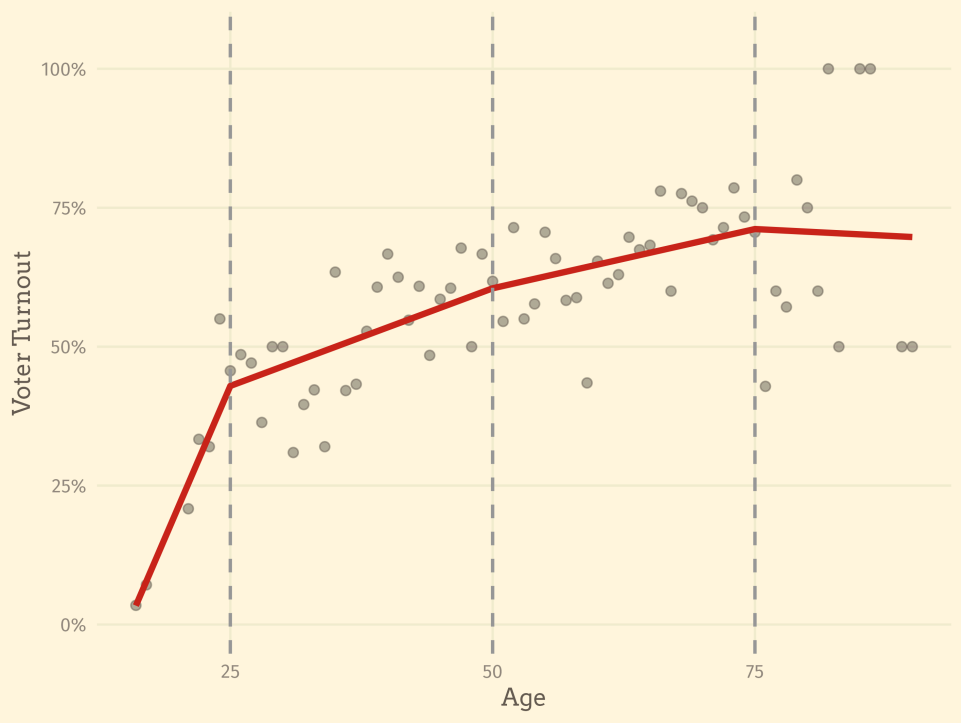
Linear Splines

Split data into smaller segments.

Fit a **straight line** through them.



Linear Splines Interpretation



Parameter	Coefficient
Intercept	-0.66888
Age < 25	0.04393
25-50	0.00703
50-75	0.00427
Age > 75	-0.00096

Linear Splines Advantages and Disadvantages

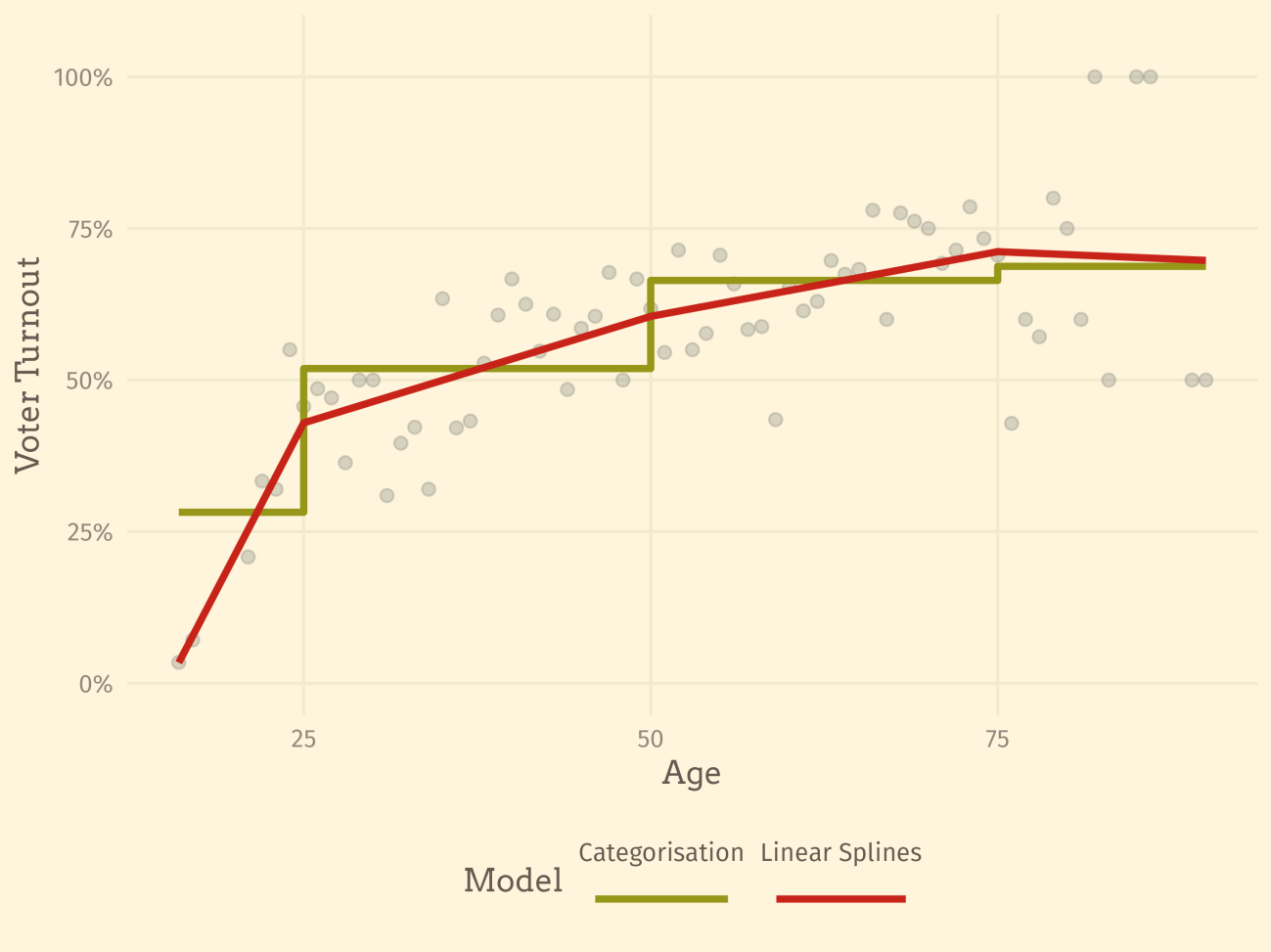
Advantages

- Can fit fairly complex nonlinear relationships.
- Easily interpretable regression coefficients.

Disadvantages

- Assumes sudden changes in relationship at cutpoints.
- Cutpoints are arbitrary.

Linear Splines Are Categorisation's Bigger Brothers



Questions?

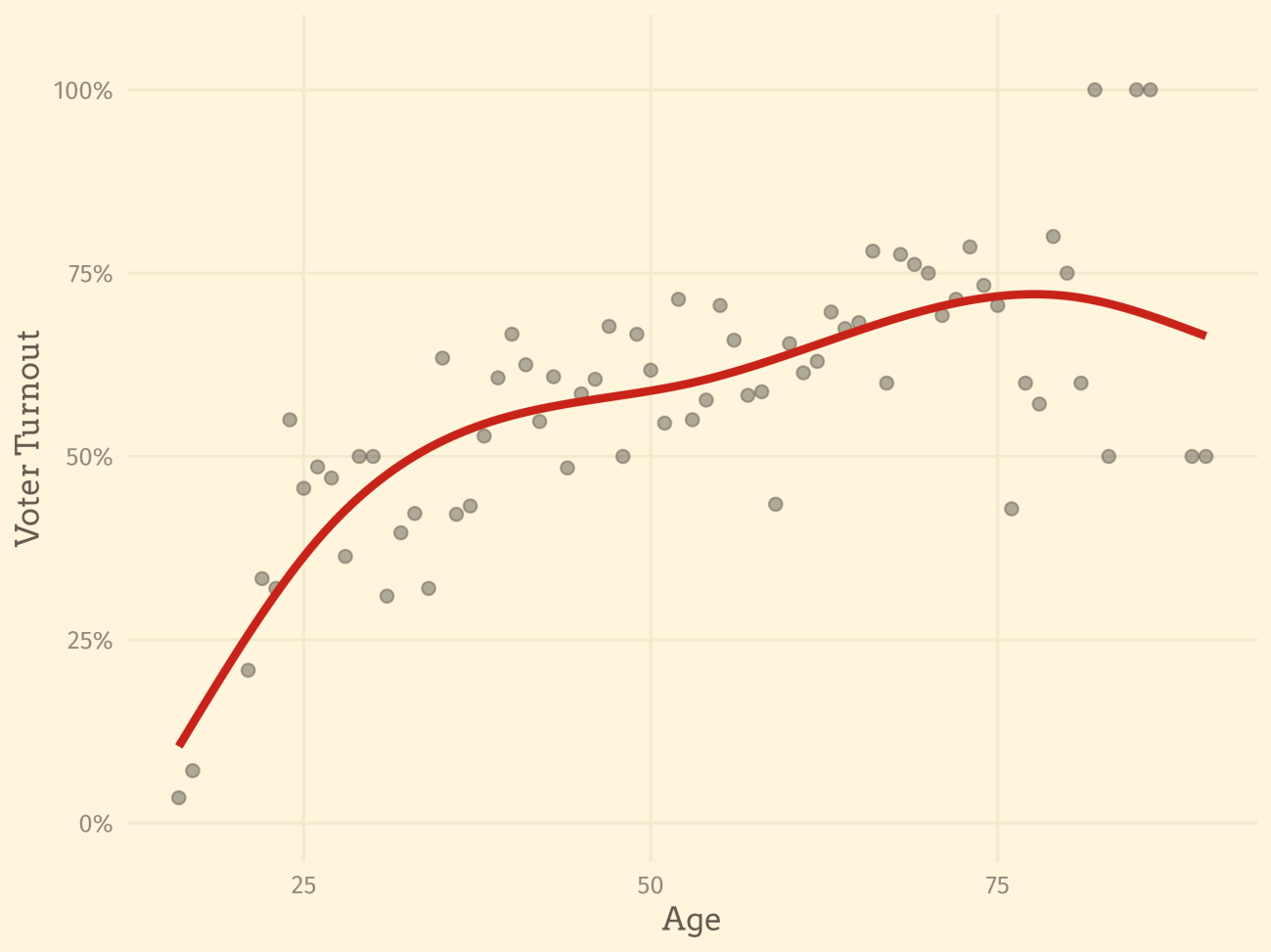
Natural Splines

Natural Splines

Also known as **restricted cubic splines**.

- We fit a **curve** inside the inner segments (using cubic polynomials) and a **straight line** inside the outer segments.
- This gives the **smoothness** of polynomials without the instability, while also being more flexible.

Natural Splines



Natural Splines Advantages and Disadvantages

Advantages

- Can fit complex nonlinear relationships.
- More stable than simple polynomials.

Disadvantages

- Regression coefficients can't be easily interpreted.

Choosing Cutpoint/Knot Location

Linear Splines are very sensitive to knot placement (similar to categorisation) — knots should be chosen based on **theory**.

Natural Splines are fairly robust, as long as the knots are spread equidistantly. The **number of knots** is more important. The optimal number can be chosen based on fit indices, e.g. adjusted R^2.

Common Natural Splines Knot Locations

Knots	Quantiles						
3		0.1	0.5	0.9			
4		0.05	0.35	0.65	0.95		
5		0.05	0.275	0.5	0.725	0.95	
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3342	0.5	0.6583	0.8167	0.975

Harrell, F. (2001). *Regression Modeling Strategies*. Springer-Verlag. https://doi.org/10.1007/978-1-4757--3462-1

Questions?

InteRmezzo!

Which Approach to Use?

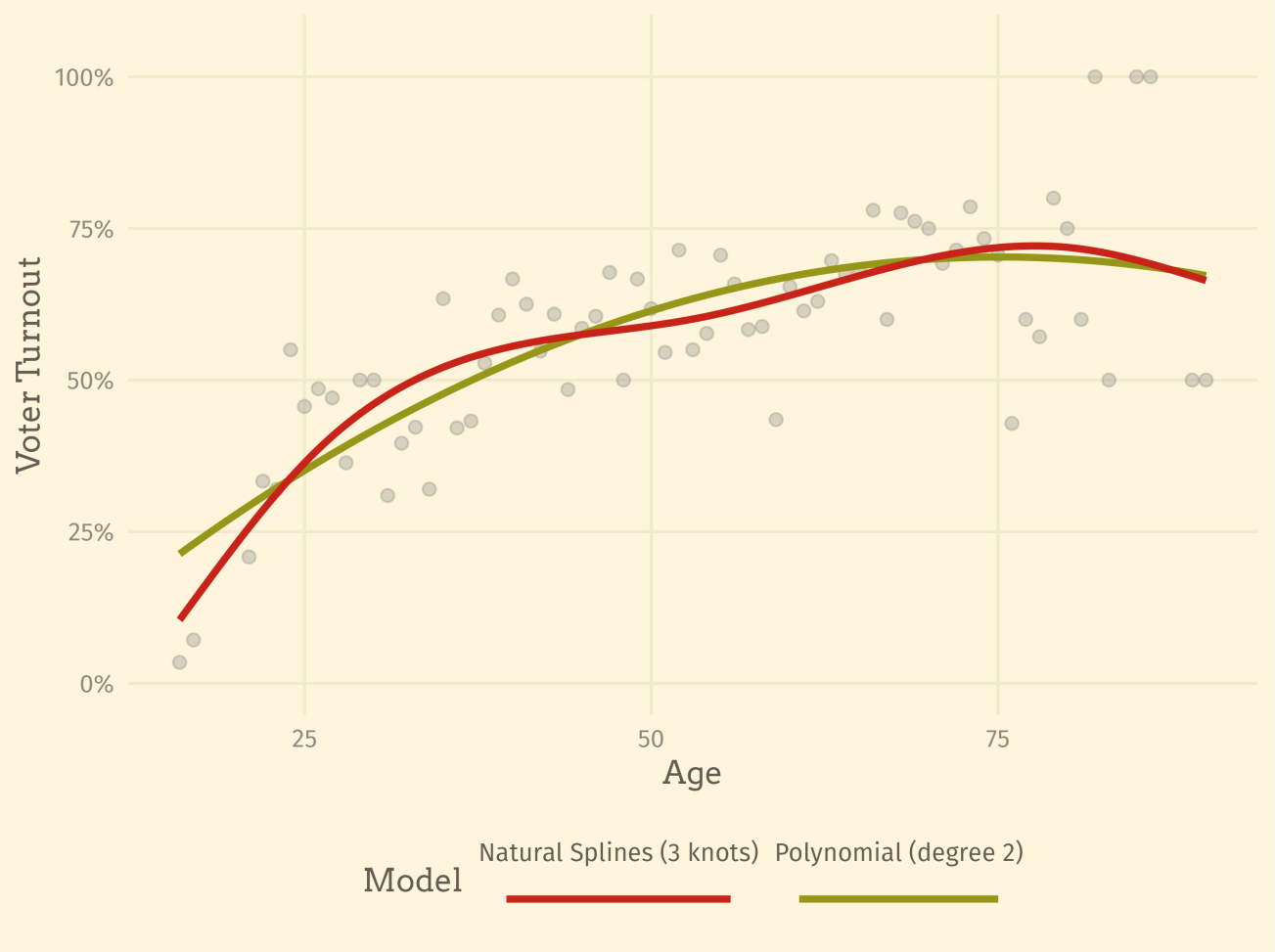
Best Approach Depends on Context

Approach	When to use
Categorisation	When it's important for nontechnical audience to understand the results.
Linear Splines	When presenting for technical audience, but you still want interpretable coefficients.
Natural Splines	When you want a good fitting model and are comfortable with marginal effects.
Simple polynomials	???

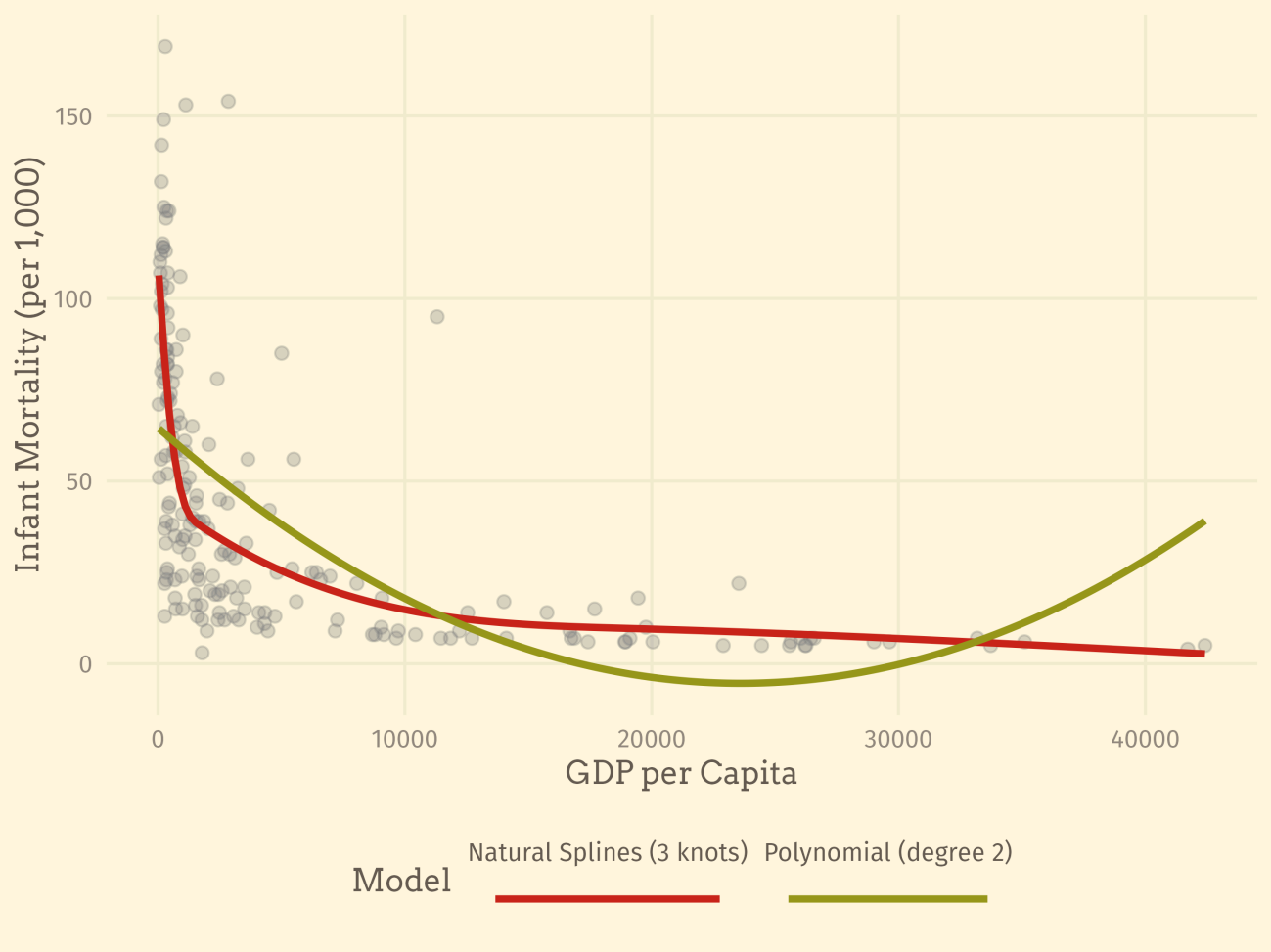
Is there still a place for simple polynomials?

Natural Splines Will Always Be At Least as Good...

Both models have the **same number of coefficients**.



… But Often Better



Best Approach Depends on Context

Approach	When to use
Categorisation	When it's important for nontechnical audience to understand the results.
Linear Splines	When presenting for technical audience, but you still want interpretable coefficients.
Natural Splines	When you want a good fitting model and are comfortable with marginal effects.
Simple polynomials	When you don't want to explain what splines are.

Questions?