Art by Amy Wolfe

# Modelling of complex, non-linear relationships in time series data while accounting for delayed effects 1

Robbie M Parks, PhD

22nd July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

# Outline from previous lecture

- Finding associations from data
- Model likelihood structures
  - Normal
  - Poisson
  - Bernoulli
  - Binomial
- Running models
- Evaluating model fit

# Outline

- Non-linear exposure-response curves
- Linear regression as an assumption
- Polynomials
- Splines
- Piecewise linear splines
- Natural splines
- Penalized splines
- Which to use?

# Linear regression

- Are any of the assumptions of linear regression relevant for this class?
- What does linearity even mean?
  - Constant $\beta$ across $X$.
- I.e., it doesn't matter where on the X distribution we are, for one unit increase in $X$ the $Y$ changes by $\beta$ units.

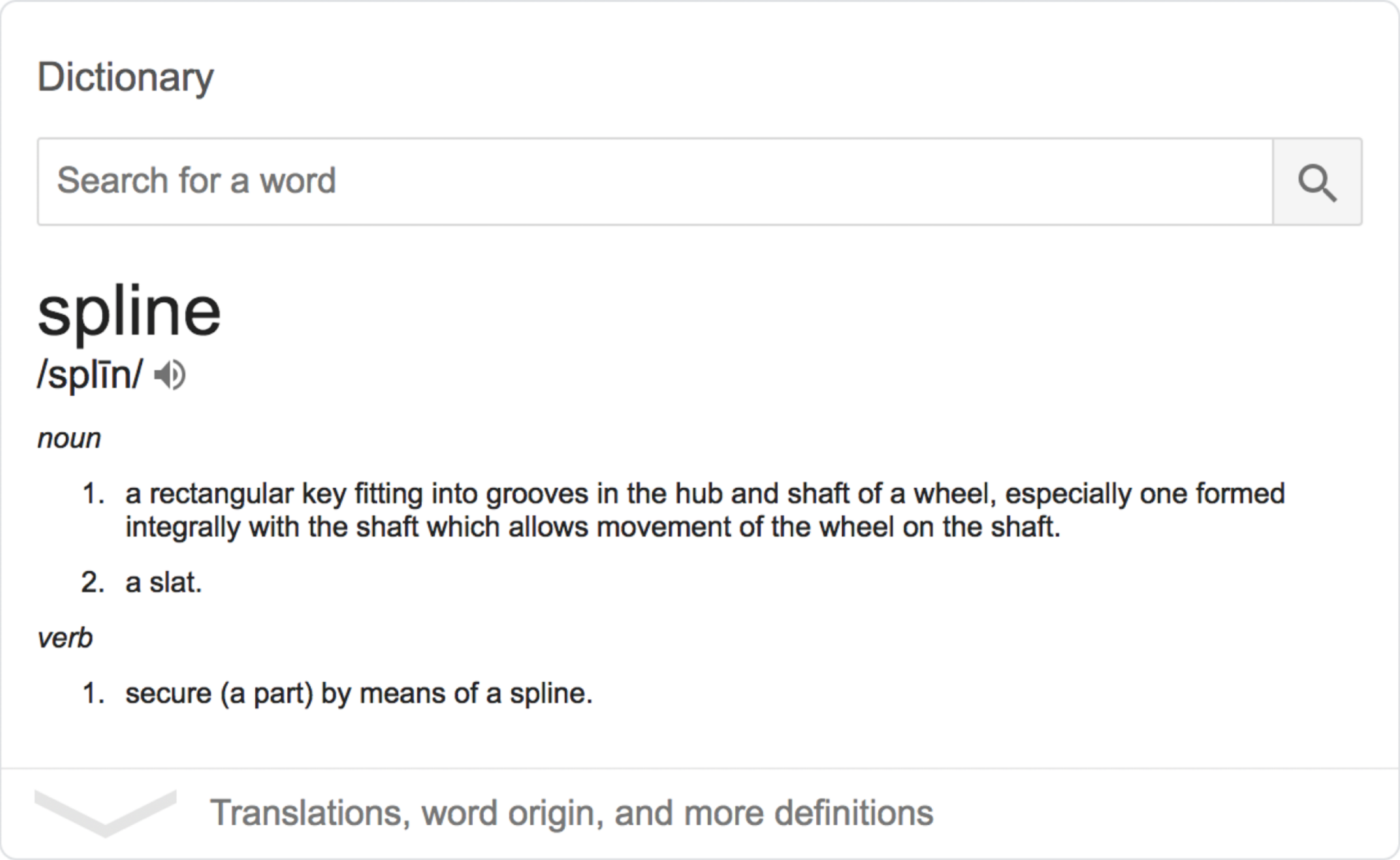
# What if not linear regression?

- Do you know of any ways to deal with this?
- Categorize my exposure $X$ into quantiles (e.g. quartiles) and use those as indicators in the model
- Is this categorization of $X$ a good idea?
  - ➢+ It is quick and easy to do
  - ➢+ Easily interpretable results
  - ➢- Assume step function exposure-response
  - ➢- Information loss (a lot...)
  - ➢- Residual confounding
- I am not a big fan, but people use it...

# Polynomials

- Another way is to add polynomial terms in the linear model
- E.g. quadratic: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$
- Why does this work?
- Are polynomials a good idea?
    - ➢ + Allows for non-linear relationships
    - ➢ + Uses information from full X distribution
    - ➢ - Strong assumption about the shape of the association
- Not used so much to model $X$
    - – There are more flexible ways
- People use it for non-linear confounders
- There are more flexible ways

# Splines



## spline

/splīn/

*noun*

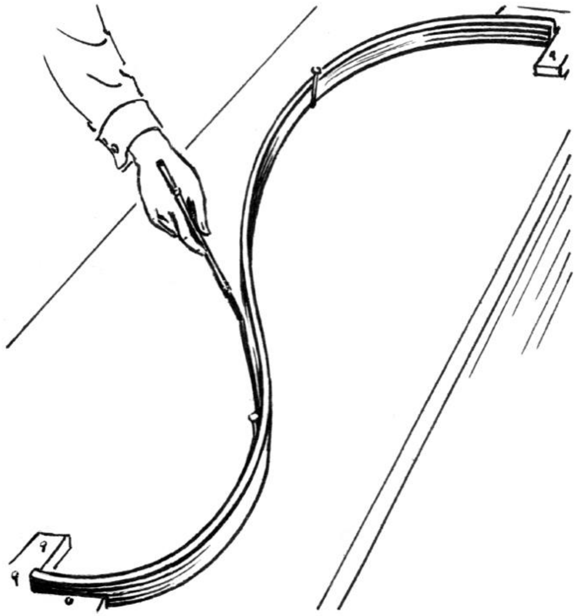
1. a rectangular key fitting into grooves in the hub and shaft of a wheel, especially one formed integrally with the shaft which allows movement of the wheel on the shaft.
2. a slat.

*verb*

1. secure (a part) by means of a spline.

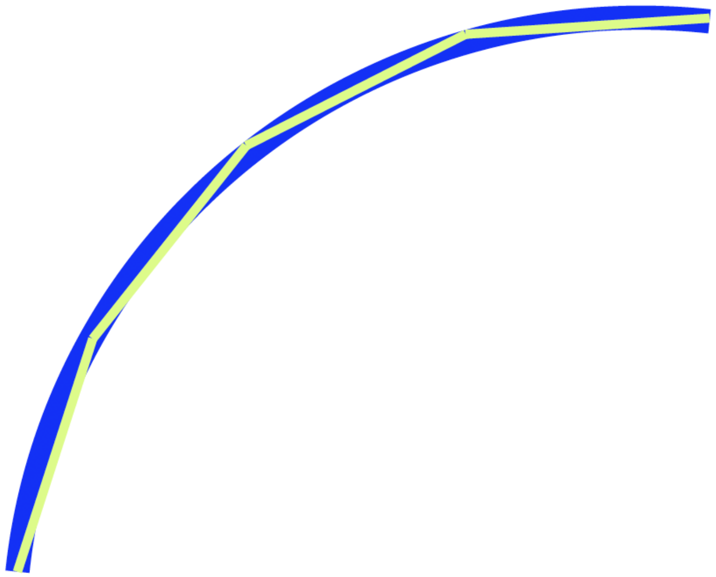
# Splines

- A long strip fixed in position at a number of points that relaxes to form and hold a smooth curve passing through those points for the purpose of transferring that curve to another material.



# Splines

- Divide range of $X$ into pieces.
- Fit a linear term in each piece.
  - How is this different from breaking into quantiles and using those indicators?
- Slopes within category.



# Piecewise linear splines

Example:

Is the $\text{PM}_{2.5}$–BMI relationship the same below and above the $\text{PM}_{2.5}$ NAAQS (12 $\mu\text{g/m}^3$)?

1. Define a new variable: $\text{PM}_{\text{high}} = (\text{PM}_{2.5}\text{-}12) \times (\text{PM}_{2.5} > 12)$
   - Knot at 12 $\mu\text{g/m}^3$
   - If $\text{PM}_{2.5} \leq 12$: $\text{PM}_{\text{high}} = 0$
   - If $\text{PM}_{2.5} > 12$: $\text{PM}_{\text{high}} = \text{PM}_{2.5} - 12$
2. $BMI = \beta_0 + \beta_1 PM_{2.5} + \beta_2 PM_{\text{high}} + \varepsilon$
   - What does $\beta_2$ tell us?

# Piecewise splines (more generally)

- Do piecewise splines need to be linear?
- Can they be more
- Yes! - polynomials between knots
- E.g. quadratic, cubic etc…
- Cubic often used because of nice mathematical properties
  - (derivative is continuous)

# Natural splines

- Piecewise splines
- Super flexible and useful
- Cubic polynomial between knots
- Linear at the ends (before first and after last knot)
  - Often at ends we don't have a lot of information
- User defines degrees of freedom (df) or knots (k)
- Knots usually at quantiles, but can also be at user-defined values of $X$

# Penalized splines

- Natural splines are great!
- Can we be more flexible?
- Can we allow our data to tell us if the relationship between $X$ and $\mathrm{Y}$ is linear?
- -> Penalized splines
  - Very flexible semi-parametric tool

# Penalized splines

- How does it work?
  - Throw many knots (default in R: k = 10)
  - Start linear
  - At first knot: do my data tell me to continue as I was going or do I need to change direction?
  - Continue like this for all knots
  - If change in direction improves fit: cubic terms between knots

# Penalized splines

- Why penalized? What is the penalty ($\lambda$)?
- Controls the level of “wiggliness"
- $\lambda$ = 0: no penalty (basically back to natural spline)
  - If k = 10: super wiggly curve! { not really interpretable
- Low penalty: still a lot of wiggliness
- As $\lambda$ increases, we get a smoother curve
- $\lambda$ : linear (complete absence of wiggliness)
- Now my *estimated* degrees of freedom are a function of the
- number of knots and this penalty term
- No longer necessarily an integer

# Penalized splines

- How do we get the penalty?
- Can be user-defined
  - But then what's the point?
- I can ask my model to estimate the best  for my data
  - Generalized Cross-validation Criterion (GCV)
  - Akaike's Information Criterion ($AIC = 2k - 2\ln(\hat{L})$). BIC/DIC/WAIC for Bayesian (day 3)
- If curve too wiggly, I can tweak the estimated penalty to
- smooth my curve
- If the best fit is linear, the model will estimate 1 df!

# Which method to use in practice?

- Different people have different preferences
- Here is what I usually do:
- Start simple, start linear - get a “feel" of my model
- Fit a penalized spline to see if there are any deviations from linearity
- If edf = 1 great! that's where I stop

# Which method to use in practice?

- If I detect non-linearity, check if my penalized spline makes biological sense
- If it does, that's what I'll use
- Sometimes though might be too wiggly (too data dependent)
  - then go to natural splines
- Tweaking the $\lambda$ feels a bit like cherry-picking
- Try natural splines with df $\in$ [2; 5] and pick the best fitting
- one (AIC & biological plausibility)

# Outline

- Non-linear exposure-response curves
- Linear regression as an assumption
- Polynomials
- Splines
- Piecewise linear splines
- Natural splines
- Penalized splines
- Which to use?

# Getting ready for the lab

- This lab will involve taking some models and concepts from the **Modelling of complex, non-linear relationships in time series data while accounting for delayed effects 1** lecture and introduce you to the way non-linear regression works:

# Application

- How can you imagine applying this learning to your data and your research questions?

# Questions

- Questions?

Art by Amy Wolfe

# Modelling of complex, non-linear relationships in time series data while accounting for delayed effects 1

Robbie M Parks, PhD

22nd July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH