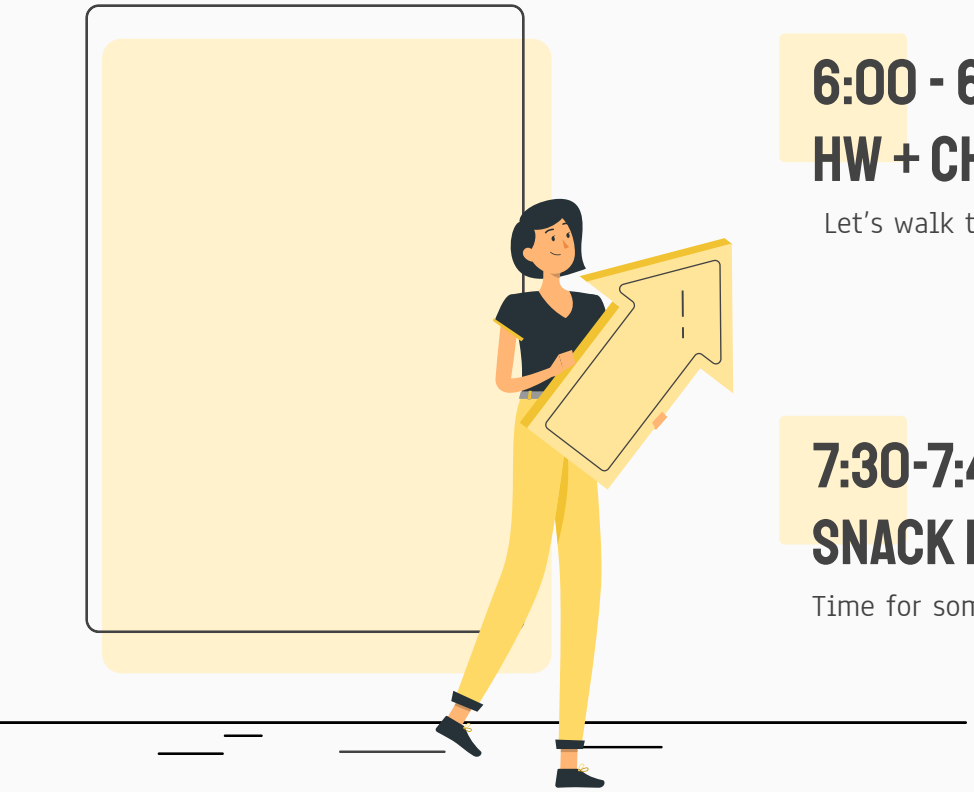
APPLIED DATA SCIENCE II

Week 6: Lions, Tigers, and Splines - oh my!

Kyle Scot Shank
WI-22



6:00 - 6:30
HW + CHECK-IN

Let's walk through it!

6:30-7:30
TOPICS + CODE!

Let's pump up our power with some **additive models and other kinds of neat stuff!** !

7:30-7:45
SNACK BREAK!

Time for some munchies

7:45 - 9:00
HANDS-ON CODE LAB

Work through stuff together

CHECK-IN

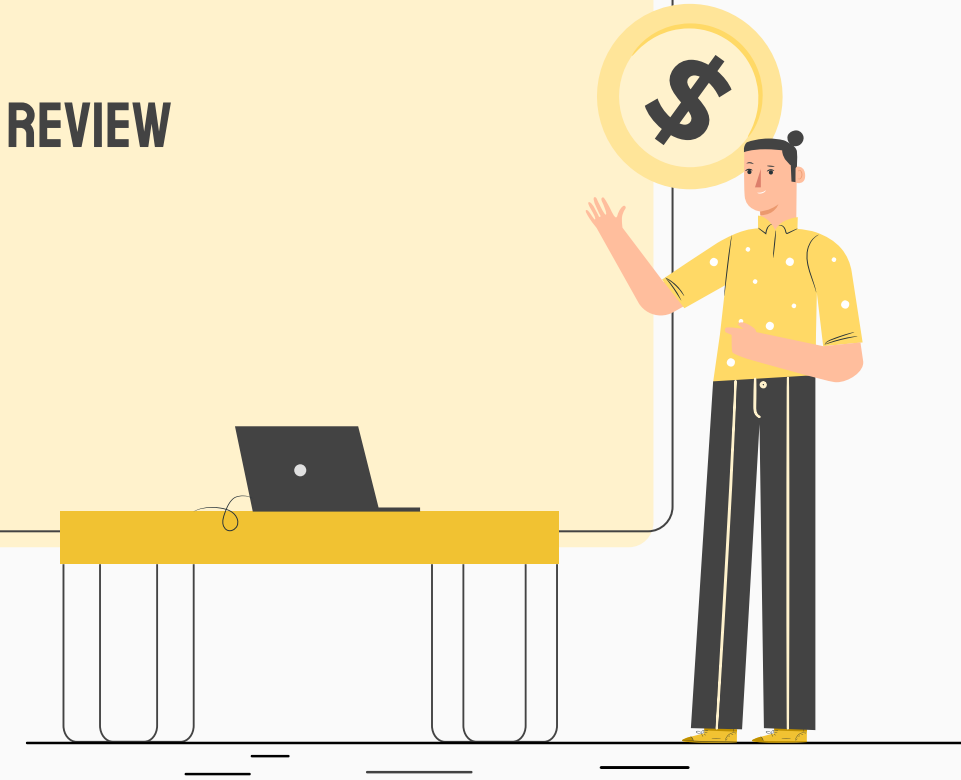
CHECKING IN!

1. *How are we feeling?*

2. *What do we want to do about class on Tuesday the 22nd?*

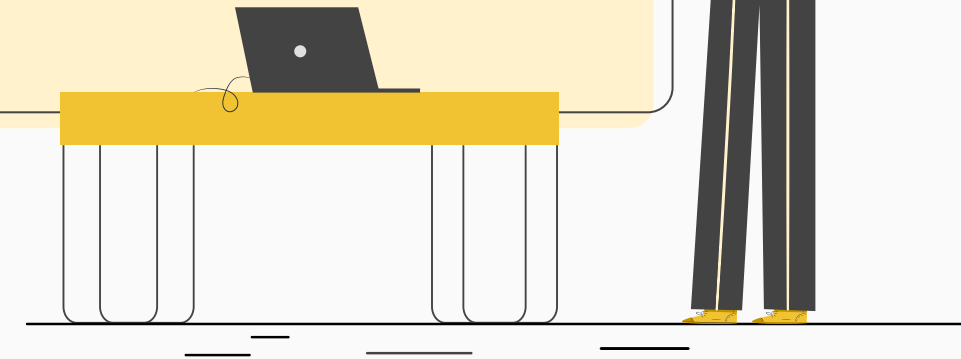
3. *Final Project Discussion*

HW REVIEW



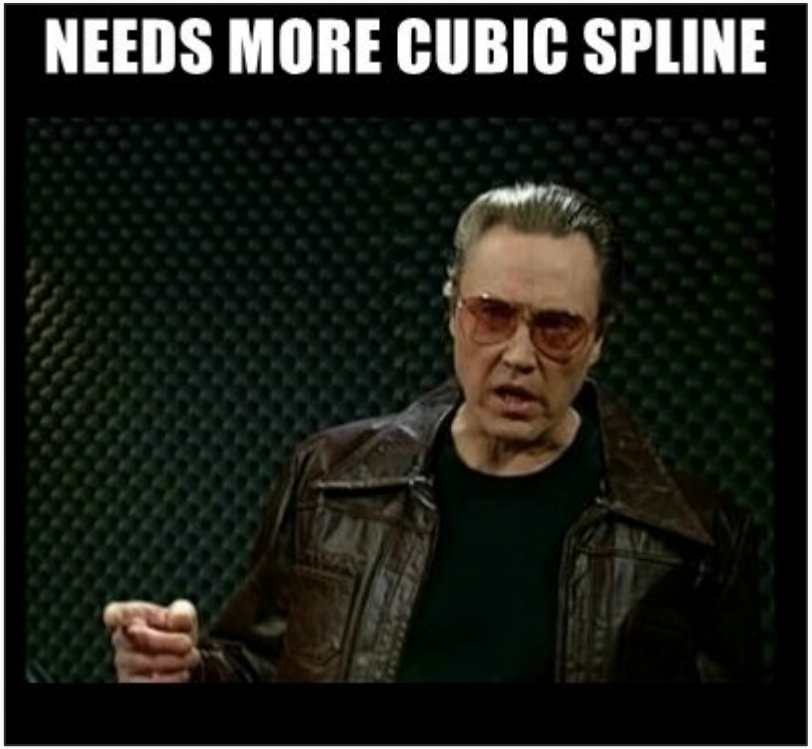
TOPIC OVERVIEW

MOVING BEYOND LINEAR MODELS!



REMEMBER THOSE NEAR LINEAR MODELS?

WHAT'S ONE OF THEIR DRAWBACKS?



MOVING BEYOND LINEAR MODELS

The truth is never linear!

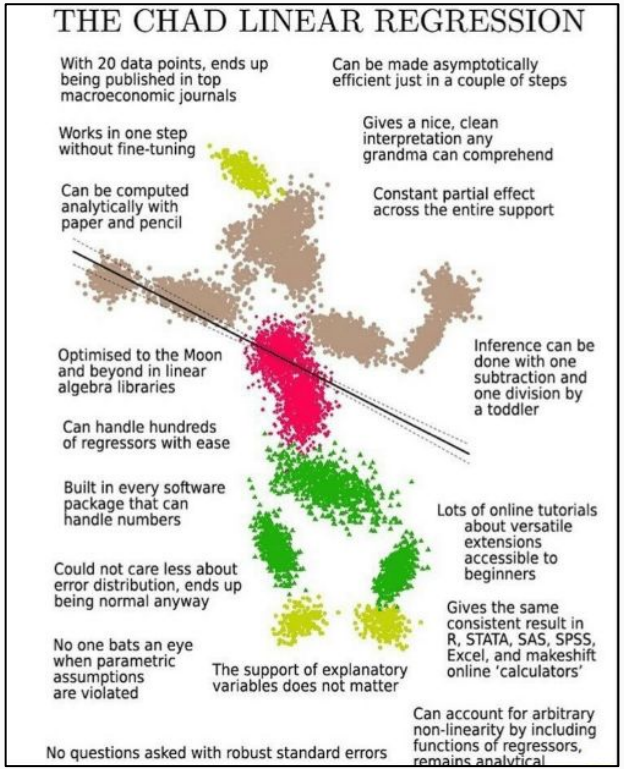
....Or almost never!

But often the linearity assumption is good enough.

When its not . . .

- *Polynomials*
- *step function*
- *Splines*
- *local regression*

And generalized additive models offer a lot of flexibility, without losing the ease and interpretability of linear models.



MOVING BEYOND LINEAR MODELS

We're going to focus on just two of these as being specifically helpful:

Spline functions + Generalized Additive Models (GAMS)

A generalised additive model is just a bunch of splines put together

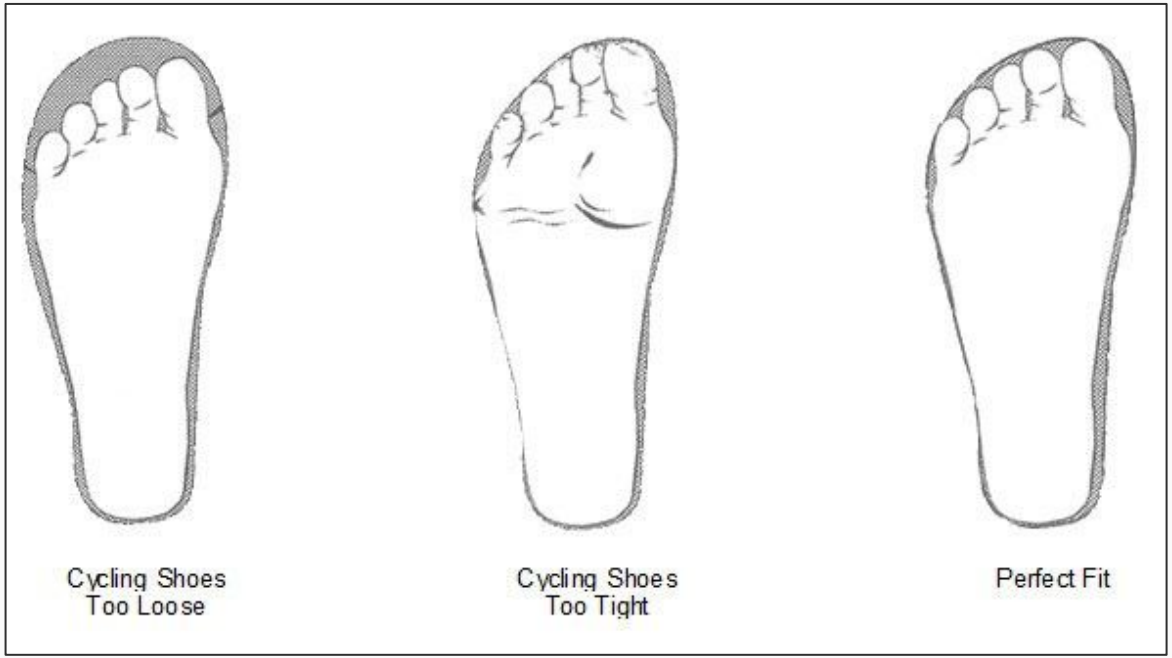
Geom smooth in ggplot fits a spline model to your data.

SPLINES!

*Let's start from the beginning: **what** in the world **is a spline?***

*Let's us an analogy of fitting a new shoe. What we've done up until this point has been fitting a **foot** to a specific **shoe** (linear models, lasso, etc.) - this let's us approximate stuff really well.*

*What splines do is fit the **shoe** to a specific **foot**.*

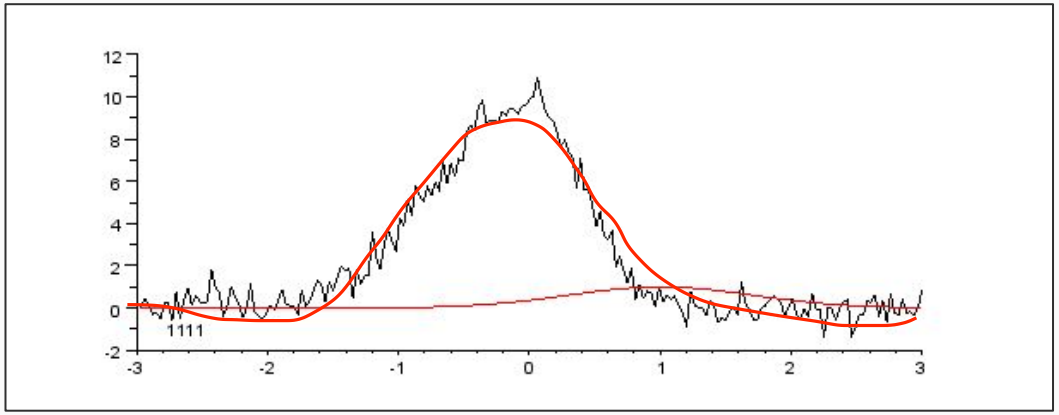


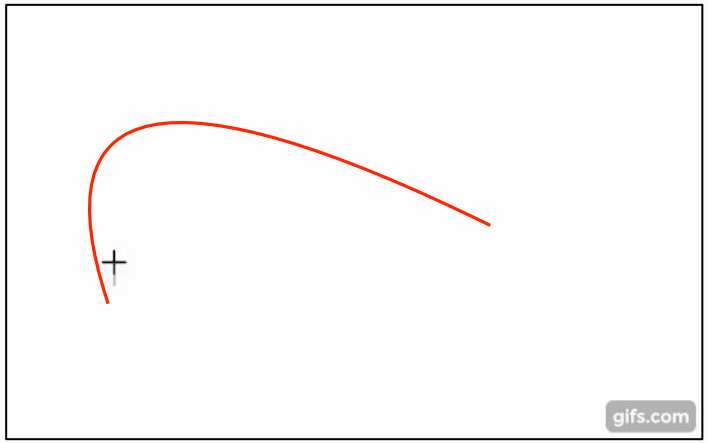
SPLINES!

This works by using things called piecewise polynomial functions.

Think of this as similar to what you do in Powerpoint when you try to use that nifty Line tool - each time you click and add a new segment, you introduce a new inflection point (or "knot") in the data.

This lets you build arbitrarily "curvy" lines to fit our data!





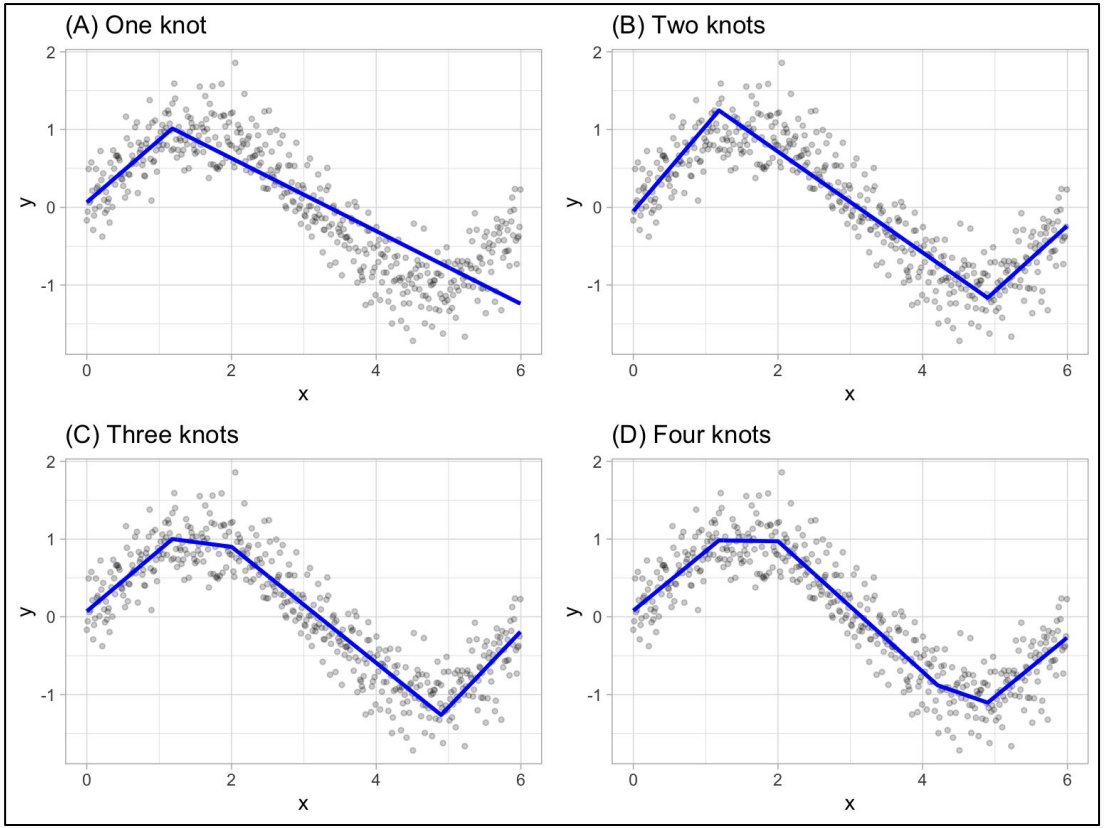
SPLINES!

How do they work?

*Spline Regression is a **non-parametric regression technique**. In this technique the dataset is divided into bins at intervals or points which we call **knots**.*

The data between *knots each get a separately fit piecewise polynomial, which are then combined. This is called interpolation (we can just call it "smoothing").*

In other words, splines are series of polynomial segments strung together, joining at knots.



SPLINES!

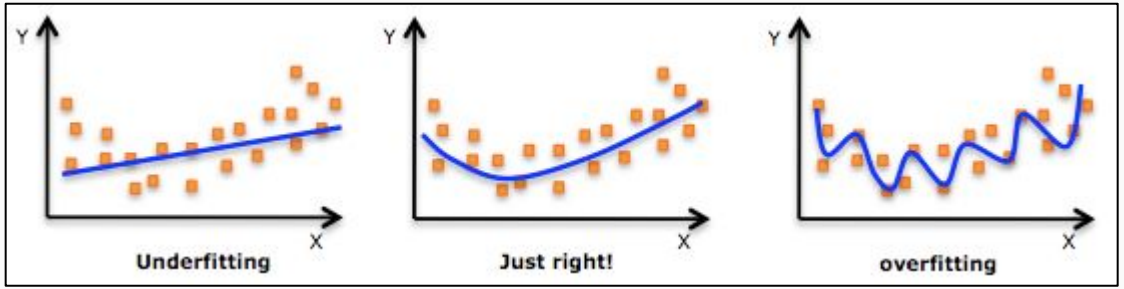
How do you, uh, pick the right number and location for the knots?

If you have the time + domain knowledge, you can do this manually!

But you shouldn't, because we're dumb and slow monkeys.

You can provide the desired degrees of freedom (which equals the number of knots) and the computer can partition the data equally.

OR - we can use cross-validation to understand the optimal value!



K-FOLD CROSS VALIDATION

- *Let's do this together!*

Open up R!

GENERALIZED ADDITIVE MODELS

Get a bunch of data scientists into a room and ask them if they've used a Random Forest model and you'll get a sea of smiles and maybe a bunch of nerdy arguments.

Ask them if they've used a GAM - and all you'll hear is crickets.

That is, until you - the incredibly savvy and VERY smart APPLIED data scientist, grab their laptop and show them what they've been missing out on.

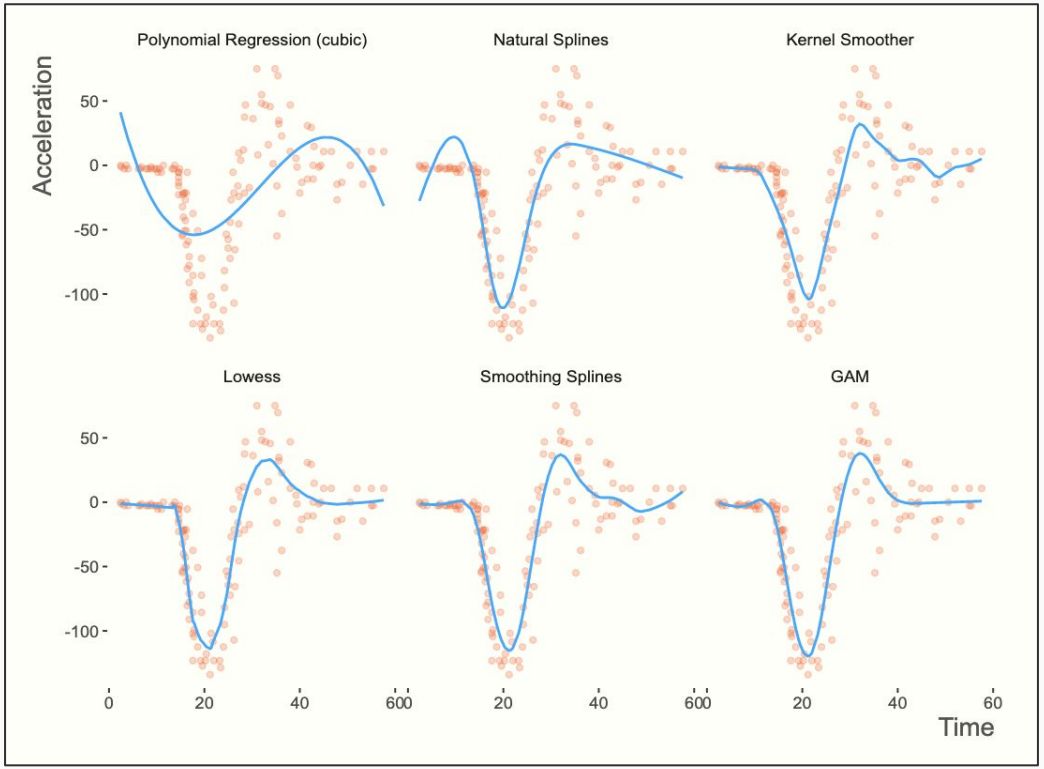
putting together lots of different types of regressions together for different ranges of data. it works by adding together splines.

GENERALIZED ADDITIVE MODELS

GAMs relax the restriction that the relationship must be a simple weighted sum, and instead assume that the outcome can be modelled by a sum of arbitrary functions of each feature.

To do this, we simply replace beta coefficients from a Linear Regression with a flexible function which allows nonlinear relationships (we'll look at the maths later).

This mysterious flexible function is called a (surprise!) spline.



GENERALIZED ADDITIVE MODELS

How do they work?

The math here can get pretty gnarly, pretty quickly, but this is the high-level answer:

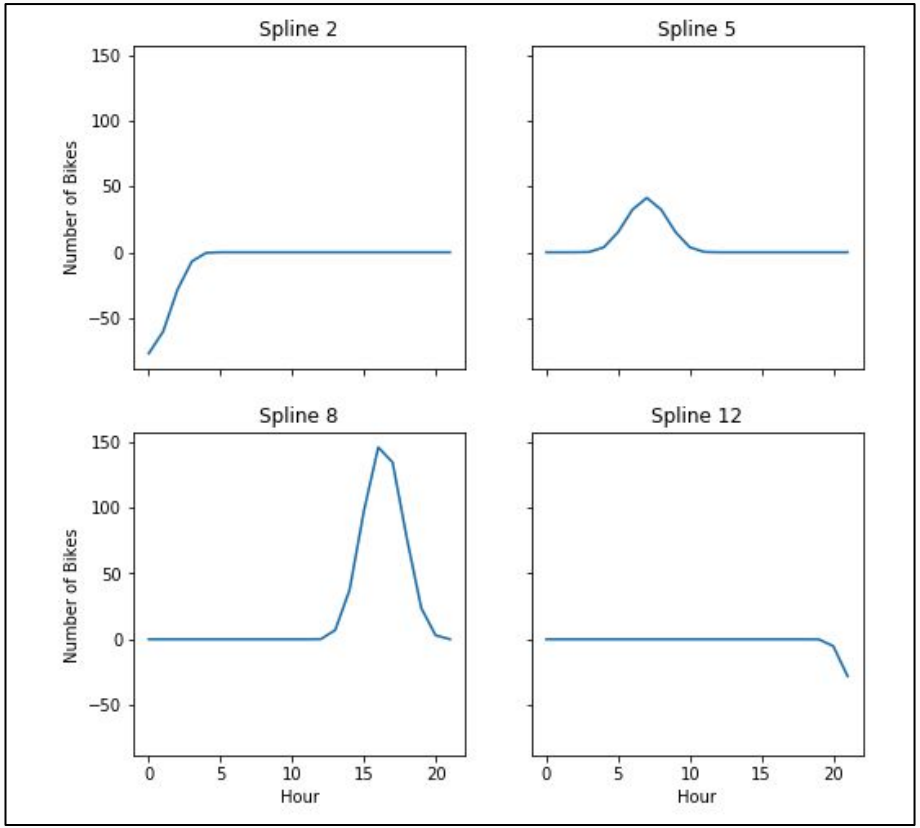
In a normal linear regression, we have something like

$$Z = \beta_0 x_0 + \beta_1 x_1 + ... + \beta_n x_n$$

In a GAM model we replace the beta coefficients with ***s****, which is a shorthand for an individual spline*

$$Z = s_0 x_0 + s_1 x_1 + ... + s_n x_n$$

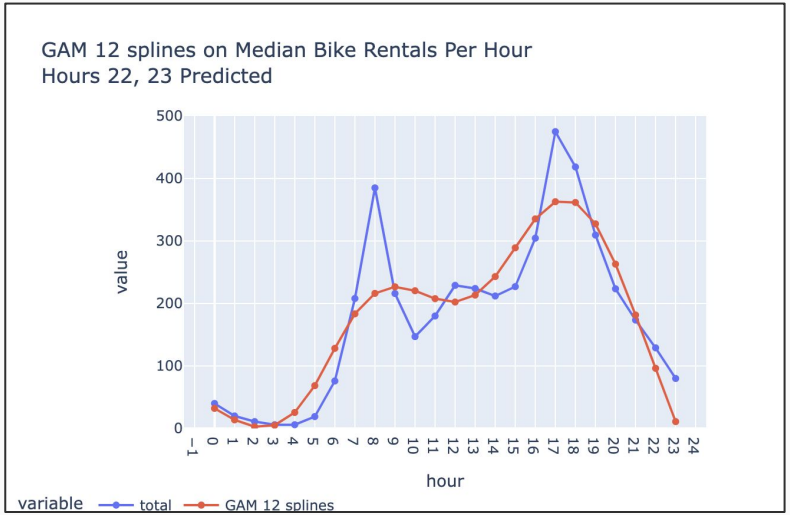
This lets us build ***really*** *complicated models that can best fit our data.*

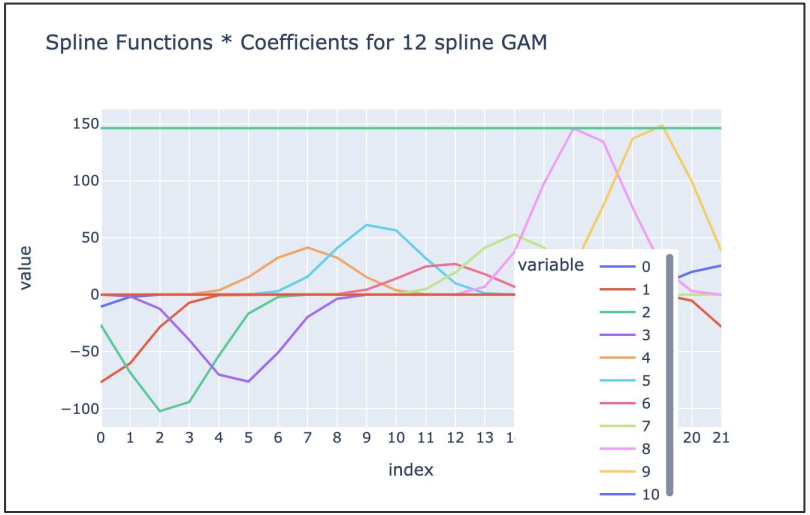


GENERALIZED ADDITIVE MODELS

How do they work?

When we've identified which Spline equations we're going to use for each variable, we can then simply add them up and this gives us the final estimate for our equation at each point in space.





GENERALIZED ADDITIVE MODELS

Why use these?

Interpretability

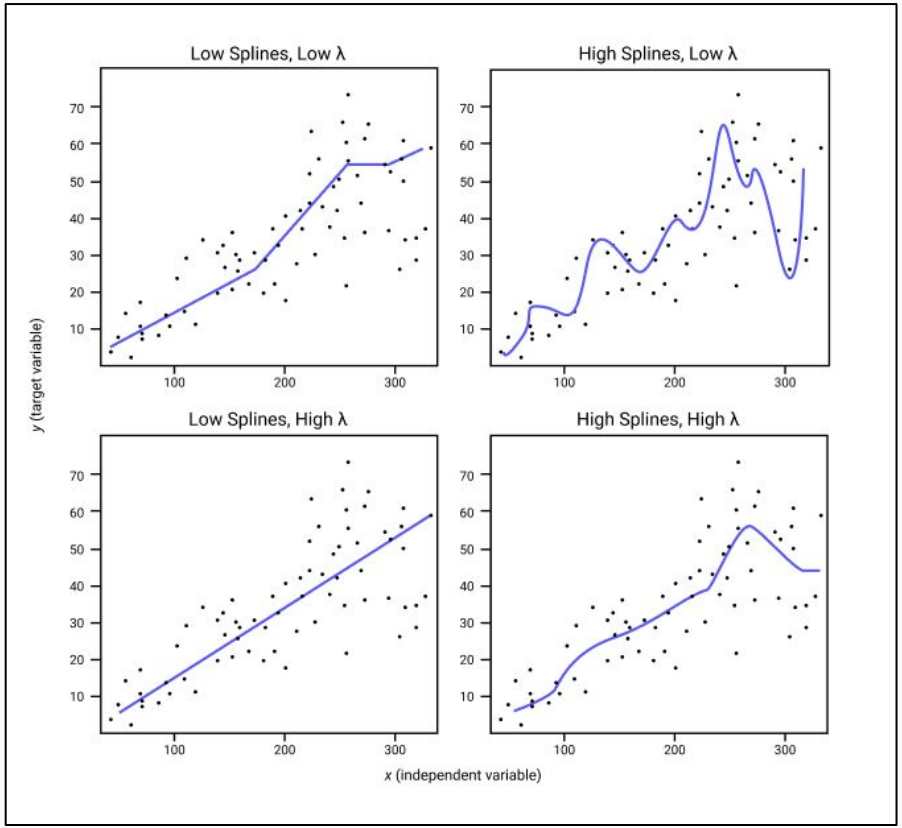
When a regression model is additive, the interpretation of the marginal impact of a single variable (the partial derivative) does not depend on the values of the other variables in the model.

Flexibility

GAMs tend to capture nonlinear effects that other models miss

Regularization

We control wiggliness (a technical term!) which is related to some very serious statistics and helps with the bias/variance trade-off.



K-NEAREST NEIGHBORS

- *Let's do this together!*

Open up R!

SNACK BREAK!

COME BACK IN 15!

CODE LAB!

OPEN UP RSTUDIO