Generalized Additive Models: Allowing for some wiggle room in your models

Sara Stoudt

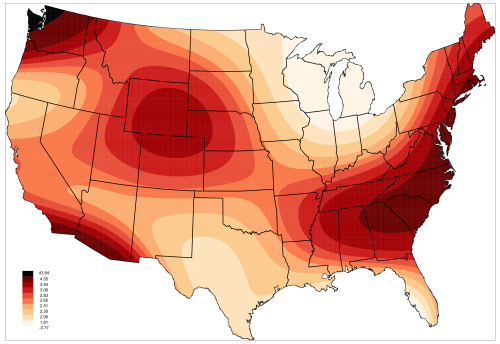
March 17, 2021

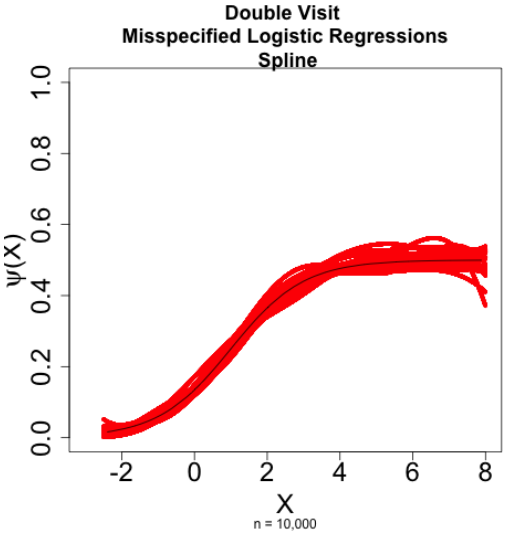
About Me

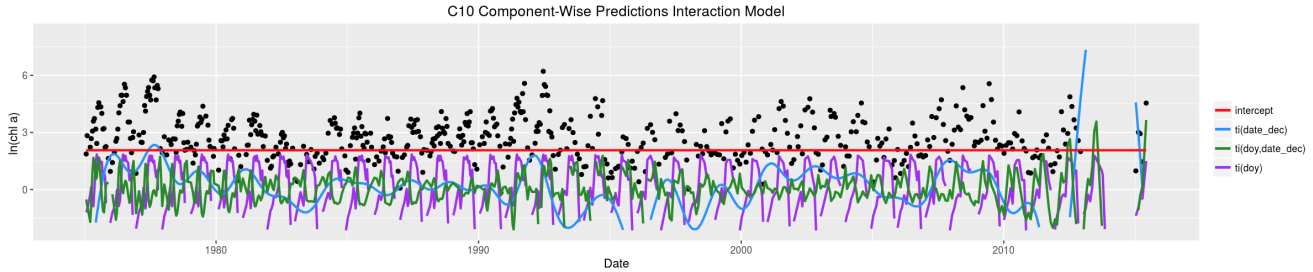
- currently teaching in the Statistical and Data Sciences Program at Smith College
- PhD in Statistics at Berkeley
  - ecology: evaluating fitness for purpose of a variety of data collection protocols for species distribution and abundance models
  - ecology: evaluating model fit in terms of community metrics for joint species distribution models
  - statistics communication: co-authored a book *Communicating with Data: The Art of Writing for Data Science* with Deborah Nolan

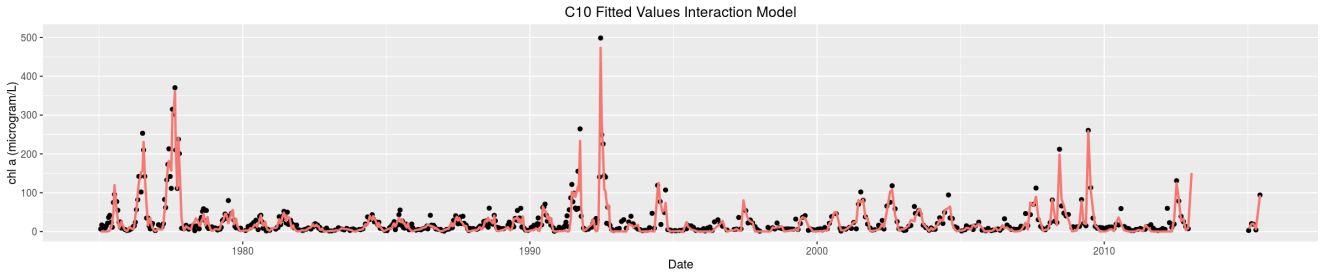
Materials here: `https://github.com/sastoudt/MZES_GAMs`

GAMs in my work

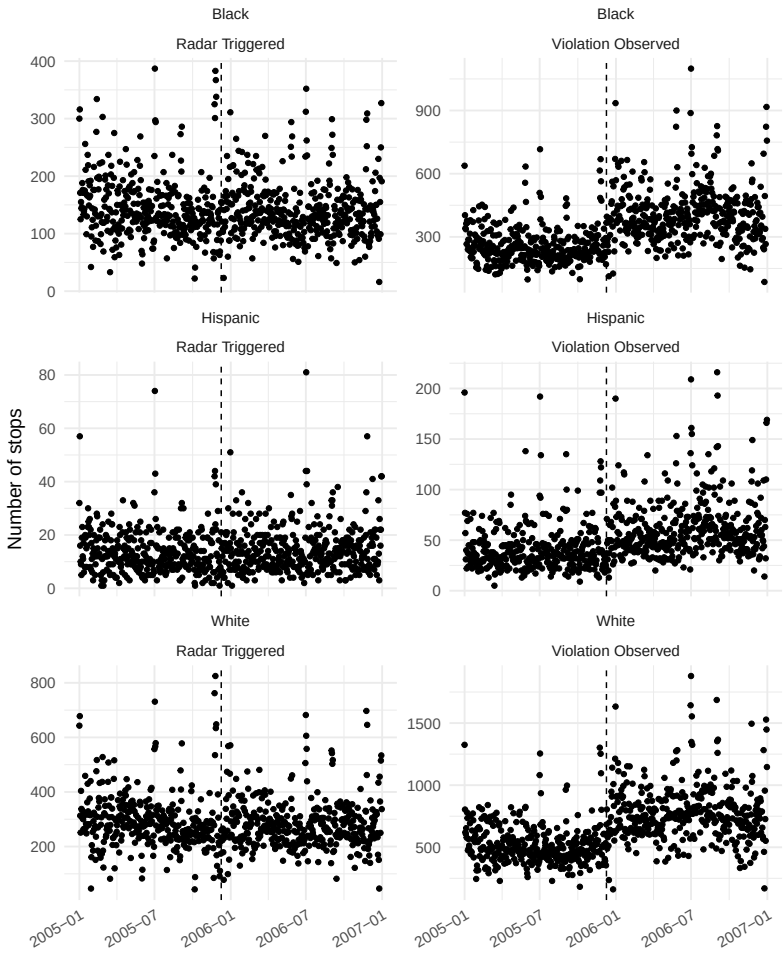








Setting the Scene



- "Using change in a seat belt law to study racially-biased policing in South Carolina" by Corinne A Riddell, Jay S Kaufman, Jacqueline M Torres, and Sam Harper
- `https://github.com/corinne-riddell/SCarolinaTrafficStops`

Linear Model - `lm`

$$Y = X\beta + \epsilon$$

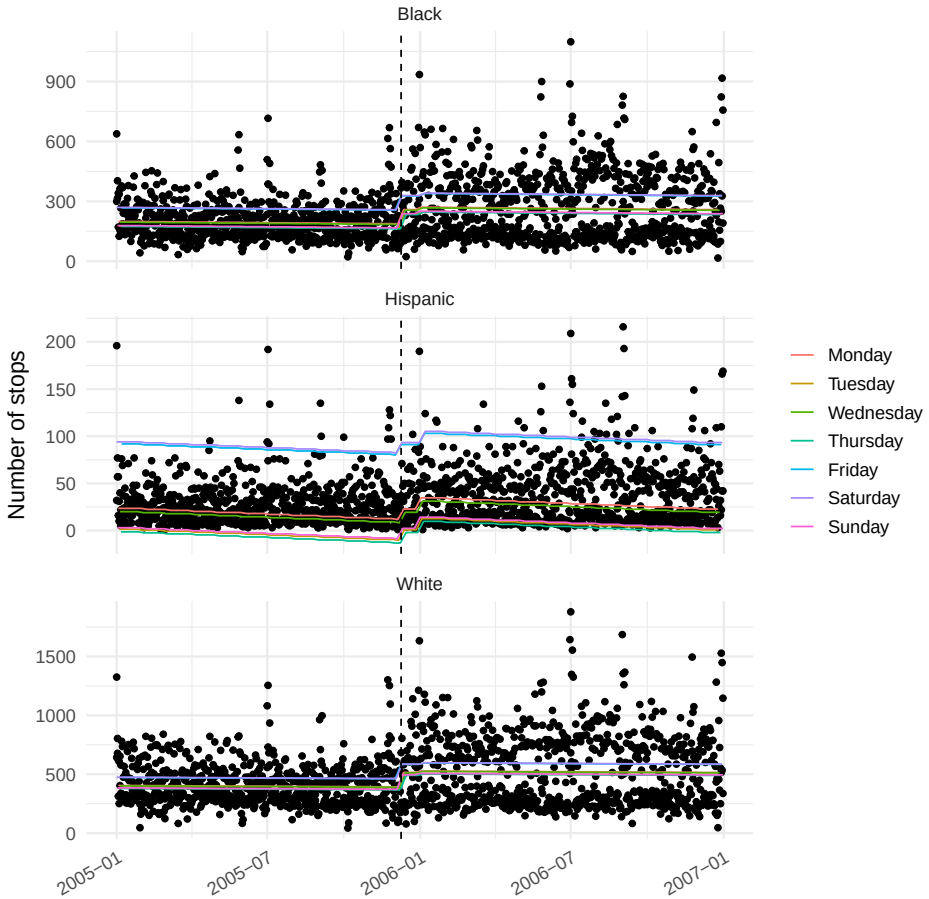
daily number of stops $\sim \beta_{driverRace} +$

$\beta_{isPostPolicy} + \beta_{driverRace,isPostPolicy} +$

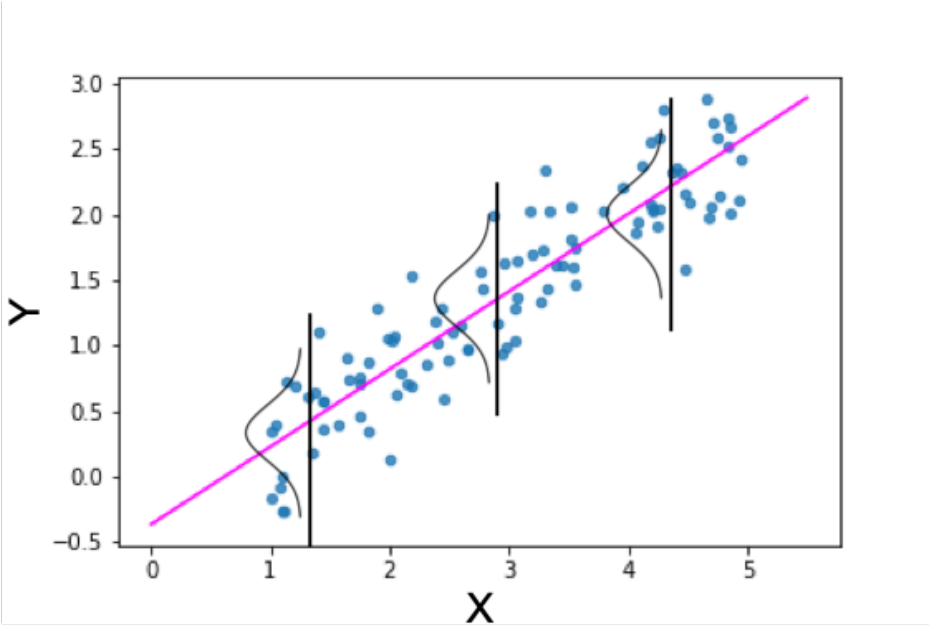
$\beta_{dayOfWeek} + \beta * \text{month}$

Choices:

- which covariates X to use



Linear Model $\rightarrow$ Generalized Linear Model



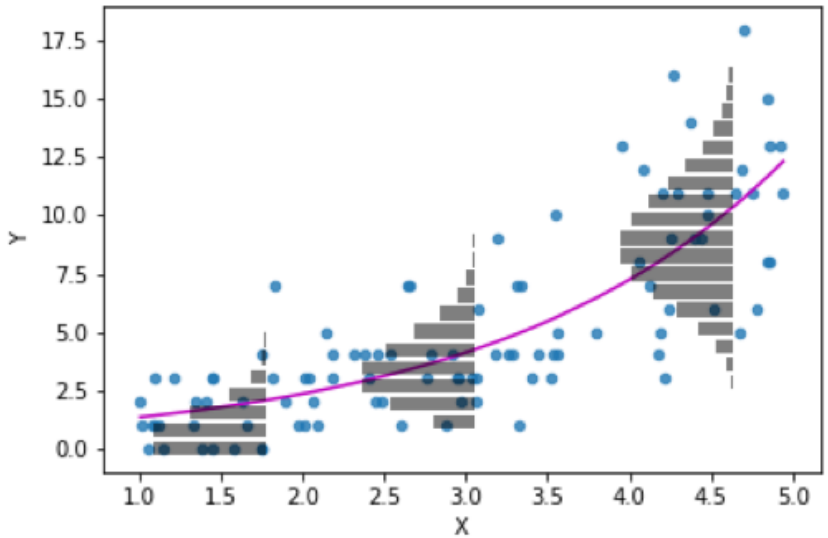


Figure Credit: `https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab`

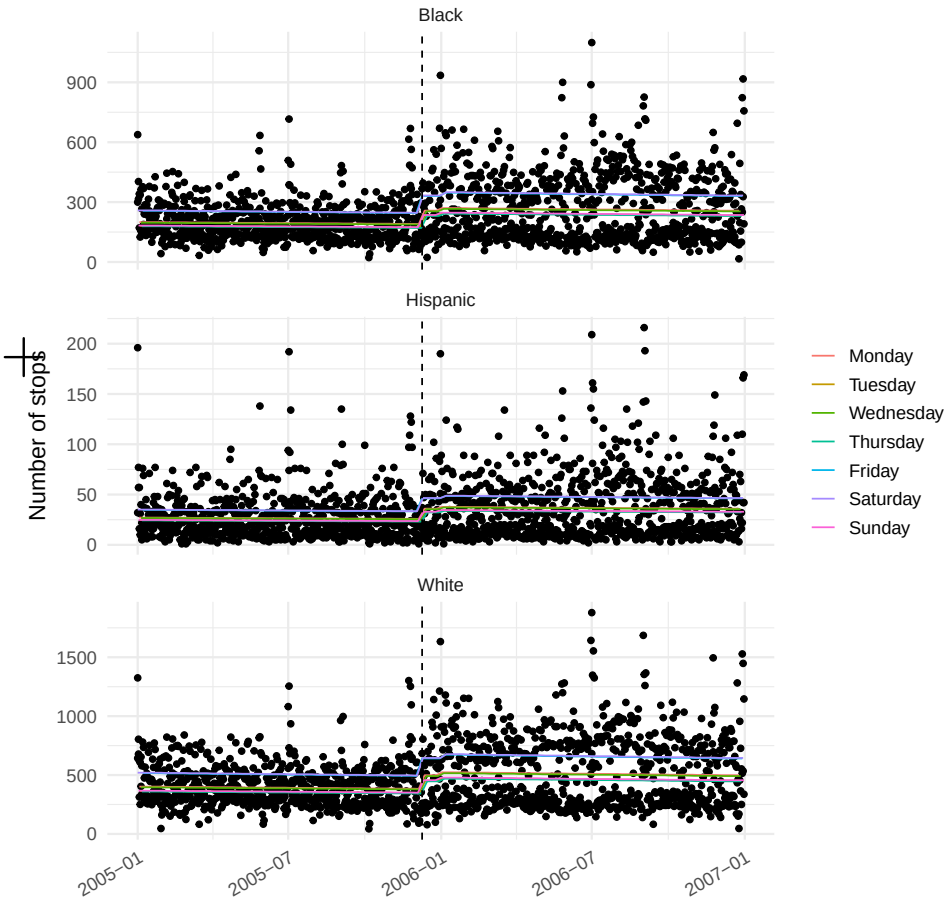
Generalized Linear Model - `glm`

$$E[Y] \sim g^{-1}(X\beta)$$
$$g(E[Y]) \sim X\beta$$

$$g(\text{daily number of stops}) \sim \beta_{driverRace} + \beta_{isPostPolicy} + \beta_{driverRace,isPostPolicy} + \beta_{dayOfWeek} + \beta * \text{month}$$

Choices:

- which covariates X to use
- response distribution (`quasipoisson`) and link function g



Generalized Additive Models: Intuition

$$g(E[Y]) = X\beta + f_1(x_{1i}) + f_2(x_{2i})$$

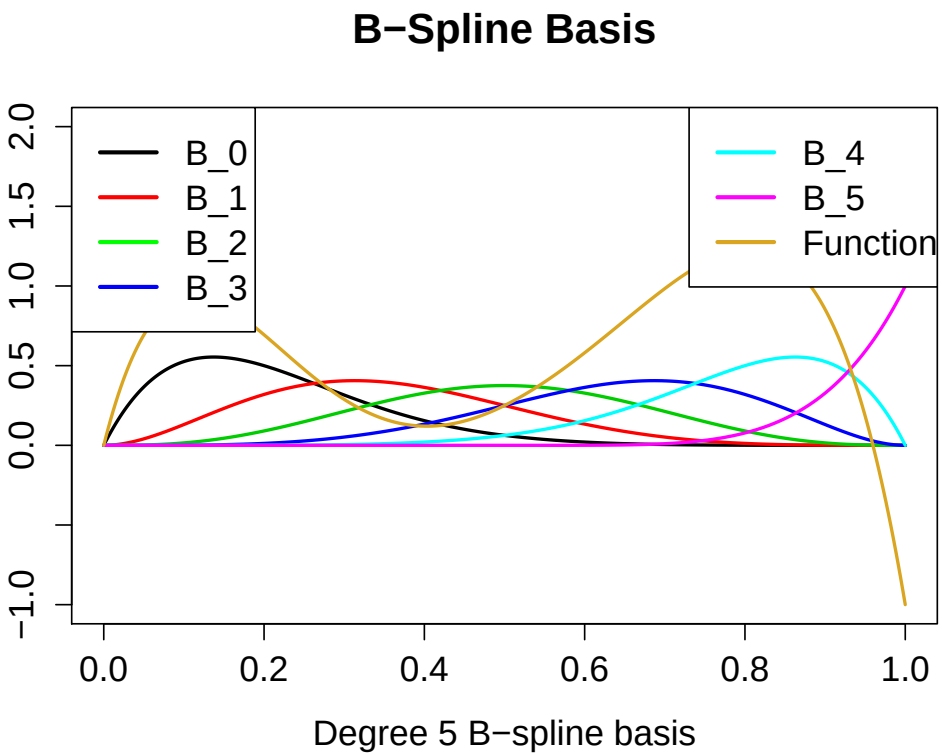
$$g(\text{daily number of stops}) \sim \beta_{driverRace} +$$
$$\beta_{isPostPolicy} + \beta_{driverRace,isPostPolicy} +$$
$$f_1(\text{dayOfWeek}) + f_2(\text{month})$$

Choices:

- which covariates X to use
- response distribution and link function g
- type of basis that defines f_i
- dimension of basis
- smoothing parameter

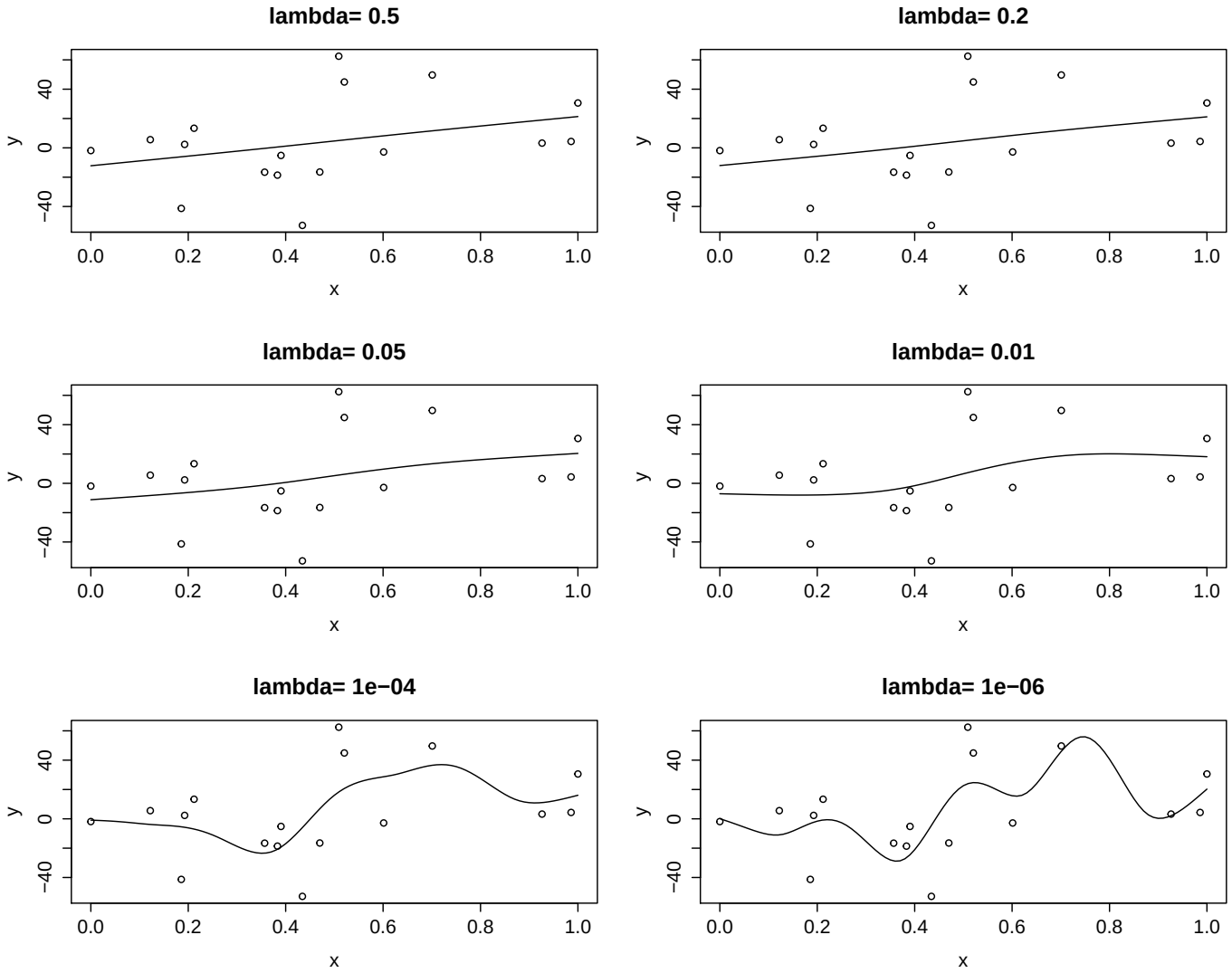
*Simon N. Wood. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC, 2017.

GAM: Parameter Intuition



Linear combination of specific curves $\approx$ general curve

GAM: Parameter Intuition



Bias-Variance Tradeoff

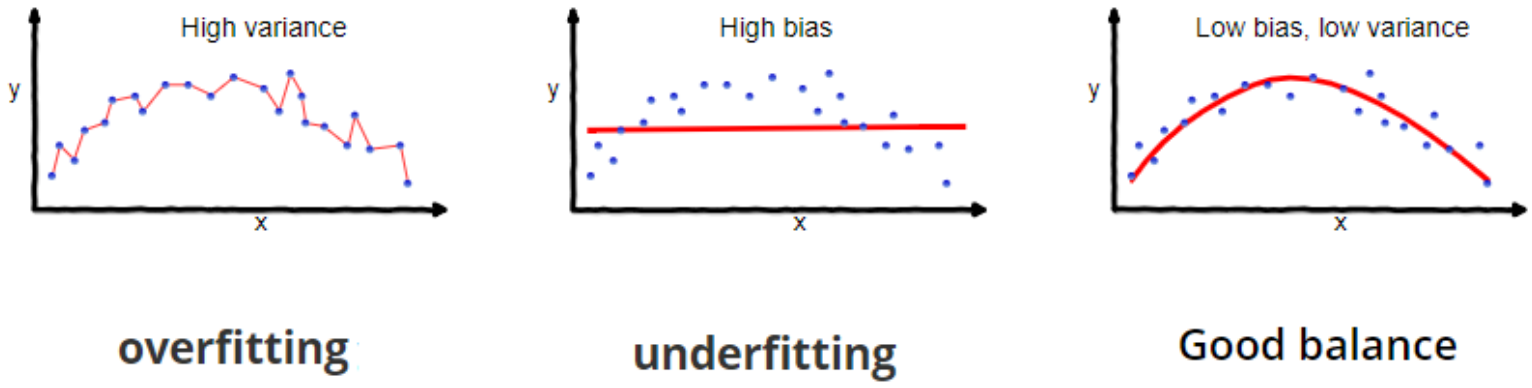
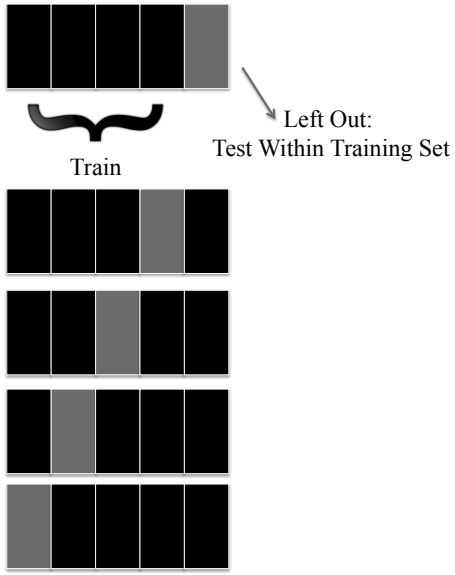
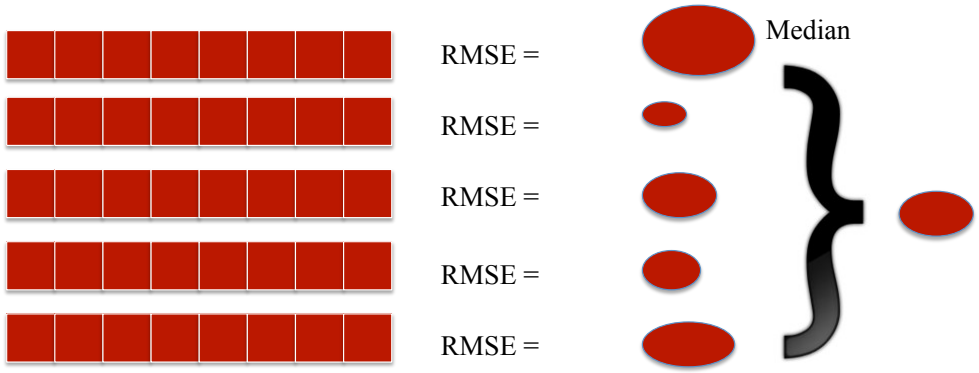


Figure Credit: `https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229`

Choosing the Smoothing Parameter: Cross Validation



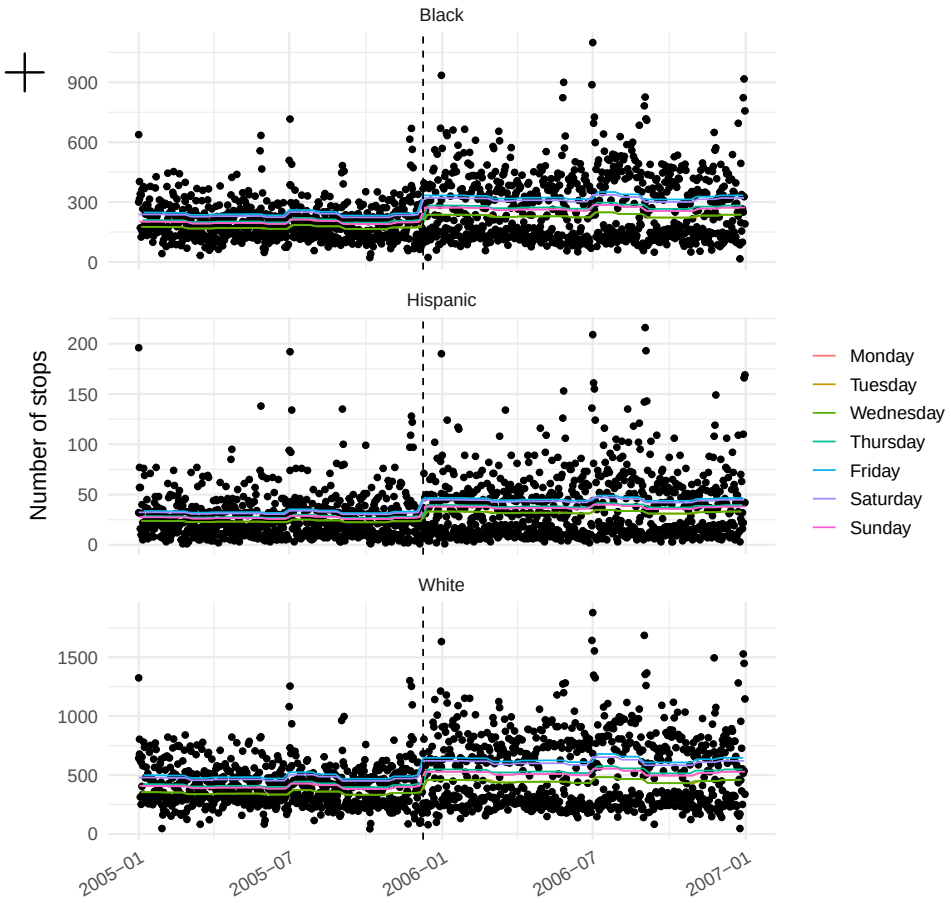


Generalized Additive Model - `mgcv::gam`

$$g(\text{daily number of stops}) \sim \beta_{driverRace} + \beta_{isPostPolicy} + \beta_{driverRace,isPostPolicy} + f_1(\text{dayOfWeek}) + f_2(\text{month})$$

Choices:

- which covariates X to use
- response distribution (`quasipoisson`) and link function g
- type of basis that defines f_i
- dimension of basis
- smoothing parameter (default: GCV)



Under the Hood - Cyclic Basis

`s(month, bs = "cc")`

- constrained to have the same beginning and end
- helpful for time components
- when you aren't in the mood to do a formal time series analysis

Under the Hood - Choosing k

$s(day\_of\_week\_num, bs = \text{"}cc\text{"}, k = 4)$

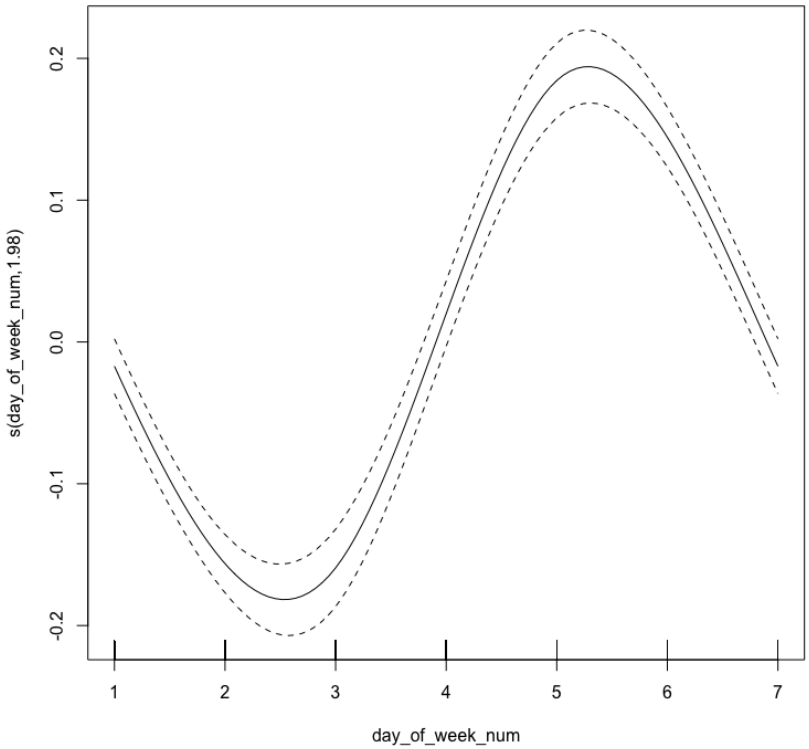
- k controls the maximum amount of flexibility
- bigger k means more computational complexity
- constrained by how much data you have (R will yell at you if you try to go too big)

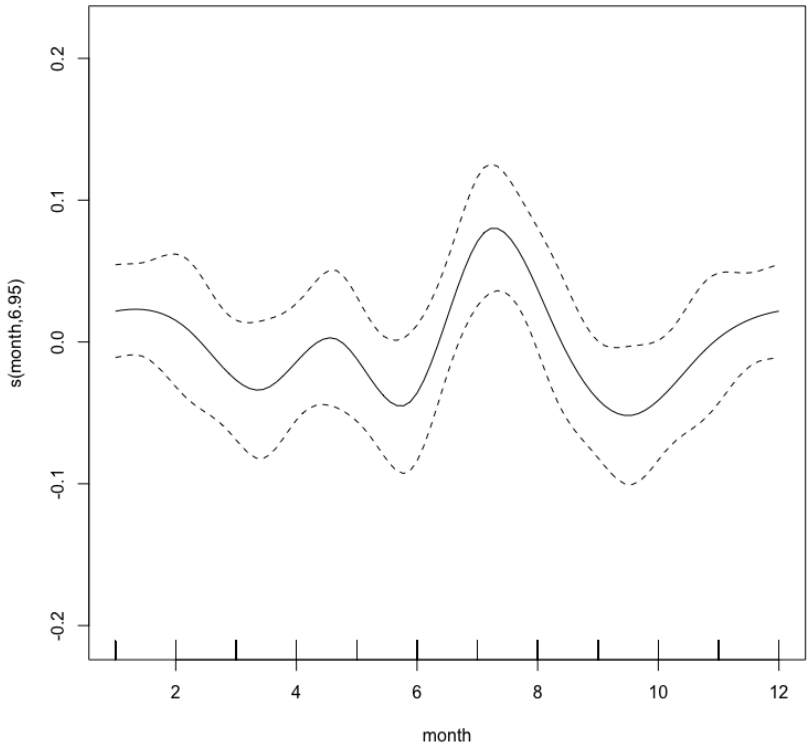
`gam.check(model)`

- there isn't one magic k, robust to choice of k as long as in a reasonable range
- but did I pick a k big enough?
- rough guide: small p-value means you could probably benefit from increasing k

Under the Hood - Seeing the Smooths

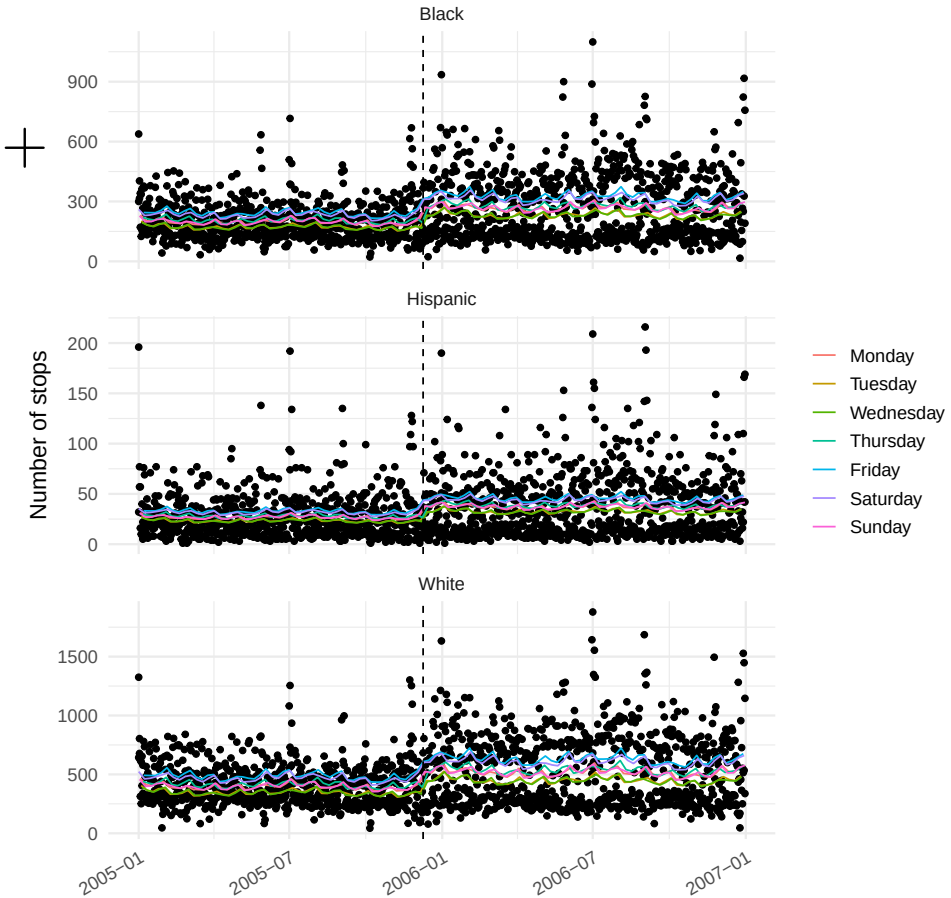
```
plot(model, rug = T)
```





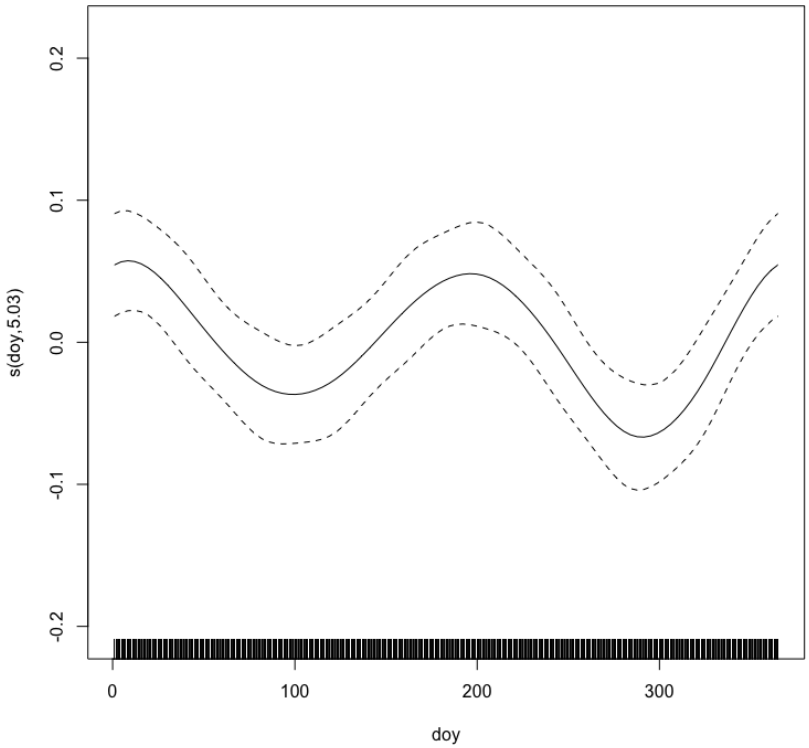
Generalized Additive Model - `mgcv::gam`

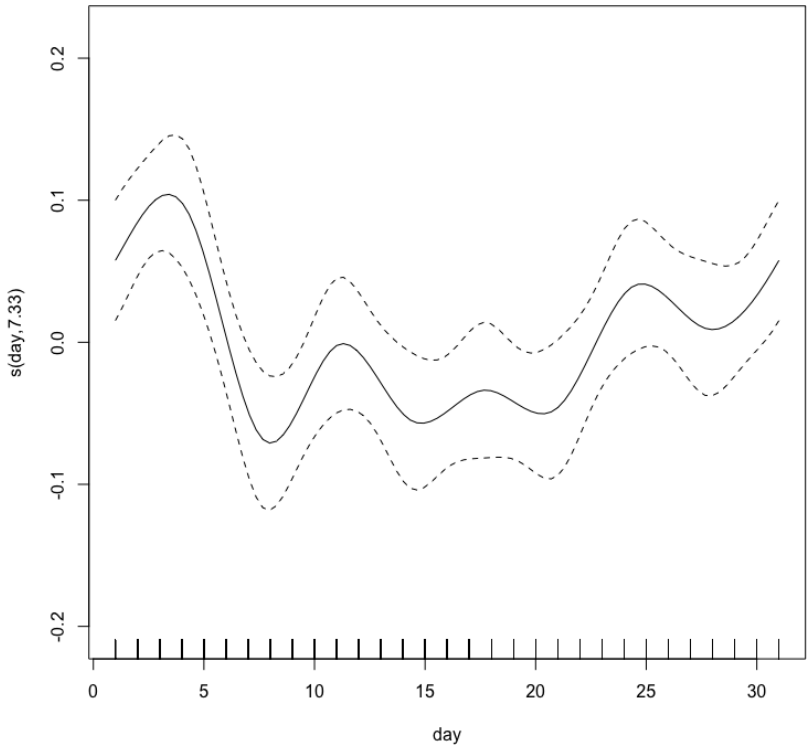
$$g(\text{daily number of stops}) \sim \beta_{driverRace} + \beta_{isPostPolicy} + \beta_{driverRace,isPostPolicy} + f_1(\text{dayOfWeek}) + f_2(\text{dayOfMonth}) + f_3(\text{dayOfYear})$$



More Smooths

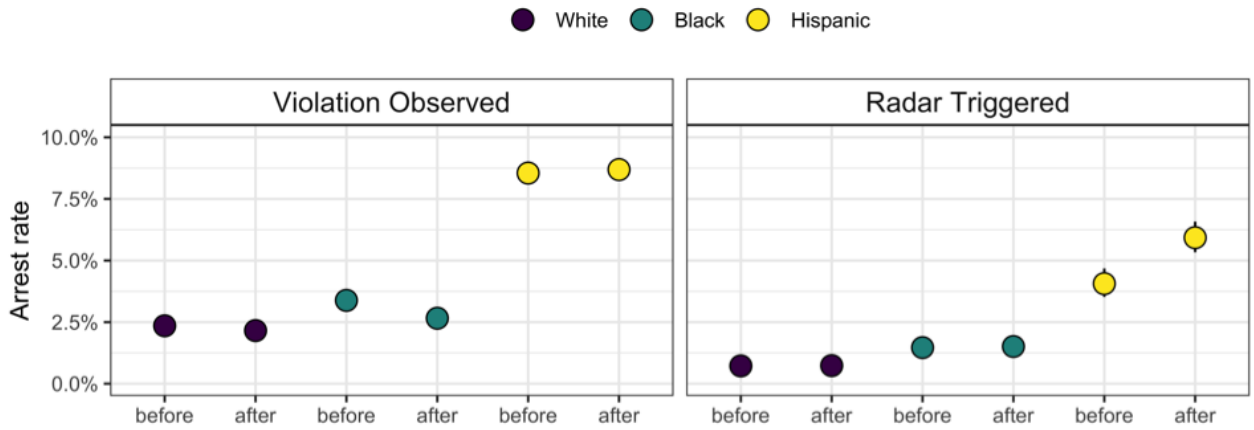
```
plot(model, rug = T)
```





Are these wiggles "real"?

Setting the Scene



- Figure credit: "Using change in a seat belt law to study racially-biased policing in South Carolina" by Corinne A Riddell, Jay S Kaufman, Jacqueline M Torres, and Sam Harper
- `https://github.com/corinne-riddell/SCarolinaTrafficStops`

Linear Model $\rightarrow$ Generalized Linear Model

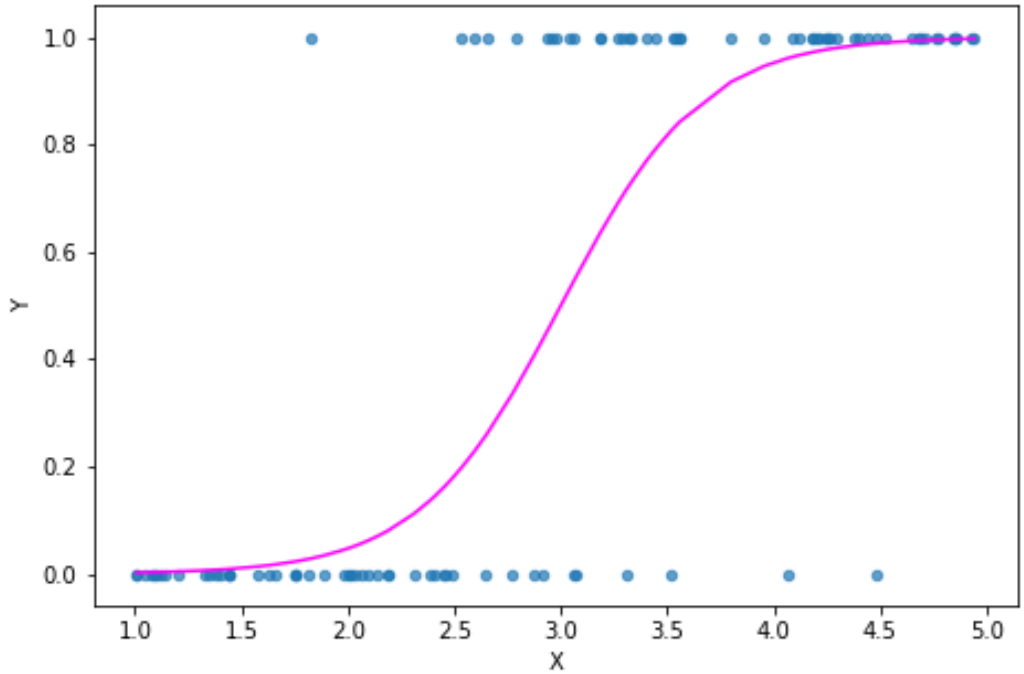


Figure Credit: `https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab`

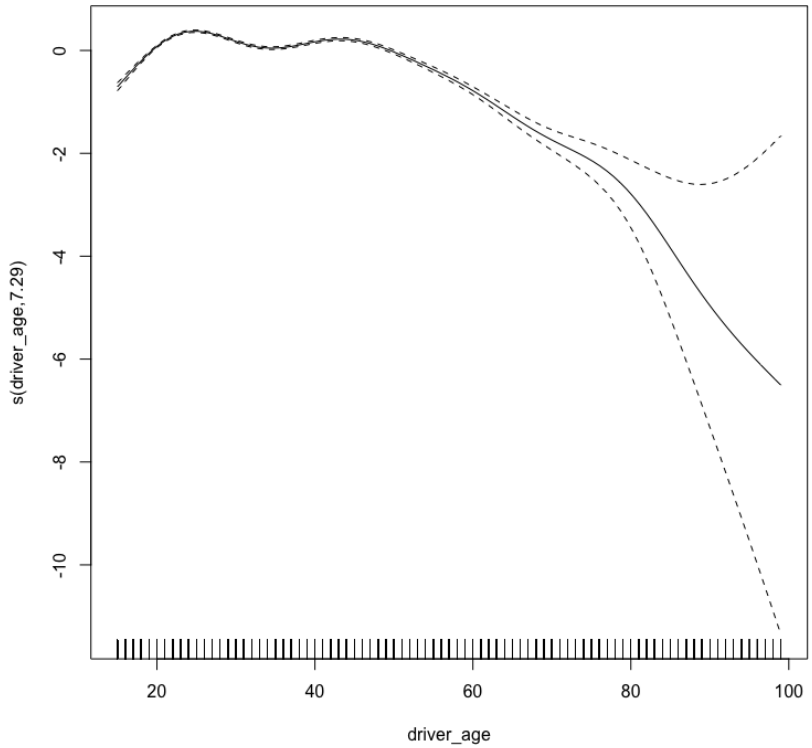
Big GAMs - `mgcv::bam`

$$g(\text{arrest}) \sim \beta_{driverRace} +$$

$$\beta_{gender} + f_1(\text{age}, bs = "cr") +$$

$$\beta_{isPostPolicy} + \beta_{driverRace,isPostPolicy}$$

- response distribution (`binomial`)
- link function g (`logit`)



Shrinkage Splines - `bs = "cs"`

$$g(\text{arrest}) \sim \beta_{driverRace} +$$
$$\beta_{gender} + f_1(\text{age}) +$$
$$\beta_{isPostPolicy} + \beta_{driverRace,isPostPolicy} +$$
$$f_1(\text{dayOfWeek}) + f_2(\text{dayOfMonth}) +$$
$$f_3(\text{dayOfYear})$$

- response distribution (`binomial`)
- link function g (`logit`)

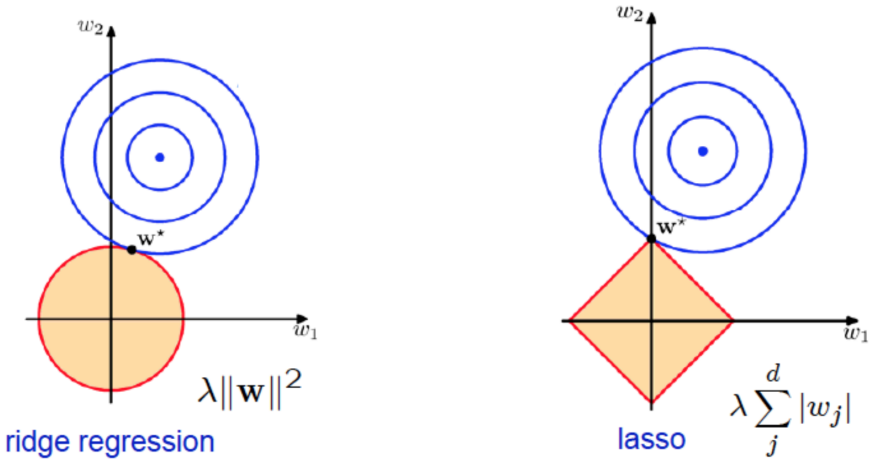
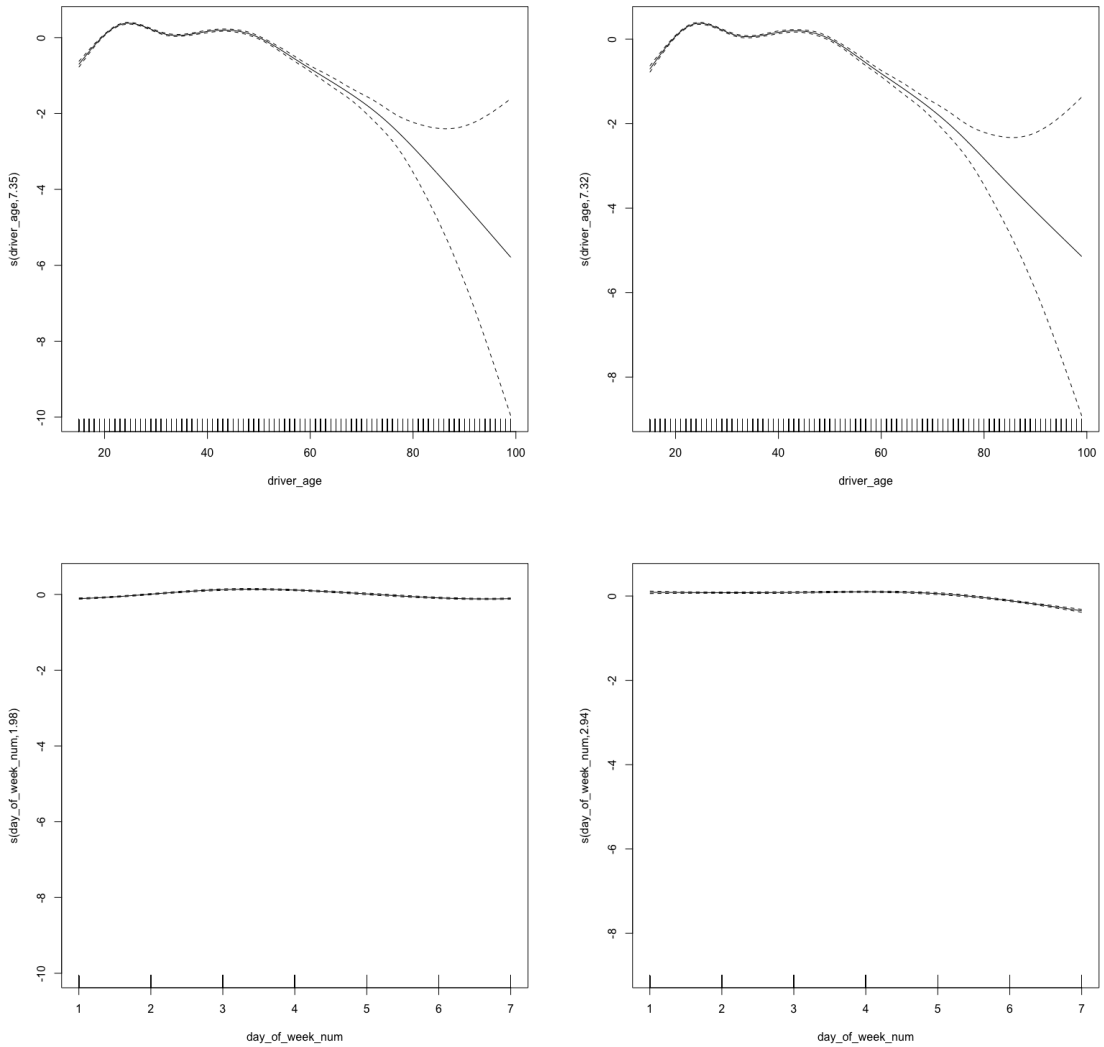
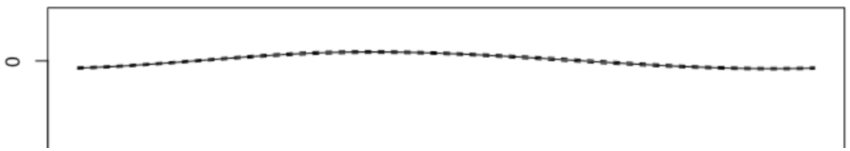


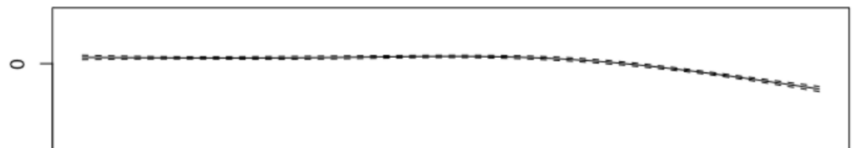
Figure credit: `http://alex.smola.org/teaching/cmu2013-10-701/slides/13_recitation_lasso.pdf`

Real wiggles?



Subtle shrinkage happening...

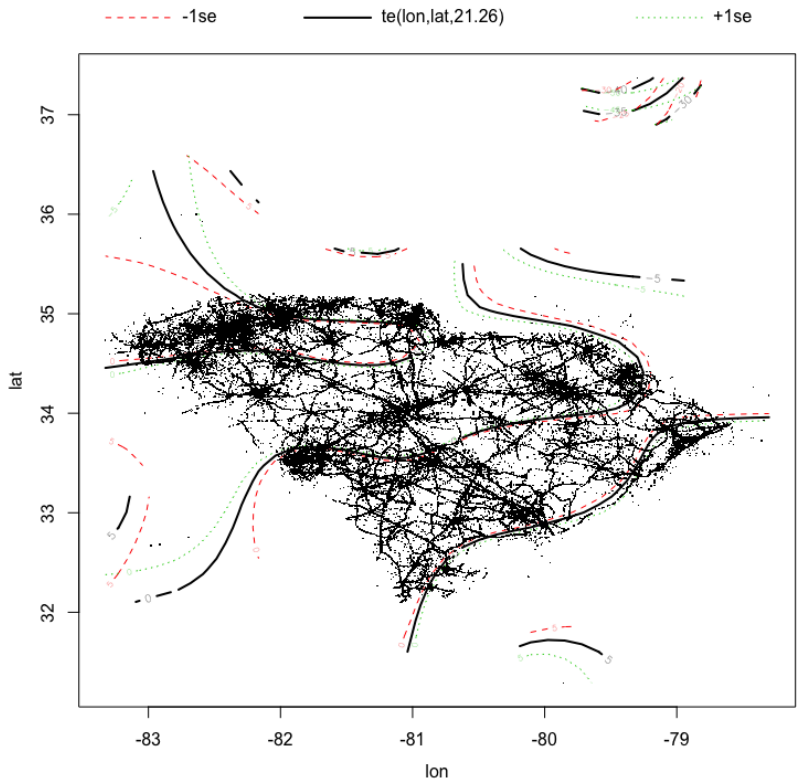




Multidimensional Smoothing

`te(lon, lat)`

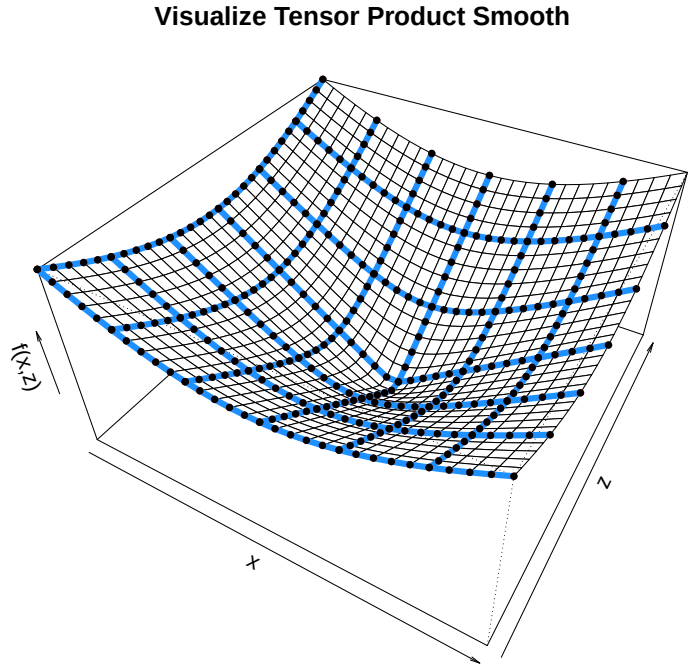
- two-dimensional smooth
- helpful for spatial components
- when you aren't in the mood to do a formal spatial analysis



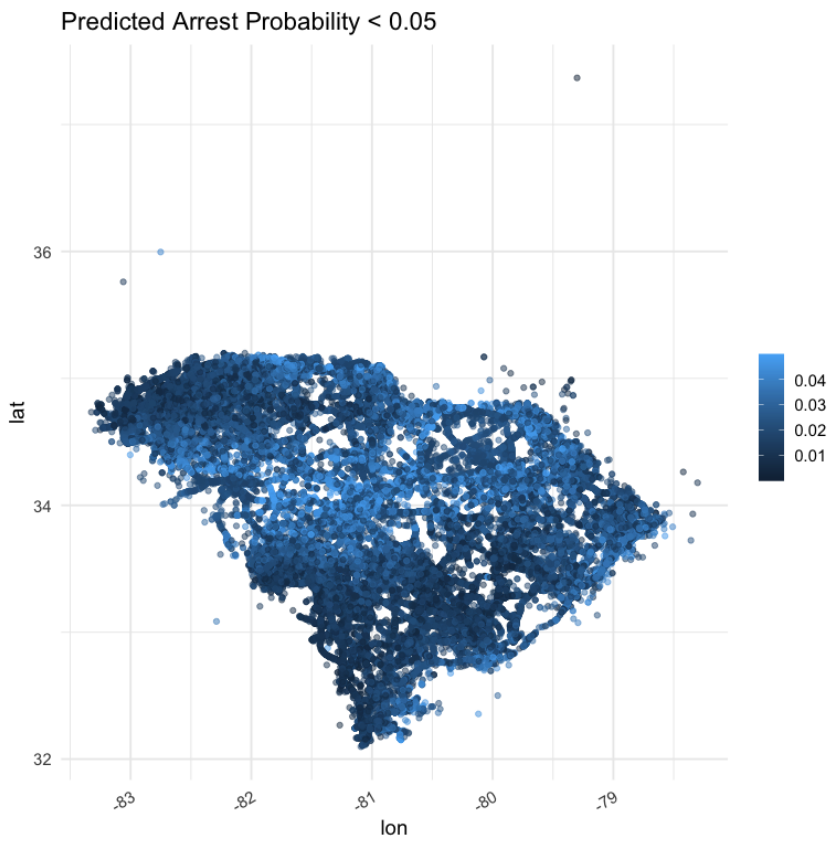
GAM: Parameter Intuition

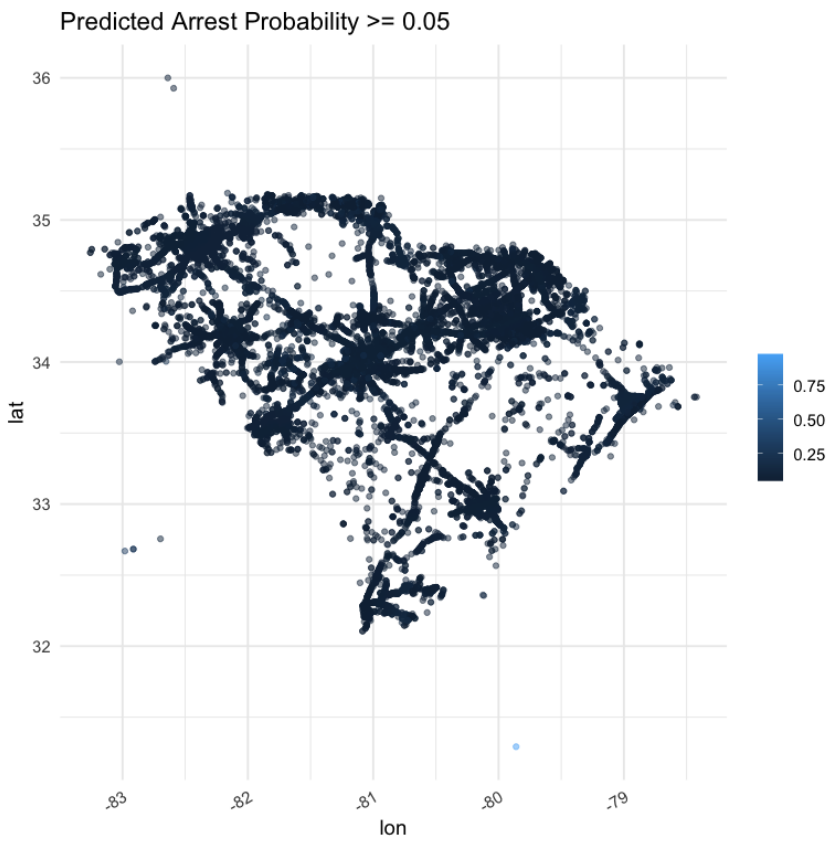
$$f(x) = \sum_{i=1}^{q} a_i(x)\alpha_k$$

$$f(x,z) = \sum \beta_i(z)a_i(x) = \sum_i \sum_j \beta_{ij} b_j(z)a_i(x)$$



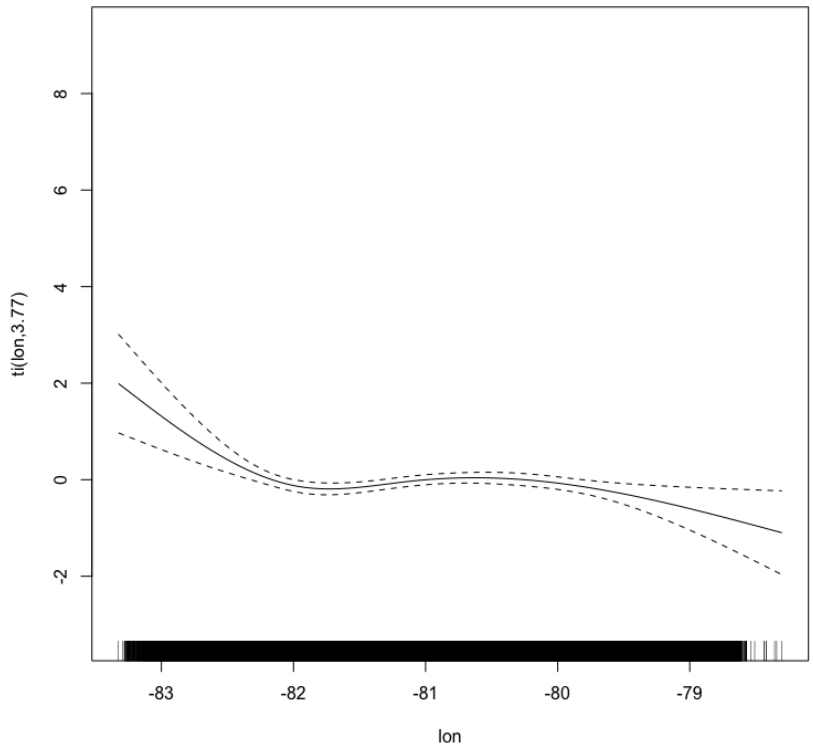
Spatial Predictions

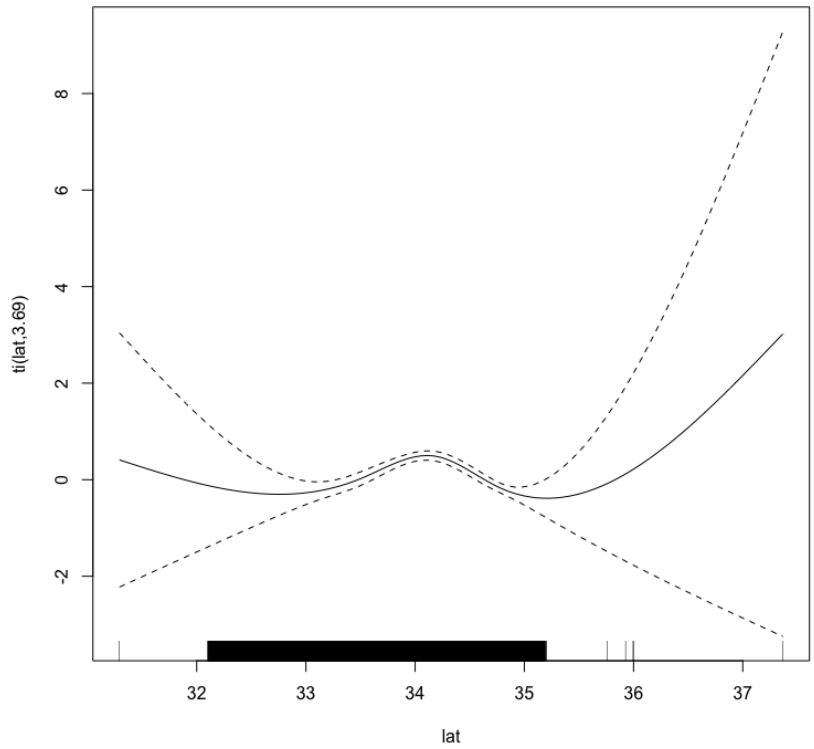




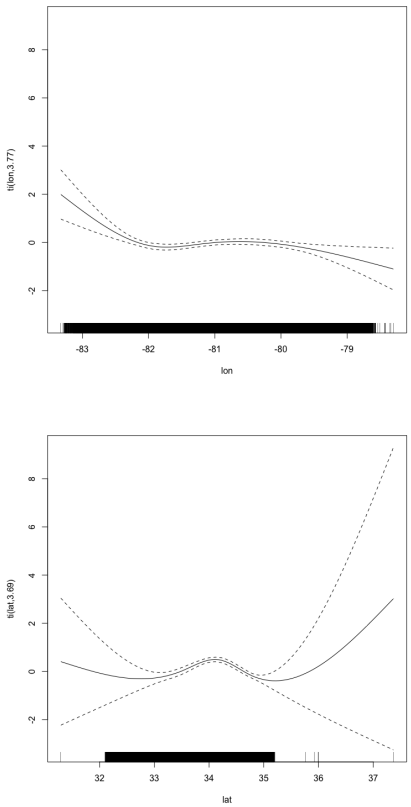
Decomposition

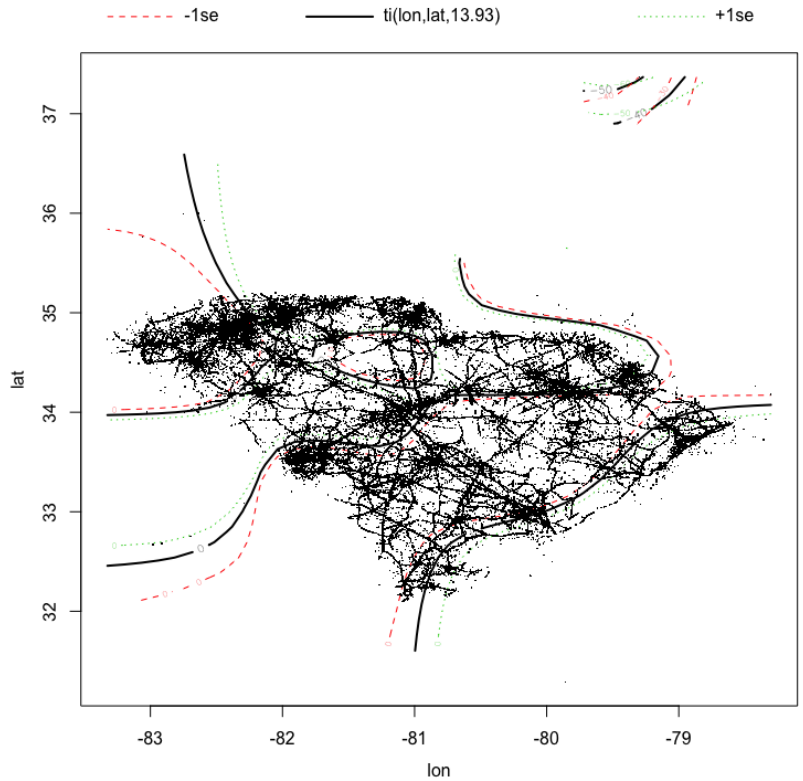
$$g(\text{arrest}) \sim \beta_{driverRace} + \beta_{gender} + f_1(\text{age}) + \beta_{isPostPolicy} + \beta_{driverRace,isPostPolicy} + ti(\text{lon}) + ti(\text{lat}) + ti(\text{lon, lat})$$





Decomposition





More Resources

- Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2017.
- `https://noamross.github.io/gams-in-r-course/`
- `http://environmentalcomputing.net/intro-to-gams/`
- `https://fromthebottomoftheheap.net/2021/02/02/random-effects-in-gams/`
- `https://www.tjmahr.com/random-effects-penalized-splines-same-thing/`
- many, many more...

Time to GAM-ify your own work?

Questions?

sstoudt@smith.edu

@sastoudt

I'll leave you with some other Wiggles...

Figure: `https://www.youtube.com/watch?v=a13WnqsRc5g`