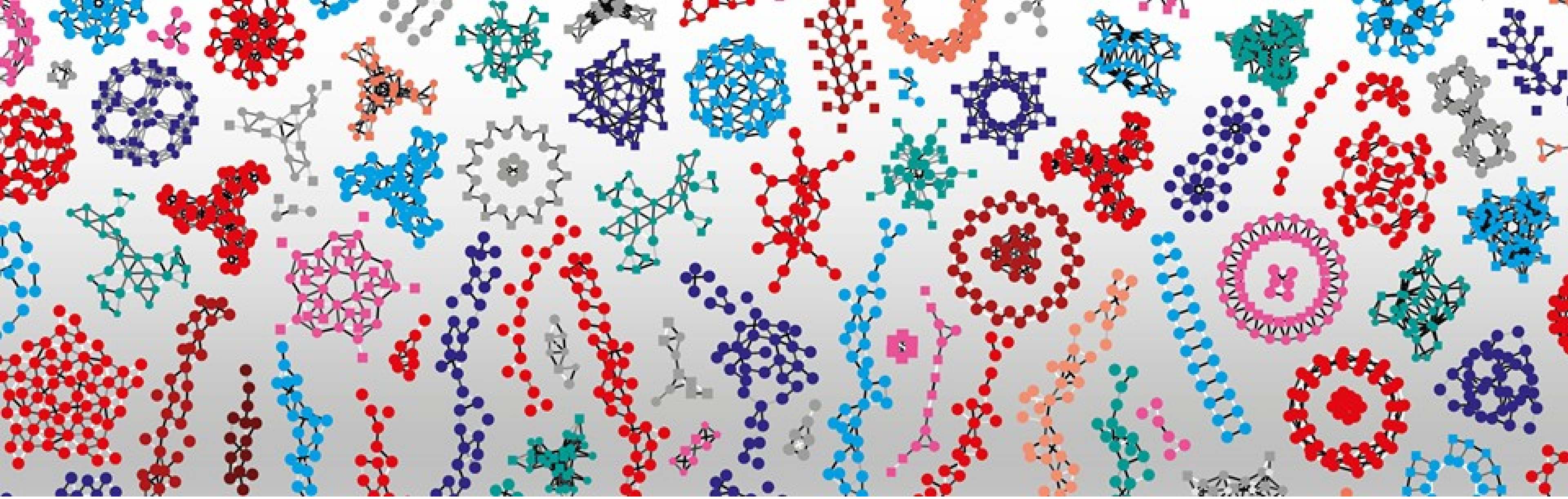


SIB
Swiss Institute of Bioinformatics

Generalized Additive Models

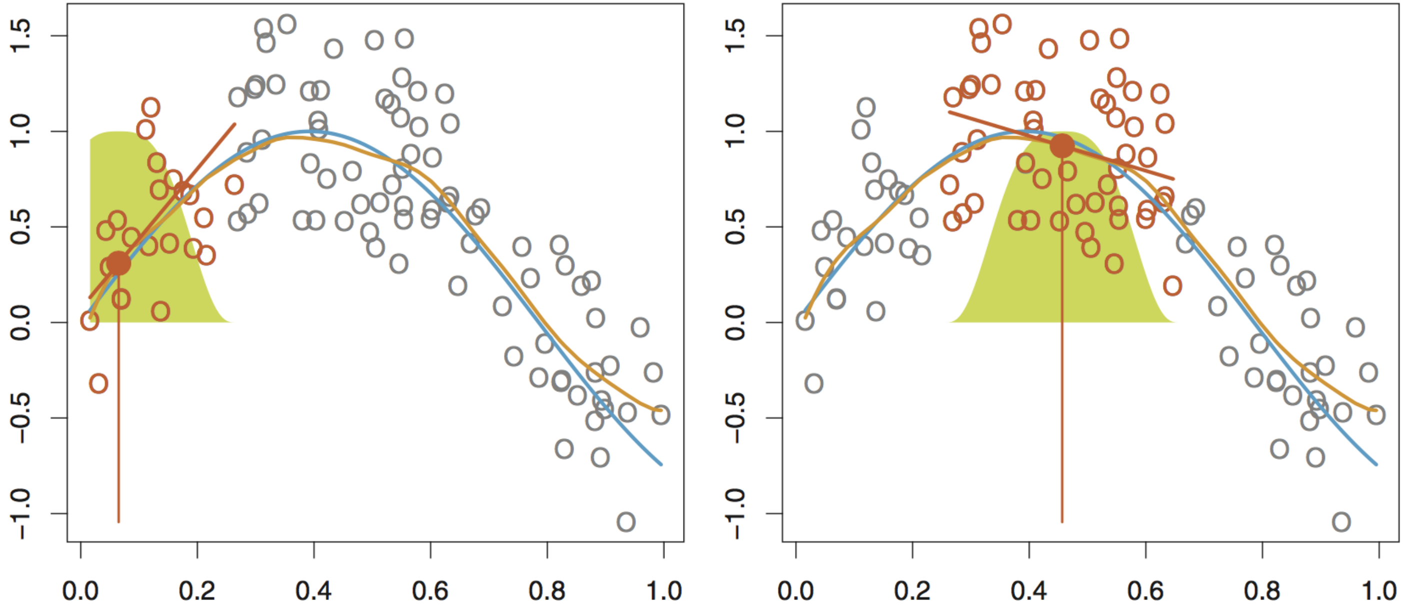
Sina Nassiri, Ph.D.

Bioinformatics Scientists @ the Bioinformatics Core Facility

Aug 20-23, 2018 - Lausanne

Local Regression

- Local regression is a different approach for fitting flexible nonlinear functions, which involves computing the fit at a target point x_0 using only the nearby training observations



Local Regression

- Local regression is a different approach for fitting flexible nonlinear functions, which involves computing the fit at a target point x_0 using only the nearby training observations
- In order to perform local regression, there are a number of choices to be made, such as:
  - how to define the weighting function
  - whether to fit a linear, constant, quadratic regression, etc.
  - The most important choice is the "span". The span plays a role like that of the tuning parameter λ in smoothing splines: it controls the flexibility of the non-linear fit
    - The smaller the value of s, the more local and wiggly will be our fit; alternatively, a very large value of s will lead to a global fit to the data using all of the training observations

Generalized Additive Models (GAMs)

- So far, we have seen a number of approaches for flexibly predicting a response Y on the basis of a single predictor X
- Here we explore the problem of predicting Y on the basis of several predictor X1, X2, …, Xp
- GAMs provide a general framework for extending a standard linear model by allowing smooth functions of each of the variables, while maintaining additivity
  - The response can be either quantitative or qualitative

Generalized Additive Models (GAMs)

- A natural way to extend the multivariable linear regression model

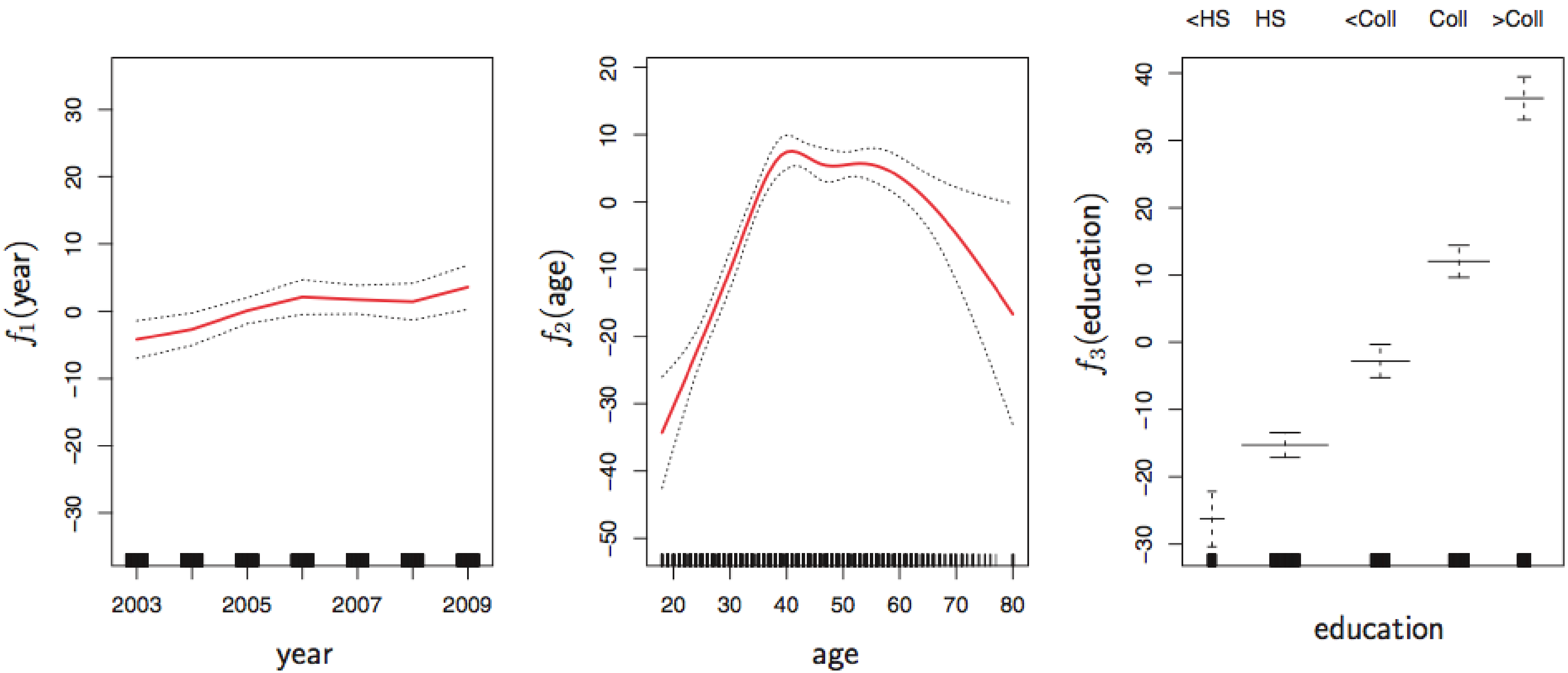
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

in order to allow for smooth relationships between each feature and the response is to replace each linear component with a smooth function:

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \varepsilon_i$$

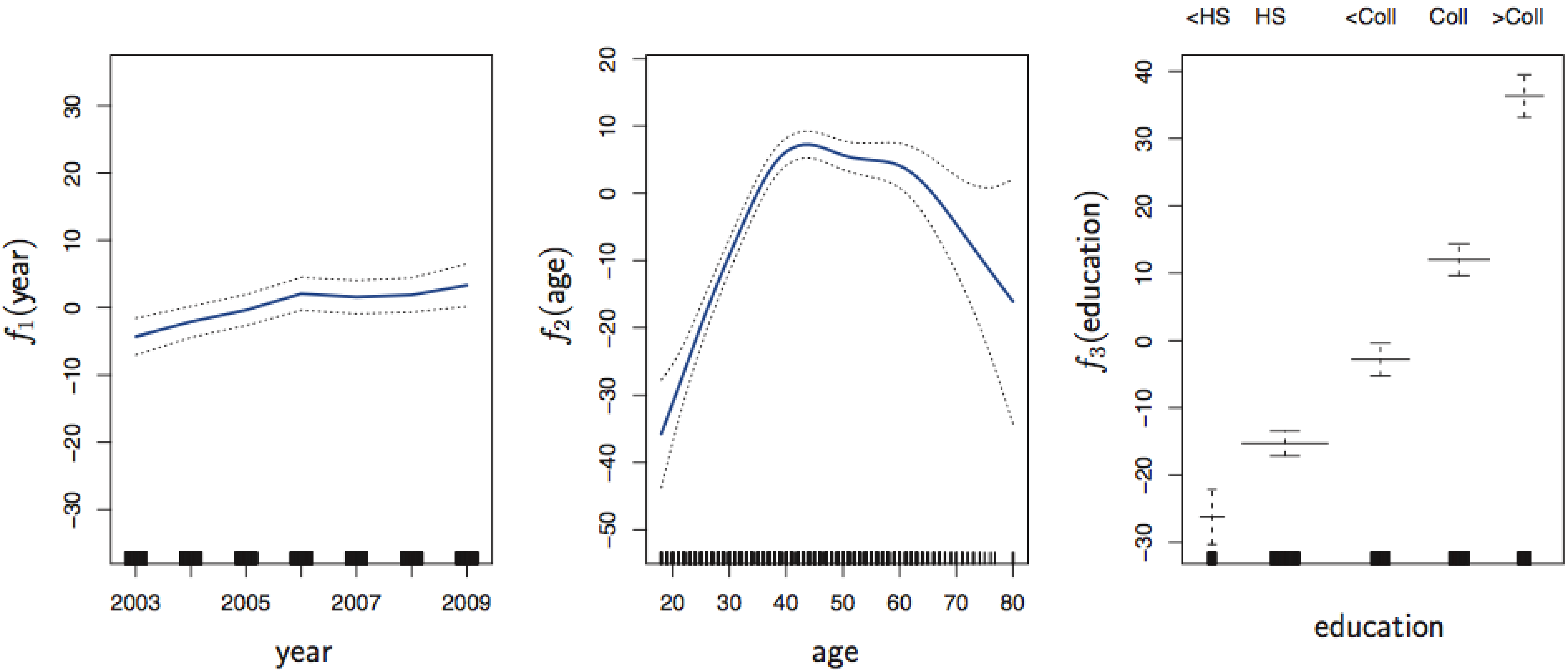
- This is an example of a GAM!
- It is called additive model because we calculate a separate f_j for each X_j, and then add together all of their contributions
- The beauty of GAMs is that we can use various smoothing methods as building blocks for fitting an additive model
  - Spline regression
  - Smoothing splines
  - Local regression (loess)
  - etc.

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$



The first two functions are **natural splines**; the third function is a step function

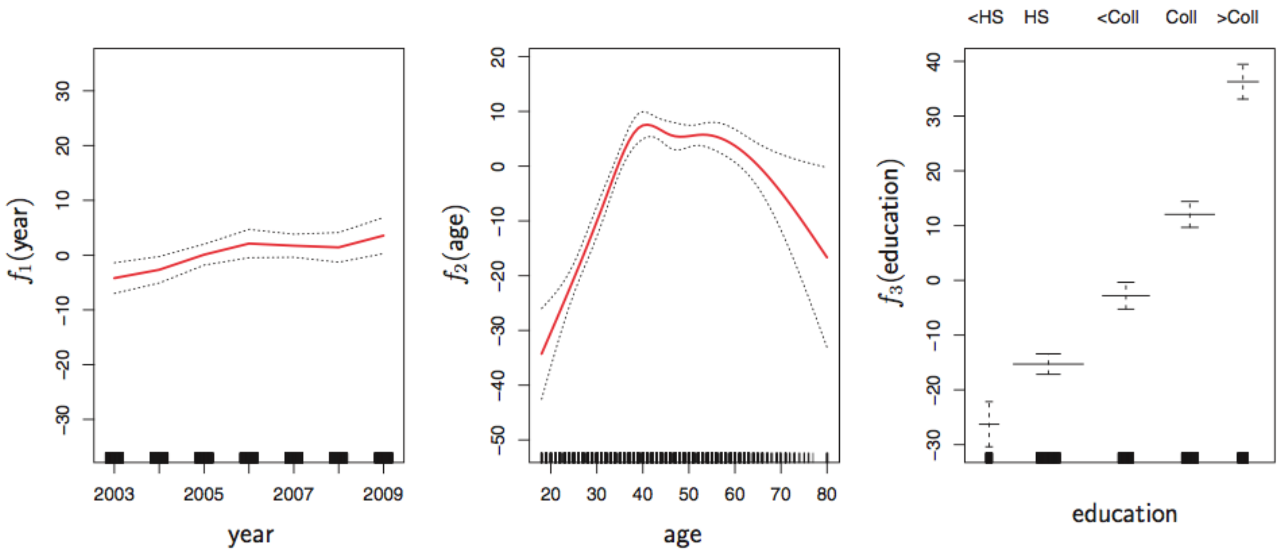
$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$



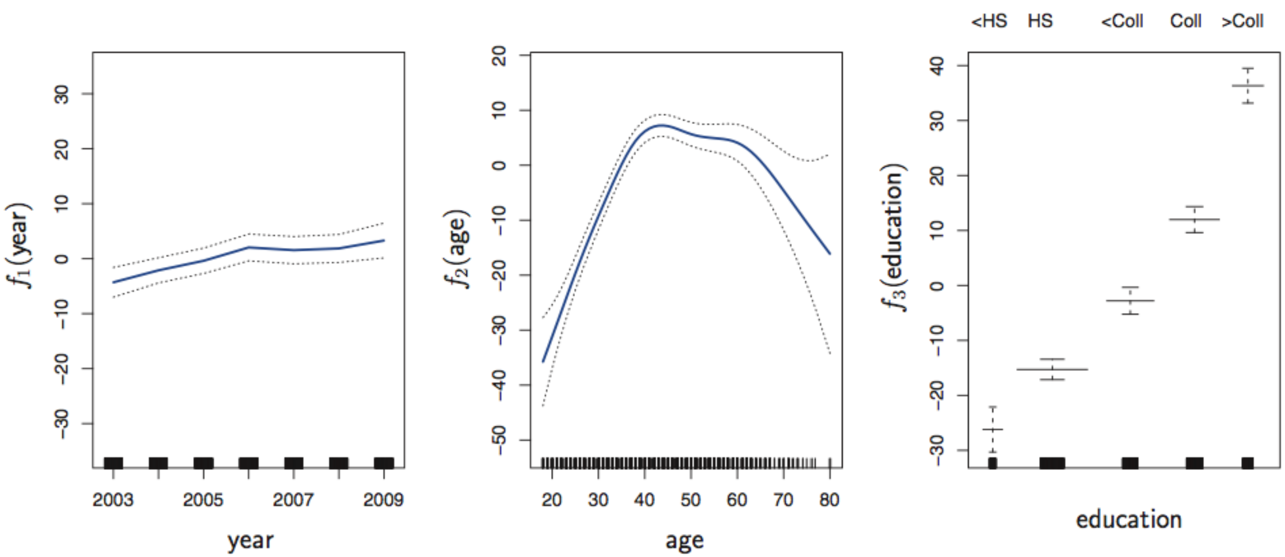
The first two functions are **smoothing splines**; the third function is a step function

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

natural splines



smoothing splines



Fitting GAMs

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

- Natural splines can be constructed using an appropriately chosen set of basis functions.
- Hence the entire model is just a big regression onto spline basis variables (for year and age) and dummy variables (for education)
- Fitting a GAM with a smoothing spline is not quite as simple as fitting a GAM with a natural spline.
- The gam package in R uses an approach known as ***backfitting***
  - Involves repeated updating of the fit for each predictor while holding others fixed
  - Each time we update a function, we simply apply the fitting method for that variable to a partial residual

A partial residual for X_3, for example, has the form $r_i = y_i - f_1(x_{i1}) - f_2(x_{i2})$

If we know f_1 and f_2, then we can fit f_3 by treating this residual as a response in a smooth regression on X_3

- The mgcv package in R uses mixed modeling framework for smoothing

Pros and Cons of GAMs

- GAMs allow us to fit a smooth f_j to each X_j, so that we can automatically model non-linear relationships that standard linear regression will miss
  - This means we do not need to manually try many different transformations on each variable independently
- The smooth fits can potentially make more accurate predictions for the response Y
- Because the model is additive, we can still examine the effect of each Xj on Y individually while holding all of the other variables fixed
- The main limitation of GAMs is that the model is restricted to be additive. With many variables, important interactions can be missed
  - However, we can manually add interaction terms by including additional parameters of the form $X_j * X_k$
  - In addition, we can add low-dimensional interaction functions of the form $f_{jk}(X_j, X_k)$ into the model; such terms can befit using two-dimensional smoothers

Smoothing Exercise: The "wage" data

- Mid-Atlantic Wage Data
  - Wage and other data for a group of 3000 make workers in the Mid-Atlantic region

References

- Semiparametric Regression; by David Ruppert, M.P. Wand, and R.J. Carroll; Cambridge University Press
- Generalized Additive Models; by T.J. Hastie and R.J. Tibshirani; Chapman & Hall/CRC
- Generalized Additive Models - An Introduction with R (2nd Edition); by Simon N. Wood; CRC Press

Lifetime Warranty ᄊ

If you have follow-up questions, need help with your future analysis, or simply want to stay in touch, feel free to contact me at:

sina.nassiri@sib.swiss