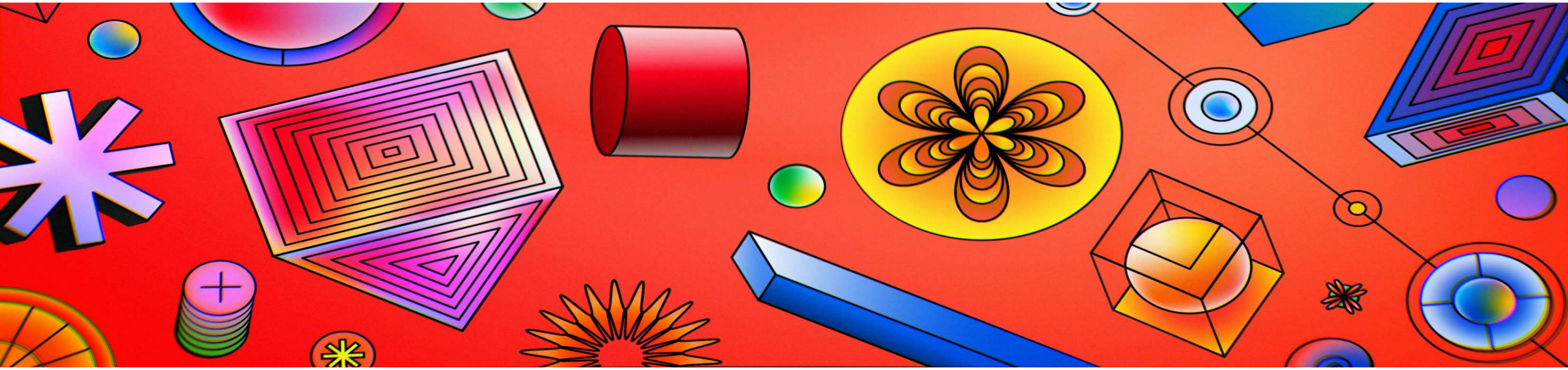
Fall 2023

BIF524/CSC463 Data Mining

Statistical Learning

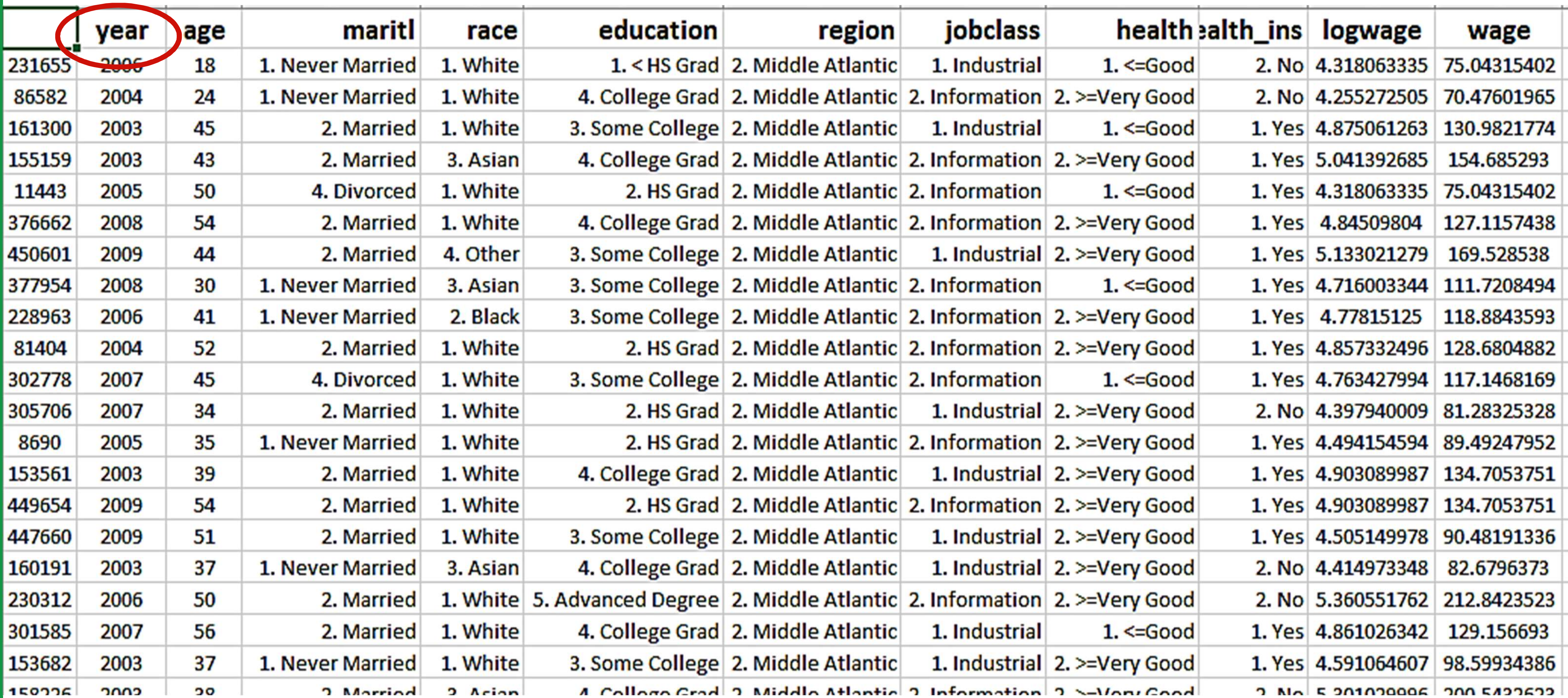
Eileen Marie Hanna, *PhD*

07/09/2023



“Wage” dataset

- Includes factors believed to be related to wages of a group of males from the Atlantic region in the US, e.g., age, education level, ..etc.



	year	age	maritl	race	education	region	jobclass	health	ealth_ins	logwage	wage
231655	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063335	75.04315402
86582	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255272505	70.47601965
161300	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061263	130.9821774
155159	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041392685	154.685293
11443	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063335	75.04315402
376662	2008	54	2. Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.84509804	127.1157438
450601	2009	44	2. Married	4. Other	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.133021279	169.528538
377954	2008	30	1. Never Married	3. Asian	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.716003344	111.7208494
228963	2006	41	1. Never Married	2. Black	3. Some College	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.77815125	118.8843593
81404	2004	52	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.857332496	128.6804882
302778	2007	45	4. Divorced	1. White	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.763427994	117.1468169
305706	2007	34	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.397940009	81.28325328
8690	2005	35	1. Never Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.494154594	89.49247952
153561	2003	39	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.903089987	134.7053751
449654	2009	54	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.903089987	134.7053751
447660	2009	51	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.505149978	90.48191336
160191	2003	37	1. Never Married	3. Asian	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.414973348	82.6796373
230312	2006	50	2. Married	1. White	5. Advanced Degree	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	5.360551762	212.8423523
301585	2007	56	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.861026342	129.156693
153682	2003	37	1. Never Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.591064607	98.59934386
158226	2003	38	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	5.301029996	200.5432623

Prediction

- Predict Y given a set of inputs X.

- Y can be predicted using:

$$\hat{Y} = \hat{f}(X)$$

resulting prediction for *Y*

estimate for *f*

Prediction – example

$$\hat{Y} = \hat{f}(X)$$

- Let $\boldsymbol{X_1, X_2, \dots, X_p}$ be the measured **characteristics of a blood sample**
  - Let $\boldsymbol{Y}$ be a variable corresponding to the **patient's risk of a severe adverse reaction to a drug**.
- In such settings, **two factors determine the accuracy of** $\boldsymbol{\widehat{Y}}$:

Prediction – example

$$\hat{Y} = \hat{f}(X)$$

- Let $\boldsymbol{X_1, X_2, \ldots, X_p}$ be the measured **characteristics of a blood sample** and let $\boldsymbol{Y}$ be a variable corresponding to the **patient's risk of a severe adverse reaction to a drug**.

- In such settings, **two factors determine the accuracy of $\boldsymbol{\widehat{Y}}$:**

  - **reducible error:** usually $\hat{f}$ is not expected to be a perfect estimate of f -> error.

    - **Reducible because we can improve the accuracy** (i.e., reduce the error) by using more appropriate learning techniques.

Prediction – example

$$\hat{Y} = \hat{f}(X)$$

- Let $\boldsymbol{X_1, X_2, \ldots, X_p}$ be the measured **characteristics of a blood sample** and let $\boldsymbol{Y}$ be a variable corresponding to the **patient's risk of a severe adverse reaction to a drug**.

- In such settings, **two factors determine the accuracy of $\boldsymbol{\hat{Y}}$:**

  - **reducible error**

  - <u>**irreducible error:**</u> there **will always be** an irreducible error introduced by ∈ because Y is also a function of ∈, which cannot be predicted by X.

    - Due to some **unmeasured or unmeasurable factors**
      - e.g., the manufacturing variation of the drug itself or the patient's wellbeing

Prediction – how?

- The goal is to **use appropriate learning techniques to minimize the reducible error**.

- Suppose that we have an estimate $\hat{f}$ and a set of predictors X leading to prediction:

$$\hat{Y} = \hat{f}(X)$$

- The expected value (average) of the square difference between the predicted and the actual value of Y is given by:

variance of error

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

$$Y = f(X) + \epsilon$$

Inference

- Understand how $\boldsymbol{Y}$ changes when $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_p}$ change.
  - i.e., how Y changes as a function of $X_1, X_2, \ldots, X_p$.
- **In such cases, $\hat{f}$ cannot be treated as a black box!**

Inference

- Inference techniques allow us to answer questions as:
  - **Which variables are associated with the response?**
    - Often, only a subset of predictors is strongly associated with Y.
  - **What is the relationship between the response and each predictor?**
    - e.g., increased predictor -> increased response (positive relationship)
  - **Can the relationship between $\boldsymbol{Y}$ and each predictor be adequately summarized as a linear equation or is it more complex?**

Inference – "Advertising" dataset

- Questions that could fall within the inference category:
  - **Which media contribute to sales?**
    - i.e., media that are strongly associated with Y.
  - **Which media lead to the biggest boost in sales?**
    - e.g., increased predictor -> increased response (positive relationship)?
  - **How much increase in sales is associated with a certain increase in TV advertising budget?**
    - adequate linear equation summarizing this relationship?

How do we estimate $\boldsymbol{f}$?

- **<u>Training data:</u> observations used to train a method how to estimate f.**

  - $\boldsymbol{n}$: number of distinct data points (observations) in a sample
  - $\boldsymbol{p}$: number of available variables (attributes)

  - $\boldsymbol{x_{ij}}$: j^{th} variable for the i^{th} observation, $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,p$
  - $\boldsymbol{y_i}$: response variable for the i^{th} observation

- In this setting, the training data consist of:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \text{ where } x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$$

Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function $\boldsymbol{f}$.

Parametric methods

Non-parametric methods

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Parametric methods

- A two-step approach:

  1. **Make an assumption about the functional form or shape of f.**

     - e.g., it is linear in X, such that:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- Here, the problem is reduced to estimating $p + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$.

Parametric methods

- A two-step approach:

1. **Make an assumption about the functional form or shape of f.**

2. **Apply a procedure that uses the training data to fit the model.** In this example, we need to estimate $\beta_0, \beta_1, \dots, \beta_p$ such that:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

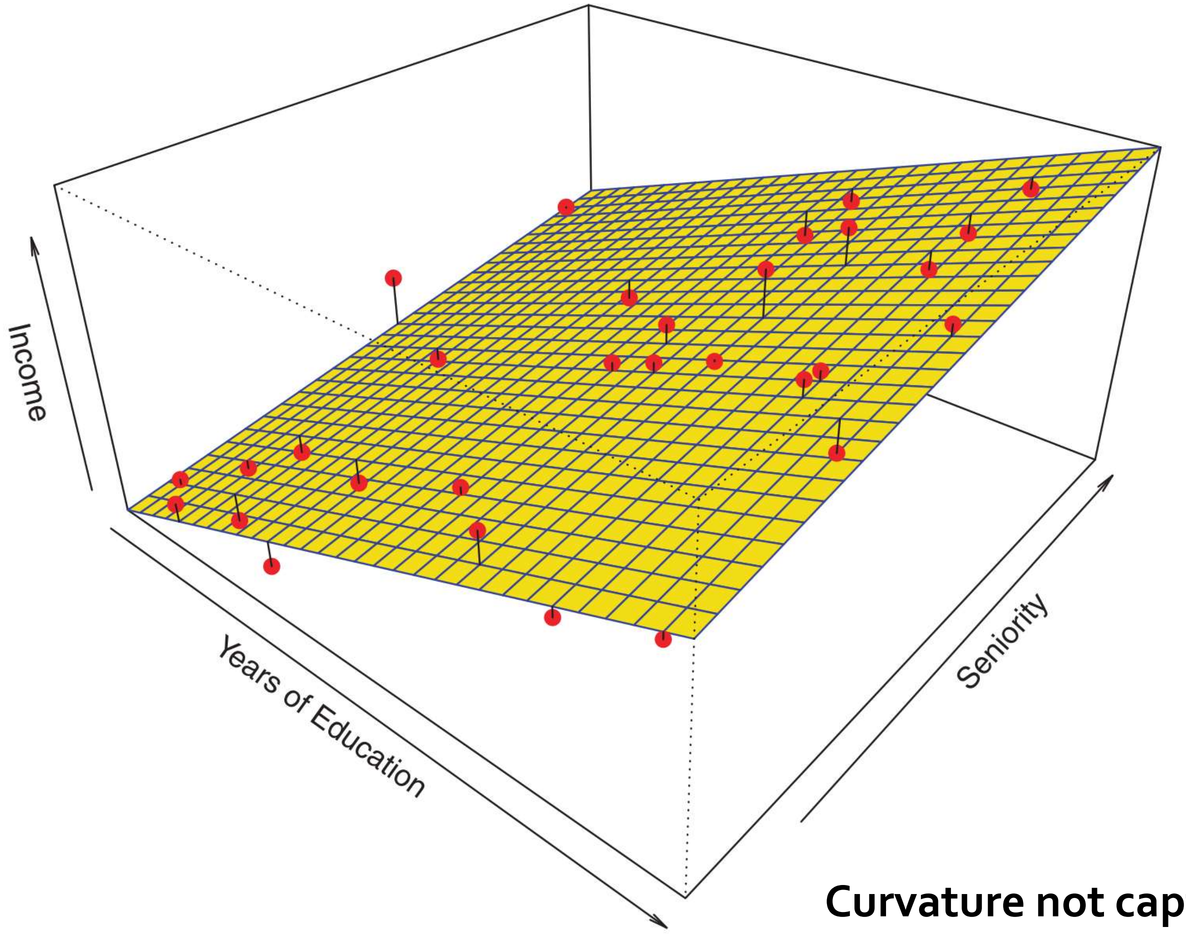
One way is to use the least squares approach – later.

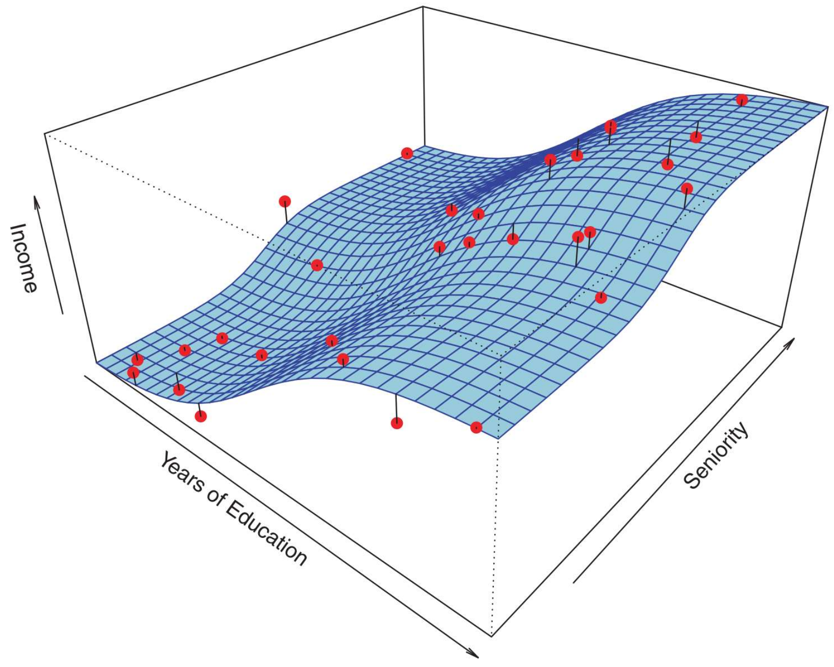
Parametric methods – comments

- **The model that we choose does not usually match the true unknown f.**
  - **The more far** the model is from true f, the **poorer the estimate**.
- One way is to **choose more flexible models** that can fit different forms of f.
  - **However**, fitting more models -> estimating more parameters -> **overfitting**, i.e., when models follow noise too closely – later.

Parametric methods – “Wages”

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}$$





Curvature not captured but the model succeeds in showing the positive relationship between years of education and income as well as the slightly positive relationship between seniority and income.

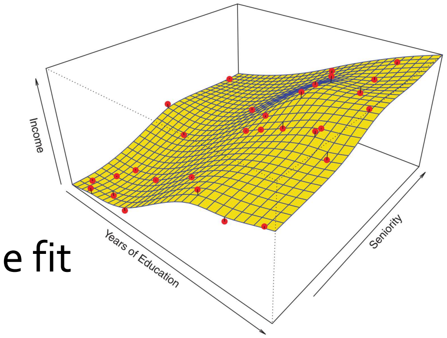
Non-parametric methods

- **No explicit assumptions about the functional form of f**
  - **potential to accurately fit a wider range of possible shapes** –advantage in contrast with parametric methods.

- Estimate f by **getting as close as possible to the data points** without being too rough or too wiggly.

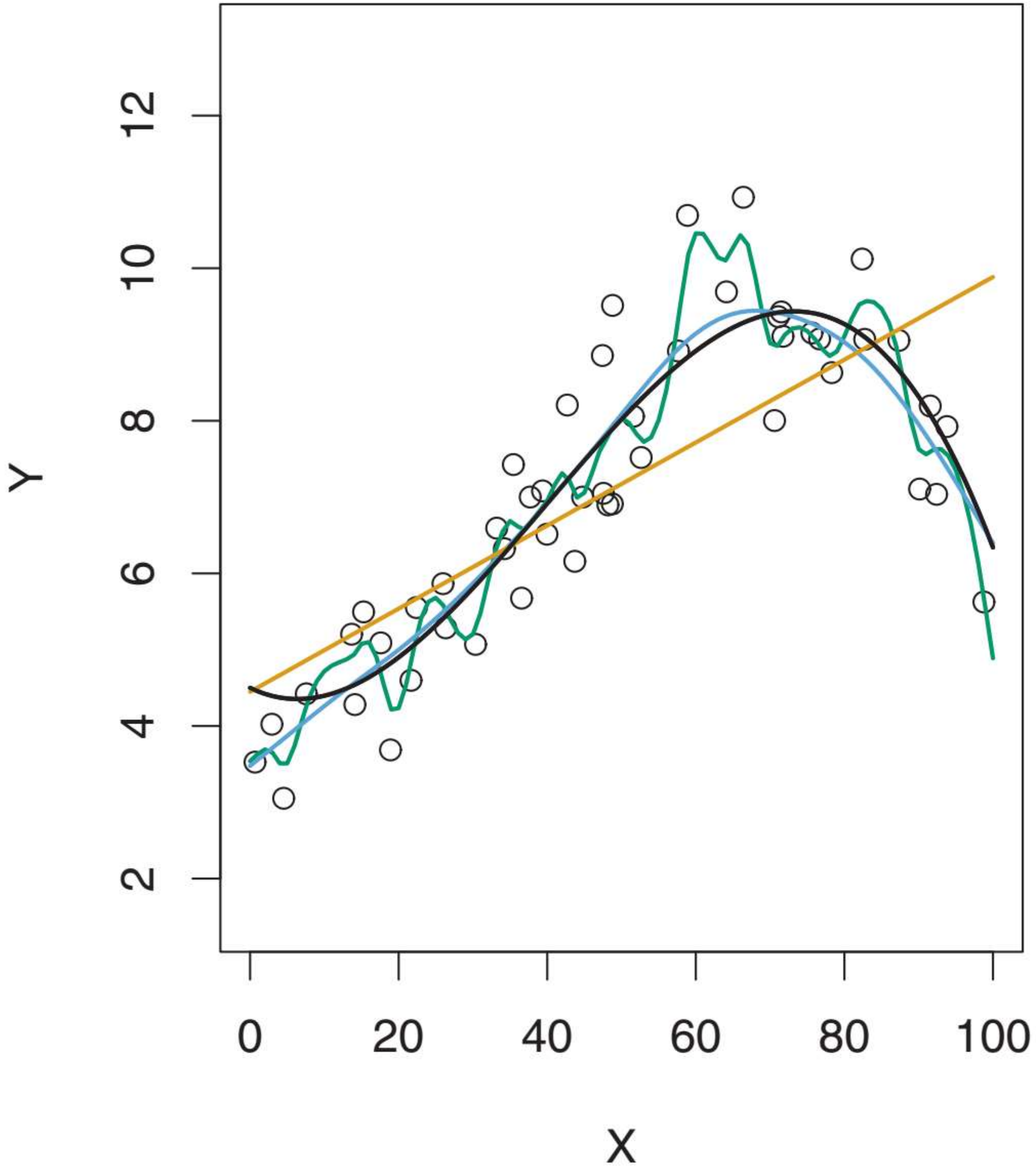
  - However, the problem is not reduced to a smaller number of parameters.

  - Such methods **require a large number of observations** in order to get an accurate estimate of f.

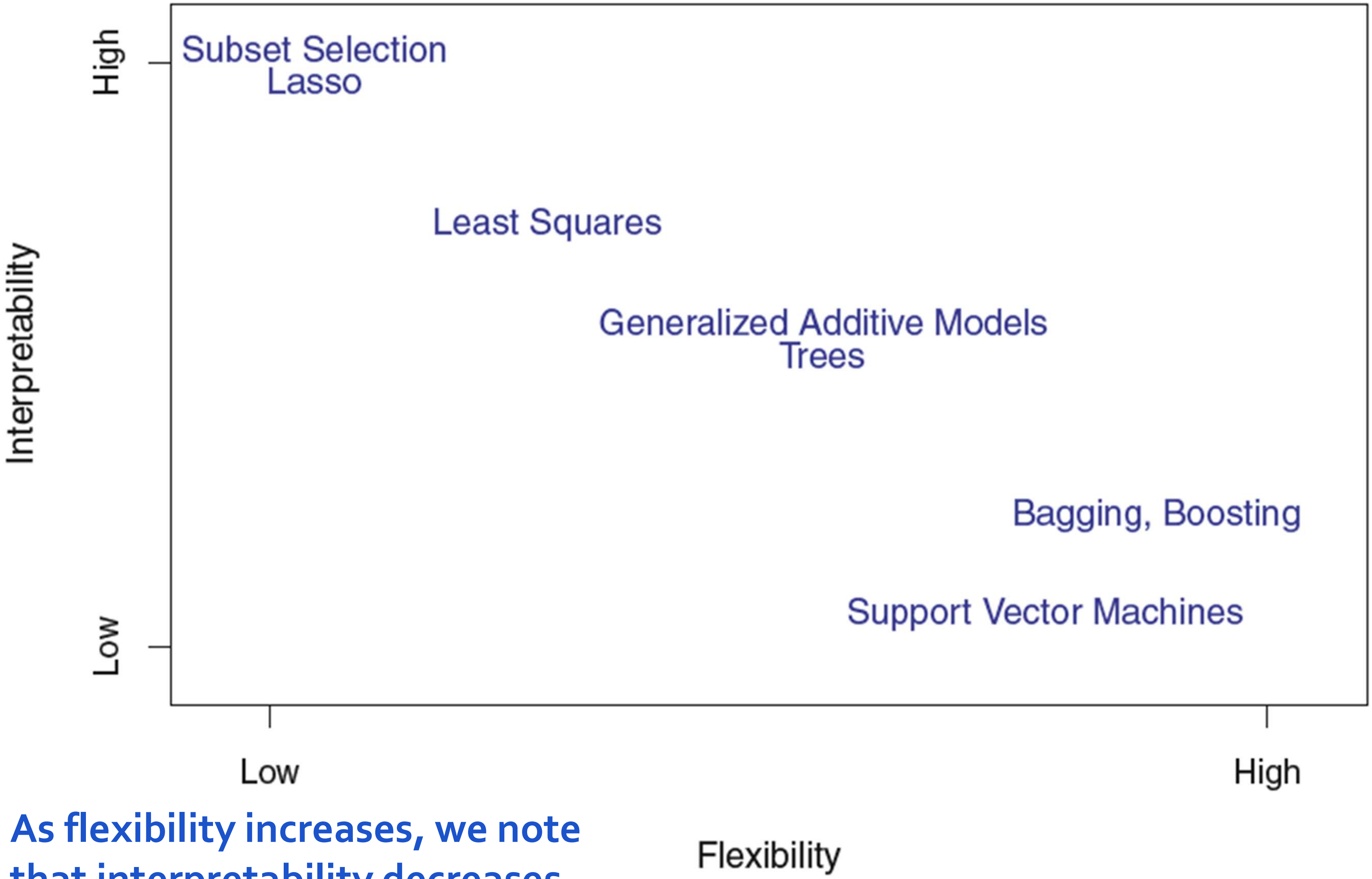


using a smooth thin-plate spline fit

Flexibility



Flexibility vs Interpretability



As flexibility increases, we note that interpretability decreases…

There is no "best" statistical method in general! Deciding on the statistical approach is the most challenging part in practice.

Why would we choose a more restrictive method over a very flexible approach?

- If we are looking for **inference**, i.e., we are interested in **knowing relationships between $\boldsymbol{Y}$ and $\boldsymbol{X_1, X_2, \ldots, X_p}$** then a **linear model** may be good enough.
  - It is restrictive but easy to understand.
  - **More flexible** approaches generate **complicated estimates of $\boldsymbol{f}$**, which makes it less straightforward to understand the relationship between predictor(s) and response.
- However, **when we want to predict output values**, we are typically less interested in interpretability and **more focused on model accuracy**.

Assessing model accuracy

- Measuring the **quality of fit**
  - **How well the predictions match the observed data?**
  - **How close (in quantity) is our predicted response** for a certain observation **to the actual true response of that observation?**

- For regression analysis, the **mean squared error (*MSE*)** is widely used.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

the prediction that $\hat{f}$ gives for the i^{th} observation

The smaller MSE is, the closer the predicted value is to the true one.

What we would really want to know is how well our model performs on previously unseen data (test data), rather than the training data. Why?

Example

- We would like our model to **accurately predict future** (tomorrow, next month,..). e.g.,
  - stock prices based on previous stock returns.

- Or, for instance, predicting the risk of a new patient to develop a certain disease based on his/her clinical measurements.

Mathematically

- **Training** observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ -> estimate $\hat{f}$

- If we compute $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$ and those values are approximately equal to $y_1, y_2, \ldots, y_n$, then the training MSE is small.

- What we are really interested in is **how well the model performs on unseen data**, e.g., whether $\hat{f}(x_0) \approx y_0$ for observation (x_0, y_0) which was **not part of the training data**.

- We are seeking the method that gives the lowest **test $\boldsymbol{MSE}$**.

- If we had a large number of testing observations, then we can compute their corresponding **average squared prediction error**:

$$\mathrm{Ave}(y_0 - \hat{f}(x_0))^2$$

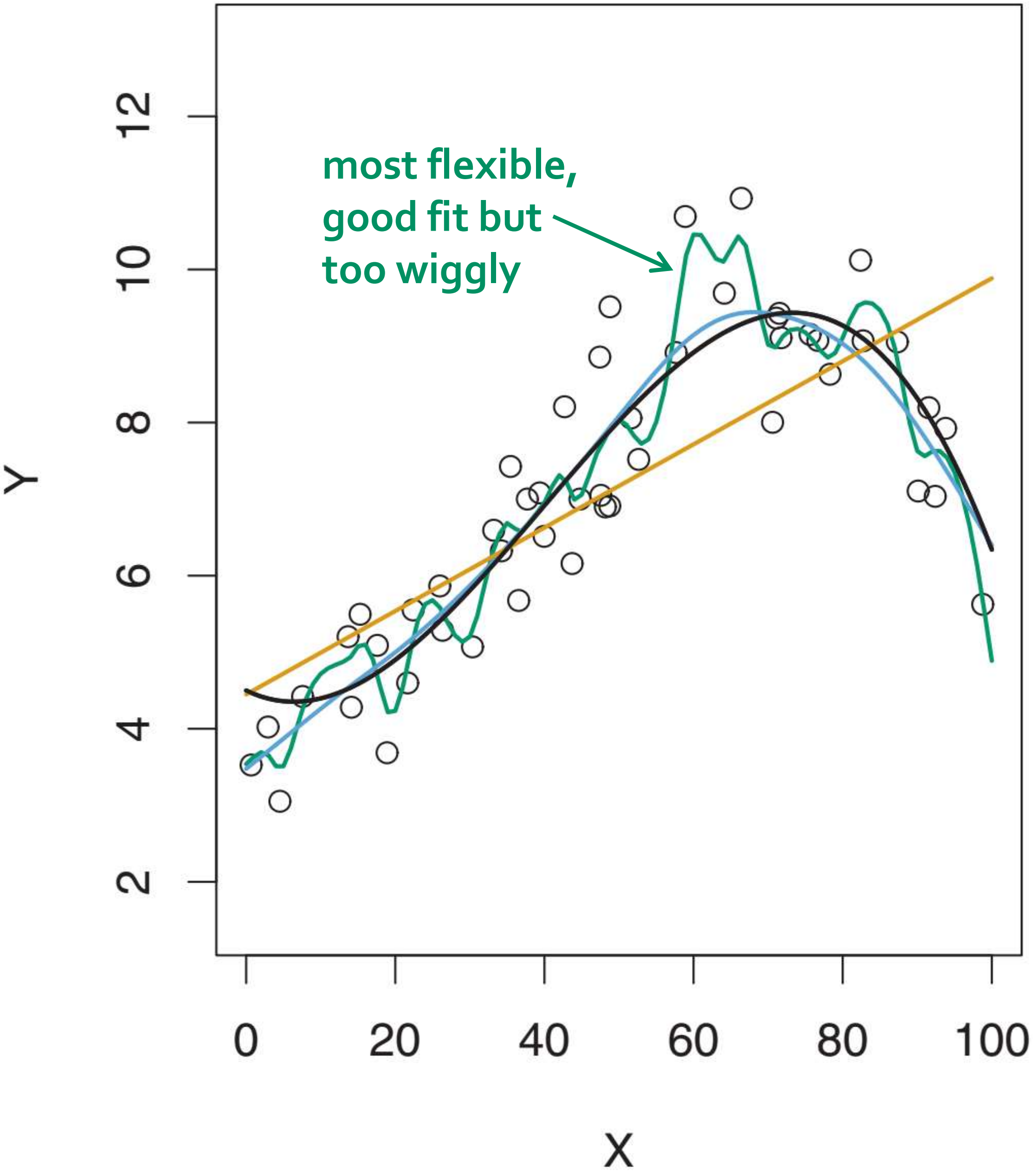
we want to minimize this measure as much as possible!

Mathematically

- **If we have a test dataset,**
  - We can repeatedly compute this measure based on different models and choose the one that gives the smallest value.

- **What if we don’t?**
  - Do we choose the method that gives the smallest training MSE? – no guarantee!

Why?

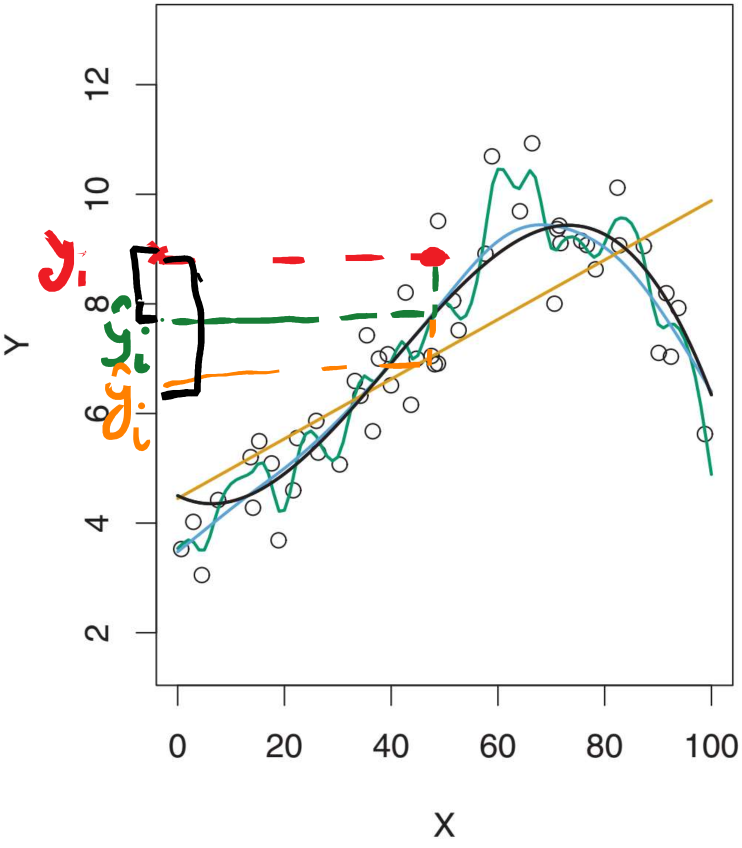


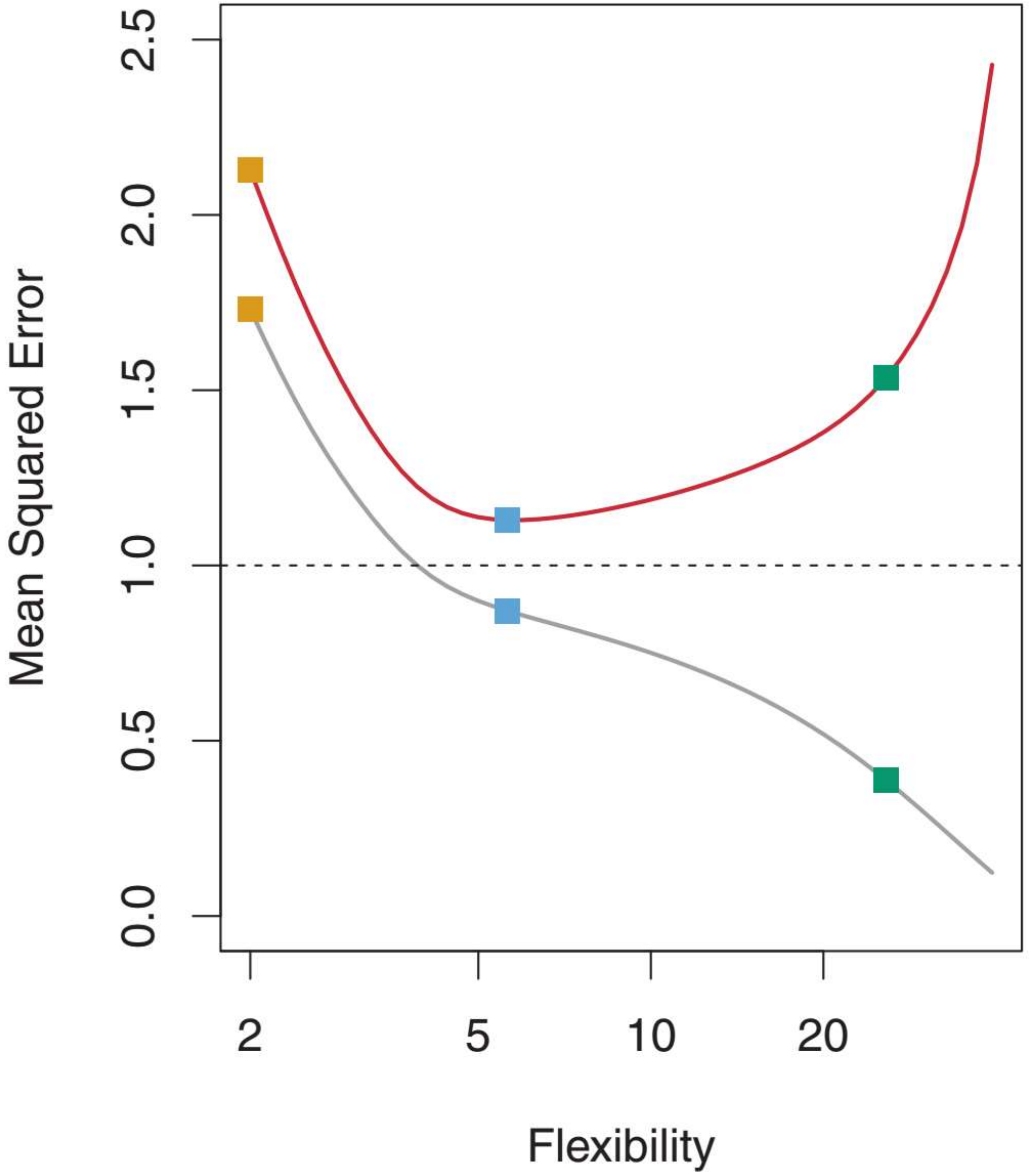
- Suppose that the **black** line represents the **true *f*** of some observations.

- The **orange**, **blue**, and **green** curves represent **estimates of *f*** obtained using different methods – linear regression and smoothing splines with different smoothness levels.

  - different levels of flexibility
  - **the higher the flexibility, the fitter the curve**.

Mathematically





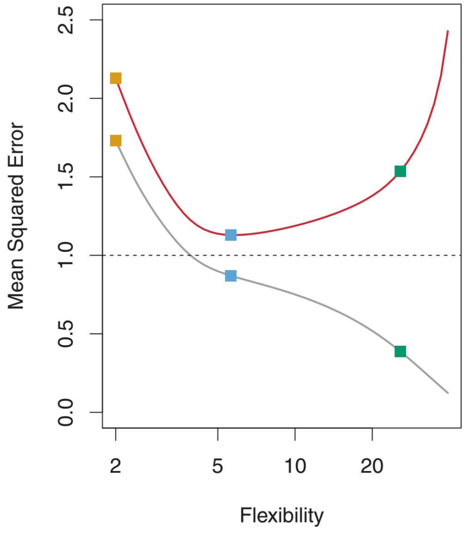
- **Training *MSE*** decreases when flexibility increases.
- Dashed line corresponds to the lowest achievable test MSE among all possible methods.

Mathematically

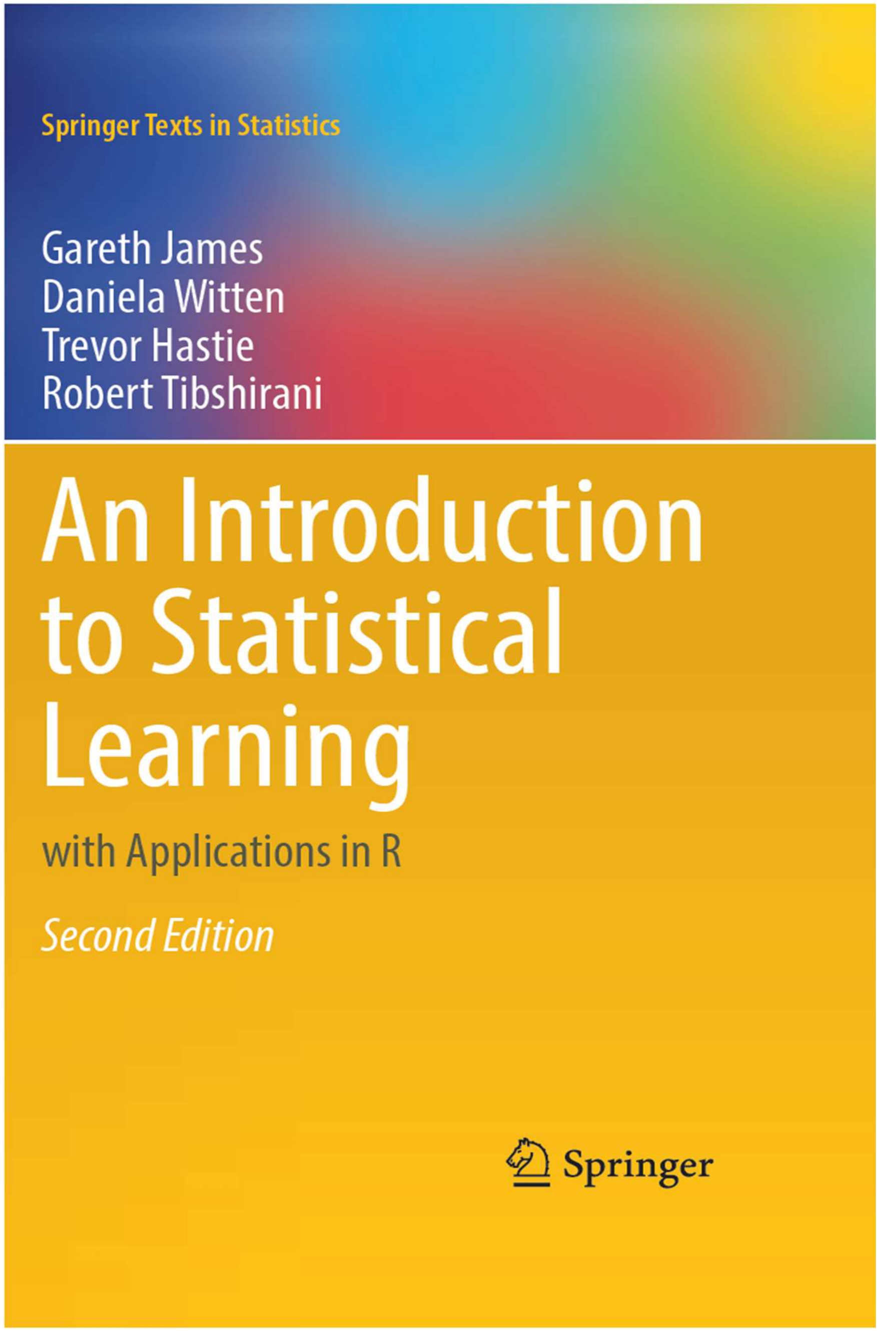
- For the test $\boldsymbol{MSE}$, we note that it initially has the same behavior as the training MSE, i.e., it **decreases as the flexibility increases**.

- However, it **levels off** at some point and **then increases** – the blue and green curves fall within this increase.

- Given that the **dashed line corresponds to Var(∈)** (the irreducible error), we can deduce that the **blue curve** is the closest to optimal.



Reference



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Second Edition

Springer