Big Data for Public Policy

Statistical Learning [Part 1]

Malka Guillot

ETH Zürich | 860-0033-00L

Table of contents

1. What is statistical learning?
2. Inference vs. predictions
3. Epilogue

Prologue

Today

- What is statistical learning?
- Statistics in social science – causality.
- Statistics in machine learning – prediction.

Next week

- Estimating f.
- Accuracy v. interpretability.
- Model accuracy.
- The bias-variance tradeoff.
- Classification

What is statistical learning?

JWHT chap 1. & 2.1

Setting

- Input variables $\mathcal{X}$
  - AKA features, independent variables, predictors
- Output variables $\mathcal{Y}$
  - AKA dependent variables, outcomes, etc.

Statistical learning theory

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

$$\mathcal{X} \in \mathbb{R}^{n \times p}, \mathcal{Y} \in \mathbb{R}^{p}$$

SL= approaches for finding a function that accurately maps the inputs $\mathcal{X}$ to outputs $\mathcal{Y}$

Statistical model

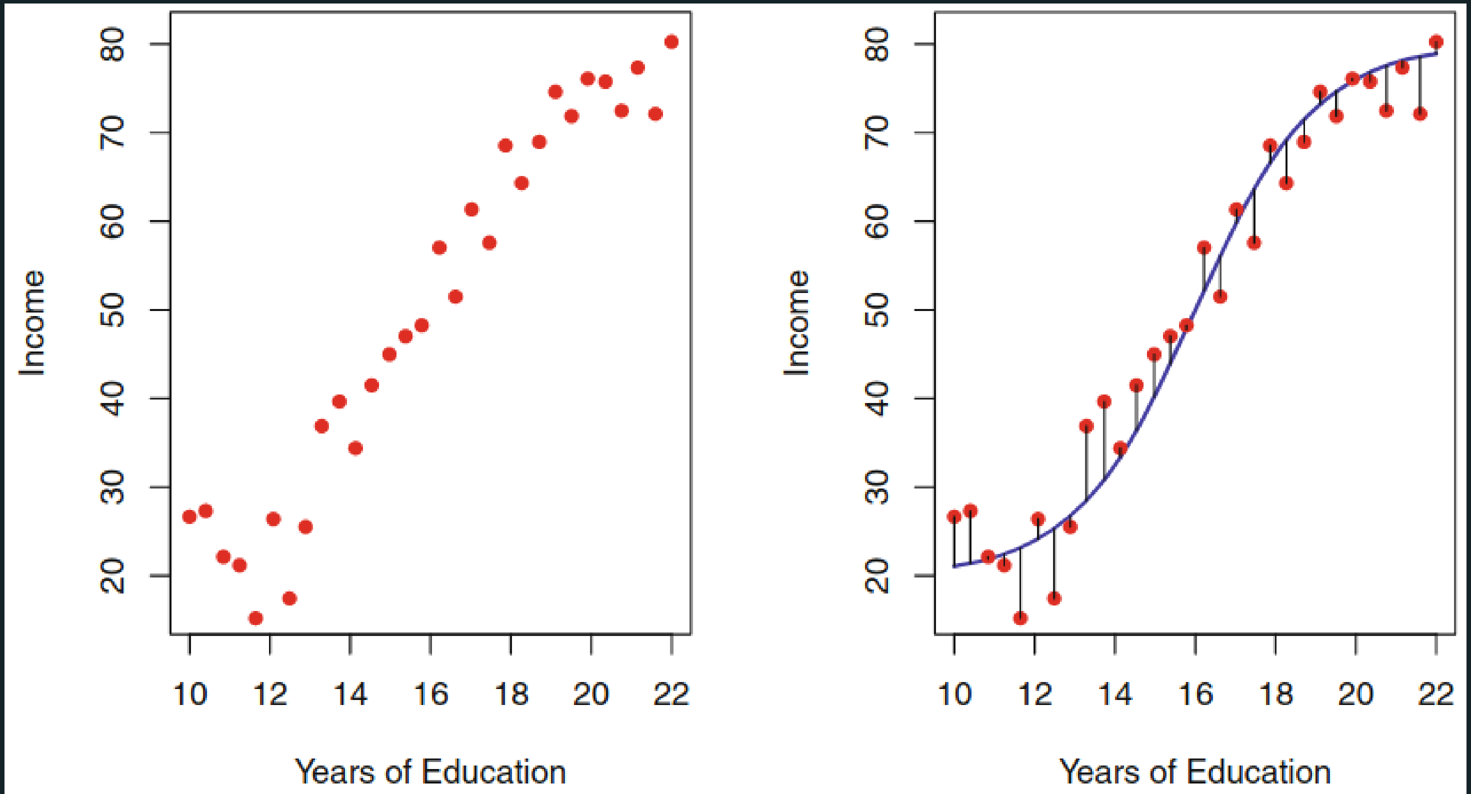
Concretely, finding $f(.)$ s.t.

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

- $f(X)$ is an unknown function of a matrix of predictors $X = (X_1, \cdots, X_p)$,
- Y: a scalar outcome variable
- an error term ϵ with mean zero.
- While X and Y are known, $f(\cdot)$ is unknown.

Goal of statistical learning: to utilize a set of approaches to estimate the “best” $f(\cdot)$ for the problem at hand.

Example: income as a function of education



Why estimate $f(X)$?

Prediction

- Predict Y by $\hat{Y} = \hat{f}(X)$
- When do we care about "pure prediction"?
  - X readily available but Y is not
- $\hat{f}$ can be a **block box**:
  - the only concern is accuracy of the prediction

Inference

- Understanding the way that Y is affected as $X_1, \ldots, X_p$ change
  - Which predictors are associated with the response?
  - What is the relationship between the response and each predictor?

$\Rightarrow \hat{f}$ is cannot be a **black box** anymore

Approach in social science

- Objective: Understanding the way that Y is affected as $X_1, \ldots, X_p$ change
- The goal not necessarily to make predictions for Y
- Often linear function to estimate Y: $f(X) = \sum_{i=1}^{p} \beta_i x_i$
- Assume $\epsilon \sim N(0, \sigma^2)$
- Parameters β are estimated by minimizing the sum of squared errors

$$Y = \sum^{p} \beta_i x_i + \epsilon$$

Approach in social science: causality

$$Y = \beta_0 + \beta_1 T + \sum_{i=1}^{p-1} \beta_i x_i + \epsilon$$

- Interested in the values of one or two parameters and whether they are **causal** or not.
- Framework to interpret statistical causality: **Rubin (1974)**
- β_1 measures the extent to which ΔX_t will affect ΔY_{t+1}

Approach in social science: causality

- Causal inference requires that $T \perp \epsilon$ or $T|X \perp \epsilon$

$\rightarrow$ can be achieved through randomization of T

- This implies that we are not really all that interested in choosing an optimal $f(.)$

Approach in machine learning: prediction

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

- Objectives:
  - find the "best" $f(\cdot)$ and the "best" set of X's which give the best predictions,$\hat{Y}$
  - **Accuracy**: find the function that **minimize the difference between *predicted* and *observed* values**

Reducible and irreducible error

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

$f(X) + \epsilon = \hat{Y}$ true function

- **Reducible error**: $\hat{f}$ is used to estimate f, but not perfect $\rightarrow$ accuracy can be improved by adding more features
- **Irreducible error**: ϵ = all other features that can be used to predict f $\rightarrow$ unobserved $\rightarrow$ irreducible

Reducible and irreducible error

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

$\Rightarrow$ **Objective:** estimating f with the aim of minimizing the reducible error

How do we estimate f?

Context

We use observations to "teach" our ML algorithm to predict outcomes

- **Training data**: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$
- Goal: use the training data to estimate the unknown function f
- 2 types of SL methods: **parameteric vs. nonparametric**

Parametric methods

Model-based approaches, 2 steps:

1. Specify a ***parametric*** **(functional) form** for $f(X)$, e.g. linear:

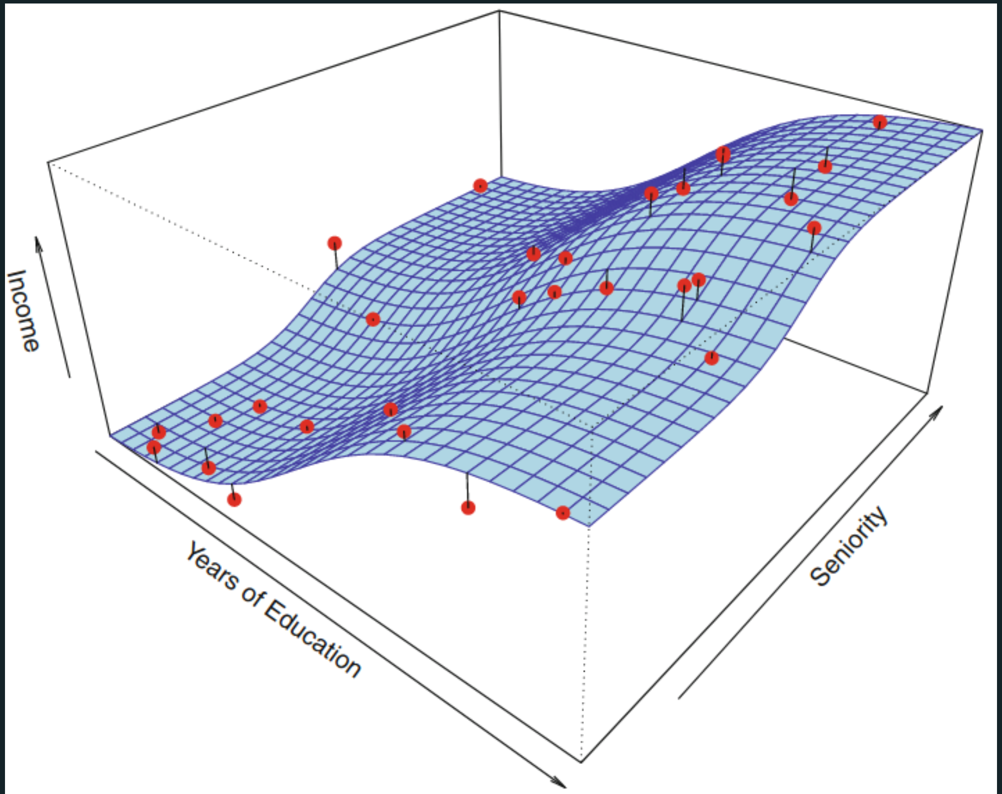
   $$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

   (Parametric means that the function depends on a finitenumber of parameters, here $p+1$).

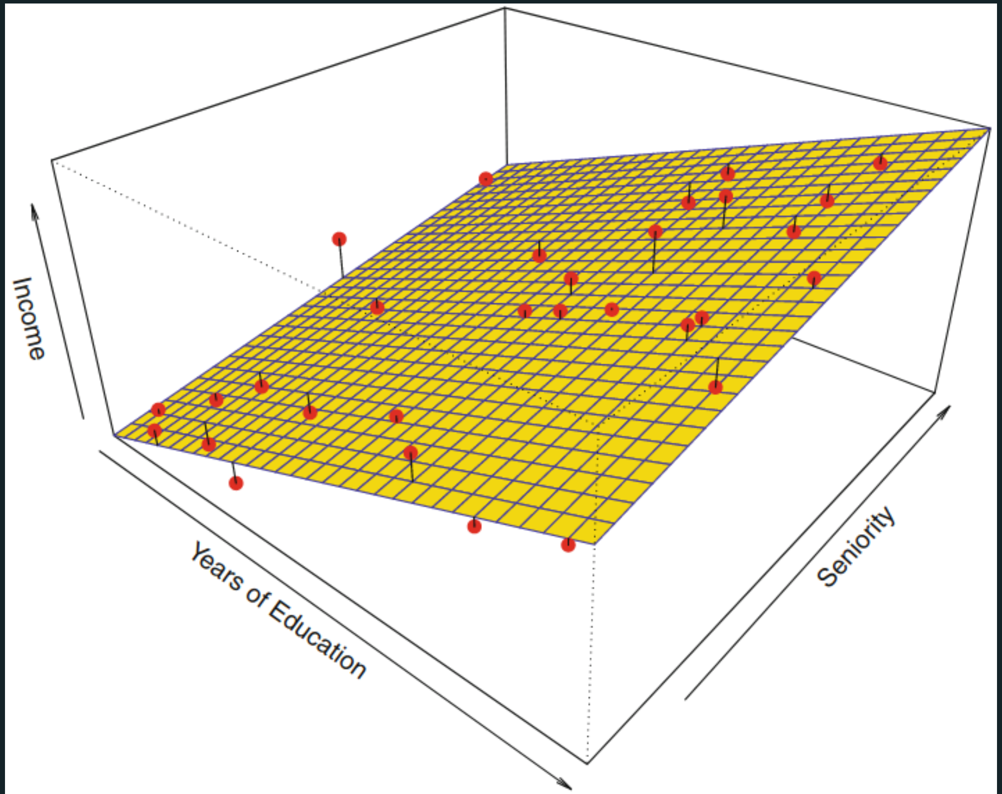
2. **Training**: Estimate the parameters by OLS and predict Y by

   $$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

True function



Linear estimate



Parametric methods -- issues

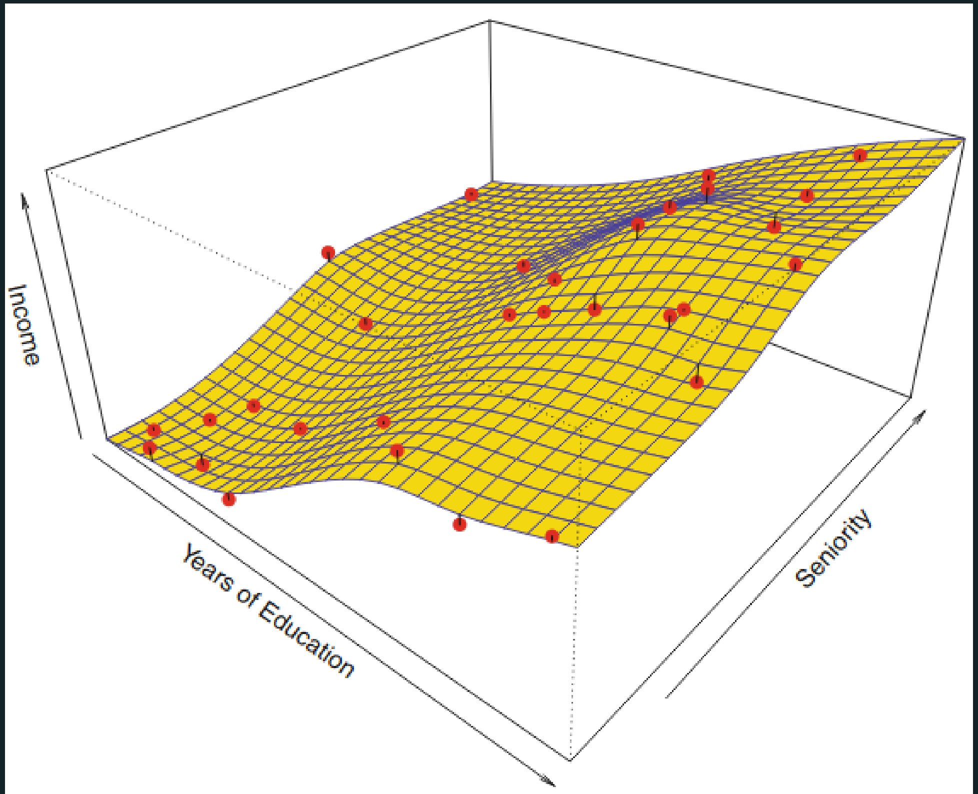
Misspecification of $f(X)$

1. Rigid models (e.g. strictly linear) may not fit the data well
2. More flexible models require more parameter estimation $\rightarrow$ **overfitting**

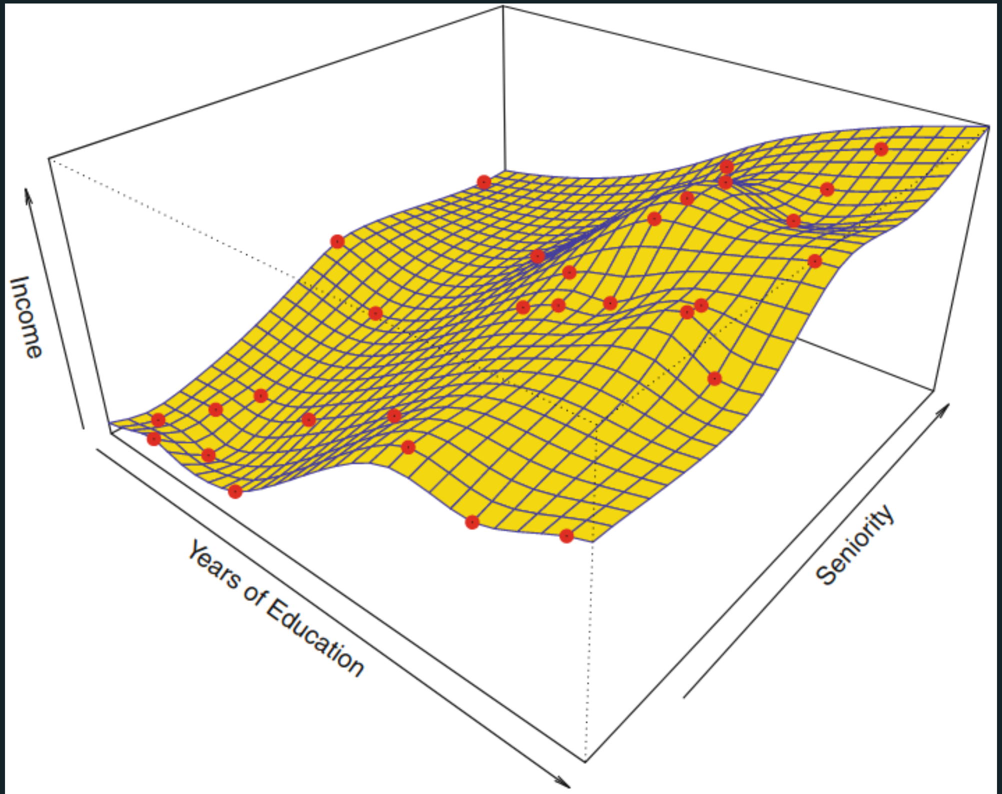
Non-parametric methods

- **No assumptions** about the functional form of f
- Estimates a function only **based on the data itself**.
- **Disadvantage**: very large number of observations is required to obtain an accurate estimate of f

“Smooth” nonlinear estimate



Rough nonlinear estimate with perfect fit $\Rightarrow$ overfit



Accuracy and interpretability tradeoffs

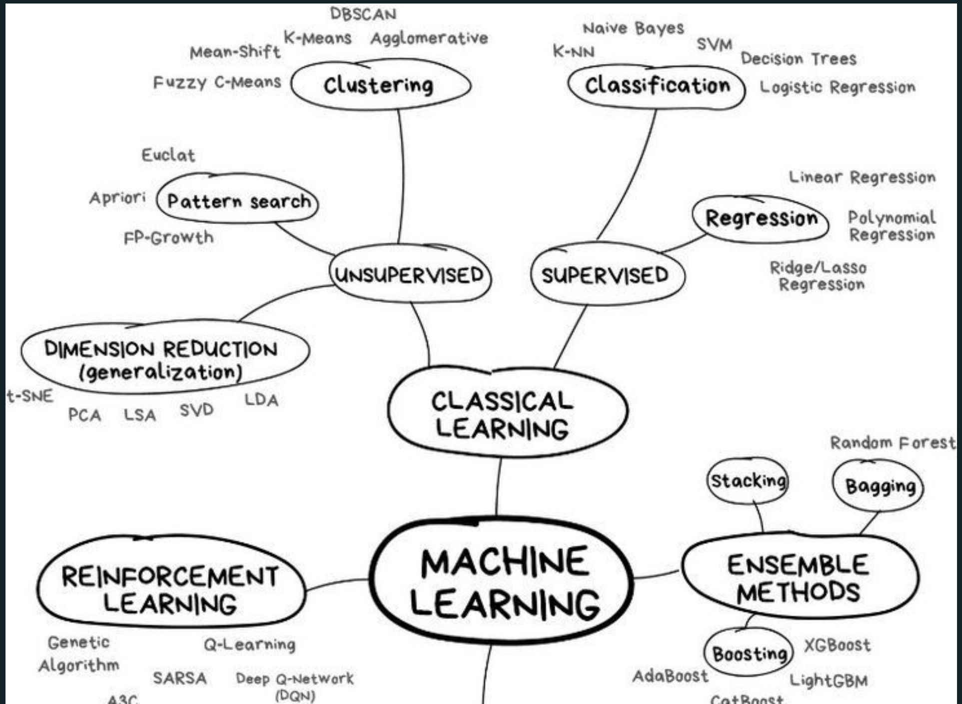
- **More accurate** models often require estimating **more parameters** and/or having more flexible models
- Models that are better at prediction generally are **less interpretable**.
- For inference, we care about interpretability.

$\longrightarrow$ More on this next week!

Supervised vs. unsupervised learning

- **Supervised learning** involves estimating functions with known observation and outcome data.
- **Unsupervised learning** involves estimating functions without the aid of outcome data.

The Machine learning landscape



Conclusion:

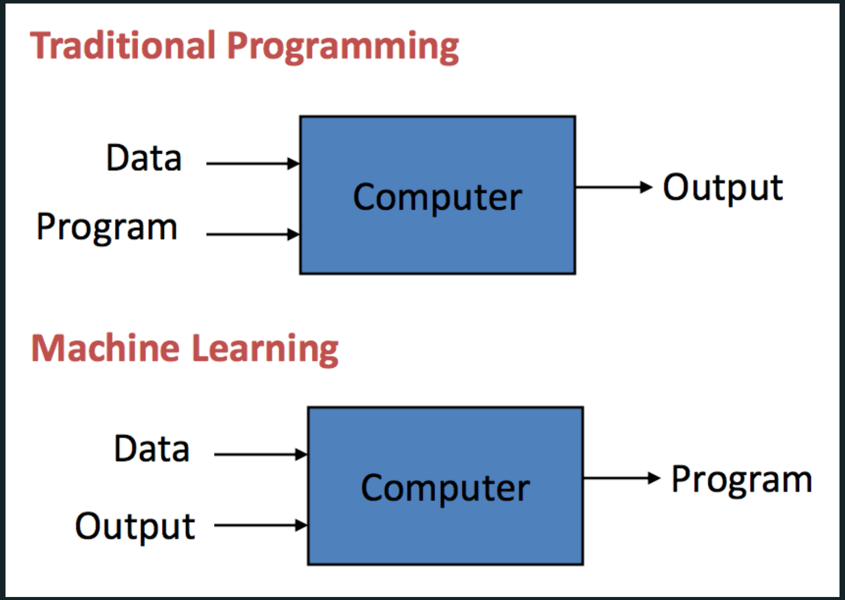
Econometrics vs. Machine Learning

Econometrics vs. Machine Learning (1)

- **Common objective**: to build a predictive model, for a variable of interest, using explanatory variables (or features)
- **Different cultures**:
  - *E*: probabilistic models designed to describe economic phenomena
  - *ML*: algorithms capable of learning from their mistakes

Charpentier A., Flachaire, E. & Ly, A. (2018). Econometrics and Machine Learning. *Economics and Statistics*, 505-506, 147–169.

Econometrics vs. Machine Learning (2)



- **Classical computer programming:** humans input the rules and the

Researcher vs. policy analysist

- The frontier can be thin
- I will sometimes be speaking from the point of view of an economist, but:
  - The model-based vs. algorithm-based problematics transfers to other social sciences
  - I try to cover a wide range of topics in the literature
  - You are welcome to propose relevant papers
- All aim at *using data to solve problems*