

# Introduction to learning, multiple and nonparametric regression


Machine Learning

Jonas Striaukas



# Course details

Basic info:

- **My email:** [js.fi@cbs.dk](mailto:js.fi@cbs.dk) or [jonas.striaukas@gmail.com](mailto:jonas.striaukas@gmail.com)
- **Lecture time:** TBA
- **Auditorium:** TBA
- **Office hours:** TBA
- **Course website:** [https://jstriaukas.github.io/ml\\_course](https://jstriaukas.github.io/ml_course) 

Exam:

- **Structure:** TBA
- **When:** TBA

What I expect from you:

- ▶ Understand the concepts we learn in the class. In particular derivations of some simple theoretical results as well as full understanding of more complex theory.
- ▶ Be creative, active during class presentations and work hard! And try **not** to miss classes...

# Machine learning, computing, etc.

“The purpose of computing is **insight**, not numbers.”

*Richard Hamming*

# Topics of the course

- Introduction to learning, multiple and nonparametric regression
  - ▶ BLAH BLAH
- High-dimensional linear regression
  - ▶ BLAH BLAH
- High-dimensional regression properties and generalized linear models (GLMs)
  - ▶ BLAH BLAH
- Prediction, loss functions and M-estimators
  - ▶ BLAH BLAH
- Introduction to deep learning
  - ▶ BLAH BLAH
- Introduction to causal machine learning
  - ▶ BLAH BLAH

# Learning, multiple and nonparametric regression

## Big data

Nowadays, Big Data are ubiquitous: from the internet, biology and medicine to government, business, economics, finance, ...

Some quotes:

- *“There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days”*, according to Eric Schmidt, the CEO of Google, in 2010.
- *“Big data is not about the data”*, according to Gary King of Harvard University.

Do we need ML or even AI to understand economics and/or finance data?

- ▶ **Yes!** ML is not that different from classical econometrics... “Black-box” deep learning is not that black box after all...

# Learning, multiple and nonparametric regression

## Big data – examples

Big data examples in economics and finance:

- ▶ high-frequency financial assets data (e.g., stocks, bonds, fx, derivatives, ...);
- ▶ large panels of economic data (e.g., 131 macroeconomics time series [FRED MD](#) database with monthly updates, [McCracken and Ng \(2016\)](#));
- ▶ spatial data (e.g., state-level data in US, euro area data);
- ▶ text-based data (e.g., newspaper articles, [GDELT project](#); [EC news data](#));
- ▶ ... .

# Learning, multiple and nonparametric regression

## Impact of Big data & dimensionality

Problems associated with Big data:

- Data are collected from various sources and populations  $\implies$  **heterogeneity**;
- typically large numbers of variables are collected  $\implies$  some variables are **heavy-tailed**, i.e. have high kurtosis which is much higher than the normal distribution;
- incidental **endogeneity** due to high-dimensionality  $\implies$  huge impact on model selection and statistical inference (**Fan and Liao, 2014**);
- computation/optimization of model parameters  $\implies$  **convexity** so far is a way out to guarantee the stability of solutions;
- **noise accumulation** and **spurious correlation** has a large impact on model selection  $\implies$  high-dimensional statistics methods.

For curious students: see **Fan, Han, and Liu (2014)** for an overview of how these features impacts the developments of big data analysis techniques.

# Learning, multiple and nonparametric regression

## Spurious correlations – examples



# Learning, multiple and nonparametric regression

Spurious correlations – some explanation

# Learning, multiple and nonparametric regression

## Statistical learning theory

The main goals of high dimensional inferences are (see Bickel, 2008):

- **Prediction:** to construct a method as effective as possible to predict future observations and;
- **(Causal) inference:** to gain insight into the relationship between features and responses for scientific purposes, as well as, hopefully, to construct an improved prediction method useful for (economic) policy.

# Multiple linear regression

## Statistical learning theory

Consider a multiple linear regression model:

$$Y = \sum_{j \in [p]} \beta_j X_j + \varepsilon, \quad (1)$$

where

- $Y$  – response or dependent variable;
- $X_j$  – variables are often called explanatory variables or covariates or independent variables;
- intercept term can be included in the model by including 1 as one of the covariates –  $\mathbf{X}_1 = 1$ ;
- $\beta_j$  – regression coefficients;
- $\varepsilon$  is the error term, some assumptions:
  - “random error”  $\varepsilon$  is often assumed has zero mean;
  - uncorrelated with covariates  $X$ , which is referred to as *exogenous* variables.

# Multiple linear regression

## Statistical learning theory

Given observed sample  $\{X_{ij}, Y_i : i \in [n], j \in [p]\}$ , where  $[p] \triangleq \{1, \dots, p\}$ , we have

$$Y_i = \sum_{j \in [p]} \beta_j X_{ij} + \varepsilon_i. \quad (2)$$

Classical estimator used to fit the model (dates back to Gauss and Legendre in 19<sup>th</sup> century): **least squares**.

Construct residuals:

$$r_i = Y_i - \sum_{j \in [p]} \beta_j X_{ij}. \quad (3)$$

Under classical assumptions, the least squares solves for  $\beta = (\beta_1, \dots, \beta_p)^\top$  by minimizing:

$$\arg \min_{\beta \in \mathbb{R}^p} \sum_{i \in [n]} r_i^2 = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i \in [n]} (Y_i - \sum_{j \in [p]} \beta_j X_{ij})^2$$

# Multiple linear regression

## Notation

Denote by:

- $\mathbf{y} = (Y_1, \dots, Y_n)^\top$  – response vector;
- $\mathbf{X}_j = (X_{1j}, \dots, X_{nj})^\top$  – covariate  $j$  vector;
- $\mathbf{X} = (\mathbf{X}_1^\top, \dots, \mathbf{X}_p^\top)$  – covariate matrix;
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$  – regression coefficient vector;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$  – regression error term vector.

Our linear model can be written in a matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (4)$$

- FAN, J., F. HAN, AND H. LIU (2014): “Challenges of big data analysis,” National science review, 1(2), 293–314.
- FAN, J., AND Y. LIAO (2014): “Endogeneity in high dimensions,” Annals of statistics, 42(3), 872.
- MCCRACKEN, M. W., AND S. NG (2016): “FRED-MD: A monthly database for macroeconomic research,” Journal of Business & Economic Statistics, 34(4), 574–589.