Econometrics & "Machine Learning"

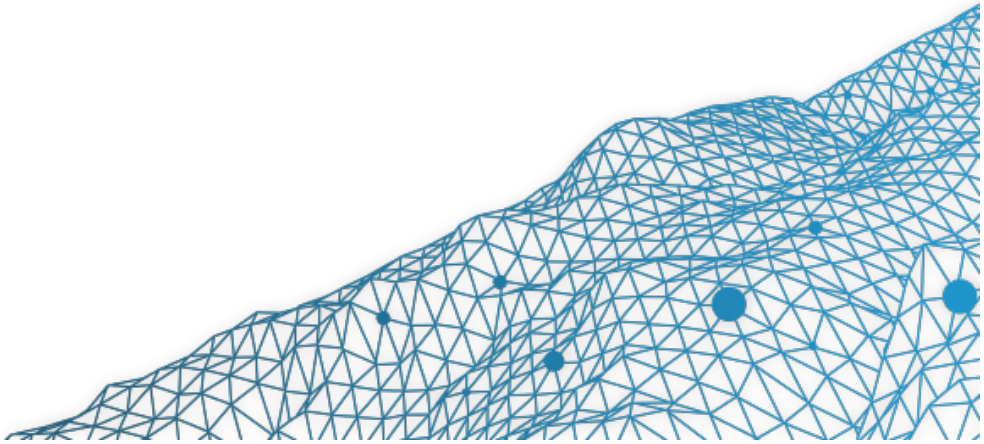
A. Charpentier (Université de Rennes)

joint work with E. Flachaire (AMSE) & A. Ly (Université Paris-Est)

https://hal.archives-ouvertes.fr/hal-01568851/

Università degli studi dell'Insubria

Seminar, May 2018.



Econometrics & "Machine Learning"

Varian (2014, The Probabilistic Approach in Econometrics)

Google

Definitions

Machine learning, data mining, predictive analytics, etc. all use data to predict some variable as a function of other variables.

- May or may not care about insight, importance, patterns
- May or may not care about *inference*---how y changes as some x changes

Econometrics: Use statistical methods for prediction, inference, *causal* modeling of economic relationships.

- Hope for some sort of insight, inference is a goal
- In particular, *causal* inference is goal for decision making

Econometrics in a "Big Data" Context

Here data means $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{X}$ a $n \times p$ matrix.

n large means " asymptotic" theorems can be invoked $(n \to \infty)$

Portnoy (1988, Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity) proved that MLE estimator is asymptotically Gaussian as long as $p^2/n \to 0$.

There might be High-dimensional issues if $p > \sqrt{n}$.

Nevertheless, there might be sparsity issues in high dimension (see Hastie, Tibshirani & Wainwright (2015, Statistical Learning with Sparsity)) : a sparse statistical model has only a small number of nonzero parameters or weights

First order condition $\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) = \boldsymbol{0}$ is based on QR decomposition (can be computationaly intensive).

Econometrics in a "Big Data" Context

In multiple regression, it is shown that least square parameter estimates can be unsatisfactory if the prediction vectors are not orthogonal. Proposed is a procedure based on adding small positive quantities to the diagonals of the normal equations to obtain estimates with smaller mean square error. [The *Science Citation Index*® (*SCI*®) and the *Social Sciences Citation Index*® (*SSCI*®) indicate that this paper has been cited in over 310 publications since 1970.]

It would be great to report that we had profound discussions on the foundations of statistics, but such was not the case. Much of the time was spent trying to find ways to do regression computations economically and to come up with solutions that made engineering sense. We were charging $90/day for our time, but had to charge $450/hour for computer time on a Univac I that had 1,000 words of memory. With this machine, it took 75 processing minutes to invert a 40 × 40 matrix through a 4 × 4 partition of 10 × 10 submatrices, using magnetic tapes for temporary storage.

This Week's Citation Classic CC/NUMBER 35 AUGUST 30, 1982

Hoerl A E & Kennard R W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12:55-67, 1970.
[University of Delaware, Newark, and E.I. du Pont de Nemours & Co., Wilmington, DE]

"In these discussions, we found that we had both encountered the same phenomenon, one that had caused some embarrassment with clients. We found that multiple linear regression coefficients computed using least squares didn't always make sense when put into the context of the process generating the data. The coefficients tended to be too large in absolute value, some would even have the wrong sign, and they could be unstable with very small changes in the data.

"Since the method proposed attacked one of the sacred cows of linear regression—least squares—there was considerable resistance. However, the solid theoretical basis and the practical usefulness of the method gradually overcame most objections.

Arthur Hoerl in 1982, back on Hoerl & Kennard (1970) on Ridge regression.

Back on the history of the "regression"

Galton (1870, Heriditary Genius, 1886, Regression towards mediocrity in hereditary stature) and Pearson & Lee (1896, On Telegony in Man, 1903 On the Laws of Inheritance in Man) studied genetic transmission of characterisitcs, e.g. the heigth.

On average the child of tall parents is taller than other children, but less than his parents.

"*I have called this peculiarity by the name of regression*", Francis Galton, 1886.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

Height of the mid-parent in inches	Height of the adult child													Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7		
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	5	5	10	4	9	2	2	45	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1886a).

Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties; he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

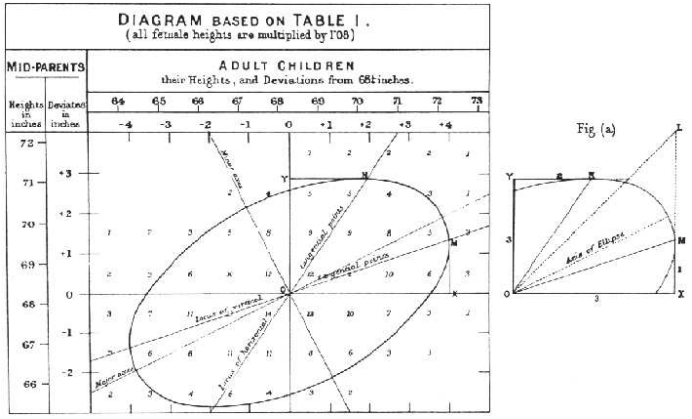
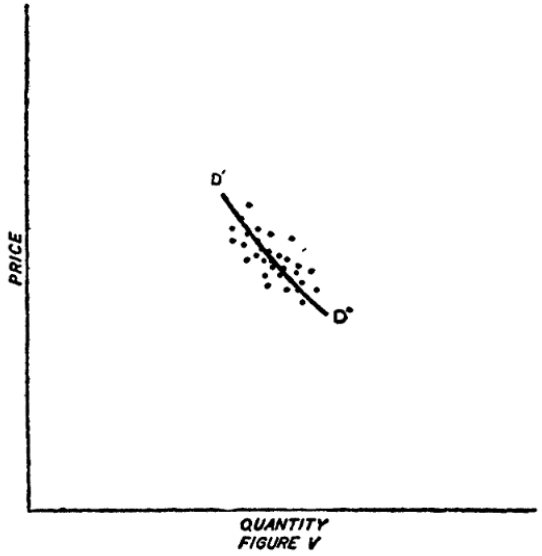


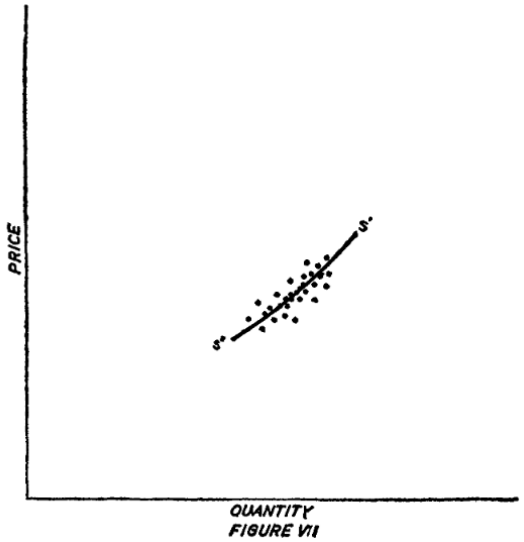
Figure 8.7. Galton's smoothed rendition of Table 8.1, with one of the "concentric and similar ellipses" drawn in. The geometric relationship of the two regression lines to the ellipse is also shown. (From Galton, 1886a.)

Economics and Statistics

Working (1925)

What Do Statistical "Demand Curves" Show?





WHAT DO STATISTICAL "DEMAND CURVES" SHOW?[1]

SUMMARY

How statistical demand curves are constructed, 213. — The theory of the demand-and-supply curve analysis applied to a period of time, 217. — Statistical curves which would result under hypothetical conditions, 218. — Data used do not necessarily reflect influence of demand more than of supply, 222. — Whether fitted curve approximates a demand or supply curve depends on the relative *variability* of demand and supply, 224. — Slope of the fitted curve may not correspond to the true demand curve, 225. — In what sense may statistical demand curves be "general" demand curves? 228. — Distinction between consumer and dealer demand, 230. — Fitted curves are "static" in the sense of showing an "average" relationship, or relationship at a "typical" point of time, 231. — Do statistical demand curves assume all other things equal? 233. — Conclusions, 234.

E. J. Working.

University of Minnesota, Minneapolis

Tinbergen (1939)

Statistical Testing of Business Cycle Theories

STATISTICAL TESTING
OF
BUSINESS-CYCLE THEORIES
A METHOD
and its Application to
INVESTMENT ACTIVITY
BY
J. TINBERGEN

Econometrics and the Probabilistic Model

Haavelno (1944, The Probabilistic Approach in Econometrics)

THE PROBABILITY APPROACH
IN ECONOMETRICS

By

Trygve Haavelmo
research associate
cowles commission for
research in economics

Supplement to Econometrica, Volume 12, July, 1944

THE ECONOMETRIC SOCIETY
The University of Chicago
Chicago 37, Illinois

CHAPTER III

Stochastical Schemes as a Basis for Econometrics

As far as is known, the scheme of probability and random variables is, at least for the time being, the only scheme suitable for formulating such theories. We may have objections to using this scheme, but among these objections there is at least one that can be safely dismissed, viz., the objection that the scheme of probability and random variables is not general enough for application to economic data. Since, however, this is apparently not commonly accepted by economists we find ourselves justified in starting our discussion in this chapter with a brief outline of the modern theory of stochastical variables, with particular emphasis on certain points that seem relevant to economics.

The more recent developments in statistical theory are based upon the so-called modernized classical theory of probability. Here "probability" is defined as an absolutely additive and nonnegative *set-function*,[1] satisfying certain formal properties.[2]

Let us first take an example to illustrate this probability concept.

Data $(y_i, \boldsymbol{x}_i)$ are seen as realizations of (iid) random varibles $(Y, \boldsymbol{X})$ on some probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$.

Mathematical Statistics

The philosophy of the classical parametric paradigm is based on the following three beliefs:

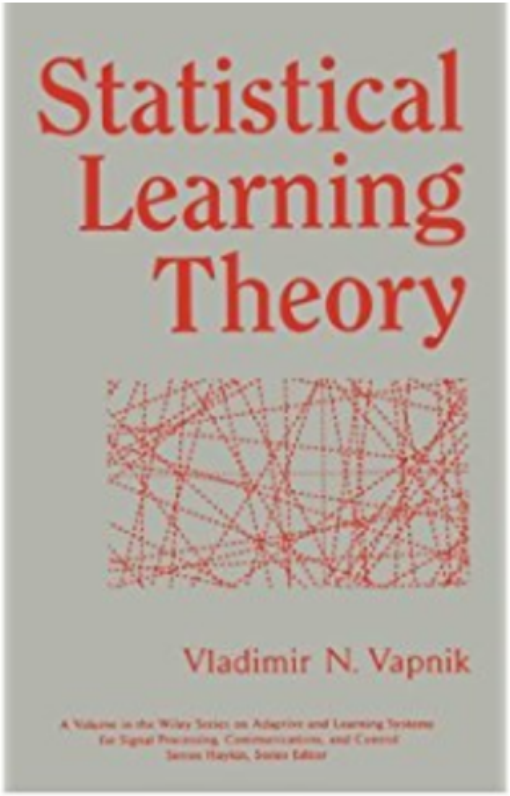
1. *To find a functional dependency from the data, the statistician is able to define a set of functions, linear in their parameters, that contain a good approximation to the desired function. The number of free parameters describing this set is small.*

2. *The statistical law underlying the stochastic component of most real-life problems is the normal law.*

3. *The induction engine in this paradigm—the maximum likelihood method —is a good tool for estimating parameters.*

Note that these three beliefs were also supported by the philosophy:

If there exists a mathematical proof that some method provides an asymptotically optimal solution, then in real life this method will provide a reasonable solution for a small number of data samples.



Vapnik (1998, Statistical Learning Theory)

Mathematical Statistics

Consider observations $\{y_1, \cdots, y_n\}$ from iid random variables $Y_i \sim F_{\boldsymbol{\theta}}$ (with "density" $f_{\boldsymbol{\theta}}$).

Likelihood is

$$\mathcal{L}(\boldsymbol{\theta}\,;\boldsymbol{y}) \mapsto \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(y_i)$$

Maximum likelihood estimate is

$$\widehat{\boldsymbol{\theta}}^{\text{mle}} \in \arg\max_{\boldsymbol{\theta}\in\Theta}\{\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})\}$$

ON AN ABSOLUTE CRITERION FOR FITTING FREQUENCY CURVES.

By *R. A. Fisher*, Gonville and Caius College, Cambridge.

theoretical curve, so the probability of any particular set of θ's is proportional to P, where

$$\log P = \sum_1^n \log f.$$

The most probable set of values for the θ's will make P a maximum.

If a continuous curve is observed—*e.g.*, the period during which a barometer is above any level during the year is a continuous function from which may be derived the relative frequency with which it stands at any height—we should use the expression

$$\log P = \int_{-\infty}^{\infty} y \log f \, dx.$$

Fisher (1912) On an absolute criterion for fitting frequency curves

Mathematical Statistics

Under standard assumptions (Identification of the model, Compactness, Continuity and Dominance), the maximum likelihood estimator is consistent $\widehat{\boldsymbol{\theta}}^{\text{mle}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}$. With additional assumptions, it can be shown that the maximum likelihood estimator converges to a normal distribution

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}^{\text{mle}} - \boldsymbol{\theta}\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1})$$

where I is Fisher information matrix (i.e. $\widehat{\boldsymbol{\theta}}^{\text{mle}}$ is asymptotically efficient).

Eg. if $Y \sim \mathcal{N}(\theta, \sigma^2)$, $\log \mathcal{L} \propto -\sum_{i=1}^{n} \left(\frac{y_i - \theta}{\sigma} \right)^2$, and $\widehat{\theta}^{\text{mle}} = \overline{y}$ (see also method of moments).

$$\max \big\{ \log \mathcal{L} \big\} \Longleftrightarrow \min \Big\{ \sum_{i=1}^{n} \varepsilon_i^2 \Big\} = \text{least squares}$$

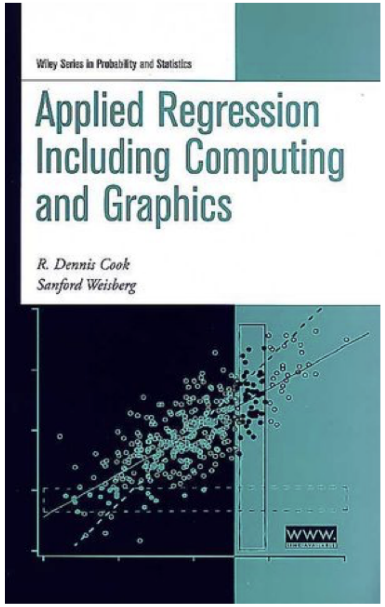
Data and Conditional Distribution

Cook & Weisberg (1999, Applied Regression Including Computing and Graphics)

CHAPTER 2

Introduction to Regression

The primary goal in a regression analysis is to understand, as far as possible with the available data, how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors. Since this is the central idea, it will be helpful to have a convenient way of referring to the response variable restricted to a subpopulation in which the predictor does not vary. We will use the notation $y \mid (x = \tilde{x})$ to indicate the response in the subpopulation where the predictor is fixed at the value $\tilde{x}$. The vertical bar in this notation stands for the word *given*. If the particular



From available data, describe the conditional distribution of Y given $\boldsymbol{X}$.

Conditional Distributions and Likelihood

We had a sample $\{y_1, \cdots, y_n\}$, with Y from a $\mathcal{N}(\theta, \sigma^2)$ distribution.

The natural extension if we had a sample $\{(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)\}$ is to assume that $Y|\boldsymbol{X} = \boldsymbol{x}$ has a $\mathcal{N}(\theta_{\boldsymbol{x}}, \sigma^2)$ distribution.

The standard linear model is obtained when $\theta_{\boldsymbol{x}}$ is linear, $\theta_{\boldsymbol{x}} = \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$. Hence

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta} \text{ and } \text{Var}[Y|\boldsymbol{X} = \boldsymbol{x}] = \sigma^2 \quad (homoskedasticity).$$

(if we center variables or if we add the constant in $\boldsymbol{x}$).

The MLE of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}^{\text{mle}} = \text{argmin}\left\{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i{}^\mathsf{T}\boldsymbol{\beta})^2\right\} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}.$$

From a Gaussian to a Bernoulli (Logistic) Regression, and GLMs

Consider $y \in \{0, 1\}$, so that Y has a Bernoulli distribution. The logarithm of the odds is linear,

$$\log\left(\frac{\mathbb{P}[Y=1|\boldsymbol{X}=\boldsymbol{x}]}{\mathbb{P}[Y\neq 1|\boldsymbol{X}=\boldsymbol{x}]}\right) = \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta},$$

i.e.

$$\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = \frac{e^{\beta_0+\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}}}{1+e^{\beta_0+\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}}} = H(\beta_0+\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}), \quad \text{où} \quad H(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)},$$

with likelihood

$$\mathcal{L}(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X}) = \prod_{i=1}^{n}\left(\frac{e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}}\right)^{1-y_i}$$

and set $\widehat{\boldsymbol{\beta}}^{\text{mle}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\{\mathcal{L}(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X})\}$ (solved numerically).

A Brief Excursion to Nonparametric Econometrics

Two strategies are usually considered to compute $m(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$

- consider a local approximation (in the neighborhood of $\boldsymbol{x}$), e.g. k-nearest neighbor, kernel regression, lowess
- consider a functional decomposition of function m in some natural basis, e.g. splines

For a kernel regression, as in Nadaraya (1964,) and Watson (1964,), consider some kernel function K and some bandwidth h

$$\widehat{m}_h(x) = \boldsymbol{s}_x{}^\intercal \boldsymbol{y} = \sum_{i=1}^{n} s_{x,i} y_i \text{ where } s_{x,i} = \frac{K_h(x - x_i)}{K_h(x - x_1) + \cdots + K_h(x - x_n)}.$$

(in a univariate problem).

A Brief Excursion to Nonparametric Econometrics

Recall - see Simonoff (1996,) that using asymtptotic approximations

$$\text{bias}[\widehat{m}_h(x)] = \mathbb{E}[\widehat{m}_h(x)] - m(x) \sim h^2 \left(\frac{C_1}{2} m''(x) + C_2 m'(x) \frac{f'(x)}{f(x)} \right)$$

$$\text{Var}[\widehat{m}_h(x)] \sim \frac{C_3}{nh} \frac{\sigma(x)}{f(x)}$$

for some constants C_1, C_2 and C_3.

Set $\text{mse}[\widehat{m}_h(x)] = \text{bias}^2[\widehat{m}_h(x)] + \text{Var}[\widehat{m}_h(x)]$ and consider the integrated version

$$\text{mise}[\widehat{m}_h] = \int \text{mse}[\widehat{m}_h(x)] dF(x)$$

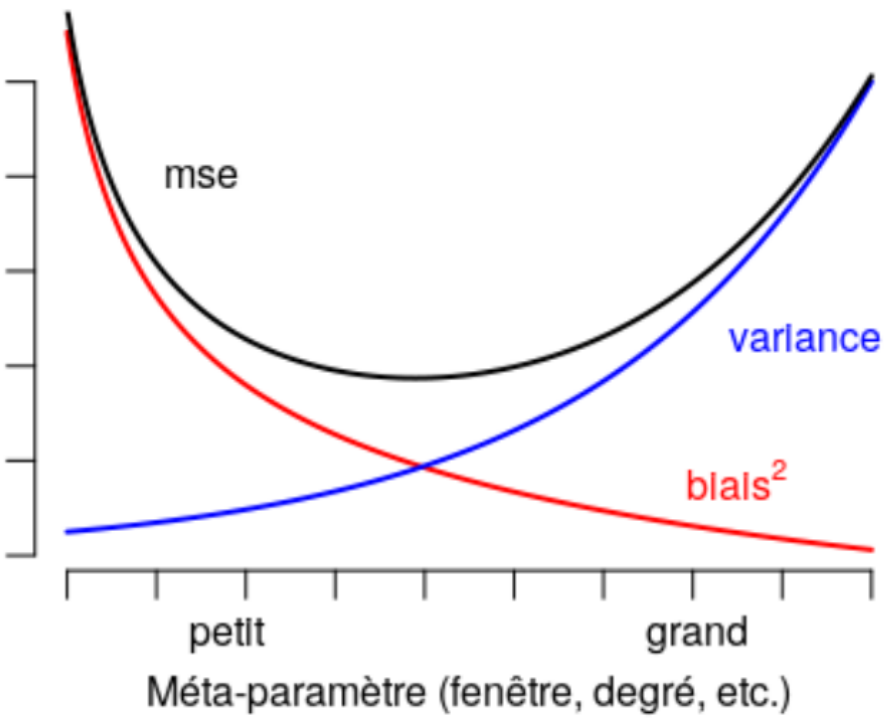
A Brief Excursion to Nonparametric Econometrics

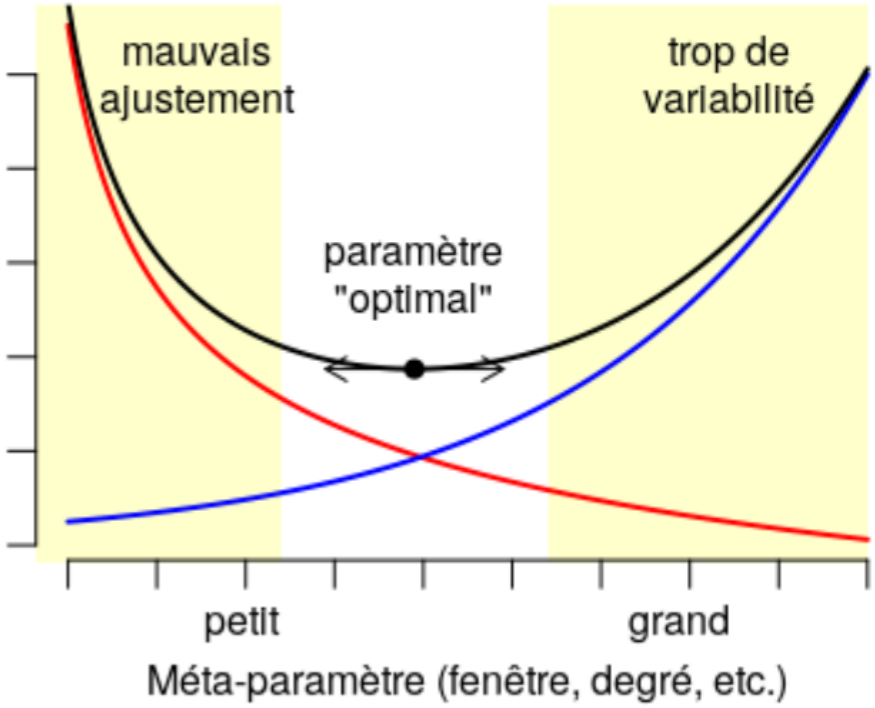
$$\text{mise}[\widehat{m}_h] \quad \sim \quad \overbrace{\frac{h^4}{4}\left(\int x^2 k(x)dx\right)^2 \int \left[m''(x) + 2m'(x)\frac{f'(x)}{f(x)}\right]^2 dx}^{\text{bias}^2}$$
$$+ \overbrace{\frac{\sigma^2}{nh}\int k^2(x)dx \cdot \int \frac{dx}{f(x)}}^{\text{variance}},$$

Thus, the optimal value is $h^\star = O(n^{-1/5})$ (see Silverman's rule of thumb).

Obtained using asymptotic theory and approximations.

A Brief Excursion to Nonparametric Econometrics





Choice of meta-parameter h : trade-off bias / variance.

Model (and Variable) Choice

Suppose that the true model is $y_i = \boldsymbol{x}_{1,i}\boldsymbol{\beta}_1 + \boldsymbol{x}_{2,i}\boldsymbol{\beta}_2 + \varepsilon_i$, but we estimate the model on $\boldsymbol{x}_1$ (only) $y_i = \boldsymbol{X}_{1,i}\boldsymbol{b}_1 + \eta_i$.

$$\begin{aligned}
\widehat{\boldsymbol{b}}_1 &= (\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\mathsf{T}Y \\
&= (\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\mathsf{T}[\boldsymbol{X}_{1,i}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2,i}\boldsymbol{\beta}_2 + \varepsilon] \\
&= (\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_1\boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_2\boldsymbol{\beta}_2 + (\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\mathsf{T}\varepsilon \\
&= \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_2\boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\mathsf{T}\varepsilon_i}_{\nu_i}
\end{aligned}$$

i.e. $\mathbb{E}(\widehat{\boldsymbol{b}}_1) = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12}$. If $\boldsymbol{X}_1^\mathsf{T}\boldsymbol{X}_2 = \boldsymbol{0}$ ($\boldsymbol{X}_1 \perp \boldsymbol{X}_2$), $\mathbb{E}(\widehat{\boldsymbol{b}}_1) = \boldsymbol{\beta}_1$

Conversely, assume that the true model is $y_i = \boldsymbol{x}_{1,i}\boldsymbol{\beta}_1 + \varepsilon_i$ but we estimated on (unnecessary) variables $\boldsymbol{X}_2$ $y_i = \boldsymbol{x}_{1,i}\boldsymbol{b}_1 + \boldsymbol{x}_{2,i}\boldsymbol{b}_2 + \eta_i$. Here estimation is unbiased $\mathbb{E}(\widehat{\boldsymbol{b}}_1) = \boldsymbol{\beta}_1$, but the estimator is not efficient...

Model (and Variable) Choice, *pluralitas non est ponenda sine necessitate*

From the variance decomposition

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{total variance}}=\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{residual variance}}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i-\bar{y})^2}_{\text{explained variance}}$$

Define the R^2 as

$$R^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

or better, the adjusted R^2

$$\bar{R}^2=1-(1-R^2)\frac{n-1}{n-p}=R^2-\underbrace{(1-R^2)\frac{p-1}{n-p}}_{\text{penalty}}.$$

Model (and Variable) Choice, *pluralitas non est ponenda sine necessitate*

One can also consider the log-likelihood, or some penalized version

$$AIC = -\log\mathcal{L} - \underbrace{2\cdot p}_{\text{penalty}} \quad \text{or} \quad BIC = -\log\mathcal{L} - \underbrace{\log(n)\cdot p}_{\text{penalty}}.$$

Objective is $\min\{-\log\mathcal{L}(\boldsymbol{\beta})\}$, and quality is measured via $-\log\mathcal{L}(\widehat{\boldsymbol{\beta}}) - 2p$.

Goodhart's law "*When a measure becomes a target, it ceases to be a good measure*"

Model (and Variable) Choice, *pluralitas non est ponenda sine necessitate*

Alternative approach on penalization : consider a linear predictor $m \in \mathcal{M}$

$$\mathcal{M} = \left\{ m : \ m(\boldsymbol{x}) = s_h(\boldsymbol{x})^\mathsf{T} \boldsymbol{y} \text{ where } S = (s(\boldsymbol{x}_1), \cdots, s(\boldsymbol{x}_n))^\mathsf{T} \text{ smoothing matrix} \right\}$$

Suppose that the true model is $y = m_0(\boldsymbol{x}) + \varepsilon$ with $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbb{I}$, so that $m_0(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$. Quadratic risk $\mathcal{R}(\widehat{m}) = \mathbb{E}\left[(Y - \widehat{m}(\boldsymbol{X}))^2\right]$ is

$$\underbrace{\mathbb{E}\left[(Y - m_0(\boldsymbol{X}))^2\right]}_{\text{error}} + \underbrace{\mathbb{E}\left[(m_0(\boldsymbol{X}) - \mathbb{E}[\widehat{m}(\boldsymbol{X})])^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[(\mathbb{E}[\widehat{m}(\boldsymbol{X})] - \widehat{m}(\boldsymbol{X}))^2\right]}_{\text{variance}}.$$

Consider empirical risk, $\widehat{\mathcal{R}}_n(\widehat{m}) = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{m}(\boldsymbol{x}_i))^2$, then one can prove that

$$\mathcal{R}(\widehat{m}) = \mathbb{E}\left[\widehat{\mathcal{R}}_n(\widehat{m})\right] + \frac{2\sigma^2}{n}\text{trace}(\boldsymbol{S}),$$

see Mallow's C_p from Mallow (1973, Some Comments on C_p)

Introduction: "Machine Learning" ?

Data Mining and Statistics: What's the Connection?

Jerome H. Friedman
Department of Statistics and
Stanford Linear Accelerator Center
Stanford University
Stanford, CA 94305
jhf@stat.stanford.edu

Abstract

Data Mining is used to discover patterns and relationships in data, with an emphasis on large observational data bases. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective it can be viewed as computer automated exploratory data analysis of (usually) large complex data sets. In spite of (or perhaps because of) the somewhat exaggerated hype, this field is having a major impact in business, industry, and science. It also affords enormous research opportu-

Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. - Zekulin.

Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. - Ferruzza.

Data mining is a decision support process where we look in large data bases for unknown and unexpected patterns of information. - Parsaye

Data Mining is ...

- Decision Trees
- Neural Networks
- Rule Induction
- Nearest Neighbors
- Genetic Algorithms

- Mehta

Friedman (1997, Datamining & Statistics, what's the connection)

Introduction: "Machine Learning" ?

Machine learning (initially) did not need any probabilistic framework.

Consider observations $(y_i, \boldsymbol{x}_i)$, and loss function ℓ and some class of models $\mathcal{M}$. The goal is to solve

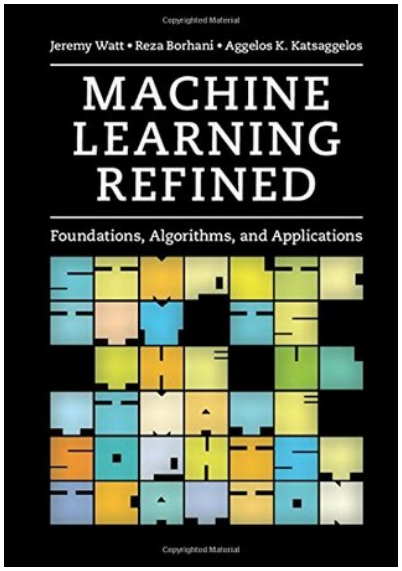
$$m^\star \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_i)) \right\}$$

4.2 The logistic regression perspective on the softmax cost

This section describes a common way of both deriving and thinking about the softmax cost function first introduced in Section 4.1.2. Here we will see how the softmax cost naturally arises as a direct approximation of the fundamental counting cost discussed in Section 4.1.5. However the major benefit of this new perspective is in adding a useful geometric viewpoint,[11] that of regression/surface-fitting, to the classification framework in general, and the softmax cost in particular.

[10] We will also see in Section 4.2 how the softmax cost can be thought of as a direct approximation of the counting cost.

[11] Logistic regression can also be interpreted from a *probabilistic* perspective (see Exercise 4.12).



Watt *et al.* (2016, Machine learning refined foundations algorithms and applications)

Probabilistic Foundations and "Statistical Learning"

Between 1960 and 1980 a revolution in statistics occurred: Fisher's paradigm, introduced in the 1920s and 1930s was replaced by a new one. This paradigm reflects a new answer to the fundamental question:

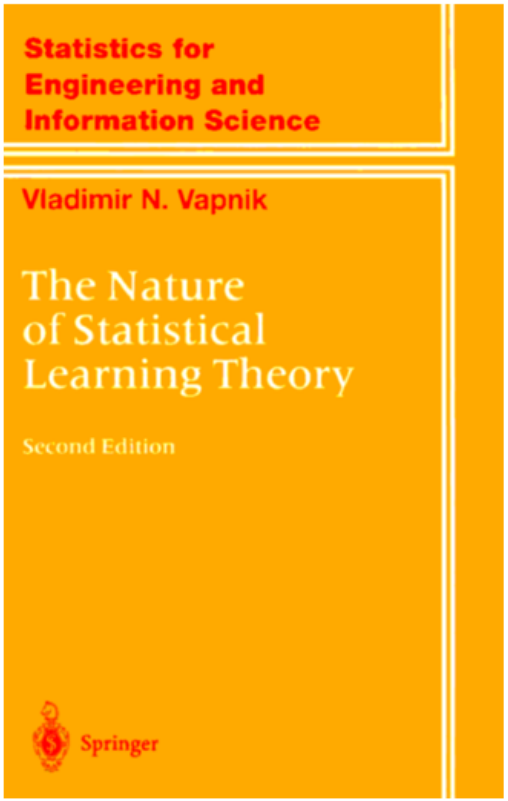
What must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations?

In writing this book I had one more goal in mind: I wanted to stress the practical power of abstract reasoning. The point is that during the last few years at different computer science conferences, I heard reiteration of the following claim:

Complex theories do not work, simple algorithms do.

One of the goals of this book is to show that, at least in the problems of statistical inference, this is not true. I would like to demonstrate that in this area of science a good old principle is valid:

Nothing is more practical than a good theory.



Vapnik (2000, The Nature of Statistical Learning Theory)

Probabilistic Foundations and "Statistical Learning"

RESEARCH CONTRIBUTIONS

Artificial Intelligence and Language Processing

David Waltz Editor

A Theory of the Learnable

L. G. VALIANT

The main contribution of this paper is that it shows that it is possible to design *learning machines* that have all three of the following properties:

1. The machines can provably learn whole classes of concepts. Furthermore, these classes can be characterized.
2. The classes of concepts are appropriate and nontrivial for general-purpose knowledge.
3. The computational process by which the machines deduce the desired programs requires a feasible (i.e., polynomial) number of steps.

More precisely we say that a class X of programs is *learnable* with respect to a given learning protocol if and only if there exists an algorithm A (the deduction procedure) invoking the protocol with the following properties:

1. The algorithm runs in time polynomial in an adjustable parameter h, in the various parameters that quantify the size of the program to be learned, and in the number of variables t.
2. For all programs $f \in X$ and all distributions D over vectors v on which f outputs 1, the algorithm will deduce with probability at least $(1 - h^{-1})$ a program $g \in X$ that never outputs one when it should not but outputs one almost always when it should. In particular, (i) for all vectors v, $g(v) = 1$ implies $f(v) = 1$, and (ii) the sum of $D(v)$ over all v such that $f(v) = 1$, but $g(v) \neq 1$ is at most h^{-1}.

Vailiant (1984, A Theory of Learnable)

Probabilistic Foundations and "Statistical Learning"

Consider a classification problem, $y \in \{-1, +1\}$. The "true" model is m_0 (target), and consider some model $m/$

Let $\mathbb{P}$ denote the (unknown) distribution of $\boldsymbol{X}$'s. The error of m can be written

$$\mathcal{R}_{\mathbb{P},m_0}(\widehat{m}) = \mathbb{P}\big[\widehat{m}(\boldsymbol{X}) \neq m_0(\boldsymbol{X})\big] = \mathbb{P}\big[\{\boldsymbol{x} : \widehat{m}(\boldsymbol{x}) \neq m_0(\boldsymbol{x})\}\big] \text{ where } \boldsymbol{X} \sim \mathbb{P}.$$

Naturally, we can assume that observations $\boldsymbol{x}_i$'s were drawn from $\mathbb{P}$. Empirical risk is

$$\widehat{\mathcal{R}}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\widehat{m}(\boldsymbol{x}_i) \neq y_i).$$

It is not possible to get a perfect model, we should seek a Probably Almost Correct model. Given ϵ, we want to find m such that $\mathcal{R}_{\mathbb{P},m_0}(\widehat{m}) \leq \epsilon$ with probability $1 - \delta$.

More precisely, we want an algorithm that might lead us to a candidate $\widehat{m}$.

Probabilistic Foundations and "Statistical Learning"

Suppose that $\mathcal{M}$ contains a finite number of models. For any ϵ, δ, $\mathbb{P}$ and m_0, if we have enough observations ($n \geq \epsilon^{-1}\log[\delta^{-1}\|\mathcal{M}\|]$), if

$$m^\star \in \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(m(\boldsymbol{x}_i) \neq y_i) \right\}$$

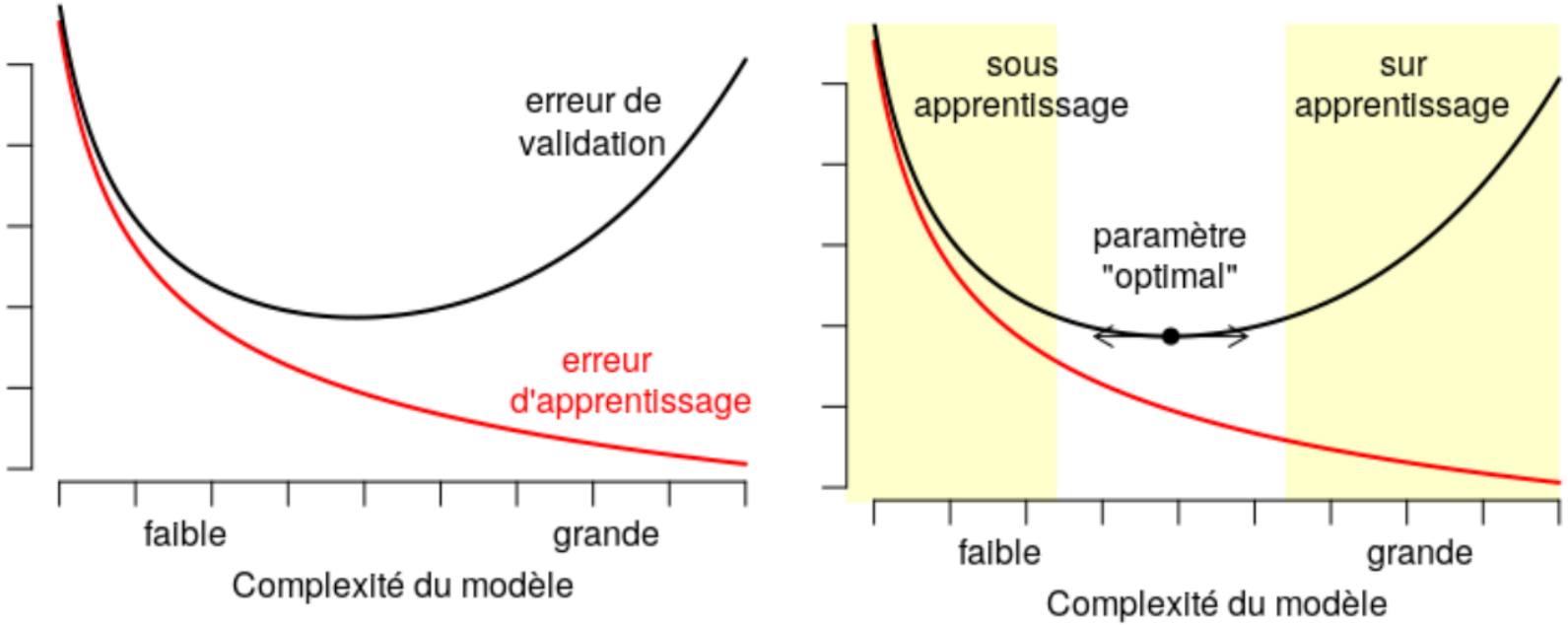
then with probability higher than $1-\delta$, $\mathcal{R}_{\mathbb{P},f}(m^\star) \leq \epsilon$.

Here $n_{\mathcal{M}}(\epsilon,\delta) = \epsilon^{-1}\log[\delta^{-1}\|\mathcal{M}\|]$ is called complexity, and $\mathcal{M}$ is PAC-learnable.

If $\mathcal{M}$ is not finite, the problem is more complicated, it is necessary to define a dimension d - so called VC-dimension - of $\mathcal{M}$, that will be a substitute to $\|\mathcal{M}\|$.

Penalization and Over/Under-Fit

The risk of $\widehat{m}$ cannot be estimated on the data used to fit m.



Alternatives are cross-validation and bootstrap

The objective function and loss functions

Consider loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ and set $m^\star \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_i)) \right\}$.

A classical loss function is the quadratic one $\ell_2(u, v) = (u - v)^2$. Recall that $\overline{y} = \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell_2(y_i, m) \right\}$, or (for the continuous version)

$$\mathbb{E}(Y) = \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \|Y - m\|_{\ell_2}^2 \right\} = \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \mathbb{E}\big[\ell_2(Y, m)\big] \right\}.$$

One can also consider the least absolute value one $\ell_1(u, v) = |u - v$. Recall that $\text{median}[y] = \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell_1(y_i, m) \right\}$. The optimisation problem

$$\widehat{m} = \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \sum_{i=1}^{n} |y_i - m(\boldsymbol{x}_i)| \right\}$$

is well know in robust econometrics.

The objective function and loss functions

More generally, since $\ell_1(y,m) = |(y-m)(1/2 - \mathbf{1}_{y\leq m})|$, can be generalized for any probability level $\tau \in (0,1)$:

$$\widehat{m}_\tau = \underset{m\in\mathcal{M}_0}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell_\tau^{\mathsf{q}}(y_i, m(\boldsymbol{x}_i)) \right\} \text{ with } \ell_\tau^{\mathsf{q}}(x,y) = (x-y)(\tau - \mathbf{1}_{x\leq y})$$

is the quantile regression, see Koenker (2005, Quantile Regression). One can also consider expectiles, with loss function $\ell_\tau^{\mathsf{e}}(x,y) = (x-y)^2 \cdot \left|\tau - \mathbf{1}_{x\leq y}\right|$, see Aigner *et al.* (1977, Formulation and estimation of stochastic frontier production function models)

Gneiting (2011, Making and Evaluating Point Forecasts) introduced the concept of ellicitable statistics: T is ellicitable if there exists $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ such that

$$T(Y) = \underset{x\in\mathbb{R}}{\text{argmin}} \left\{ \int_{\mathbb{R}} \ell(x,y) dF(y) \right\} = \underset{x\in\mathbb{R}}{\text{argmin}} \left\{ \mathbb{E}\big[\ell(x,Y)\big] \text{ where } Y \overset{\mathcal{L}}{\sim} F \right\}.$$

Thus, the mean, the median, any quantiles are ellicitable.

Weak and iterative learning (Gradient Boosting) We want to solve

$$m^\star = \text{argmin}\big\{\mathbb{E}\big[(Y - m(\boldsymbol{X}))^2\big]\big\}$$

The heuristics is simple: we consider an iterative process where we keep modeling the errors.

Fit model for $\boldsymbol{y}$, $h_1(\cdot)$ from $\boldsymbol{y}$ and $\boldsymbol{X}$, and compute the error, $\boldsymbol{\varepsilon}_1 = \boldsymbol{y} - h_1(\boldsymbol{X})$.

Fit model for $\boldsymbol{\varepsilon}_1$, $h_2(\cdot)$ from $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{X}$, and compute the error, $\boldsymbol{\varepsilon}_2 = \boldsymbol{\varepsilon}_1 - h_2(\boldsymbol{X})$, etc. Then set

$$m_k(\cdot) = \underbrace{h_1(\cdot)}_{\sim y} + \underbrace{h_2(\cdot)}_{\sim \varepsilon_1} + \underbrace{h_3(\cdot)}_{\sim \varepsilon_2} + \cdots + \underbrace{h_k(\cdot)}_{\sim \varepsilon_{k-1}}$$

Hence, we consider an iterative procedure, $m^{(k)}(\cdot) = m^{(k-1)}(\cdot) + h_k(\cdot)$.

Weak and iterative learning (Gradient Boosting)

$$m^{(k)} = m^{(k-1)} + \underset{h\in\mathcal{H}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(\underbrace{y_i - m^{(k-1)}(\boldsymbol{x}_i)}_{\varepsilon_{k,i}}, h(\boldsymbol{x}_i)) \right\}$$

Equivalently, start with $m^{(0)} = \underset{m\in\mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m) \right\}$, and solve iteratively

$$m^{(k)} = m^{(k-1)} + \underset{h\in\mathcal{H}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m^{(k-1)}(\boldsymbol{x}_i) + h(\boldsymbol{x}_i)) \right\}$$

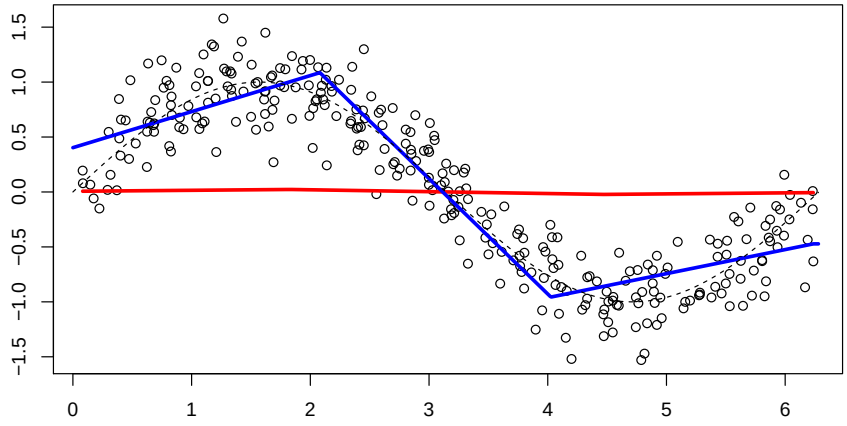
Weak and iterative learning (Gradient Boosting)

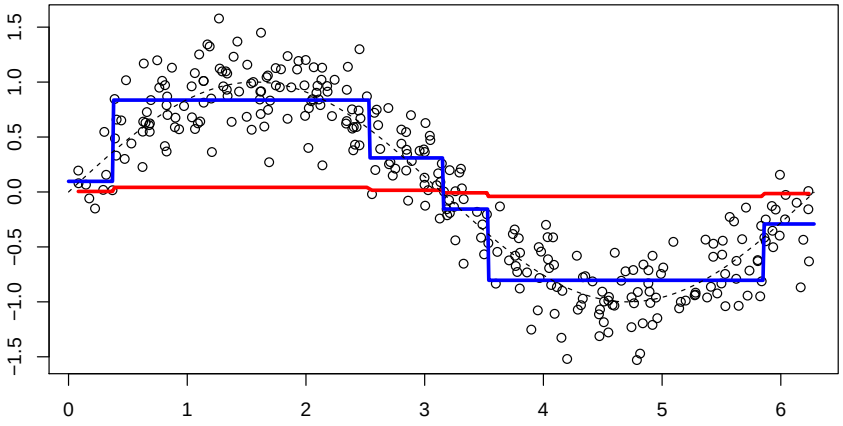
if $\mathcal{H}$ is a set of differentiable functions,

$$m^{(k)} = m^{(k-1)} - \gamma_k \sum_{i=1}^{n} \nabla_{m^{(k-1)}} \ell(y_i, m^{(k-1)}(\boldsymbol{x}_i)),$$

where

$$\gamma_k = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \ell\left(y_i, m^{(k-1)}(\boldsymbol{x}_i) - \gamma \nabla_{m^{(k-1)}} \ell(y_i, m^{(k-1)}(\boldsymbol{x}_i))\right).$$





Penalization and Over/Under-Fit

Consider some penalized objective function, given $\lambda \geq 0$,

$$(\widehat{\beta}_{0,\lambda}, \widehat{\boldsymbol{\beta}}_\lambda) = \text{argmin} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\| \right\}$$

with centered and scaled variables, for some penalty $\|\boldsymbol{\beta}\|$.

Note that there is some Bayesian interpretation

$$\underbrace{\mathbb{P}[\boldsymbol{\theta}|\boldsymbol{y}]}_{\text{a posteriori}} \propto \underbrace{\mathbb{P}[\boldsymbol{y}|\boldsymbol{\theta}]}_{\text{likelihood}} \cdot \underbrace{\mathbb{P}[\boldsymbol{\theta}]}_{\text{a priori}} \quad \text{i.e.} \quad \log \mathbb{P}[\boldsymbol{\theta}|\boldsymbol{y}] = \underbrace{\log \mathbb{P}[\boldsymbol{y}|\boldsymbol{\theta}]}_{\text{log likelihood}} + \underbrace{\log \mathbb{P}[\boldsymbol{\theta}]}_{\text{penalization}} .$$

Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

Define

$$\|\boldsymbol{a}\|_{\ell_0} = \sum_{i=1}^{d} \mathbf{1}(a_i \neq 0), \quad \|\boldsymbol{a}\|_{\ell_1} = \sum_{i=1}^{d} |a_i| \quad \text{and} \quad \|\boldsymbol{a}\|_{\ell_2} = \left(\sum_{i=1}^{d} a_i^2 \right)^{1/2}, \text{ for } \boldsymbol{a} \in \mathbb{R}^d.$$

	constrained optimization	penalized optimization	
ℓ_0	$\underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_0}\leq s}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \right\}$	$\underset{\boldsymbol{\beta},\lambda}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{\ell_0} \right\}$	$(\ell 0)$
ℓ_1	$\underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_1}\leq s}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \right\}$	$\underset{\boldsymbol{\beta},\lambda}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{\ell_1} \right\}$	$(\ell 1)$
ℓ_2	$\underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_2}\leq s}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \right\}$	$\underset{\boldsymbol{\beta},\lambda}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{\ell_2} \right\}$	$(\ell 2)$

Assume that ℓ is the quadratic norm.

Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

The two problems $(\ell 2)$ are equivalent : $\forall(\boldsymbol{\beta}^\star, s^\star)$ solution of the left problem, $\exists\lambda^\star$ such that $(\boldsymbol{\beta}^\star, \lambda^\star)$ is solution of the right problem. And conversely.

The two problems $(\ell 1)$ are equivalent : $\forall(\boldsymbol{\beta}^\star, s^\star)$ solution of the left problem, $\exists\lambda^\star$ such that $(\boldsymbol{\beta}^\star, \lambda^\star)$ is solution of the right problem. And conversely. Nevertheless, if there is a theoretical equivalence, there might be numerical issues since there is not necessarily unicity of the solution.

The two problems $(\ell 0)$ are not equivalent : if $(\boldsymbol{\beta}^\star, \lambda^\star)$ is solution of the right problem, $\exists s^\star$ such that $\boldsymbol{\beta}^\star$ is a solution of the left problem. But the converse is not true.

More generally, consider a ℓ_p norm,

- sparsity is obtained when $p \leq 1$
- convexity is obtained when $p \geq 1$

Using the ℓ_0 penalty

Foster & George (1994, the risk inflation criterion for multiple regression) tried to solve directly the penalized problem of $(\ell 0)$.

But it is a complex combinatorial problem in high dimension (Natarajan (1995) sparse approximate solutions to linear systems proved that it was a NP-hard problem)

One can prove that if $\lambda \sim \sigma^2 \log(p)$, alors

$$\mathbb{E}\big([\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_0]^2\big) \leq \underbrace{\mathbb{E}\big([\boldsymbol{x}_{\mathcal{S}}{}^\mathsf{T}\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_0]^2\big)}_{=\sigma^2 \# \mathcal{S}} \cdot \big(4\log p + 2 + o(1)\big).$$

In that case

$$\widehat{\boldsymbol{\beta}}^{\text{sub}}_{\lambda,j} = \begin{cases} 0 \text{ si } j \notin \mathcal{S}_\lambda(\boldsymbol{\beta}) \\ \widehat{\boldsymbol{\beta}}^{\text{ols}}_j \text{ si } j \in \mathcal{S}_\lambda(\boldsymbol{\beta}), \end{cases}$$

where $\mathcal{S}_\lambda(\boldsymbol{\beta})$ is the set of non-null values in solutions of $(\ell 0)$.

Using the ℓ_2 penalty

With the ℓ_2-norm, $(\ell 2)$ is the Ridge regression

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\mathbb{I})^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})\widehat{\boldsymbol{\beta}}^{\text{ols}}.$$

see Tikhonov (1943, On the stability of inverse problems). Hence

$$\text{biais}[\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}] = -\lambda[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}+\lambda\mathbb{I}]^{-1}\,\widehat{\boldsymbol{\beta}}^{\text{ols}}, \text{Var}[\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}] = \sigma^2[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}+\lambda\mathbb{I}]^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}+\lambda\mathbb{I}]^{-1}$$

With orthogonal variables (i.e. $\boldsymbol{X}^\mathsf{T}\boldsymbol{X} = \mathbb{I}$), we get

$$\text{biais}[\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}] = \frac{\lambda}{1+\lambda}\,\widehat{\boldsymbol{\beta}}^{\text{ols}} \text{ and } \text{Var}[\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}] = \frac{\sigma^2}{(1+\lambda)^2}\mathbb{I} = \frac{\text{Var}[\widehat{\boldsymbol{\beta}}^{\text{ols}}]}{(1+\lambda)^2}.$$

Note that$\text{Var}[\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}] < \text{Var}[\widehat{\boldsymbol{\beta}}^{\text{ols}}]$. Further, $\text{mse}[\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}]$ is minimal when $\lambda^\star = p\sigma^2/\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}$.

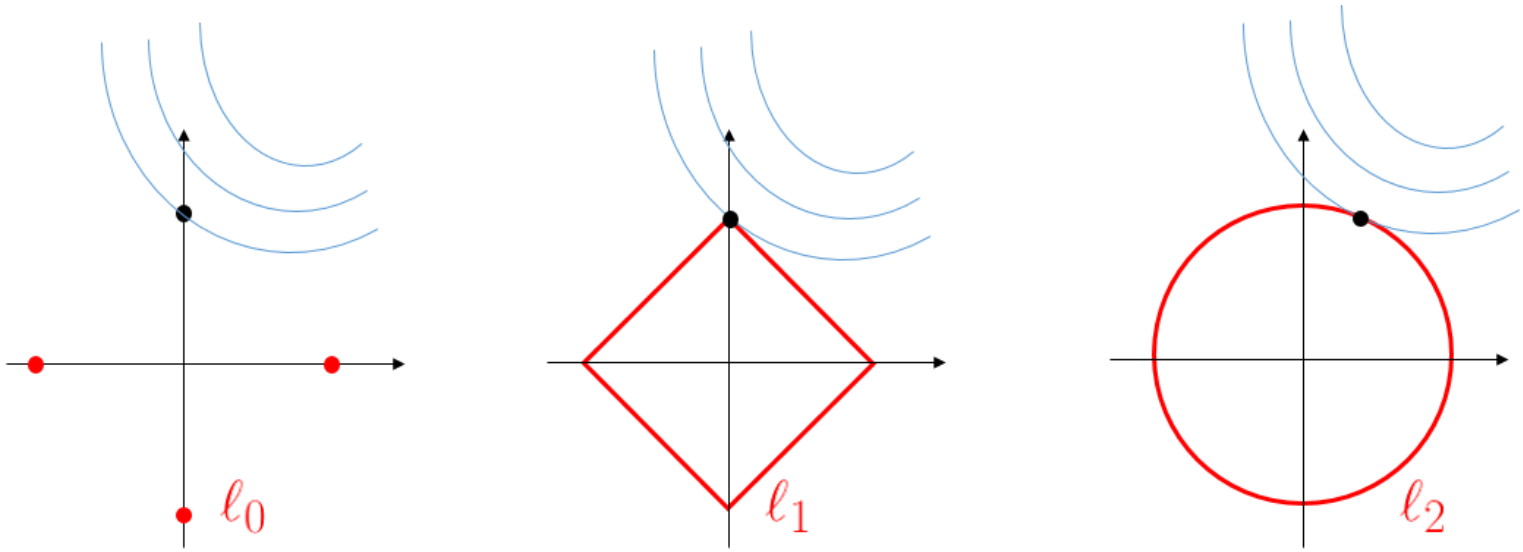
Using the ℓ_1 penalty

If ℓ is no longer the quadratic norm but ℓ_1, problem $(\ell 1)$ is not alway strictly convex, and optimum is not always unique (e.g. if $\boldsymbol{X}^\top\boldsymbol{X}$ is singular).

But in the quadratic case, ℓ is strictly convex, and at least $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is unique.

Further, note that solutions are necessarily coherent (signs of coefficients) : it is not possible to have $\widehat{\beta}_j < 0$ for one solution and $\widehat{\beta}_j > 0$ for another one.

In many cases, problem $(\ell 1)$ yields a corner-type solution, which can be seen as a "best subset" solution - like in $(\ell 0)$.



Using the ℓ_1 penalty

Consider a simple regression $y_i = x_i\beta + \varepsilon$, with ℓ_1-penality and a ℓ_2-loss fuction. $(\ell 1)$ becomes

$$\min\left\{\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{y}^\mathsf{T}\boldsymbol{x}\beta + \beta\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\beta + 2\lambda|\beta|\right\}$$

First order condition can be written $-2\boldsymbol{y}^\mathsf{T}\boldsymbol{x} + 2\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\widehat{\beta}\pm 2\lambda = 0$. (the sign in $\pm$ being the sign of $\widehat{\beta}$). Assume that $\boldsymbol{y}^\mathsf{T}\boldsymbol{x} > 0$, then solution is

$\widehat{\beta}_\lambda^{\text{lasso}} = \max\left\{\dfrac{\boldsymbol{y}^\mathsf{T}\boldsymbol{x} - \lambda}{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}}, 0\right\}$. (we get a corner solution when λ is large).

In higher dimension, see Tibshirani & Wasserman (2016, a closer look at sparse regression) or Candès & Plan (2009, Near-ideal model selection by ℓ_1 minimization).

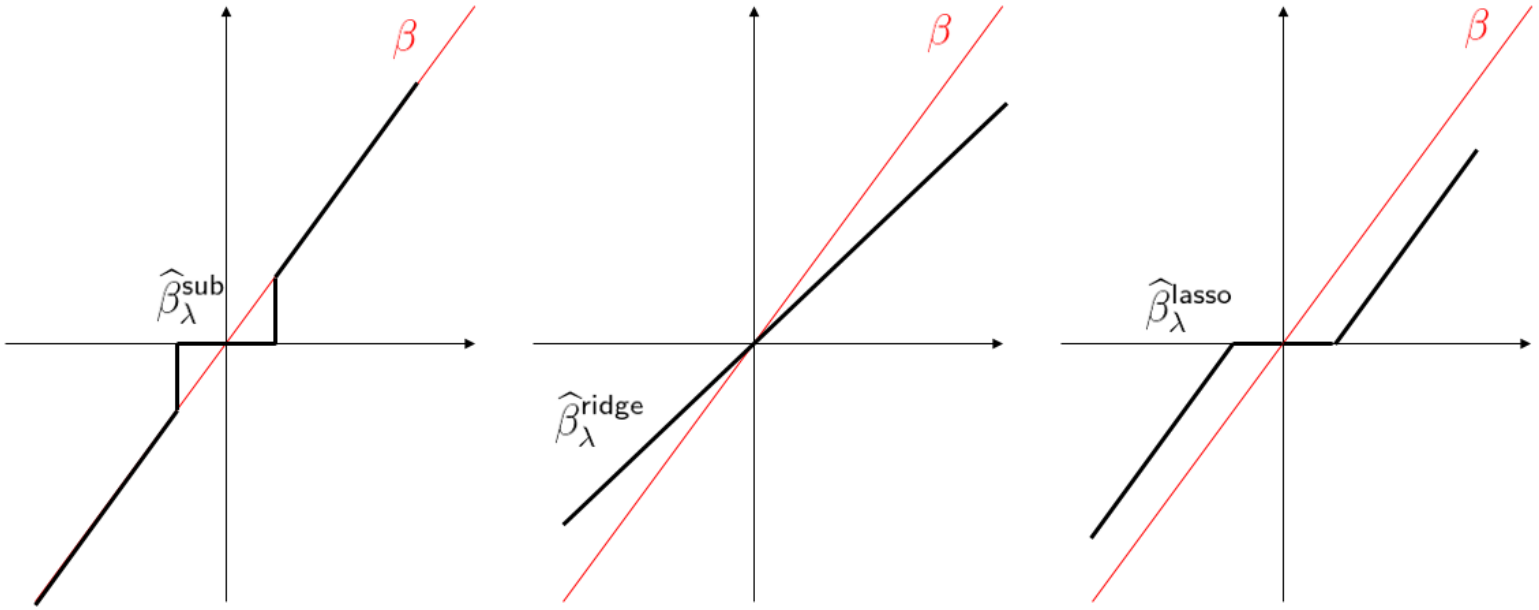
With some additional technical assumption, that LASSO estimator is "sparsistent" in the sense that the support of $\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}}$ is the same as $\boldsymbol{\beta}$

Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

Thus, LASSO can be used for variable selection (see Hastie *et al.* (2001 The Elements of Statistical Learning).

With orthonormal covariance, one can prove that

$$\widehat{\beta}^{\text{sub}}_{\lambda,j} = \widehat{\beta}^{\text{ols}}_j \mathbf{1}_{|\widehat{\beta}^{\text{sub}}_{\lambda,j}|>b}, \quad \widehat{\beta}^{\text{ridge}}_{\lambda,j} = \frac{\widehat{\beta}^{\text{ols}}_j}{1+\lambda} \quad \text{and} \quad \widehat{\beta}^{\text{lasso}}_{\lambda,j} = \text{sign}[\widehat{\beta}^{\text{ols}}_j] \cdot (|\widehat{\beta}^{\text{ols}}_j| - \lambda)_+.$$

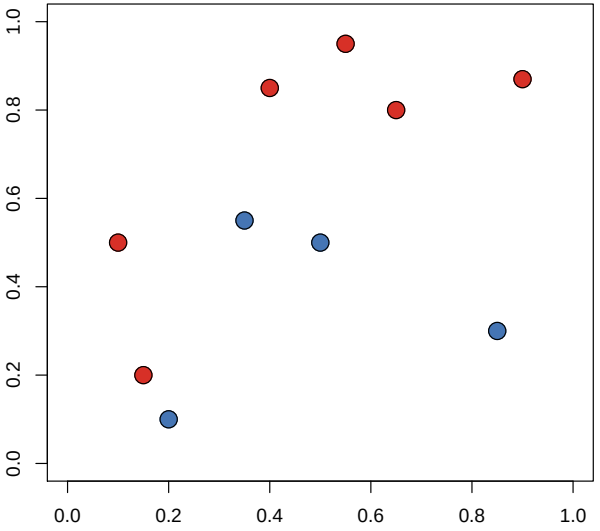


Econometric Modeling

Data $\{(y_i, \boldsymbol{x}_i)\}$, for $i = 1, \cdots, n$, with $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^p$ and $y_i \in \mathcal{Y}$.

A model is a $m : \mathcal{X} \mapsto \mathcal{Y}$ **mapping**

- regression, $\mathcal{Y} = \mathbb{R}$ (but also $\mathcal{Y} = \mathbb{N}$)
- classification, $\mathcal{Y} = \{0, 1\}$, $\{-1, +1\}$, $\{\bullet, \bullet\}$ (binary, or more)



Classification models are based on two steps,

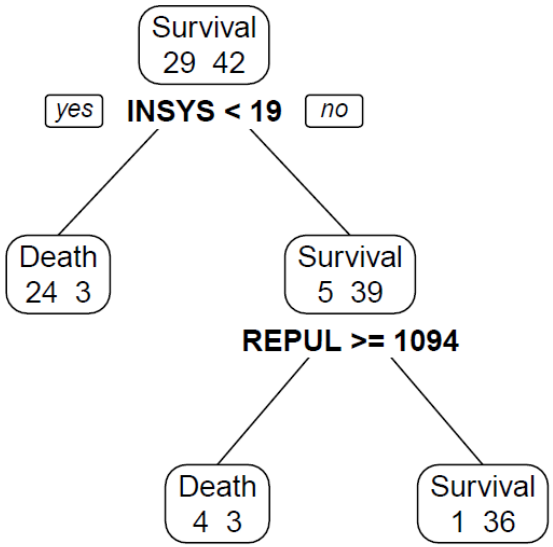
- **score** function, $s(\boldsymbol{x}) = \mathbb{P}(Y = 1 | \boldsymbol{X} = \boldsymbol{x}) \in [0, 1]$

- **classifier** $s(\boldsymbol{x}) \to \widehat{y} \in \{0, 1\}$.

Classification Trees

To split $\{N\}$ into two $\{N_L, N_R\}$, consider

$$\mathcal{I}(N_L, N_R) = \sum_{x\in\{L,R\}} \frac{n_x}{n}\mathcal{I}(N_x)$$



e.g. **Gini index** (used originally in CART, see Breiman *et al.* (1984, Classification and Regression Trees)

$$\text{gini}(N_L, N_R) = -\sum_{x\in\{L,R\}} \frac{n_x}{n} \sum_{y\in\{0,1\}} \frac{n_{x,y}}{n_x}\left(1-\frac{n_{x,y}}{n_x}\right)$$

and the **cross-entropy** (used in C4.5 and C5.0)

$$\text{entropy}(N_L, N_R) = -\sum_{x\in\{L,R\}} \frac{n_x}{n} \sum_{y\in\{0,1\}} \frac{n_{x,y}}{n_x}\log\left(\frac{n_{x,y}}{n_x}\right)$$

Model Selection & ROC Curves

Given a scoring function $m(\cdot)$, with $m(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$, and a threshold $s \in (0, 1)$, set

$$\widehat{Y}^{(s)} = \mathbf{1}[m(\boldsymbol{x}) > s] = \begin{cases} 1 \text{ if } m(\boldsymbol{x}) > s \\ 0 \text{ if } m(\boldsymbol{x}) \leq s \end{cases}$$

Define the confusion matrix as $\boldsymbol{N} = [N_{u,v}]$

	$Y = 0$	$Y = 1$	
$\widehat{Y}_s = 0$	TN_s	FN_s	TN_s+FN_s
$\widehat{Y}_s = 1$	FP_s	TP_s	FP_s+TP_s
	TN_s+FP_s	FN_s+TP_s	n

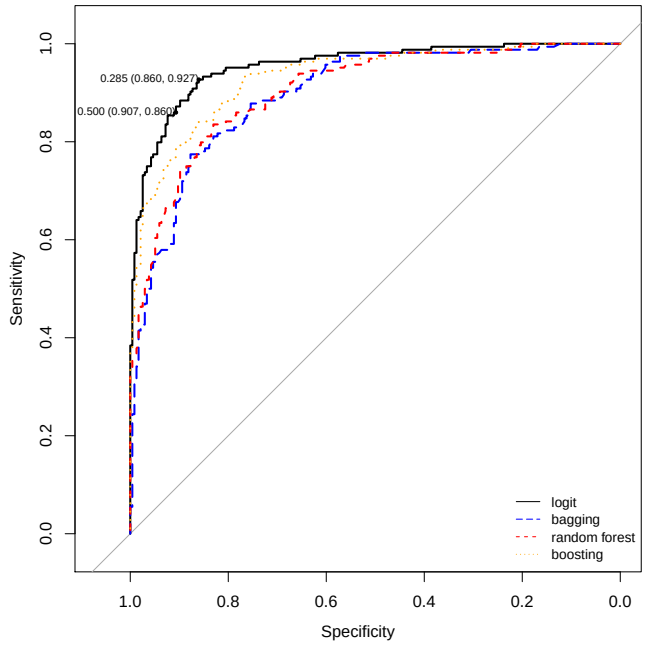
ROC curve is

$$\text{ROC}_s = \left(\frac{\text{FP}_s}{\text{FP}_s + \text{TN}_s}, \frac{\text{TP}_s}{\text{TP}_s + \text{FN}_s} \right) \text{ with } s \in (0, 1)$$

Machine Learning Tools versus Econometric Techniques

Car seats sales, used in James *et al.* (2014)
An Introduction to Statistical Learning

	AUC
logit	0.9544
bagging	0.8973
random forest	0.9050
boosting	0.9313

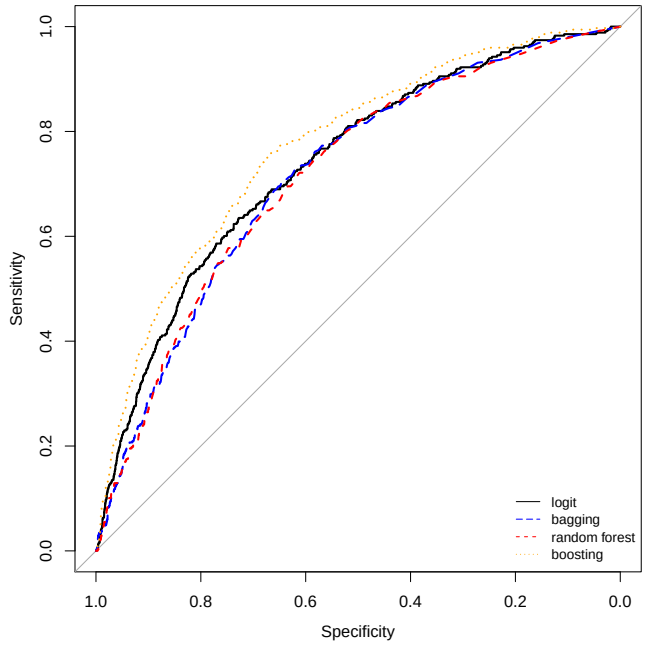


see Section 5.1 in C. *et al.* (2018, Econométrie et Machine Learning)

Machine Learning Tools versus Econometric Techniques

Caravan sales, used in James *et al.* (2014)
An Introduction to Statistical Learning

	AUC
logit	0.7372
bagging	0.7198
random forest	0.7154
boosting	0.7691



see Section 5.2 in C. *et al.* (2018, Econométrie et Machine Learning)

Machine Learning Tools versus Econometric Techniques

Credit database, used in Nisbet *et al.* (2001, Handbook of Statistical Analysis and Data Mining)

Stepwise	AIC	Random Forest	Gini	Lasso
checking_statusA14	1112.1730	checking_statusA14	30.818197	checking_statusA14
credit_amount(4e+03,Inf]	1090.3467	installment_rate	20.786313	credit_amount(4e+03,Inf]
credit_historyA34	1071.8062	residence_since	19.853029	credit_historyA34
installment_rate	1056.3428	duration(15,36]	11.377471	duration(36,Inf]
purposeA41	1044.1580	credit_historyA34	10.966407	credit_historyA31
savingsA65	1033.7521	credit_amount	10.964186	savingsA65
purposeA43	1023.4673	existing_credits	10.482961	housingA152
housingA152	1015.3619	other_payment_plansA143	10.469886	duration(15,36]
other_payment_plansA143	1008.8532	telephoneA192	10.217750	purposeA41
personal_statusA93	1001.6574	age	10.071736	installment_rate
savingsA64	996.0108	savingsA65	9.547362	property_magnitudeA124
other_partiesA103	991.0377	checking_statusA12	9.502445	age(25,Inf]
checking_statusA13	985.9720	housingA152	8.757095	checking_statusA13
checking_statusA12	982.9530	jobA173	8.734460	purposeA43
employmentA74	980.2228	personal_statusA93	8.715932	other_partiesA103
age(25,Inf]	977.9145	property_magnitudeA123	8.634527	employmentA72
purposeA42	975.2365	personal_statusA92	8.438480	savingsA64
duration(15,36]	972.5094	purposeA43	8.362432	employmentA74
duration(36,Inf]	966.7004	employmentA73	8.225416	purposeA46
purposeA49	965.1470	employmentA75	8.089682	personal_statusA93
purposeA410	963.2713	duration(36,Inf]	8.029945	personal_statusA92
credit_historyA31	962.1370	purposeA42	8.025749	savingsA63
purposeA48	961.1567	property_magnitudeA122	7.908813	telephoneA192

Conclusion

Breiman (2001, The Two Cultures)
Econometrics and Machine Learning are focusing on the same problem, but with different cultures

- probabilistic interpretation of econometric models (unfortunately sometimes misleading, e.g. p-value) can deal with non-i.id data (time series, panel, etc)
- machine learning is about predictive modeling and generalization with algorithmic tools, based on bootstrap (sampling and sub-sampling), cross-validation, variable selection, nonlinearities, cross effects, etc