Online Complement – History and Foundations of Econometric and Machine Learning Models

Econometrics and the Probabilistic Model

The importance of probabilistic models in economics is rooted in Working's (1927) questions and the attempts to answer them in Tinbergen's two volumes (1939). The latter have subsequently generated a great deal of reflexions, as recalled by Duo (1993) in his book on the foundations of econometrics, and more particularly in the first chapter "*The Probability Foundations of Econometrics*". Trygve Haavelmo, who was awarded the Nobel Prize in Economics in 1989 for his "*clarification of the foundations of the probabilistic theory of econometrics*" recognized that influence. More specifically, Haavelmo (1944), which initiated a profound change in econometric theory (as recalled in the Chapter 8 of Morgan (1990)), stated that econometrics is fundamentally based on a Probabilistic Model, for two main reasons. First, the use of statistical quantities (or "measures") such as means, standard errors and correlation coefficients for inferential purposes can only be justified if the process generating the data can be expressed in terms of a probabilistic model. Second, the probability approach is relatively general, and is particularly well suited to the analysis of "dependent" and "non-homogeneous" observations, as they are often found on economic data. In this framework, we will assume that there is a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$ such that observations (y_i, x_i) are seen as realizations of random variables (Y_i, X_i). In practice, we are not very interested by the joint distribution of the pair (Y, X) : the law of X is unknown (actually all the analysis is done conditional on observations x_i) and it is the law of Y conditional on X that we will be interested in. In the following, we will denote x a single observation, $\boldsymbol{x}$ a vector of observations, X a random variable, and $\boldsymbol{X}$ a random vector. Abusively, $\boldsymbol{X}$ may also designate the matrix of individual observations (denoted $\boldsymbol{x}_i$), depending on the context.

Foundations of Mathematical Statistics

As recalled in Vapnik (1998)'s introduction, inference in parametric statistics is based on the following belief: the statistician knows the problem to be analyzed well, in particular, he knows the physical law that generates the stochastic properties of the data, and the function to be found is written via a finite number of parameters[1]. To find these parameters, the maximum likelihood method is usually considered. There are many theoretical justification for this approach. We will see that in learning, philosophy is very different, since we do not have a priori reliable information on the statistical law underlying the problem, nor even on the function we would like to approach (we will then propose methods to construct an approximation from the data at our disposal, as in Vapnik (1998)). A "golden age" of parametric inference, from 1930 to 1960, laid the foundations for mathematical statistics, which can be found in all statistical textbooks, including those used nowadays as references in many courses.

As Vapnik (1998) states, the classical parametric paradigm is based on the following three beliefs:
- To find a functional relationship from the data, the statistician is able to define a set of functions, linear in their parameters, that contain a good approximation of the desired function. The number of parameters describing this set is small.
- The statistical law underlying the stochastic component of most real-life problems is the normal law. This belief has been supported by reference to the central limit theorem, which stipulates that under large conditions the sum of a large number of random variables can be approximated by the normal distribution.
- The maximum likelihood method is a good tool for estimating parameters.

In this section we will come back to the construction of the econometric paradigm, directly inspired by that of classical inferential statistics.

Conditional Distributions and Likelihood

Linear econometrics has been constructed under the assumption of individual data, which amounts to assuming independent[2] variables (Y_i, X_i). More precisely, we will assume that, conditionally to the explanatory variables X_i, variables Y_i are independent. We will also assume that these conditional distributions remain in the same parametric family, but that the parameter is a function of x. In the Gaussian linear model it is assumed that:

$$(Y|X = x) \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu(x), \sigma^2) \text{ where } \mu(x) = \beta_0 + x^T\beta, \text{and } \beta \in \mathbb{R}^p. \quad (1)$$

[1] This approach can be compared to structural econometrics, as presented for example in Kean (2010).
[2] It is possible to consider temporal observations, then we would have time series (Y_t, X_t), but we will not discuss those in this article.

It is usually called a 'linear' model since $\mathbb{E}[Y|X=x]=\beta_0+x^T\beta$ is a linear combination of the covariates. It is said to be a homoscedastic model if $\text{Var}[Y|X=x]=\sigma^2$, where σ^2 is a positive constant. To estimate the parameters, the traditional approach is to use the Maximum Likelihood estimator, as initially suggested by Ronald Fisher. In the case of the Gaussian linear model, log-likelihood is written:

$$\log\mathcal{L}(\beta_0,\beta,\sigma^2|y,x)=-\frac{n}{2}\log[2\pi\sigma^2]-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-x_i{}^T\beta)^2.$$

Note that the term on the right, measuring a distance between the data and the model, will be interpreted as deviance in generalized linear models. Then we will set:

$$\left(\hat{\beta}_0,\hat{\beta},\hat{\sigma}^2\right)=argmax\{\log\mathcal{L}(\beta_0,\beta,\sigma^2|y,x)\}$$

From the right term, the maximum likelihood estimator is obtained by minimizing the sum of the error squares (the so-called "least squares" estimator) that we will find in the "machine learning" approach. The first order conditions allow to find the normal equations, whose matrix writing is $X^T[y-X\hat{\beta}]=0$, which can also be written $(X^TX)\,\hat{\beta}=X^Ty$. If X is a full (column) rank matrix, then we find the classical estimator:

$$\hat{\beta}=(X^TX)^{-1}X^Ty=\beta+(X^TX)^{-1}X^T\varepsilon \qquad (2)$$

using residual-based writing (as often in econometrics), $y=\mathrm{x}^T\beta+\varepsilon$. Gauss Markov's theorem ensures that this estimator is the unbiased linear estimator with minimum variance. It can then be shown that $\hat{\beta}\overset{\mathcal{L}}{\sim}\mathcal{N}(\beta,\sigma^2[X^TX]^{-1})$, and in particular, if we simply need the first two moments:

$$\mathbb{E}[\hat{\beta}]=\beta \text{ and } \text{Var}[\hat{\beta}]=\sigma^2[\mathrm{X^TX}]^{-1}$$

In fact, the normality hypothesis makes it possible to make a link with mathematical statistics, but it is possible to construct this estimator given by equation (2) without that Gaussian assumption. Hence, if we assume that $Y|X=x\overset{\mathcal{L}}{\sim}x^T\beta+\varepsilon$, where ε have the same distribution, with $\mathbb{E}[\varepsilon]=0$, $\text{Var}[\varepsilon]=\sigma^2$, and $\text{Cov}[\varepsilon,X_j]=0$ for all j, then $\hat{\beta}$ is an unbiased estimator of β with smallest variance among unbiased linear estimators. Furthermore, if we cannot get normality at finite distance, asymptotically this estimator is Gaussian, with $\sqrt{n}(\hat{\beta}-\beta)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma)$ as $n\to\infty$, for some matrix Σ.

The condition of having a full rank X matrix can be (numerically) strong in large dimensions. If it is not satisfied, $\hat{\beta}=(X^TX)^{-1}X^Ty$ does not exist. If $\mathbb{I}$ denotes the identity matrix, however, it should be noted that $(X^TX+\lambda\mathbb{I})^{-1}X^Ty$ always exists, whatever $\lambda>0$. This estimator is called the ridge estimator of level \lambda (introduced in the 1960s by Hoerl (1962), and associated with a regularization studied by Tikhonov, 1963). This estimator naturally appears in a Bayesian econometric context.

Residuals

It is not uncommon to introduce the linear model from the distribution of the residuals, as we mentioned earlier. Also, equation (1) is written as often:

$$y_i=\beta_0+x_i{}^T\beta+\varepsilon_i \qquad (3)$$

where ε_i's are realizations of independent and identically distributed random variables (i.i.d.) from some $\mathcal{N}(0,\sigma^2)$. distribution. With a vector notation, we will write $\varepsilon\overset{\mathcal{L}}{\sim}\mathcal{N}(0,\sigma^2\mathbb{I})$; the estimated residuals are defined as:

$$\hat{\varepsilon}_i=y_i-[\hat{\beta}_0+x_i{}^T\hat{\beta}]$$

Those (estimated) residuals are basic tools for diagnosing the relevance of the model. An extension of the model described by equation (1) has been proposed to take into account a possible heteroscedastic feature:

$$(Y|X=x)\overset{\mathcal{L}}{\sim}\mathcal{N}(\mu(x),\sigma^2(x))$$

where $\sigma^2(x)$ is a positive function of the explanatory variables. In that case, this model can be rewritten as :

$$y_i=\beta_0+x_i{}^T\beta+\sigma^2(\mathrm{x_i})\cdot\varepsilon_i$$

where residuals are always i.i.d., with unit variance:

$$\varepsilon_i = \frac{y_i - [\beta_0 + {x_i}^T \beta]}{\sigma(x_i)}$$

While equations based on residuals are popular in linear econometrics when the dependent variable is continuous, it is no longer popular with counting models, or logistic regression. However, writing using an error term (as in equation (3)) raises many questions about the representation of an economic relationship between two quantities. For example, it can be assumed that there is a relationship (linear to begin with) between the quantities of a traded good, q and its price p. This allows us to imagine a supply equation:

$$q_i = \beta_0 + \beta_1 p_i + u_i$$

(u_i being an error term) where the quantity sold depends on the price, but in an equally legitimate way, one can imagine that the price depends on the quantity produced (what one could call a demand equation):

$$p_i = \alpha_0 + \alpha_1 q_i + v_i$$

(v_i denoting another error term). Historically, the error term in equation (3) could be interpreted as an idiosyncratic error on the variable y, the so-called explanatory variables being assumed to be fixed, but this interpretation often makes the link between an economic relationship and a complicated economic model difficult, the economic theory speaking abstractly about a relationship between a magnitude, the econometric model imposing a specific shape (what magnitude is y and what magnitude is x) as shown in more detail in Morgan (1990) Chapter 7.

Geometric Properties of this Linear Model

Let's define the scalar product in $\mathbb{R}^d$, $\langle a, b\rangle = a^T b$, and let's note $\|\cdot\|$ the associated Euclidean standard, $\|a\| = \sqrt{a^T a}$ (denoted $\|\cdot\|_{\ell_2}$ in the following sections). Note $\mathcal{E}_X$ the space generated by all linear combinations of the x components (including the constant). If the explanatory variables are linearly independent, X is a full (column) rank matrix and $\mathcal{E}_X$ is a space of dimension $p+1$. Let's assume from now on that the variables x and y are centered here. Note that no distributional hypothesis is made in this section, the geometric properties are derived from the properties of expectation and variance in the set of finite variance variables.
With this notation, it should be noted that the linear model is written $m(\mathrm{x}) = \langle \mathrm{x}, \beta\rangle$. The space $\mathcal{H}_z = \{x \in \mathbb{R}^k : m(x) = z\}$ is a hyperplane (affine) that separates the space in two. Let's define the orthogonal projection operator on $\mathcal{H}_0$, $\Pi_X = X[X^T X]^{-1} X^T$. Thus, the forecast that can be made for y is:

$$\hat{y} = X[X^T X]^{-1} X^T y = \Pi_X \mathrm{y}. \tag{4}$$

As $\hat{\varepsilon} = y - \hat{y} = (\mathbb{I} - \Pi_X)y = \Pi_{X^\perp} y$, we note that $\hat{\varepsilon} \perp x$, which will be interpreted as meaning that residuals are a term of innovation, unpredictable in the sense that $\Pi_X\, \hat{\varepsilon} = 0$.

The Pythagorean theorem is written here:

$$\|y\|^2 = \|\Pi_X y\|^2 + \|\Pi_{X^\perp} y\|^2 = \|\Pi_X y\|^2 + \|y - \Pi_X y\|^2 = \|\,\hat{y}\,\|^2 + \|\,\hat{\varepsilon}\,\|^2$$

which is classically translated in terms of the sum of squares:

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

The coefficient of determination R^2 is then interpreted as the square of the cosine of the angle θ between y and $\Pi_X \mathrm{y}$:

$$R^2 = \frac{\|\Pi_X y\|^2}{\|y\|^2} = 1 - \frac{\|\Pi_{X^\perp} y\|^2}{\|\mathrm{y}\|^2} = \cos^2(\theta).$$

An important application was obtained by Frish & Waugh (1933), when the explanatory variables are divided into two groups, $X = [X_1|X_2]$, so that the regression becomes:

$$y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$$

Frish & Waugh (1933) showed that two successive projections could be considered. Indeed, if $y_2^\star = \Pi_{x_1^\perp} y$ and $X_2^\star = \Pi_{x_1^\perp} X_2$, we can show that :

$$\hat{\beta}_2 = [X_2^{\star T} X_2^\star]^{-1} X_2^{\star T} y_2^\star$$

In other words, the overall estimate is equivalent to the combination of independent estimates of the two models if $X_2^\star = X_2$, i.e. $X_2 \in \mathcal{E}_{X_1}^\perp$, which can be noted $x_1 \perp x_2$. We obtain here the Frisch-Waugh theorem which guarantees that if the explanatory variables between the two groups are orthogonal, then the overall estimate is equivalent to two independent regressions, on each of the sets of explanatory variables. This is a theorem of double projection, on orthogonal spaces. Many results and interpretations are obtained through geometric interpretations (fundamentally related to the links between conditional expectation and the orthogonal projection in space of variables of finite variance).

This geometric interpretation might help to get a better understanding of the problem of under-identification, i.e. the case where the real model would be $y_i = \beta_0 + {x_1}^T\beta_1 + {x_2}^T\beta_2 + \varepsilon_i$, but the estimated model is $y_i = \beta_0 + {x_1}^T b_1 + \eta_i$. The maximum likelihood estimator of b_1 is :

$$\begin{aligned}
\hat{b}_1 &= (X_1^T X_1)^{-1} X_1^T y \\
&= (X_1^T X_1)^{-1} X_1^T [X_{1,i}\beta_1 + X_{2,i}\beta_2 + \varepsilon] \\
&= (X_1^T X_1)^{-1} X_1^T X_1 \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \varepsilon \\
&= \beta_1 + \underbrace{(X_1{}' X_1)^{-1} X_1^T X_2 \beta_2}_{\beta_{12}} + \underbrace{(X_1^T X_1)^{-1} X_1^T \varepsilon}_{\nu_i}
\end{aligned}$$

so that $\mathbb{E}[\hat{b}_1] = \beta_1 + \beta_{12}$, the bias ($\beta_{12}$) being null only in the case where $X_1^T X_2 = 0$ (i.e. $X_1 \perp X_2$): we find here a consequence of the Frisch-Waugh theorem.

On the other hand, over-identification corresponds to the case where the real model would be $y_i = \beta_0 + {x_1}^T\beta_1 + \varepsilon_i$, but the estimated model is $y_i = \beta_0 + {x_1}^T b_1 + {x_2}^T b_2 + \eta_i$. In this case, the estimate is unbiased, in the sense that $\mathbb{E}(\hat{b}_1) = \beta_1$ but the estimator is not efficient. Later on, we will discuss an effective method for selecting variables (and avoid over-identification).

From parametrics to non-parametrics

We can rewrite equation (4) in the form:

$$\hat{y} = X\,\hat{\beta} = X[X^T X]^{-1} X^T y = \Pi_X \mathrm{y}$$

which helps us to see the forecast directly as a linear transformation of the observations. More generally, a linear predictor can be obtained by considering $m(x) = {s_\mathrm{x}}^T y$, where $\mathrm{s_x}$ is a weight vector, which depends on x, interpreted as a smoothing vector. Using the vectors $\mathrm{s}_{\mathrm{x}_i}$, calculated from the observations x_i, we obtain a $n \times n$ matrix S so that $\hat{y} = Sy$. In the case of the linear regression described above, $s_x = X[X^T X]^{-1} x$, $S = X[X^T X]^{-1} X$ and in that case $trace(S)$ is the number of columns in the X matrix (the number of explanatory variables). In this context of more general linear predictors, trace(S) is often seen as equivalent to the number of parameters (or complexity, or dimension, of the model), and $\nu = n - trace(S)$ is then the number of degrees of freedom (see Ruppert *et al*., 2003; Simonoff, 1996). The principle of parsimony says that we should minimize this dimension (the trace of the matrix S) as much as possible. But in the general case, this dimension is more to derive, explicitely.

The estimator introduced by Nadaraya (1964) and Watson (1964), in the case of a simple non-parametric regression, is also written in this form since:

$$\hat{m}_h(x) = {\mathrm{s}_x}^T \mathrm{y} = \sum_{i=1}^n s_{x,i}\, y_i \text{ with } s_{x,i} = \frac{K_h(x - x_i)}{K_h(x - x_1) + \cdots + K_h(x - x_n)},$$

where $K(\cdot)$ is a kernel function, which assigns a value that is lower the closer x_i is to x, and $h > 0$ is the bandwidth. The introduction of this meta parameter h is an important issue, as it should be chosen wisely. Using asymptotic developments, we can show that if X has density f,

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

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

for some constants that can be estimated (see Simonoff (1996) for a discussion). These two functions evolve inversely with h, as shown in Figure C1-I (where the metaparameter is actually h^{-1}), so a trade-off is necessary. The natural idea is then to try to minimize the mean square error, the MSE, defined as, $bias[\hat{m}_h(x)]^2 + \text{Var}[\hat{m}_h(x)]$, and them integrate over x, which gives an optimal value for h of the form $h^\star = O(n^{-1/5})$, and reminds us of Silverman's rule – see Silverman (1986). In larger dimensions, for continuous x variables, a multivariate kernel with matrix bandwidth H can be used, and :

$$\mathbb{E}[\hat{m}_H(x)] \sim m(\text{x}) + \frac{C_1}{2} trace(H^T m''(x) H) + C_2 \frac{m'(x)^T H H^T \nabla f(x)}{f(x)}$$
$$\text{and } \text{Var}[\hat{m}_H(x)] \sim \frac{C_3}{n \det(H)}\frac{\sigma(x)}{f(x)}$$

If H is a diagonal matrix, with the same term h on the diagonal, then $h^\star = O(n^{-1/(4+\dim(x))})$]. However, in practice, there will be more interest in the integrated version of the quadratic error,

$$MISE(\hat{m}_h) = \mathbb{E}[MSE(\hat{m}_h(X))] = \int MSE(\hat{m}_h(x))dF(x),$$

and we can prove that :

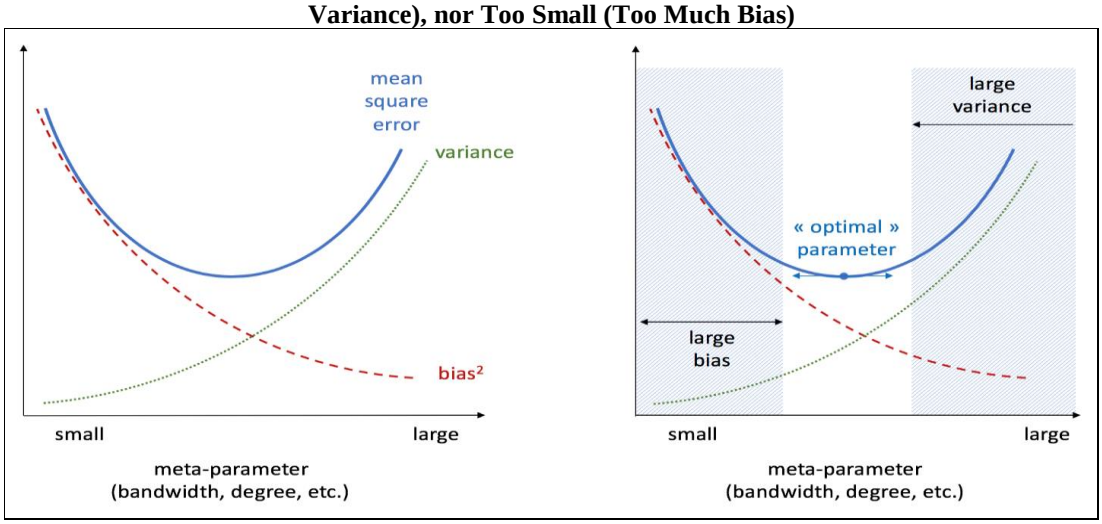
$$MISE[\hat{m}_h] \sim \overbrace{\frac{h^4}{4}\left(\int x^2 k(x)dx\right)^2 \int [m''(x) + 2m'(x)\frac{f'(x)}{f(x)}]^2 dx}^{bias^2} + \overbrace{\frac{\sigma^2}{nh}\int k^2(x)dx \cdot \int \frac{dx}{f(x)}}^{variance},$$

as $n \to \infty$ and $nh \to \infty$. Here we find an asymptotic relationship that again recalls Silverman's (1986) order of magnitude,

$$h^\star = n^{-\frac{1}{5}}\left(\frac{C_1 \int \frac{dx}{f(x)}}{C_2 \int [m''(x) + 2m'(x)\frac{f'(x)}{f(x)}]dx}\right)^{\frac{1}{5}}$$

The main problem here, in practice, is that many of the terms in the expression above are unknown. Automatic learning offers computational techniques, when the econometrician used to search for asymptotic (mathematical) properties.

Figure C1-I
Optimal Meta-Parameter (or Goldilocks' Problem): It Should Be neither Too Large (Too Much Variance), nor Too Small (Too Much Bias)



Exponential Family and Linear Models

The Gaussian linear model is a special case of a large family of linear models, obtained when the conditional distribution of Y (given the covariates) belongs to the exponential family:

$$f(y_i|\theta_i,\phi,x_i\ ) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right) \text{ with } \theta_i = \psi(x_i^T\beta)$$

Functions a, b and c are specified according to the type of exponential law (studied extensively in statistics since Darmoix (1935), as Brown (1986) reminds us), and ψ is a one-to-one mapping that the user must specify. Log-likelihood has then a simple expression:

$$\log\mathcal{L}(\theta,\phi|y) = \prod_{i=1}^{n}\log f(y_i|\theta_i,\phi) = \frac{\sum_{i=1}^{n} y_i\,\theta_i - \sum_{i=1}^{n} b\,(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c\,(y_i,\phi)$$

and the first order condition is then written:

$$\frac{\partial\log\mathcal{L}(\theta,\phi|y)}{\partial\beta} = X^TW^{-1}[y-\hat{y}] = 0$$

based on Müller's (2011) notations, where W is a weight matrix (which depends on β). Given the link between θ and the expectation of Y, instead of specifying the function $\psi(\cdot)$, we will tend to specify the link function $g(\cdot)$ defined by:

$$\hat{y} = m(x) = \mathbb{E}[Y|X=x] = g^{-1}(x^T\beta)$$

For the Gaussian linear regression, we consider an identity link, while for the Poisson regression, the natural link (called canonical) is the logarithmic link. Here, as W depends on β (with $W = \text{diag}(\nabla g(\hat{y})\text{Var}[y])$) there is generally no explicit formula for the maximum likelihood estimator. But an iterative algorithm makes it possible to obtain a numerical approximation. By setting:

$$z = g(\hat{y}) + (y-\hat{y})\cdot\nabla g(\hat{y})$$

corresponding to the error term of a Taylor development in order 1 of g, we obtain an algorithm of the form:

$$\hat{\beta}_{k+1} = [X^TW_k^{-1}X]^{-1}X^TW_k^{-1}z_k$$

By iterating, we will define $\hat{\beta} = \hat{\beta}_\infty$, and we can show that – with some additional technical assumptions (detailed in Müller (2011)) – this estimator is asymptotically Gaussian, with:

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\beta)^{-1}),$$

where numerically $I(\beta) = \phi\cdot[X^TW_\infty^{-1}X]$.

From a numerical point of view, the computer will solve the first-order condition, and actually, the law of Y does not really intervene. For example, one can estimate a "Poisson regression" even when $y\in\mathbb{R}_+$, not necessarily $y\in\mathbb{N}$. In other words, the distribution of Y is only an interpretation here, and the algorithm could be introduced in a different way (as we will see later on), without necessarily having an underlying probabilistic model.

Logistic Regression

Logistic regression is the generalized linear model obtained with a Bernoulli's distribution, and a link function which is the quantile function of a logistic law (which corresponds to the canonical link in the sense of the exponential family). Considering the form of Bernoulli's law, econometrics proposes a model for $y_i\in\{0,1\}$, in which the logarithm of the odds follows a linear model:

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

or:

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

is the cumulative distribution function of a logistic variable. The estimation of (β_0, β) is performed by maximizing the likelihood:

$$\mathcal{L} = \prod_{i=1}^{n}\left(\frac{e^{\beta_0+x_i^T\beta}}{1+e^{\beta_0+x_i^T\beta}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+x_i^T\beta}}\right)^{1-y_i}$$

It is said to be a linear model because iso-probability curves here are the parallel hyperplanes $b_0 + x^T\beta$. Rather than this model, popularized by Berkson (1944), some will prefer the probit model (see Berkson, 1951), introduced by Bliss (1934). In this model:

$$\mathbb{E}[Y|X=x] = \mathbb{P}[Y=1|X=x] = \Phi(\beta_0+x^T\beta)$$

where Φ denotes the distribution function of the reduced centered normal distribution. This model has the advantage of having a direct link with the Gaussian linear model, since :

$$y_i = 1(y_i^\star > 0) \text{ with } y_i^\star = \beta_0 + x^T\beta + \varepsilon_i$$

where the residuals are Gaussian, $\mathcal{N}(0,\sigma^2)$. An alternative is to have centered residuals of unit variance, and to consider a latent modeling of the form $y_i = 1(y_i^\star > \xi)$ (where ξ will be fixed). As we can see, these techniques are fundamentally linked to an underlying stochastic model. In the body of the article, we present several alternative techniques – from the learning literature – for this classification problem (with two classes, here 0 and 1).

Regression in High Dimension

As we mentioned earlier, the first order condition $X^T(X\hat{\beta} - \mathrm{y}) = 0$ is solved numerically by performing a QR decomposition, at a cost which consists in $O(np^2)$ operations (where p is the rank of X^TX). Numerically, this calculation can be long (either because p is large or because - to a lesser extent - because n is large), and a simpler strategy may be to sub-sample. Let $n_s \ll n$, and consider a sub-sample size n_s from $\{1,\cdots,n\}$. Then $\hat{\beta}_s = ({X_s}^T X_s)^{-1}{X_s}^T y_s$ is a good approximation of $\hat{\beta}$ as shown by Dhillon *et al.* (2014). However, this algorithm is dangerous if some points have a high leverage (i.e. $L_i = x_i(X^TX)^{-1}{x_i}^T$). Tropp (2011) proposes to transform the data (in a linear way), but a more popular approach is to do non-uniform sub-sampling, with a probability related to the influence of observations (defined by $I_i = \hat{\varepsilon}_i\, L_i/(1-L_i)^2$, and which unfortunately can only be calculated once the model is estimated).

In general, we will talk about massive data when the data table of size does not fit in the RAM memory of the computer. This situation is often encountered in statistical learning nowadays with very often $p \ll n$. This is why, in practice, many libraries of algorithms assimilated to machine learning use iterative methods to solve the first-order condition. When the parametric model to be calibrated is indeed convex and semi-differentiable, it is possible to use, for example, the stochastic gradient descent method as suggested by Bottou (2010). This last one allows to free oneself at each iteration from the calculation of the gradient on each observation of our learning base. Rather than making an average descent at each iteration, we start by drawing (without replacement) an observation x_i among the n available. The model parameters are then corrected so that the prediction made from x_i is as close as possible to the true value y_i. The method is then repeated until all the data have been reviewed. In this algorithm there is therefore as much iteration as there are observations. Unlike the gradient descent algorithm (or Newton's method) at each iteration, only one gradient vector is calculated (and no longer n). However, it is sometimes necessary to run this algorithm several times to increase the convergence of the model parameters. If the objective is, for example, to minimize a loss function ℓ between the model $m_\beta(x)$ and y (like the quadratic loss function, as in the Gaussian linear regression) the algorithm can be summarized as follows:

- Step 0: Mix the data
- Iteration step: For $t \geq 0$, we draw $i \in \{1,\cdots,n\}$ without replacement, and we set:

$$\beta^{t+1} = \beta^t - \gamma_t \frac{\partial \ell(y_i, m_{\beta^t}(x_i))}{\partial \beta}$$

This algorithm can be repeated several times as a whole depending on the user's needs. The advantage of this method is that at each iteration, it is not necessary to calculate the gradient on all observations (more sum). It is

therefore suitable for large databases. This algorithm is based on a convergence in probability towards a neighborhood of the optimum (and not the optimum itself).

Goodness of fit and model selection

In the Gaussian linear model, the determination coefficient – noted R^2 – is often used as a measure of fit quality. It is based on the variance decomposition formula discussed previously:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}_{total\ variance}=\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{residual\ variance}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{explained\ variance} \quad (5)$$

The R^2 is defined as the ratio of explained variance and total variance, another interpretation of the coefficient that we had introduced from the geometry of the least squares:

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

The sums of the error squares in this writing can be rewritten as a log-likelihood. However, it should be remembered that, up to one additive constant (obtained with a saturated model) in generalized linear models, deviance is defined (up to an additive constant) by:

$$Deviance(\beta)=-2\log[\mathcal{L}]$$

which can also be denoted $\text{Deviance}(\hat{y})$. A null deviance can be defined as the one obtained without using the explanatory variables x, so that $\hat{y}_i=\overline{y}$. It is then possible to define, in a more general context (with a non-Gaussian distribution for *Y*):

$$R^2=\frac{Deviance(\overline{y})-Deviance(\hat{y})}{Deviance(\overline{y})}=1-\frac{Deviance(\hat{y})}{Deviance(\overline{y})}.$$

However, this measure cannot be used to choose a model, if one wishes to have a relatively simple model in the end, because it increases artificially with the addition of explanatory variables without significant effect. We will then tend to prefer the adjusted R^2 :

$$\overline{R}^2=1-(1-R^2)\frac{n-1}{n-p}=R^2-\underbrace{(1-R^2)\frac{p-1}{n-p}}_{penalty},$$

where p is the number of parameters of the model. Here, the quality of fit will be penalized overly too complex models.

A similar idea will be found in the Akaike criterion, where $AIC=Deviance+2\cdot p$ or in (bayesian) Schwarz criterion $BIC=Deviance+\log(n)\cdot p$. In large dimensions (typically $p>\sqrt{n}$), we will tend to use a corrected AIC defined as:

$$AICc=Deviance+2\cdot p\cdot\frac{n}{n-p-1}$$

These criterias are used in so-called "stepwise" methods. In the "forward" method, we start by regressing to the constant, then we add one variable at a time, retaining the one that lowers the *AIC* criterion the most, until adding a variable increases the *AIC* criterion of the model. In the "backward" method, we start by regressing on all variables, then we remove one variable at a time, removing the one that lowers the *AIC* criterion the most, until removing a variable increases the *AIC* criterion from the model.

Another justification for this notion of penalty (we will come back to this idea in machine learning) can be the following. Let us consider an estimator in the class of linear predictors:

$$\mathcal{M}=\{m:\ m(x)=s_h(\mathrm{x})^T y \text{ where } S=\left(s(x_1),\cdots,s(x_n)\right)^T \text{ is a smoothing matrix}\}$$

and assume that $y = m_0(x) + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$ and $\text{Var}[\varepsilon] = \sigma^2\mathbb{I}$, so that $m_0(x) = \mathbb{E}[Y|X = x]$. From a theoretical point of view, the quadratic risk, associated with an estimated model $\hat{m}$, $\mathbb{E}[(Y - \hat{m}(X))^2]$ is written:

$$\mathcal{R}(\hat{m}) = \underbrace{\mathbb{E}[(Y - m_0(X))^2]}_{error} + \underbrace{\mathbb{E}[(m_0(X) - \mathbb{E}[\hat{m}(X)])^2]}_{bias} + \underbrace{\mathbb{E}[(\mathbb{E}[\hat{m}(X)] - \hat{m}(X))^2]}_{variance}$$

if m_0 is the true model. The first term is sometimes called "Bayes error", and does not depend on the estimator selected $\hat{m}$.

The empirical quadratic risk, associated with a model m, is here:

$$\hat{\mathcal{R}}_n(m) = \frac{1}{n}\sum_{i=1}^{n}(y_i - m(x_i))^2 = \frac{1}{n}\| y - m(x) \|^2$$

(by convention). We recognize here the mean square error, "mse", which will more generally give the "risk" of the model m when using another loss function (as we will discuss later on). It could be proved that:

$$\mathbb{E}[\hat{\mathcal{R}}_n(m)] = \frac{1}{n}\| m_0(x) - m(x) \|^2 + \frac{1}{n}\mathbb{E}(\| y - m_0(x) \|^2)$$

and :

$$n\mathbb{E}[\hat{\mathcal{R}}_n(\hat{m})] = \mathbb{E}(\| y - \hat{m}(x) \|^2) = \| (\mathbb{I} - S)m_0 \|^2 + \sigma^2 \| \mathbb{I} - S \|^2$$

so that the (real) risk of $\hat{m}$ is:

$$\mathcal{R}_n(\hat{m}) = \mathbb{E}[\hat{\mathcal{R}}_n(\hat{m})] + 2\frac{\sigma^2}{n}trace(\text{S}).$$

So, if $\text{trace}(\text{S}) \geq 0$, (which is not a too strong assumption), the empirical risk underestimates the true risk of the estimator. Actually, we recognize here the number of degrees of freedom of the model, the right-hand term corresponding to Mallow's C_p introduced in Mallows (1973) using not the deviance but the R^2).

Machine Learning Philosophy

In parallel with these tools developed by, and for economists, a whole literature has been developed on similar issues, centered on the problems of prediction and forecasting. For Breiman (2001a), a first difference comes from the fact that the statistic has developed around the principle of inference (or to explain the relationship linking y to variables x) while another culture is primarily interested in prediction. In a discussion that follows the article, David Cox states very clearly that in statistic (and econometrics) "*predictive success* […] *is not the primary basis for model choice*". We will get back here on the roots of automatic learning techniques. The important point, as we will see, is that the main concern of machine learning is related to the generalization properties of a model, i.e. its performance – according to a criterion chosen a priori – on new data, and therefore on non-sample tests.

A Learning Machine

Nowadays, we speak of "machine learning" to describe a whole set of techniques, often computational, as alternatives to the classical econometric approach. Before characterizing them as much as possible, it should be noted that historically other names have been given. For example, Friedman (1997) proposes to make the link between statistics (which closely resemble econometric techniques – hypothesis testing, ANOVA, linear regression, logistics, GLM, etc.) and what was then called "data mining" (which then included decision trees, methods from the closest neighbors, neural networks, etc.). The bridge between those two cultures corresponds to "statistical learning" techniques described in Hastie *et al.* (2009). But one should keep in mind that machine learning is a very large field of research.

The so-called "natural" learning (as opposed to machine learning) is that of children, who learn to speak, read and play. Learning to speak means segmenting and categorizing sounds, and associating them with meanings. A child also learns simultaneously the structure of his or her mother tongue, and acquires a set of words describing the world around him or her. Several techniques are possible, ranging from rote learning, generalization, discovery, more or less supervised or autonomous learning, etc. The idea in artificial intelligence is to take inspiration from the functioning of the brain to learn, to allow "artificial" or "automatic" learning, by a machine. A first application was to teach a machine to play a game (tic-tac-toe, chess, go, etc.). An essential step is to explain the objective it must achieve to win. One historical approach has been to teach the machine the rules of the game. If it allows a machine to play, it will not help the machine to play well. Assuming that the machine knows the rules of the game, and that it has a choice between several dozen possible moves, which one should it choose? The classical approach in artificial intelligence uses the so-called min-max algorithm using an evaluation function: in this algorithm, the

machine searches forward in the possible moves tree, as far as the calculation resources allow (about ten moves in chess, for example). Then, it calculates different criteria (which have been previously indicated to her) for all positions (number of pieces taken, or lost, occupancy of the center, etc. in our example of the chess game), and finally, the machine plays the move that allows it to maximize its gain. Another example may be the classification and recognition of images or shapes. For example, the machine must identify a number in a handwritten handwriting (checks, ZIP code on envelopes, etc.). It is a question of predicting the value of a variable y, knowing that a priori $y \in \{0,1,2,\cdots,8,9\}$. A classical strategy is to provide the machine with learning samples, in other words here millions of labelled (identified) images of handwritten numbers. A simple (and natural) strategy is to use a decision criterion based on the closest neighbors whose labels are known (using a predefined metric).

The method of the closest neighbors ("*k-nearest neighbors"*) can be described as follows: we consider (as in the previous part) a set of n observations, i. e. pairs (y_i, x_i) with $x_i \in \mathbb{R}^p$. Let us consider a distance Δ on $\mathbb{R}^p$ (the Euclidean distance or the Mahalanobis distance, for example). Given a new observation $x \in \mathbb{R}^p$ the Euclidean distance or the Mahalanobis distance, for example). Given a new observation x_i and x, in the sense that $\Delta(x_1, x) \leq \Delta(x_2, x) \leq \cdots \leq \Delta(x_n, x)$ then we can consider as prediction for y the average of the k nearest neighbors:

$$m_k(x) = \frac{1}{k}\sum_{i=1}^{k} y_i.$$

Learning here works by induction, based on a sample (called the learning – or training – sample).

Automatic learning includes those algorithms that give computers the ability to learn without being explicitly programmed (as Arthur Samuel defined it in 1959). The machine will then explore the data with a specific objective (such as searching for the nearest neighbors in the example just described). Tom Mitchell proposed a more precise definition in 1998: a computer program is said to learn from experience E in relation to a task T and a performance measure P, if its performance on T, measured by P, improves with experience E. Task T can be a defect score for example, and performance P can be the percentage of errors made. The system learns if the percentage of predicted defects increases with experience.

As we can see, machine learning is basically a problem of optimizing a criterion based on data (from now on called learning). Many textbooks on machine learning techniques propose algorithms, without ever mentioning any probabilistic model. In Watt *et al*. (2016) for example, the word "*probability*" is mentioned only once, with this footnote that will surprise and make smile any econometricians, "*the logistic regression can also be interpreted from a probabilistic perspective*" (p. 86). But many recent books offer a review of machine learning approaches using probabilistic theories, following the work of Vaillant and Vapnik. By proposing the paradigm of "probably almost correct" learning (PAC), a probabilistic flavor has been added to the previously very computational approach, by quantifying the error of the learning algorithm (usually in a classification problem).

The probabilistic formalism in the 1980's

We have a training sample, with observations (x_i, y_i) where the variable y is in a set $\mathcal{Y}$. In the case of classification, $\mathcal{Y} = \{-1, +1\}$, but a relatively general set can be considered[3]. A predictor m taking values in $\mathcal{Y}$, used to label (or classify) future new observations, using some features that lie in a set $\mathcal{X}$. It is assumed that the labels are produced by an (unknown) classifier f called target. For a statistician, this function would be the real model. Naturally, we want to build m as close as possible to f. Let $\mathbb{P}$ be a (unknown) distribution on $\mathcal{X}$. The error of m with respect to target f is defined as $\mathcal{R}_{\mathbb{P},f}(m) = \mathbb{P}[m(X) \neq f(X)]$, where $X \sim \mathbb{P}$ or equivalently, $\mathcal{R}_{\mathbb{P},f}(m) = \mathbb{P}[\{x \in \mathcal{X}: m(x) \neq f(x)\}]$. To obtain our "optimal" classifier, it becomes necessary to assume that there is a link between the data in our sample and the pair $(\mathbb{P}, f)$, i.e. a data generation model. We will then assume that the x_i are obtained by independent draws according to $\mathbb{P}$, and that then $y_i = f(x_i)$.

We can define the empirical risk of a classifier m, $\hat{\mathcal{R}}(m) = \frac{1}{n}\sum_{i=1}^{n} 1(m(x_i) \neq y_i)$.
It is important to recognize that a perfect model cannot be found, in the sense that $\mathcal{R}_{\mathbb{P},f}(m) = 0$. Indeed, if we consider the simplest case, with $\mathcal{X}$=$\{x_1, x_2\}$ such that $\mathbb{P}(\{x_1\}) = p$ and $\mathbb{P}(\{x_2\}) = 1 - p$. The probability of never observing $\{x_2\}$ among the n observations is $(1-p)^n$, and if $1 < 1/n$,, it is quite likely never to observe $\{x_2\}$ so it can never be predicted. We cannot therefore hope to have a zero risk whatever $\mathbb{P}$. And more generally, it is also

[3] Econometricians will always prefer {0,1}, because of connections with the Bernoulli distribution (and corresponds to lower and upper bound of probabilities).

possible to observe $\{x_1\}$ and $\{x_2\}$, and despite everything, to make mistakes on the labels. Also, instead of looking for a perfect model, we can try to have an "approximately correct" model. We will then try to find mmm such that $\mathcal{R}_{\mathbb{P},f}(m) \leq \epsilon$, where ϵ is an a priori specified threshold. But even this condition is too strong, and cannot be fulfilled. Thus, we will usually as to have $\mathcal{R}_{\mathbb{P},f}(m) \leq \epsilon$ with some probability $1-\delta$. Hence, we will try to be "probably approximately correct" (PAC), allowing to make a mistake with a probability δ, again fixed a priori.

The interpretation is that since we cannot learn (in the PAC sense) about all the functions m we will then force m to belong to a particular class, noted $\mathcal{M}$. Let us suppose, to start with, that $\mathcal{M}$ contains a finite number of possible models. We can then show that for all ϵ and δ, for all $\mathbb{P}$ and f, if we have enough observations (more precisely $n \geq \epsilon^{-1}\log[\delta^{-1}|\mathcal{M}|]$, then with a greater probability than $1-\delta$, then $\mathcal{R}_{\mathbb{P},f}(m^\star) \leq \epsilon$ where:

$$m^\star \in \underset{m\in\mathcal{M}}{argmin}\{\frac{1}{n}\sum_{i=1}^{n} 1\,(m(x_i) \neq y_i)\}$$

in other words $m^\star$ is a model in class $\mathcal{M}$ that minimizes empirical risk.

We can go a little further, staying in the case where $\mathcal{Y} = \{-1, +1\}$. A $\mathcal{M}$ class of classifiers will be called PAC-learnable if there is a function $n_{\mathcal{M}}: [0,1]^2 \to \mathbb{N}$ such that, for all ϵ, δ, $\mathbb{P}$ and if it is assumed that the target f also belongs to class $\mathcal{M}$, then using $n > n_{\mathcal{M}}(\epsilon, \delta)$ observations x_i drawn from $\mathbb{P}$, labelled y_i by f, then there exists $m \in \mathcal{M}$ such that, with probability $1-\delta$, $\mathcal{R}_{\mathbb{P},f}(m) \leq \epsilon$. Function $n_{\mathcal{M}}$ s then called "sample complexity to learn". In particular, we have seen that if $\mathcal{M}$ contains a finite number of classifiers, then $\mathcal{M}$ is PAC-learnable with complexity $n_{\mathcal{M}}(\epsilon, \delta) = \epsilon^{-1}\log[\delta^{-1}|\mathcal{M}|]$.

Naturally, we would like to have a more general result, especially if $\mathcal{M}$ is not finite. To do this, the VC dimension of Vapnik-Chervonenkis must be used, which is based on the idea of shattering points (for a binary classification). Consider k points $\{x_1, \cdots x_k\}$, and the set $\mathcal{E}_k = \{(m(x_1), \cdots, m(x_k)) \text{ for } m \in \mathcal{M})\}$. Observe that elements of $\mathcal{E}_k$ belong to $\{-1, +1\}^k$. In other words, $|\mathcal{E}_k| \leq 2^k$. We will say that $\mathcal{M}$ shatter all the points if all the combinations are possible, i. e. $|\mathcal{E}_k| = 2^k$. Intuitively, the labels of the set of points do not provide enough information on target f, because any situation is possible. The VC dimension of $\mathcal{M}$ is then:

$$VC(\mathcal{M}) = \sup\{k \text{ such that } \mathcal{M} \text{ shatters } \{x_1, \cdots x_k\}\}$$

For example, if $\mathcal{X} = \mathbb{R}^k$, and consider separations by hyperplanes passing through the origin (we will say homogeneous), in the sense that[4] $m_{\mathrm{w}}(x) = 1_{\pm}(w^T x \geq 0)$. It can be shown that no set of $k+1$ points can be shattered by these two homogeneous spaces in $\mathbb{R}^k$, and therefore $VC(\mathcal{M}) = k$. If we add a constant, in the sense that $m_{\mathrm{w},b}(\mathrm{x}) = 1_{\pm}(w^T x + b \geq 0)$, we can show that no set of $k+2$ points can be sprayed by these two (non-homogeneous) spaces in $\mathbb{R}^k$, and therefore $VC(\mathcal{M}) = k+1$. In econometrics, $k+1$ was the trace of the hat matrix, or complexity (or dimension) of the model.

From this dimension VC, we deduce the so-called fundamental theorem of learning: if $\mathcal{M}$ is a class of dimension $d = VC(\mathcal{M})$, then there are positive constants $\underline{C}$ and $\overline{C}$ such as the sample complexity for $\mathcal{M}$ to be PAC-learnable satistfies:

$$\underline{C}\epsilon^{-1}(d + \log[\delta^{-1}]) \leq n_{\mathcal{M}}(\epsilon, \delta) \leq \overline{C}\epsilon^{-1}(d\log[\epsilon^{-1}] + \log[\delta^{-1}])$$

The link between the notion of learning (as defined in Vailiant (1984)) and the VC dimension was clearly established in Blumer *et al.* (1989).

Nevertheless, while the work of Vladimir Vapnik and Alexey Chervonenkis is considered to be the foundation of statistical learning, Thomas Cover's work in the 1960s and 1970s should also be mentioned, in particular Cover (1965) on the capacities of linear models, and Cover & Hart (1967) on learning in the context of the algorithm of the k-nearest neighbors. These studies have linked learning, information theory (with the textbook Cover & Thomas (1991)), complexity and statistics. Other authors have subsequently brought the two communities closer together, in terms of learning and statistics. For example, Halbert White proposed to see neural networks in a statistical context in White (1989), going so far as to state that "*learning procedures used to train artificial neural*

[4] Where the indicator $1_{\pm}$ does not take values 0 or 1 (like the classical 1 indicator function), but -1 and $+1$ ("negative" and "positive", as in the decision testing framework).

networks are inherently statistical techniques. It follows that statistical theory can provide considerable insight into the properties, advantages, and disadvantages of different network learning methods". This turning point in the late 1980s will anchor learning theory in a probabilistic context.

Objective and Loss Function

These choices (the objective and the loss function) are essential, and very dependent on the problem under consideration. Let us begin by describing a historically important model, Rosenblatt's (1958) "perceptron", introduced into classification problems, where $y \in \{-1, +1\}$, inspired by McCulloch & Pitts (1943). We have data $\{(y_i, x_i)\}$, and we will iteratively build a set of $m_k(\cdot)$, models, where at each step, we will learn from the errors of the previous model. In the perceptron, a linear model is considered so that :

$$m(x) = 1_{\pm}(\beta_0 + x^T\beta \geq 0) = \begin{cases} +1 \text{ if } \beta_0 + x^T\beta \geq 0 \\ -1 \text{ if } \beta_0 + x^T\beta < 0 \end{cases}$$

Here β coefficients are often interpreted as "weights" assigned to each of the explanatory variables. We give ourselves initial weights $(\beta_0^{(0)}, \beta^{(0)})$, which we will update incorporating the prediction error made, between y_i and its prediction $\hat{y}_i^{(k)}$:

$$\hat{y}_i^{(k)} = m^{(k)}(x_i) = 1_{\pm}(\beta_0^{(k)} + x^T\beta^{(k)} \geq 0),$$

with, in the case of the perceptron:

$$\beta_j^{(k+1)} = \beta_j^{(k)} + 1(y \neq \hat{y}^{(k)})^T x_j$$

Here $\ell(y, y') = 1(y \neq y')$ is a loss function, which will allow to give a price to an error made, by predicting $y' = m(x)$ and observing y. For a regression problem, we can consider a quadratic error ℓ_2, such that $\ell(y, m(x)) = (y - m(x))^2$ or the absolute value ℓ_1, with $\ell(y, m(x)) = |y - m(x)|$. Here, for our classification problem, we used a mis-qualification indicator (we could discuss the symmetry of this loss function, suggesting that a false positive costs as much as a false negative). Once this loss function has been specified, we recognize in the problem previously described a gradient descent, and we see that we are trying to solve:

$$m^\star(\mathrm{x}) = \underset{m \in \mathcal{M}}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, m(x_i)\right)\right\} \tag{6}$$

for a predefined set of predictors $\mathcal{M}$. Here, the machine learning problem is mathematically formulated as an optimization problem, whose solution determines a set of model parameters (if the $\mathcal{M}$ family is described by a set of parameters – which can be coordinates in a functional database). We can note $\mathcal{M}_0$ the space of the hyperplanes of $\mathbb{R}^p$ in the sense that:

$$m \in \mathcal{M}_0 \text{ means } m(x) = \beta_0 + \beta^T x \text{ for some } \beta \in \mathbb{R}^p$$

generating the class of linear predictors. We will then have the estimator that minimizes the empirical risk. Some of the recent work in statistical learning aims to study the properties of the estimator $\hat{m}^\star$, known as "oracle", in a family of $\mathcal{M}$:

$$\hat{m}^\star = \underset{\hat{m} \in \mathcal{M}}{argmin}\{\mathcal{R}(\hat{m}, m)\}.$$

This estimator is, of course, impossible to obtain because it depends on m the real model, obviously unknown.

But let's come back a little bit more to these loss functions. A loss function ℓ is a symmetric function $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, that satisfies the triangular inequality, and such that $\ell(x, y) = 0$ if and only if $x = y$. The associated norm $\|\cdot\|$ satisfies $\ell(x, y) = \| x - y \| = \ell(x - y, 0)$ (using the fact that $\ell(x, y + z) = \ell(x - y, z)$ - we will review this fundamental property later).

For a quadratic loss function, it should be noted that we can have a particular interpretation of this problem, since:

$$\overline{y} = \underset{m \in \mathbb{R}}{argmin}\left\{\sum_{i=1}^{n} \frac{1}{n}[y_i - m]^2\right\} = \underset{m \in \mathbb{R}}{argmin}\left\{\sum_{i=1}^{n} \ell_2\,(y_i, m)\right\},$$

where ℓ_2 is the usual quadratic distance If we assume – as we did in econometrics – that there is an underlying probabilistic model, and observe that:

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

it should be noted that what we are trying to obtain here, by solving the problem (6) by taking the ℓ_2 norm, is an approximation (in a given functional space, $\mathcal{M}$) of the conditional expectation $x \mapsto \mathbb{E}[Y|X=x]$. Another particularly interesting loss function is the loss ℓ_1, $\ell_1(y,m) = |y-m|$. It should be recalled that:

$$median(y) = \underset{m\in\mathbb{R}}{argmin}\left\{\sum_{i=1}^{n} \ell_1(y_i, m)\right\}.$$

The optimization problem:

$$\hat{m}^\star = \underset{m\in\mathcal{M}_0}{argmin}\left\{\sum_{i=1}^{n} |y_i - m(x_i)|\right\}$$

is obtained in econometrics by assuming that the conditional law of Y follows a (shifted) Laplace law centered on $m(x)$, and by maximizing the likelihood (log) (the sum of the absolute values of the errors corresponds to the log-reasonableness of a Laplace law). It should also be noted that if the conditional law of Y is symmetrical with respect to 0, the median and the mean coincide if this loss function is rewritten:

$\ell_1(y,m) = |(y-m)(1/2 - 1_{y\leq m})|$,

a generalization can be obtained for $\tau \in (0,1)$:

$$\hat{m}_\tau^\star = \underset{m\in\mathcal{M}_0}{argmin}\left\{\sum_{i=1}^{n} \ell_\tau^{\mathrm{q}}(y_i, m(x_i))\right\} \text{ with } \ell_\tau^q(x,y) = (x-y)(\tau - 1_{x\leq y})$$

is then the quantile regression of level τ (Koenker, 2003 ; Haultefœuille & Givord, 2014). Another loss function, introduced by Aigner *et al.* (1977) and analyzed in Waltrup *et al.* (2014), is the function associated with the notion of expectations:

$$\ell_\tau^{\mathrm{e}}(x,y) = (x-y)^2 \cdot |\tau - 1_{x\leq y}|$$

with $\tau \in [0,1]$. We see the parallel with the quantile function:

$$\ell_\tau^q(x,y) = |x-y| \cdot |\tau - 1_{x\leq y}|$$

Koenker & Machado (1999) and Yu & Moyeed (2001) also noted a link between this condition and the search for maximum likelihood when Y's conditional law follows an asymmetric Laplace law. In connection with this approach, Gneiting (2011) introduced the notion of "*ellicitable statistics*" – or "*ellicitable measurement*" in its probabilistic (or distributional) version: a statistic T will be said to be "*ellicitable*" if there is a loss function $\ell: \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ such that:

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

The mean (mathematical expectation) is thus ellicitable by the quadratic distance, ℓ_2, while the median is ellicitable by the distance ℓ_1. According to Gneiting (2011), this property is essential for obtain predictions and forecasts. There may then be a strong link between measures associated with probabilistic models and loss functions. Finally, Bayesian statistics provide a direct link between the form of the a priori law and the loss function, as studied by Berger (1985) and Bernardo & Smith (2000). We will come back to the use of these different norms in the section on penalization.

Boosting and Sequential Learning

As we have seen before, modelling here is based on solving an optimization problem, and solving the problem described by equation (6) is all the more complex because the functional space $\mathcal{M}$ is large. The idea of boosting, as introduced by Shapire & Freund (2012), is to learn, slowly, from the errors of the model, in an iterative way. In the first step, we estimate a model m_1 for y, from X, which will give an error ε_1. In the second step, we estimate a model m_2 for ε_1, from X, which will give an error ε_2, etc. We will then retain as a model, after k iterations:

$$m^{(k)}(\cdot) = \underset{\sim \mathbf{y}}{m_1}(\cdot) + \underset{\sim \boldsymbol{\varepsilon}_1}{m_2}(\cdot) + \underset{\sim \boldsymbol{\varepsilon}_2}{m_3}(\cdot) + \cdots + \underset{\sim \boldsymbol{\varepsilon}_{k-1}}{m_k}(\cdot) = m^{(k-1)}(\cdot) + m_k(\cdot) \tag{7}$$

Here, the error ε is seen as the difference between y and the model $m(x)$, but it can also be seen as the gradient associated with the quadratic loss function. Formally, ε can be seen as $\nabla \ell$ in a more general context (here we find an interpretation that reminds us of residuals in generalized linear models).
Equation (7) can be seen as a descent of the gradient, but written in a dual way. The problem will then be rewritten as an optimization problem:

$$m^{(k)} = m^{(k-1)} + \underset{h \in \mathcal{H}}{argmin} \left\{ \sum_{i=1}^{n} \ell \left(y_i - m^{(k-1)}(\mathrm{x}_i), h(x_i) \right) \right\} \tag{8}$$

where the trick is to consider a relatively simple space $\mathcal{H}$ (we will speak of "weak learner"). Classically, functions in set $\mathcal{H}$ are step-functions (which will be found in classification and regression trees) called "stumps". To ensure that learning is indeed slow, it is not uncommon to use a shrinkage parameter, and instead of setting, for example, $\varepsilon_1 = y - m_1(x)$, we will set $\varepsilon_1 = y - \alpha \cdot m_1(x)$ with $\alpha \in [0,1]$. It should be noted that it is because a non-linear space is used for $\mathcal{H}$ and learning is slow, that this algorithm works well. In the case of the Gaussian linear model, remember that the residuals $\hat{\varepsilon} = y - X\hat{\beta}$ are orthogonal to the explanatory variables, X, and it is then impossible to learn from our errors. The main difficulty is to stop in time, because after too many iterations, it is no longer the m function that is approximated, but the noise. This problem is called over-fitting.

This presentation has the advantage of having a heuristic reminiscent of an econometric model, by iteratively modelling the residuals by a (very) simple model. But this is often not the presentation used in the learning literature, which places more emphasis on an optimization algorithm heuristic (and gradient approximation). The function is learned iteratively, starting from a constant value,

$$m^{(0)} = \underset{m \in \mathbb{R}}{argmin} \left\{ \sum_{i=1}^{n} \ell \left(y_i, m \right) \right\}$$

then we consider the following learning procedure:

$$m^{(k)} = m^{(k-1)} + \underset{h \in \mathcal{H}}{argmin} \sum_{i=1}^{n} \ell \left(y_i, m^{(k-1)}(x_i) + h(x_i) \right), \tag{9}$$

which can be written, if $\mathcal{H}$ is a set of differentiable functions:

$$m^{(k)} = m^{(k-1)} - \gamma_k \sum_{i=1}^{n} \nabla_{m^{(k-1)}} \ell \left(y_i, m^{(k-1)}(x_i) \right), \tag{10}$$

where :

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

To better understand the relationship with the approach described above, at step k, pseudo-residuals are defined by setting

$$r_{i,k} = -\frac{\partial \ell(y_i, m(x_i))}{\partial m(x_i)}\Big|_{m(\mathrm{x}) = m^{(k-1)}(x)} \quad \text{where } i = 1, \cdots, n.$$

A simple model is then sought to explain these pseudo-residuals according to the explanatory variables x_i, i.e. $r_{i,k} = h^\star(\mathrm{x}_i)$, where $h^\star \in \mathcal{H}$. In a second step, we look for an optimal multiplier by solving:

$$\gamma_k = \underset{\gamma \in \mathbb{R}}{argmin} \left\{ \sum_{i=1}^{n} \ell \left(y_i, m^{(k-1)}(x_i) + \gamma h^\star(x_i) \right) \right\}$$

then update the model by setting $m_k(x) = m_{k-1}(x) + \gamma_k h^\star(x)$. More formally, we move from equation (8) – which clearly shows that we are building a model on residuals – to equation (9) – which will then be translated as a gradient calculation problem – noting that $\ell(y, m + h) = \ell(y - m, h)$. Classically, class $\mathcal{H}$ of functions consists in regression trees. It is also possible to use a form of penalty by setting $m_k(x) = m_{k-1}(x) + \nu\gamma_k h^\star(x)$, with $\nu \in (0,1)$. But let's go back a little further – in our next post – on the importance of penalization before discussing the numerical aspects of optimization.

Penalization and Variable Selection

One important concept in econometrics is Ockham's razor – also known as the law of parsimony (*lex parsimoniae*) – which can be related to abductive reasoning. Akaike's criterion was based on a penalty of likelihood taking into account the complexity of the model (the number of explanatory variables retained). If in econometrics, it is customary to maximize the likelihood (to build an asymptotically unbiased estimator), and to judge the quality of the ex-post model by penalizing the likelihood, the strategy here will be to penalize ex-ante in the objective function, even if it means building a biased estimator. Typically, we will build:

$$\left(\hat{\beta}_{0,\lambda}, \hat{\beta}_{\lambda}\right) = argmin\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right) + \lambda\, penalty(\beta)\right\}, \tag{11}$$

where the penalty function will often be a standard $\|\cdot\|$ chosen *a priori*, and a penalty parameter λ (we find in a way the distinction between AIC and BIC if the penalty function is the complexity of the model – the number of explanatory variables retained). In the case of the ℓ_2, standard, we find the ridge estimator, and for the ℓ_1, standard, we find the lasso estimator (« *Least Absolute Shrinkage and Selection Operator* »)[5]. The penalty previously used involved the number of degrees of freedom of the model, so it may seem surprising to use $\|\beta\|_{\ell_2}$ as in the ridge regression. However, we can envisage a Bayesian vision of this penalty. It should be recalled that in a Bayesian model:

$$\underbrace{\mathbb{P}[\theta|y]}_{posterior} \propto \underbrace{\mathbb{P}[y|\theta]}_{likelihood} \cdot \underbrace{\mathbb{P}[\theta]}_{prior} \text{ or } \log\mathbb{P}[\theta|y] \propto \underbrace{\log\mathbb{P}[y|\theta]}_{log-likelihood} + \log \underbrace{\mathbb{P}[\theta]}_{penalty}$$

In a Gaussian linear model, if we assume that the a priori law of θ follows a centred Gaussian distribution, we find a penalty based on a quadratic form of the components of θ.

Before going back in detail to these two estimators, obtained using the $\ell_1$or ℓ_2 norms, let us return for a moment to a very similar problem: the best choice of explanatory variables. Classically (and this will be even more true in large dimension), we can have a large number of explanatory variables, p, but most are just noise, in the sense that $\beta_j = 0$ for a large number of j's. This perperty is called « sparsity ». Let s be the number of (really) relevant covariates, $s = \#\mathcal{S}$ with $\mathcal{S} = \{j = 1, \cdots, p; \beta_j \neq 0\}$. If $X_{\mathcal{S}}$ denotes the matrix composed of the relevant variables (in columns), then we assume that the real model is of the form $y = x_{\mathcal{S}}{}^T\beta_{\mathcal{S}} + \varepsilon$. ntuitively, an interesting estimator would then be $\hat{\beta}_{\mathcal{S}} = [x_{\mathcal{S}}{}^T X_{\mathcal{S}}]^{-1} X_{\mathcal{S}} y$, but this estimator is only theoretical because the set $\mathcal{S}$ is unknown, here. This estimator can actually be seen as the oracle estimator mentioned above. One may then be tempted to solve:

$$\left(\hat{\beta}_{0,s}, \hat{\beta}_{s}\right) = argmin\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right)\right\}, \text{subject to } \#\mathcal{S} = s$$

This problem was introduced by Foster & George (1994) using the ℓ_0 notation. More precisely, let us define here the following three norms, where:

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

Table C1-I
Constrained Optimization and Regularization

	Contrained optimization	Penalty	
Best subgroup	$\underset{\beta;\|\beta\|_{\ell_0}\leq s}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right)\right\}$	$\underset{\beta,\lambda}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right) + \lambda\|\beta\|_{\ell_0}\right\}$	$(\ell 0)$
Lasso	$\underset{\beta;\|\beta\|_{\ell_1}\leq s}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right)\right\}$	$\underset{\beta,\lambda}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right) + \lambda\|\beta\|_{\ell_1}\right\}$	$(\ell 1)$
Ridge	$\underset{\beta;\|\beta\|_{\ell_2}\leq s}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right)\right\}$	$\underset{\beta,\lambda}{argmin}\left\{\sum_{i=1}^{n} \ell\left(y_i, \beta_0 + x^T\beta\right) + \lambda\|\beta\|_{\ell_2}\right\}$	$(\ell 2)$

Let us consider the optimization problems in Table C1-1. If we consider the classical problem where the quadratic norm is used for ℓ, the two problems of the equation $(\ell 1)$ of Tableau C1-1 are equivalent, in the sense that, for any solution$(\beta^\star, s)$ to the left problem, there is $\lambda^\star$ such that $(\beta^\star, \lambda^\star)$ is the solution of the right problem; and vice versa.

[5] Term « lasso » can be seen as a reference to Breiman (1995), which suggested to use a "*garrote"* for variable selection.

The result is also true for problems $(\ell 2)$[6]. These are indeed convex problems. On the other hand, the two problems $(\ell 0)$ are not equivalent: if for $(\beta^\star, \lambda^\star)$ solution of the right problem, there is a $s^\star$ such that $\beta^\star$ is solution of the left problem, the reverse is not true. More generally, if one wants to use a ℓ_p norm, sparsity is obtained if $p \leq 1$ whereas you need $p \geq 1$ to have the convexity of the optimization program.

One may be tempted to resolve the penalized program $(\ell 0)$ directly, as suggested by Foster & George (1994). Numerically, it is a complex combinatorial problem in large dimension (Natarajan, 1995, notes that it is a NP-difficult problem), but it is possible to show that if $\lambda \sim \sigma^2 \log(p)$, then:

$$\mathbb{E}([\mathrm{x}^T\,\hat{\beta} - \mathrm{x}^T\beta_0]^2) \leq \underbrace{\mathbb{E}(\mathrm{x}_S{}^T\,\hat{\beta}_S - \mathrm{x}^T\beta_0]^2)}_{=\sigma^2 \# S} \cdot (4\log p + 2 + o(1)).$$

Observe that in this case

$$\hat{\beta}_{\lambda,j}^{\text{sub}} = \begin{cases} 0 \text{ if } j \notin \mathcal{S}_\lambda \\ \hat{\beta}_j^{\text{ ols}} \text{ if } j \in \mathcal{S}_\lambda, \end{cases}$$

where $\mathcal{S}_\lambda$ refers to all non-zero coordinates when solving $(\ell 0)$.

Problem $(\ell 2)$ s strictly convex if ℓ is the quadratic norm, in other words, the Ridge estimator is always well defined, with in addition an explicit form for the estimator:

$$\hat{\beta}_\lambda^{ridge} = (X^T X + \lambda \mathbb{I})^{-1} X^T y = (X^T X + \lambda \mathbb{I})^{-1} (X^T X)\, \hat{\beta}^{\,ols}$$

Therefore, it can be deduced that:

$$bias[\hat{\beta}_\lambda^{ridge}] = -\lambda [X^T X + \lambda \mathbb{I}]^{-1}\ \hat{\beta}^{\,ols} \text{ and } \mathrm{Var}[\hat{\beta}_\lambda^{ridge}] = \sigma^2 [X^T X + \lambda \mathbb{I}]^{-1} X^T X [X^T X + \lambda \mathbb{I}]^{-1}.$$

With a matrix of orthonormal explanatory variables (i.e. $X^T X = \mathbb{I}$), the expressions can be simplified, and a shrinkage can clearly be observed:

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

Observe further that $\mathrm{Var}[\hat{\beta}_\lambda^{ridge}] < \mathrm{Var}[\hat{\beta}^{\,ols}]$. And because:

$$mse[\hat{\beta}_\lambda^{ridge}] = \frac{k\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\,\beta^T\beta$$

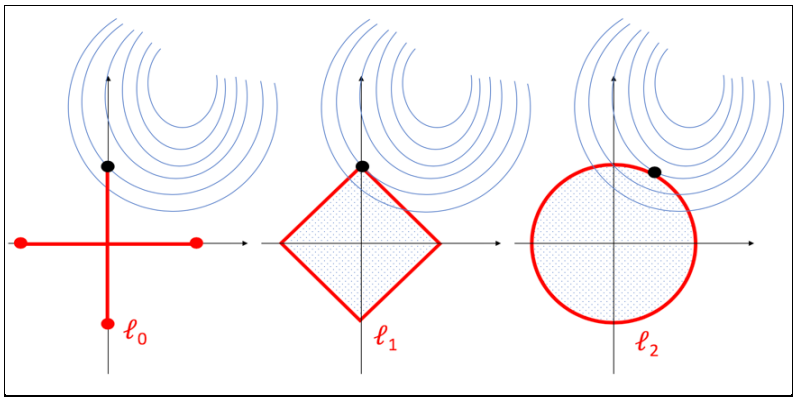
we obtain an optimal value forλ: $\lambda^\star = k\sigma^2/\beta^T\beta$.

On the other hand, if ℓ is no longer the quadratic norm but the ℓ_1 norm, the problem $(\ell 1)$ is not always strictly convex, and in particular, the optimum is not always unique (for example if $\mathrm{X}^T\mathrm{X}$ is singular). But if ℓ is strictly convex, then predictions $X\hat{\beta}$ will be unique. It should also be noted that two solutions are necessarily consistent in terms of sign of coefficients: it is not possible to have $\hat{\beta}_j < 0$ for one solution and $\hat{\beta}_j > 0$ for another. From a heuristic point of view, the program $(\ell 1)$ is interesting because it allows to obtain in many cases a corner solution, which corresponds to a problem resolution of type $(\ell 0)$ – as shown visually on Figure C1-II.

Tibshirani & Wasserman (2016) goes back at length on the geometry of the solutions) but as Candès & Plan (2009) notes, under minimal assumptions guaranteeing that the predictors are not strongly correlated, the Lasso obtains a quadratic error almost as good as if we had an oracle providing perfect information on the set of β_j's that are not zero. With some additional technical hypotheses, it can be shown that this estimator is "sparsistant" in the sense that the support of $\hat{\beta}_\lambda^{lasso}$ is that of β, in other words Lasso has made it possible to select variables (more discussions on this point can be obtained in Hastie *et al.*, 2016).
More generally, it can be shown that $\hat{\beta}_\lambda^{lasso}$ is a biased estimator, but may be of sufficiently low variance that the mean square error is lower than the one we had using least squares.

[6] For $(\ell 1)$, if there is a theoretical equivalence, numerical issues could be experience because of non-unicity of solutions.

Figure C1-II
Penalty Based on Norms ℓ_0, ℓ_1 and ℓ_2 (of β) Respectively (from Hastie *et al.*, 2016)



In-Sample, Out-of-Sample and Cross-Validation

These techniques seem intellectually interesting, but we have not yet discussed the choice of the penalty parameter λ. But this problem is actually more general, because comparing two parameters $\hat{\beta}_{\lambda_1}$ and $\hat{\beta}_{\lambda_2}$ is actually comparing two models. In particular, if we use a Lasso method, with different thresholds λ we compare models that do not have the same dimension. Previously, we have addressed the problem of model comparison from an econometric perspective (by penalizing overly complex models). In the learning literature, judging the quality of a model on the data used to construct it does not make it possible to know how the model will behave on new data. This is the so-called "generalization" problem. The traditional approach then consists in separating the sample (size n) into two parts: a part that will be used to train the model (the training database, in-sample, size m) and a part that will be used to test the model (the testing database, out-of-sample, size $n-m$). The latter then makes it possible to measure a real predictive risk. Suppose that the data are generated by a linear model $y_i = x_i{}^T\beta_0 + \varepsilon_i$ where ε_i are independent and centered law achievements. The empirical quadratic risk in-sample is here:

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left([x_i{}^T\hat{\beta} - x_i{}^T\beta_0]^2\right) = \mathbb{E}([x_i{}^T\hat{\beta} - x_i{}^T\beta_0]^2)$$

for any observation i. Assuming that residuals ε have a Gaussian distribution, then we can show that this risk is worth $\sigma^2 trace(\Pi_X)/m$ is $\sigma^2 p/m$. On the other hand, the empirical out-of-sample quadratic risk is here:

$$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2)$$

where x is a new observation, independent of the others. It can be obtained that:

$$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2|x) = \sigma^2 x^T(X^TX)^{-1}x$$

and by integrating with respect to x:

$$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2) = \mathbb{E}(\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2|x)) = \sigma^2 trace(\mathbb{E}[xx^T]\mathbb{E}[(X^TX)^{-1}])$$

The expression is then different from that obtained in-sample, and using the Groves & Rothenberg (1969) increase, we can show that:

$$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2) \geq \sigma^2\frac{p}{m},$$

which is pretty intuitive, when we start thinking about it. Except in some simple cases, there is no simple (explicit) formula. Note, however, that if $X \sim \mathcal{N}(0, \sigma^2\mathbb{I})$, then x^Tx follows a Wishart law, and it can be shown that

$$\mathbb{E}([x^T\hat{\beta} - x^T\beta_0]^2) = \sigma^2\frac{p}{m-p-1}$$

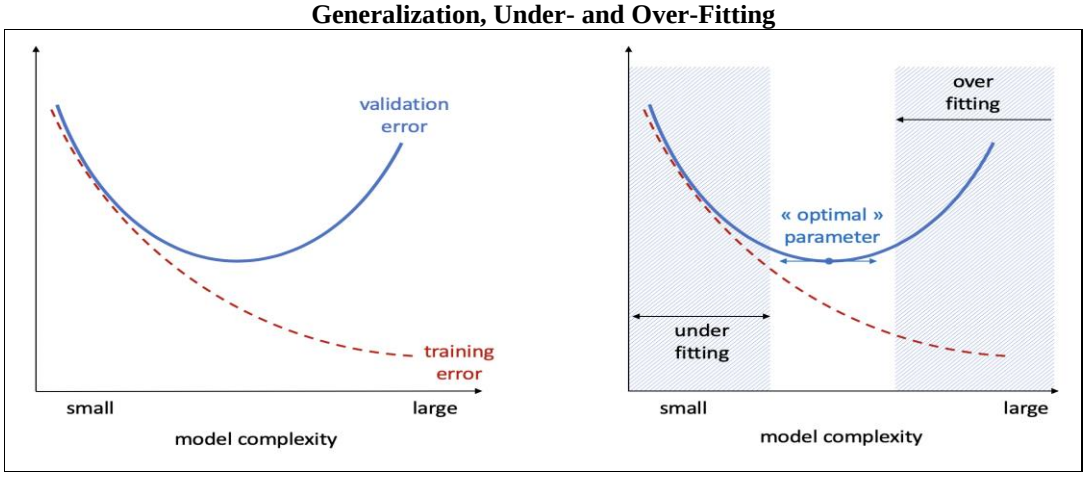
If we now look at the empirical version: if $\hat{\beta}$ is estimated on the first m observations,

$$\hat{\mathcal{R}}^{\text{IS}} = \sum_{i=1}^{m} [\, y_i - x_i{}^T \hat{\beta}]^2 \text{ and } \hat{\mathcal{R}}^{OS} = \sum_{i=m+1}^{n} [\, y_i - x_i{}^T \hat{\beta}]^2$$

and as Leeb (2008) noted, $\hat{\mathcal{R}}^{\text{IS}} - \hat{\mathcal{R}}^{\text{OS}} \approx 2 \cdot \nu$ where ν represents the number of degrees of freedom, which is not unlike the penalty used in the Akaike test.

Figure C1-IV shows the respective evolution of $\hat{\mathcal{R}}^{\text{IS}}$ and $\hat{\mathcal{R}}^{OS}$ according to the complexity of the model (number of degrees in a polynomial regression, number of nodes in splines, etc). The more complex the model, the more $\hat{\mathcal{R}}^{\text{IS}}$ will decrease (this is the red botted curve, below). But that's not what we're interested in here: we want a model that predicts well on new data (i. e. out-of-sample). As Figure C1-III shows, if the model is too simple, it does not predict well (as it does with in-sample data). But what we can see is that if the model is too complex, we are in a situation of "overlearning": the model will start to model the noise.

Figure C1-III
Generalization, Under- and Over-Fitting



Instead of splitting the database in two, with some of the data that will be used to calibrate the model and some to study its performance, it is also possible to use cross-validation. To present the general idea, we can go back to the "jackknife", introduced by Quenouille (1949) (and formalized by Quenouille (1956) and Tukey (1958)) relatively used in statistics to reduce bias. Indeed, if we assume that $\{y_1, \cdots, y_n\}$ is a sample drawn according to a law F_θ, and that we have an estimator $T_n(y) = T_n(y_1, \cdots, y_n)$, , but that this estimator is biased, with $\mathbb{E}[T_n(Y)] = \theta + O(n^{-1})$, it is possible to reduce the bias by considering:

$$\tilde{T}_n\,(\mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n} T_{n-1}\,(\mathrm{y}_{(i)}) \text{ where } y_{(i)} = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n).$$

It can then be shown that $\mathbb{E}[\tilde{T}_n\,(Y)] = \theta + O(n^{-2})$. The idea of cross-validation is based on the idea of building an estimator by removing an observation. Since we want to build a predictive model, we will compare the forecast obtained with the estimated model, and the missing observation:

$$\hat{\mathcal{R}}^{\text{CV}} = \frac{1}{n}\sum_{i=1}^{n} \ell\,(y_i, \hat{m}_{(i)}\,(x_i))$$

We will speak here of the "leave-one-out" (loocv) method.

This technique reminds us of the traditional method used to find the optimal parameter in exponential smoothing methods for time series. In simple smoothing, we will construct a forecast from a time series as ${}_t\hat{y}_{t+1} = \alpha \cdot {}_{t-1}\hat{y}_t + (1-\alpha) \cdot y_t$, where $\alpha \in [0,1]$, and we will consider as "optimal":

$$\alpha^\star = \underset{\alpha\in[0,1]}{argmin}\left\{\sum_{t=2}^{T} \ell\,({}_{t-1}\hat{y}_t, y_t)\right\},$$

as described in Hyndman *et al.* (2009).

The main problem with the leave-one-out method is that it requires calibration of n models, which can be problematic in large dimensions. An alternative method is cross validation over k-blocks (called « *k-fold cross validation* ») which consists in using a partition of $\{1,\cdots,n\}$ in k groups (or blocks) of the same size, $\mathcal{I}_1,\cdots,\mathcal{I}_k$, and let us note $\mathcal{I}_{\overline{j}} = \{1,\cdots,n\} \setminus \mathcal{I}_j$. By noting $\hat{m}_{(j)}$ built on the sample $\mathcal{I}_{\overline{j}}$, we then set:

$$\hat{\mathcal{R}}^{k-\text{CV}} = \frac{1}{k}\sum_{j=1}^{k} \mathcal{R}_j \text{ where } \mathcal{R}_j = \frac{k}{n}\sum_{i\in\mathcal{I}_j} \ell\,(y_i, \hat{m}_{(j)}\,(\mathrm{x}_i)).$$

Standard cross-validation, where only one observation is removed each time ("leave-one-out"), is a special case, with $k = n$. Using $k = 10$ has a double advantage over $k = n$: (1) the number of estimates to be made is much smaller, 10 rather than n ; (2) the samples used for estimation are less similar and therefore less correlated to each other, which tends to avoid excess variance, as recalled by James *et al.* (2013).

Another alternative is to use boosted samples. Let $\mathcal{I}_b$ be a sample of size n obtained by drawing with replacement in $\{1,\cdots,n\}$ to know which observations(y_i, x_i) will be kept in the learning population (at each draw). Note $\mathcal{I}_{\overline{b}} = \{1,\cdots,n\} \setminus \mathcal{I}_b$. If $\hat{m}_{(b)}$ is built on sample $\mathcal{I}_b$, we then set:

$$\hat{\mathcal{R}}^{\text{B}} = \frac{1}{B}\sum_{b=1}^{B} \mathcal{R}_b \text{ where } \mathcal{R}_b = \frac{n_{\overline{b}}}{n}\sum_{i\in\mathcal{I}_{\overline{b}}} \ell\,(y_i, \hat{m}_{(b)}\,(x_i)),$$

where $n_{\overline{b}}$ is the number of observations that have not been kept in $\mathcal{I}_b$. It should be noted that with this technique, on average $e^{-1} \sim 36.7\%$ of the observations do not appear in the boosted sample, and we find an order of magnitude of the proportions used when creating a calibration sample, and a test sample. In fact, as Stone (1977) had shown, the minimization of AIC is to be compared to the cross-validation criterion, and Shao (1997) showed that the minimization of BIC corresponds to k-fold cross-validation, with $k = n/\log n$.

Bibliography

Ahamada, I. & Flachaire, E. (2011). *Non-Parametric Econometrics*. Oxford: Oxford University Press.
Aigner, D., Lovell, C. A. J & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1), 21–37.
https://doi.org/10.1016/0304-4076(77)90052-5
Aldrich, J. (2010). The Econometricians' Statisticians, 1895-1945. *History of Political Economy*, 42(1), 111–154.
https://doi.org/10.1215/00182702-2009-064
Altman, E., Marco, G. & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience).*Journal of Banking & Finance*, 18(3), 505–529.
https://doi.org/10.1016/0378-4266(94)90007-8
Angrist, J. D. & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, 114(2), 533–575.
https://doi.org/*10.1162/003355399556061*
Angrist, J. D. & Pischke, J. S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspective*, 24(2), 3–30.
https://doi.org/10.1257/jep.24.2.3
Angrist, J. D. & Pischke, J. S. (2015). Mastering Metrics. Princeton University Press.
Angrist, J. D. & Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4), 979–1014.
https://doi.org/10.2307/2937954
Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, 177–187.
https://doi.org/10.1007/978-3-7908-2604-3_16
Bajari, P., Nekipelov, D., Ryan, S. P. & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, 105(5), 481–485.
https://doi.org/10.1257/aer.p20151021
Bazen, S. & Charni, K. (2015). Do earnings really decline for older workers? AMSE, *Working Paper* 2015-11.
https://halshs.archives-ouvertes.fr/halshs-01119425
Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.
Belloni, A., Chernozhukov, V. & Hansen, C. (2010). Inference for High-Dimensional Sparse Econometric Models. *Advances in Economics and Econometrics*, 245–295
https://doi.org/10.1017/CBO9781139060035.008
Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, 80(6), 2369–2429.
https://doi.org/10.3982/ECTA9626

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1), 289–300.
https://www.jstor.org/stable/2346101
Berger, J. O. (1985). *Statistical decision theory and Bayesian Analysis* (2nd ed.). New York, Berlin: Springer-Verlag.
Berk, R. A. (2008). *Statistical Learning from a Regression Perspective*. New York: Springer Verlag.
Berkson, J. (1944). Applications of the Logistic Function to Bioassay. *Journal of the American Statistical Association*, 39(227), 357–365.
https://doi.org/10.1080/01621459.1944.10500699
Berkson, J. (1951). Why I Prefer Logits to Probits. *Biometrics*, 7(4), 327–339.
https://doi.org/10.2307/3001655
Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian Theory*. New York: John Wiley.
Berndt, E. R. (1990). *The Practice of Econometrics: Classic and Contemporary*. Reading, Mass: Addison Wesley.
Bickel, P. J., Gotze, F. & van Zwet, W. (1997). Resampling Fewer than n Observations: Gains, Losses and Remedies for Losses. *Statistica Sinica*, 7, 1–31.
http://www3.stat.sinica.edu.tw/statistica/oldpdf/a7n11.pdf
Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York: Springer Verlag.
Blanco, A. Pino-Mejias, M., Lara, J. & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, 40(1), 356–364.
https://doi.org/10.1016/j.eswa.2012.07.051
Bliss, C. I. (1934). The method of probits. *Science*, 79(2037), 38–39.
https://doi.org/10.1126/science.79.2037.38
Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M.K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4), 929–965.
https://doi.org/10.1145/76359.76371
Breiman, L. Fiedman, J., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Londres: Chapman & Hall/CRC.
Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4), 373–384.
https://doi.org/10.1080/00401706.1995.10484371
Breiman, L. (2001a). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), 199–231.
https://doi.org/10.1214/ss/1009213726
Breiman, L. (2001b). Random forests. *Machine learning*, 45(1), 5–32.
https://doi.org/10.1023/A:1010933404324
Brown, L. D. (1986). *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Hayworth, CA, USA: Institute of Mathematical Statistics.
Bühlmann, P. & van de Geer, S. (2011). *Statistics for High Dimensional Data: Methods, Theory and Applications*. Heidelberg, New York: Springer Verlag.
Candès, E. & Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5), 2145–2177.
https://doi.org/10.1214/08-AOS653
Clarke, B. S., Fokoué, E. & Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. New York: Springer Verlag.
Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297.
https://doi.org/10.1023/A:1022627411411
Cover, T. M. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers*, 14(3), 326–334.
https://doi.org/10.1109/PGEC.1965.264137
Cover, T. M. & Hart, P. (1965). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.
https://doi.org/10.1109/TIT.1967.1053964
Cover, T. M. & Thomas, J. (1991). *Elements of Information Theory*. Wiley.
Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function 1989 *Mathematics of Control, Signals, and Systems*, 2(4), 303–314.
https://doi.org/10.1007/BF02551274
Darmois, G. (1935). Sur les lois de probabilités à estimation exhaustive. *Comptes Rendus de l'Académie des Sciences, Paris*, 200, 1265–1266.
Daubechies, I., Defrise, M. & De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11), 1413–1457
https://doi.org/10.1002/cpa.20042
Davison, A. C. (1997). *Bootstrap*. Cambridge: Cambridge University Press.
Davidson, R. & MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
Davidson, R. & MacKinnon, J. G. (2003). *Econometric Theory and Methods*. Oxford: Oxford University Press.
Duo, Q. (1993). *The Formation of Econometrics*. Oxford: Oxford University Press
Debreu, G. (1986). Theoretic Models: Mathematical Form and Economic Content. *Econometrica*, 54(6), 1259–1270.
https://doi.org/10.2307/1914299
Dhillon, P., Lu, Y. Foster, D. P. & Ungar, L. H. (2014). New Subsampling Algorithms for Fast Least Squares Regression. In: Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 26*. New York: Curran Associates.

http://papers.nips.cc/paper/5105-new-subsampling-algorithms-for-fast-least-squares-regression.pdf
Efron, B. & Tibshirani, R. (1993). *Bootstrap*. Londre : Chapman Hall CRC.
Engel, E. (1857). Die Productions- und Consumtionsverhältnisse des Königreichs Sachsen. *Statistisches Bureau des Königlich Sächsischen Ministeriums des Innern*.
Feldstein, M. & Horioka, C. (1980). Domestic Saving and International Capital Flows. *Economic Journal*, 90(358), 314–329. https://doi.org/10.2307/2231790
Flach, P. (2012). *Machine Learning*. Cambridge: Cambridge University Press.
Foster, D. P. & George, E. I. (1994). The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics*, 22(4), 1947–1975.
https://doi.org/10.1214/aos/1176325766
Friedman, J. H. (1997). Data Mining and Statistics: What's the Connection. *Proceedings of the 29th Symposium on the Interface between Computer Science and Statistics*.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.374.7489&rep=rep1&type=pdf
Frisch, R. & Waugh, F.V. (1933). Partial Time Regressions as Compared with Individual Trends. *Econometrica*. 1(4), 387–401.
https://doi.org/10.2307/1907330
Galton, Edgeworth, Frish, and prospects for quantile regression in Econometrics (1998). Conference on Principles of Econometrics, Madison.
Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.
https://doi.org/10.1198/jasa.2011.r10138
Givord, P. (2010). Méthodes économétriques pour l'évaluation de politiques publiques. *Economie & Prévision*, 204-205, 1–28.
https://www.cairn.info/revue-economie-et-prevision-2014-1-page-1.htm
Grandvalet, Y., Mariéthoz, J., & Bengio, S. (2005). Interpretation of SVMs with an application to unbalanced classification. *Advances in Neural Information Processing Systems*, 18.
https://papers.nips.cc/paper/2763-a-probabilistic-interpretation-of-svms-with-an-application-to-unbalanced-classification
Groves, T. & Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika*, 56(3), 690–691.
https://doi.org/10.1093/biomet/56.3.690
Haavelmo, T. (1944). The probability approach in econometrics, *Econometrica*, 12, iii–vi, 1–115.
https://doi.org/10.2307/1906935
Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Londres: Chapman & Hall/CRC.
Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer Verlag.
Hastie, T., Tibshirani, W. & Wainwright, M. (2015). *Statistical Learning with Sparsity*. Londres: Chapman CRC.
Hastie, T., Tibshiriani, R. & Tibshiriani, R. J. (2016). Extended Comparisons of Best Subset Selection, Forward Stepwise Selection and the Lasso.
https://arxiv.org/abs/1707.08692
Haultefœuille, X. (d') & Givord, P. (2014). La régression quantile en pratique. *Économie & Statistiques*, 471, 85–111. https://www.persee.fr/doc/estat_0336-1454_2014_num_471_1_10484
Hebb, D. O. (1949). The organization of behavior. New York: Wiley.
Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), 153–161.
https://doi.org/10.2307/1912352
Heckman, J. J., Tobias, J .L. & Vytlacil, E. (2003). Simple Estimators for Treatment Parameters in a Latent-Variable Framework. *The Review of Economics and Statistics*, 85(3), 748–755.
https://doi.org/10.1162/003465303322369867
Hendry, D F. & Krolzig, H.-M. (2001). Automatic Econometric Model Selection. London: Timberlake Press.
Herbrich, R., Keilbach, M., Graepel, T. Bollmann-Sdorra, P. & Obermayer, K. (1999). Neural Networks in Economics. In: Brenner, T. (Ed.), *Computational Techniques for Modelling Learning in Economics*, pp. 169–196. Boston, MA: Springer Verlag,.
https://doi.org/10.1007/978-1-4615-5029-7_7
Hoerl, A. E. (1962). Applications of ridge analysis to regression problems. *Chemical Engineering Progress*, 58(3), 54–59.
Hoerl, A. E. & Kennard, R. W. (1981). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *This Week's Citation Classic*, ISI.
Holland, P. (1986). *Statistics and causal inference. Journal of the American Statistical Association*, 81(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354
Hyndman, R., Koehler, A. B., Ord, J. K. & Snyder, R. D. (2009). Forecasting with Exponential Smoothing. Springer Verlag.
James, G., D. Witten, T. Hastie, & R. Tibshirani (2013). *An introduction to Statistical Learning*. Springer Series in Statistics.
Khashman, A. (2011). Credit risk evaluation using neural networks: Emotional versus conven- tional models. *Applied Soft Computing*, 11(8), 5477–5484.
https://doi.org/10.1016/j.asoc.2011.05.011
Kean, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1), 3–20.
https://doi.org/10.1016/j.jeconom.2009.09.003
Kleiner, A., Talwalkar, A., Sarkar, P. & Jordan, M. (2012). *The Big Data Bootstrap*.
https://arxiv.org/abs/1206.6415
Koch, I. (2013). *Analysis of Multivariate and High-Dimensional Data*. Cambridge: Cambridge University Press.
Koenker, R. (2003). *Quantile Regression*. Cambridge: Cambridge University Press.

Koenker, R. & Machado, J. (1999). Goodness of fit and related inference processes for quantile regression *Journal of the American Statistical Association*, 94(448), 1296–1309.
https://doi.org/10.1080/01621459.1999.10473882
Kolda, T. G. & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3), 455–500.
https://doi.org/10.1137/07070111X
Koopmans, T. C. (1957). *Three Essays on the State of Economic Science*. New York: McGraw-Hill.
Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. Springer Verlag.
Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
https://doi.org/10.2307/2529310
LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep Learning. *Nature*, 521, 436–444.
https://doi.org/10.1038/nature14539
Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, 14(3), 661–690.
https://doi.org/10.3150/08-BEJ127
Lemieux, T. (2006). The « Mincer Equation » Thirty Years After Schooling, Experience, and Earnings. In: Grossbard, S. (Ed.), *Jacob Mincer A Pioneer of Modern Labor Economics*, pp. 127–145; Springer Verlag.
https://doi.org/10.1007/0-387-29175-X_11
Li, J. & J. S. Racine (2006). *Nonparametric Econometrics*. Princeton: Princeton University Press.
Li, C., Li, Q., Racine, J. & Zhang, D. (2017). Optimal Model Averaging Of Varying Coefficient Models. *Department of Economics Working Papers 2017-01*, McMaster University.
https://doi.org/10.5705/ss.202017.0034
Lin, H. W., Tegmark, M. & Rolnick, D. (2016). Why does deep and cheap learning work so well?
https://arxiv.org/abs/1608.08225
Lucas, R. E. (1976). Econometric Policy Evaluation: A Critique. *Carnegie-Rochester Conference Series on Public Policy*, 19–46.
https://doi.org/10.1016/S0167-2231(76)80003-6
Mallows, C.L. (1973). Some Comments on C_p. *Technometrics*, 15(4), 661–675.
https://doi.org/10.2307/1267380
McCullogh, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4), 115–133.
https://doi.org/10.1007/BF02478259
Mincer, J. (1974). *Schooling, experience and earnings*. New York: Columbia University Press.
Mitchell, T. (1997). *Machine Learning*. New York: McGraw-Hill.
Morgan, J. N. & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58(302), 415–434.
https://doi.org/10.1080/01621459.1963.10500855
Morgan, M. S. (1990). *The history of econometric ideas*. Cambridge: Cambridge University Press.
Mohri, M., Rostamizadeh, A. & Talwalker, A. (2012). Foundations of Machine Learning. Cambridge, Mass: MIT Press.
Mullainathan, S. & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.
https://doi.org/10.1257/jep.31.2.87
Müller, M. (2011). Generalized Linear Models In*:* Gentle, J. E., Härdle, W. K. & Mori, Y. (Eds.), *Handbook of Computational* Statistics. Springer Verlag.
Murphy, K. R. (2012). *Machine Learning: a Probabilistic Perspective*. Cambridge, Mass: MIT Press.
Murphy, K. M. & Welch, F. (1990). Empirical Age-Earnings Profiles. *Journal of Labor Economics*, 8(2), 202–229.
https://doi.org/10.1086/298220
Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and its Applications*, 9(1), 141–2.
https://doi.org/10.1137/1109020
Natarajan, B. K. (1995). Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing* (SICOMP), 24(2), 227–234.
https://doi.org/10.1137/S0097539792240406
Nevo, A. & Whinston, M. D. (2010). Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference. *Journal of Economic Perspective*, 24(2), 69–82.
https://doi.org/10.1257/jep.24.2.69
Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles : Essai des principes. Mémoire de master, republié dans *Statistical Science*, 5(4), 463–472.
https://doi.org/10.1214/ss/1177012031
Nisbet, R., Elder, J. & Miner, G. (2011). *Handbook of Statistical Analysis and Data Mining Applications*. New York: Academic Press.
Okun, A. (1962). Potential GNP: Its measurement and significance. *Proceedings of the Business and Economics Section of the American Statistical Association*, 98–103.
https://milescorak.files.wordpress.com/2016/01/okun-potential-gnp-its-measurement-and-significance-p0190.pdf
Orcutt, G. H. (1952). Toward a partial redirection of econometrics. *Review of Economics and Statistics*, 34(3), 195–213.
https://doi.org/10.2307/1925626

Pagan, A. & Ullah, A. (1999). Nonparametric Econometrics. Themes in Modern Econometrics. Cambridge: Cambridge University Press.
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11), 559–572. https://doi.org/10.1080/14786440109462720
Platt, J. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 10, 61–74.
http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639
Portnoy, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity. *Annals of Statistics*, 16(1), 356–366.
https://doi.org/10.1214/aos/1176350710
Quenouille, M. H. (1949). Problems in Plane Sampling. *The Annals of Mathematical Statistics*, 20(3), 355–375.
https://doi.org/10.2307/2332914
Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*, 43(3-4), 353–360.
https://doi.org/10.2307/2332914
Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
https://doi.org/10.1007/BF00116251
Reiersøl, O. (1945). Confluence analysis of means of instrumental sets of variables. *Arkiv. for Mathematik, Astronomi Och Fysik*, 32.
Rosenbaum, P. & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, 41–55.
https://doi.org/10.21236/ada114514
Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
https://doi.org/10.1037/h0042519
Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5), 688–701.
https://doi.org/10.1037/h0037350
Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 44(1).
https://doi.org/10.1147/rd.33.0210
Schultz, H. (1930). *The Meaning of Statistical Demand Curves*. Chicago: University of Chicago.
Shai, S. S. & Shai, B. D. (2014). *Understanding Machine Learning From Theory to Algorithms*. Cambridge: Cambridge University Press.
Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 88(422), 486–494.
https://doi.org/10.2307/2290328
Shalev-Shwartz, S. & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press.
Shao, J. (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*, 7, 221–264.
Shapire, R.E. & Freund, Y. (2012). *Boosting*. Cambridge, Mass: MIT Press.
Silverman, B.W. (1986). *Density Estimation*. London: Chapman & Hall.
Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer.
Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B*, 39(1), 44–47.
https://doi.org/10.1111/j.2517-6161.1977.tb01603.x
Tam, K. Y. & Kiang, M. Y. (1992). Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. *Management Science*, 38(7), 926–947.
https://doi.org/10.1287/mnsc.38.7.926
Tan, H. (1995). *Neural-Network model for stock forecasting*. MSc Thesis, Texas Tech. University. *https://bit.ly/2UplmYu*
Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B.*, 58(1), 267–288.
https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
Tibshirani, R. & Wasserman, L. (2016). *A Closer Look at Sparse Regression*.
http://bit.ly/2FrGQ32
Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*, 4(4), 1035–1038.
Tinbergen, J. (1939). *Statistical Testing of Business Cycle Theories. Vol. 1: A Method and its Application to Investment activity; Vol. 2: Business Cycles in the United States of America*, 1919–1932. Geneva: League of Nations.
Tobin, J. (1958). Estimation of Relationship for Limited Dependent Variables. *Econometrica*, 26(1), 24–36.
https://doi.org/10.2307/1907382
Tropp, (2011). Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(1), 115–126.
https://doi.org/10.1142/S1793536911000787
Tseng, P. (2001). Convergence of a block coordinate descent for nondifferentiable minization. *Journal of Optimization Theory and Applications*, 109(3), 475–494.

https://doi.org/10.1023/A:1017501703105
Tufféry, S. (2001). *Data Mining and Statistics for Decision Making*. New York: Wiley Interscience.
Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, 29(2), 614–623. https://doi.org/10.1214/aoms/1177706647
Vailiant, L.G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134–1142. https://doi.org/10.1145/1968.1972
Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.
Vapnik, C. & Chervonenkis, A. (1971). On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*, 16(2), 264–280. https://doi.org/10.1137/1116025
Varian, H.R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3–28. https://doi.org/10.1257/jep.28.2.3
Vert, J. P. (2017). *Machine learning in computational biology*. ENSAE.
Waltrup, L. S., Sobotka, F., Kneib, T. & Kauermann, G. (2014). Expectile and quantile regression—David and Goliath? *Statistical Modelling*, 15, 433 – 456. https://doi.org/10.1177/1471082X14561155
Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26(4), 359–372. https://www.jstor.org/stable/25049340
Watt, J., Borhani, R. & Katsaggelos, A. (2016). *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge: Cambridge University Press.
White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. *Neural Computation*, 1(4), 425–464. https://doi.org/10.1162/neco.1989.1.4.425
Widrow, B. & Hoff, M. E. Jr. (1960). Adaptive Switching Circuits. *IRE WESCON Convention Record*, 4, 96–104. https://apps.dtic.mil/dtic/tr/fulltext/u2/241531.pdf
Wolpert, D. H. & Macready, W. G. (1997). No Free Lunch Theorems for Optimization, *IEEE Transactions on Evolutionary Computation*, 1(1), 67. https://doi.org/10.1109/4235.585893
Wolpert, David (1996). The Lack of A Priori Distinctions between Learning Algorithms, *Neural Computation*, 1341-1390. https://doi.org/10.1162/neco.1996.8.7.1341
Working, E. J. (1927). What Do Statistical "Demand Curves" Show? *The Quarterly Journal of Economics*, 41(2), 212–35. https://doi.org/10.2307/1883501
Yu, K. & Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4), 437–447. https://doi.org/10.1016/S0167-7152(01)00124-9
Zinkevich M. A., Weimer, M., Smola, A. & Li, L. (2010). Parallelized Stochastic Gradient. *Advances in neural information processing systems*, 2595–2603. https://papers.nips.cc/paper/4006-parallelized-stochastic-gradient-descent.pdf