Lecture notes on ordinary least squares

François Portier

January 10, 2018

This document is a summary of the 3 lectures given on nov. 29, dec. 13 and 20 for the course SD204. A proposition with a star (as Proposition* 1) indicates that it is secondary or too advanced regarding the main purpose of the course. This document is a first version. It would be much appreciated if you could inform me of any typos or mistakes you will find while reading it. A special page is dedicated to that in the course website (site pédagogique) but you can also send me an email at "francois.portier@gmail.com". Thank you in advance.

Contents

Notation

- $< \cdot, \cdot >$ is the usual inner product in $\mathbb{R}^d$. $\|\cdot\|$ is the Euclidean norm. The elements forming the canonical basis of $\mathbb{R}^d$ are denoted by $e_0, \ldots, e_{d-1}$.
- If $A \in \mathbb{R}^{n \times d}$ is a matrix, $A^T \in \mathbb{R}^{d \times n}$ is the transpose matrix, $\ker(A) = \{u \in \mathbb{R}^d \, : \, Au = 0\}$.
- For any set of vectors $(u_1, \ldots, u_d)$ in $\mathbb{R}^n$, $\text{span}(u_1, \ldots, u_d) = \{\sum_{k=1}^d \alpha_k u_k \, : \, (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d\}$. When A is a matrix $\text{span}(A)$ stands for the linear subspace generated by its columns.
- When A is a square invertible matrix, the inverse is denoted by A^{-1}. The Moore–Penrose inverse is denoted by A^+. The trace of A is given by $\text{tr}(A)$.
- The identity matrix in $\mathbb{R}^{d \times d}$ is I_d.
- When two random variables X and Y have the same distribution we write $X \sim Y$.
- When X_n is a sequence of random variables that converges in distribution (resp. in probability) to X, we write $X_n \rightsquigarrow X$ (resp. $X_n \xrightarrow{p} X$).

1 Definition of the OLS

We are interested in a regression problem with n observations and p covariates. Our goal is to predict an output variable with a linear combination of the p covariates. For $i = 1, \dots, n$, we observe $x_i = (x_{i,0}, \dots, x_{i,p}) \in \mathbb{R}^{p+1}$, the covariates, and $y_i \in \mathbb{R}$, the output. For notational convenience we will suppose that $x_{i,0} = 1$. This is to model the intercept of the regression in the same way as the parameters associated to the covariates. The OLS estimator is the vector $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^{p+1}$ such that

$$\hat{\boldsymbol{\theta}}_n \in \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (y_i - x_i^T \boldsymbol{\theta})^2. \tag{1}$$

It is useful to introduce the notations

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{1,0} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,0} & \dots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}, \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Also we will use the index $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\theta}}_{n,0}, \dots, \hat{\boldsymbol{\theta}}_{n,p})^T$. Then (1) becomes

$$\hat{\boldsymbol{\theta}}_n \in \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \|Y - X\boldsymbol{\theta}\|^2,$$

where $\|\cdot\|$ stands for the Euclidean norm. With the above formulation, the OLS has a nice geometric interpretation : $\hat{Y} = X\hat{\boldsymbol{\theta}}_n$ is the closest point to Y in the linear subspace $\operatorname{span}(X) \subset \mathbb{R}^n$ (where $\operatorname{span}(A)$ stands for the linear subspace generated by the columns of A). Using the Hilbert projection theorem ($\mathbb{R}^n$ is a Hilbert space, $\operatorname{span}(X)$ is a (closed) linear subspace of $\mathbb{R}^n$), $\hat{Y}$ is unique and is characterized by the normal equation:

$$< X, (Y - \hat{Y}) >= 0.$$

The vector $\hat{\boldsymbol{\theta}}_n$ is then such that

$$X^T X \, \hat{\boldsymbol{\theta}}_n = X^T Y. \tag{2}$$

Note that in contrast with $\hat{Y}$, which always exists and is unique, the vector $\hat{\boldsymbol{\theta}}_n$ is not uniquely defined without further assumption. For instance, take $u \in \ker(X)$ then $\hat{\boldsymbol{\theta}}_n + u$ verifies (2).

Definition 1. *The matrix $\hat{G}_n = X^T X/n$ is called the Gram matrix. Let $\hat{H}_X$ denote the orthogonal projector on* $\operatorname{span}(X)$.

When the Gram matrix is invertible, the OLS is well-defined. When it is not the case, then we have an infinity of solution for $\hat{\boldsymbol{\theta}}_n$.

Proposition 1. *The OLS estimator always exists. It is either*

(i) uniquely defined. This happens if and only if the Gram matrix is invertible, which is equivalent to $\ker(X) = \ker(X^T X) = \{0\}$. *In this case, the OLS has the following expression:*

$$\hat{\boldsymbol{\theta}}_n = (X^T X)^{-1} X^T Y.$$

(ii) or not unique, with an infinite number of solution. This happens if and only if $\ker(X) \neq \{0\}$. *In this case, the set of solution writes* $\boldsymbol{\theta} + \ker(X)$ *where* $\boldsymbol{\theta}$ *is a particular solution.*

Proof. The existence has already been shown using the Hilbert projection theorem. The linear system (2) has therefore a unique solution or an infinite number of solutions whether the Gram matrix is invertible or not. Hence it remains to show that that $\ker(X) = \ker(X^T X)$ which follows easily noting that when $u \in \ker(X^T X)$, $\|Xu\|^2 = 0$. □

When the OLS is not unique, the solution traditionally considered is

$$\hat{\boldsymbol{\theta}}_n = (X^T X)^+ X^T Y,$$

where $(X^T X)^+$ denotes the Moore–Penrose inverse of $X^T X$, which always exists. For a symmetric matrix with eigenvectors u_i and corresponding eigenvalues $\lambda_i \geq 0$, the Moore–Penrose inverse is given by $\sum_i \lambda_i^{-1} u_i u_i^T 1_{\{\lambda_i > 0\}}$.

Another consequence of the Hilbert projection theorem is that $\hat{Y} = \hat{H}_X Y$. This formula permits the important observation that any invertible transformation on the covariate, i.e. X is replaced by XA with A invertible, does not change the prediction $\hat{Y}$. The projector $\hat{H}_X$ can be written as $X(X^T X)^+ X^T$. This is because $\hat{H}_X^2 = \hat{H}_X$, $\hat{H}_X = \hat{H}_X^T$, verifying that $\hat{H}_X X = X$ and that $\hat{H}_X u = 0$ for any u orthogonal to X.

2 Statistical model

In the previous section, we have defined the OLS estimator based on the observed data. When assuming that the observation are independent realizations of some random variables, we can rely on probability theory to further study the behaviour of the OLS. In the following we describe different approaches to model linearly the explanatory variable.

2.1 The fixed-design model

The fixed design model takes the form:

$$Y_i = x_i^T \boldsymbol{\theta}^\star + \epsilon_i, \qquad \text{for all } i = 1, \dots, n,$$

where (x_i) is a deterministic sequence of points in $\mathbb{R}^{p+1}$ and (ϵ_i) is a random sequence of random variables in $\mathbb{R}$ such that

$$\mathbb{E}[\epsilon] = 0, \qquad \text{var}(\epsilon) = \sigma^2 I_n, \qquad \text{with } \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

For instance, (ϵ_i) can be an identically distributed and independent sequence of centered random variable with variance σ^2. The level of noise σ of course reflects the difficulty of the problem.

The fixed-design model is appropriate when the (x_i) is chosen by the analyst, e.g., in a physics laboratory experiment, one can fix some variables such as the temperature, or in a clinical survey one can give to patients a determine quantity of some serum. In contrast, the random design (see Section 2.3) model is appropriate when the covariates are unpredictable as for instance the wind speed observed in the nature or the age of some individuals in a survey.

Based on this model, we can derive some statistical properties (given in the following). These properties are concerned with different types of error related to the estimation of $\boldsymbol{\theta}^\star$ by $\hat{\boldsymbol{\theta}}_n$ and will be obtained under the assumption that the dimension of $\text{span}(X)$ is $p+1$, i.e., $\ker(X) = \{0\}$, i.e., $\hat{\boldsymbol{\theta}}_n$ is unique. We therefore implicitly assume that $n \geq p+1$. We can now state a useful decomposition:

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star = (X^T X)^{-1} X^T \epsilon. \tag{3}$$

2.1.1 Bias, variance and risk

Definition 2. *An estimator $\boldsymbol{\theta}(X, Y)$ is said to be unbiased if for all $(X, \epsilon, \boldsymbol{\theta}^\star)$ used to generate Y according to the model, it holds that $\mathbb{E}[\boldsymbol{\theta}(X, Y)] = \boldsymbol{\theta}^\star$.*

When unique, the OLS estimator is unbiased i.e., it holds that $\mathbb{E}[\hat{\boldsymbol{\theta}}_n] = \boldsymbol{\theta}^\star$. Its variance is given by

$$\mathrm{var}(\hat{\boldsymbol{\theta}}_n) = (X^T X)^{-1}\sigma^2.$$

The quadratic risk associated to $\hat{\boldsymbol{\theta}}_n$ estimating $\boldsymbol{\theta}^\star$ is $R_{\mathrm{quad}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^\star) = \mathbb{E}[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\|^2]$. We have that

$$R_{\mathrm{quad}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^\star) = \mathrm{tr}((X^T X)^{-1})\sigma^2.$$

Hence whenever the smallest eigenvalue of $\hat{G}_n$ is larger than b, positive and independent of n, the quadratic risk of the OLS decreases with the rate $1/n$, which is the classical estimation rate in statistics, e.g., empirical average estimating the expectation.

In contrast with the quadratic risk defined on the regression coefficients β, the prediction risk takes care of the prediction error, i.e., the error when predicting y. It is define as

$$R_{\mathrm{pred}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^\star) = \mathbb{E}[\|Y^\star - \hat{Y}\|^2]/n,$$

where $Y^\star$ is the prediction we would make if we knew the true regression vector, i.e., $Y^\star = X\boldsymbol{\theta}^\star$. We have that

$$R_{\mathrm{pred}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^\star) = (p+1)\sigma^2/n.$$

2.1.2 Best linear unbiased estimator (BLUE)

This section is dedicated to the so called Gauss-Markov theorem which asserts that the OLS is BLUE.

We introduce the following partial order (reflexivity, anti-symmetry and transitivity) on the set of symmetric matrix. Let $V_1 \in \mathbb{R}^{d\times d}$ and $V_2 \in \mathbb{R}^{d\times d}$ be two symmetric matrices. We write $V_1 \leq V_2$ whenever $u^T V_1 u \leq u^T V_2 u$ for every $u \in \mathbb{R}^d$. This partial order is particularly useful to compare the covariance matrices of estimators. Indeed if $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimators with respective covariance V_1 and V_2. Then, $V_1 \leq V_2$ if and only if the linear combinations of $\hat{\beta}_1$ have smaller variances than the linear combinations of $\hat{\beta}_2$.

Definition 3. *An estimator is said to be linear if, for any dataset (Y, X), it writes as AY, where $A \in \mathbb{R}^{(p+1)\times n}$ depends only on X.*

Proposition 2 (Gauss-Markov)**.** *Under the fixed design model, among all the unbiased linear estimators AY, $\hat{\boldsymbol{\theta}}_n$ is the one with minimal variance, i.e.,*

$$\mathrm{cov}(\hat{\boldsymbol{\theta}}_n) \leq \mathrm{cov}(AY),$$

with equality if and only if $A = (X'X)^{-1}X'$.

Proof. First note that AY is unbiased if and only if $(A-(X'X)^{-1}X')X\boldsymbol{\theta}^\star = 0$ for all $\boldsymbol{\theta}^\star$, equivalently, $BX = 0$ with $B = (A - (X'X)^{-1}X')$. Consequently, $\mathrm{cov}(BY, \hat{\boldsymbol{\theta}}_n) = 0$. Then, just write

$$\begin{aligned}\mathrm{cov}(AY) &= \mathrm{cov}(BY + \hat{\boldsymbol{\theta}}_n)\\ &= \mathrm{cov}(BY) + \mathrm{cov}(\hat{\boldsymbol{\theta}}_n)\\ &= \sigma^2 BB' + \mathrm{cov}(\hat{\boldsymbol{\theta}}_n) \geq \mathrm{cov}(\hat{\boldsymbol{\theta}}_n).\end{aligned}$$

The previous inequality is an equality if and only if $B = 0$. □

2.1.3 A concentration inequality

We can now provide an additional guarantee for the OLS estimator. It consists in an upper bound on the probability that the estimation error exceeds any given $t > 0$. The upper bound naturally depends on p, n, t, and the smallest eigenvalue of $\hat{G}_n$.

Proposition* 1. *Denote by λ_n the smallest eigenvalue of $\hat{G}_n$ and suppose that $\hat{\lambda}_n > 0$ for all $n \geq 1$. Suppose that (ϵ_i) is a sequence of independent random variables bounded by $c > 0$. Then, for any $k \in \{0, \ldots, p\}$,*

$$\mathbb{P}\left(\left|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k\right| > t\right) \leq 2\exp(-t^2 n/2c^2\hat{s}^2_{n,k}),$$

where $\hat{s}_{n,k} = e_k^T \hat{G}_n^{-1} e_k$. Moreover,

$$\mathbb{P}\left(\max_{k=0,\ldots,p}\left|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k\right| > t\right) \leq 2(p+1)\exp(-t^2 n\hat{\lambda}_n/2c^2).$$

Proof. Apply Hoeffding inequality to the sequence $\sum_{i=1}^n (u^T\tilde{X}_i)\epsilon_i$ to obtain that

$$\mathbb{P}\left(\left|\sum_{i=1}^n (\tilde{X}_i^T u)\epsilon_i\right| > t\right) \leq 2\exp(-2t^2/\sum_{i=1}^n (b_i - a_i)^2),$$

where (a_i, b_i) which must be such as $a_i \leq (u^T\tilde{X}_i)\epsilon_i \leq b_i$, can be chosen as follows $b_i = -a_i = c|\tilde{X}_i^T u|$. Then applying this with $\tilde{X}_i = \hat{G}_n^{-1}X_i$, $u = e_k$ and using (3), one find the first inequality. The second inequality follows from $\hat{s}^2_{n,k} \leq \hat{\lambda}_n^{-1} = \max_{\|u\|=1} |u^T\hat{G}_n^{-1}u|$ and the union bound:

$$\begin{aligned}\mathbb{P}\left(\max_{k=0,\ldots,p}\left|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k\right| > t\right) &= \mathbb{P}\left(\bigcup_{k=0,\ldots,p}\left\{|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k| > t\right\}\right)\\ &\leq \sum_{k=0,\ldots,p}\mathbb{P}\left(\left|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k\right| > t\right).\end{aligned}$$

□

Remark 1. *The first inequality of Proposition 1 is important as it shows that each coordinate might not behaves similarly depending on the associated diagonal element $\hat{G}_n$. For instance, for the intercept, the bound just becomes $2\exp(-t^2 n/2c^2)$. The quantity $\hat{s}_{n,k}$ is also important in practice as it influences the size of the confidence interval (see section 3).*

Remark 2. *Proposition 1 suggests that the value of the smallest eigenvalue $\hat{\lambda}_n$ of $\hat{G}_n$ plays an important role on the accuracy of the estimation. The smaller $\hat{\lambda}_n$ the worst the estimation accuracy.*

2.1.4 Noise estimation

Providing only an estimate $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}^\star$ is often not enough as it does not give any clue on the accuracy of the estimation. When possible one should also furnish an estimation of the error σ^2. If we knew the error ϵ_i, one would take the empirical variance of $\epsilon_1, \ldots, \epsilon_n$, but this is not possible. Alternatively, one can take

$$\tilde{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Because of the first normal equations expressed in (2), we have $\sum_{i=1}^n (Y_i - \hat{Y}_i) = 0$. Consequently, $\tilde{\sigma}_n^2$ is the empirical variance estimate of the residuals vector $Y_i - \hat{Y}_i$. Noting that $\tilde{\sigma}_n^2 = n^{-1}\|(I_n - \hat{H}_X)\epsilon\|^2$ one can compute the expectation:

$$\mathbb{E}[\tilde{\sigma}_n^2] = \sigma^2(n - p - 1)/n.$$

The unbiased version (which should be used in practice) is then

$$\hat{\sigma}_n = \tilde{\sigma}_n^2 \left(\frac{n}{n - p - 1} \right),$$

where from now on we assume that $n > p + 1$. In the case when $n = p + 1$ and X has rank $p + 1$, we obtain that $Y_i = \hat{Y}_i$ for all $i = 1, \ldots, n$.

2.2 The Gaussian model

Here we introduce the Gaussian model as a submodel of the fixed design model where the distribution of the noise sequence (ϵ_i) is supposed to be Gaussian with mean 0 and variance σ^2. The Gaussian model can then be formulated as follows:

$$Y_i \overset{i.i.d.}{\sim} \mathcal{N}(x_i^T \boldsymbol{\theta}^\star, \sigma^2), \qquad \text{for all } i = 1, \ldots, n,$$

where (x_i) is non-random sequence of vector in $\mathbb{R}^{p+1}$. We keep assuming that $\ker(X) = \{0\}$ in the following. The Student's t-distribution with p degrees of freedom is defined as the distribution of the random variable $X/\sqrt{Z/p}$, where X (resp. Z) has standard normal distribution (resp. chi-square distribution with p degrees of freedom).

Proposition 3. *Under the Gaussian model, if* $\ker(X) = \{0\}$ *and* $n > p+1$, *it holds that*

- $\hat{\boldsymbol{\theta}}_n$ *and* $\hat{\sigma}_n^2$ *are independent,*
- $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star) \sim \mathcal{N}(0, n\sigma^2(X^TX)^{-1})$,
- $(n-p-1)(\hat{\sigma}_n^2/\sigma^2) \sim \chi^2_{n-p-1}$,
- *if* $\hat{s}^2_{n,k}$ *is the* k*-th term in the diagonal of* $\hat{G}_n^{-1}$, *then*

$$(n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)(\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k) \sim \mathcal{T}_{n-p-1},$$

where T_{n-p-1} *is the Student's t-distribution with* $n-p-1$ *degrees of freedom.*

Proof. For the first point, remark that $X^T\epsilon$ and $(I - \hat{H}_X)\epsilon$ are two independent Gaussian vector:

$$\operatorname{cov}(X^T\epsilon, (I-\hat{H}_X)\epsilon) = \mathbb{E}[X^T\epsilon\epsilon^T(I-\hat{H}_X)] = 0.$$

Then writing

$$\begin{aligned}(n-p-1)\hat{\sigma}^2 &= \|Y - \hat{Y}\|^2 = \|(I-\hat{H}_X)Y\|^2 = \|(I-\hat{H}_X)\epsilon\|^2\\ \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star &= (X^TX)^{-1}X^T\epsilon,\end{aligned}$$

we see that $\hat{\boldsymbol{\theta}}_n$ and $\hat{\sigma}^2$ are measurable transformations of two independent Gaussian vector. They then are independent. We can use for instance the following characterisation of independence, say for random variables ξ_1 and ξ_2 : for any f_1 and f_2 positive and measurable, $\mathbb{E}[f_1(\xi_1)f_2(\xi_2)] = \mathbb{E}[f_1(\xi_1)]\mathbb{E}[f_2(\xi_2)]$.

For the second point, as ϵ is Gaussian, one just has to compute the variance.

For the third point, let $V \in \mathbb{R}^{n\times n}$ be an orthogonal matrix such that $V = (V_1, V_2)$ where V_1 is a basis of $\operatorname{span}(X)$, and note that $V_1^T(I-\hat{H}_X) = 0$ and $V_2^T(I-\hat{H}_X) = V_2^T$. As the norm is invariant by orthogonal transformation, one has

$$(n-p-1)\hat{\sigma}^2 = \|(I-\hat{H}_X)\epsilon\|^2 = \|V^T(I-\hat{H}_X)\epsilon\|^2 = \|V_2^T\epsilon\|^2.$$

Consequently,

$$(n-p-1)(\hat{\sigma}^2/\sigma^2) = \sum_{i=1}^{n-p-1} \tilde{\epsilon}_i^2,$$

with $\tilde{\epsilon} = V_2^T \epsilon/\sigma$. It remains to show that $\tilde{\epsilon}$ is a Gaussian vector with covariance I_{n-p-1}.

For the fourth point, use the second point to obtain that

$$(n^{1/2}/\hat{s}_{n,k}\sigma)(\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_k^\star) \sim \mathcal{N}(0,1).$$

Then $(n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)(\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_k^\star)$ writes as the quotient of two independent random variables: a Gaussian and the square root of a chi-square. This is a Student's t-distribution with $n-p-1$ degrees of freedom.

□

A direct application of the previous proposition gives us the following equality, which is informative on the estimation error, for any $k = 1, \ldots, p+1$,

$$\mathbb{P}(|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_k^\star| \geq t) = 2S_{T_{n-p-1}}(tn^{1/2}/\hat{s}_{n,k}\sigma^2),$$

where $S_{T_{n-p-1}}$ is the survival function of the the distribution T_{n-p-1}.

2.3 The random design model

In the random design model, we suppose that (Y_i, X_i) is a sequence of independent and identically distributed random vectors defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The aim is to estimate the best linear approximation of Y_1 made up with X_1 in terms of L_2-risk, i.e., to find $\boldsymbol{\theta}$ that minimizes $\mathbb{E}[(Y_1 - X_1^T\boldsymbol{\theta}^*)^2]$. Such a minimizer can be characterized with the help of the normal equation.

Proposition* 2. *Suppose that for all* $k = 1, \ldots, p$, $\mathbb{E}[X_{1,k}^2] < \infty$ *and* $\mathbb{E}[Y_1^2] < \infty$, *then*

$$\inf_{\boldsymbol{\theta}} \mathbb{E}[(Y_1 - X_1^T\boldsymbol{\theta})^2] = \mathbb{E}[(Y_1 - X_1^T\boldsymbol{\theta}^*)^2],$$

if and only if

$$\mathbb{E}[X_1 X_1^T]\boldsymbol{\theta}^* = \mathbb{E}[X_1 Y_1].$$

Proof. Note that the minimization problem of interest is equivalent to

$$\inf_{Z_1 \in \mathcal{F}} \mathbb{E}[(Y_1 - Z_1)^2],$$

where $\mathcal{F}$ is the linear subspace of the Hilbert space $L_2(\Omega, \mathcal{A}, \mathbb{P})$ generated by $X_{1,0}, \ldots, X_{1,p}$. As $\mathcal{F}$ is a closed linear subspace (because it has a finite dimension), the minimizer is unique and characterized by the normal equations. □

Remark 3. *A more general regression problem can be formulated without specifying a linear link : the regression function f^* is any measurable function that minimizes the risk*

$$R(f) = \mathbb{E}[(Y_1 - f(X_1))^2].$$

When $\mathbb{E}[Y_1^2] < \infty$, the minimizer is unique and coincides, in $L^2(\Omega, \mathcal{A}, \mathbb{P})$, with the conditional expectation of Y given X_1 : $f^*(X_1) = \mathbb{E}[Y_1|X_1]$.

The previous proposition does not imply that $\boldsymbol{\theta}^*$ is unique. In fact we are facing a similar situation as in Proposition 1 : either θ^* is unique, which is equivalent to $\mathbb{E}[X_1 X_1^T]$ is invertible, or $\boldsymbol{\theta}^*$ is not uniquely defined, in which case one might take $\boldsymbol{\theta}^* = \mathbb{E}[X_1 X_1^T]^+ \mathbb{E}[X_1 Y_1]$. The case where $\boldsymbol{\theta}^*$ is not unique happens as soon as we add the constant variable or as soon as one variable is a combination of the others. Some asymptotic properties are available. They will be useful to run some statistical tests. We consider the following definition, valid for any $n \geq 1$,

$$\hat{\boldsymbol{\theta}}_n = (X^T X)^+ X^T Y.$$

Proposition* 3. *Suppose that $\mathbb{E}[X_1 X_1^T]$ and $\mathbb{E}[Y_1^2]$ exist and that $\mathbb{E}[X_1 X_1^T]$ is invertible. Then*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \rightsquigarrow \mathcal{N}(0, \sigma^2 G^{-1}),$$

where $\sigma^2 = \mathrm{var}(Y_1 - X_1^T \boldsymbol{\theta}^*)$ and $G = \mathbb{E}[X_1 X_1^T]$. Moreover

$$\hat{\sigma}_n^2 \to \sigma^2, \textit{in probability.}$$

In particular, $(n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)(\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_k^\star) \rightsquigarrow \mathcal{N}(0,1)$.

Proof. Note that

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = n^{1/2}(X^T X)^+ X^T \epsilon + n^{1/2}((X^T X)^+ (X^T X) - I_{p+1})\boldsymbol{\theta}^*.$$

It suffices to show that the term in the right converges to 0 in probability and that the term in the left converges in distribution to the stated limit. The first point is a consequence of the continuity of the determinant. The second point is a consequence of Slutsky's theorem using the fact that the Moore-Penrose inverse is a continuous operation.

The convergence of $\hat{\sigma}_n^2$ is obtained by the decomposition

$$\begin{aligned}\hat{\sigma}_n^2 &= (n-p+1)^{-1}\|(I-\hat{H}_X)\epsilon\|_2^2 \\ &= (n-p+1)^{-1}\left(\|\epsilon\|^2 - \epsilon^T X(X^TX)^+X^T\epsilon\right).\end{aligned}$$

Invoking the law of large number, we only need to show that the term in the write goes to 0 in probability. We have

$$\epsilon^T X(X^TX)^+X^T\epsilon = \left(n^{-1/2}\sum_{i=1}^n X_i\epsilon_i\right)^T \hat{G}_n^+ \left(n^{-1/2}\sum_{i=1}^n X_i\epsilon_i\right)$$

Because $\hat{G}_n^+ \to G^{-1}$ and $n^{-1/2}\sum_{i=1}^n X_i\epsilon_i \rightsquigarrow \mathcal{N}(0,G)$, we get that

$$\epsilon^T X(X^TX)^+X^T\epsilon \rightsquigarrow \|\mathcal{N}(0,\sigma^2 I_{p+1})\|^2 = \sigma^2\chi^2_{p+1}.$$

When divided by $(n-p+1)$ the previous term goes to 0. □

3 Confidence intervals and hypothesis testing

3.1 Confidence intervals

From a practical perspective, building confidence intervals is often an inevitable step as it furnishes to the analyst an idea on the quality of the estimation. The construction of confidence intervals follows the estimation step. The accuracy/quality of the estimation is measured by the size of an interval, which is called confidence interval. It is a region (based on the observed data) in which the parameter of interest is most likely to lie. As we shall see, it is based on the variability of the estimation procedure.

We consider a regression model with n observed data points (Y, X) and we focus on the task of building confidence intervals for the k-th coordinate $\boldsymbol{\theta}_k^\star$ of the regression vector, $k = 0, \dots, p$.

Definition 4. *A confidence interval of level* $1-\alpha$ *is an interval* $\hat{I}_n(Y, X) \subset \mathbb{R}$ *satisfying*

$$\mathbb{P}(\boldsymbol{\theta}_k^\star \in \hat{I}_n(Y, X)) \geq 1 - \alpha.$$

Confidence interval can be obtained easily when the assumption on the model allows to know the distribution of the quantity $\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_k^\star$. This is the case for instance in the popular Gaussian model in virtue of Proposition 3.

Proposition 4. *In the Gaussian model, if* $\ker(X) = \{0\}$ *and* $n > p + 1$,

$$\hat{\boldsymbol{\theta}}_{n,k} + \left[-\left(\frac{\hat{s}_{n,k}\hat{\sigma}_n}{n^{1/2}} \right) Q_{n-p-1}(1-\alpha/2)\,,\, \left(\frac{\hat{s}_{n,k}\hat{\sigma}_n}{n^{1/2}} \right) Q_{n-p-1}(1-\alpha/2) \right],$$

where Q_{n-p-1} *is the quantile function of* $\mathcal{T}_{n-p-1}$*, is a confidence interval of level* $1-\alpha$*.*

When the noise distribution is not Gaussian, the previous confidence interval has no reason to be valid. In this case, there are basically two techniques permitting the construction of confidence interval:

- Concentration inequalities. This usually produces pessimistic (too large) confidence interval.
- Asymptotics. This only produces asymptotically valid confidence interval.

We start by giving confidence intervals for using two concentration inequality: Markov and Hoeffding.

Proposition* 4. *In the fixed design model, suppose that (for clarity) $X'X = nI_n$ and that (ϵ_i) is an identically distributed sequence of centered random variables with variance σ^2, then the interval*

$$\hat{\boldsymbol{\theta}}_{n,k} + \left[-\sqrt{\sigma^2/(n\alpha)}\,,\ \sqrt{\sigma^2/(n\alpha)}\right],$$

is a confidence interval of level $1-\alpha$. If moreover, $|\epsilon_i| \leq c$ for all $i = 1, \ldots, n$,

$$\hat{\boldsymbol{\theta}}_{n,k} + \left[-\sqrt{2c\log(2/\alpha)c/n}\,,\ \sqrt{2c\log(2/\alpha)/n}\right],$$

is a confidence interval of level $1-\alpha$.

Proof. We have using (3), $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star = X'\epsilon/n$. Applying the Markov inequality

$$\begin{aligned}\mathbb{P}(|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k| \geq t) &\leq t^{-2}\mathbb{E}[(\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k)^2] \\ &\leq \sigma^2 \sum_{i=1}^n X_{i,k}^2/(t^2n^2) \\ &= \sigma^2/(t^2n),\end{aligned}$$

leading to the first confidence interval.

Applying Hoeffding inequality with the sequence (ϵ_i), one has

$$\mathbb{P}(|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k| \geq t) \leq 2\exp\left(-2(nt)^2/\sum_{i=1}^n (b_i - a_i)^2\right),$$

where $a_i \leq \epsilon_i \leq b_i$. Choosing $a_i = -c$, $b_i = c$, we get that

$$\mathbb{P}(|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k| \geq t) \leq 2\exp(-t^2n/2c),$$

leading to the second confidence interval. □

Note that the first confidence interval based on Markov inequality is very pessimistic (i.e., very large) compared to the second one, based on Hoeffding's inequality. This is because $\log(1/\alpha) << 1/\alpha$ when $\alpha \to 0$.

Proposition* 5. *In the random design model, suppose that $\mathbb{E}[X_{1,k}^2] < \infty$ and $\mathbb{E}[Y_1^2] < \infty$, then*

$$\hat{\boldsymbol{\theta}}_{n,k} + \left[-\left(\frac{\hat{s}_{n,k}\hat{\sigma}_n}{n^{1/2}}\right)\Phi^-(1-\alpha/2)\,,\ \left(\frac{\hat{s}_{n,k}\hat{\sigma}_n}{n^{1/2}}\right)\Phi^-(1-\alpha/2)\right],$$

where Φ^- is the quantile function of $\mathcal{N}(0,1)$, is, asymptotically, a confidence interval of level $1-\alpha$, i.e.,

$$\liminf_{n\to\infty} \mathbb{P}(\boldsymbol{\theta}^\star_k \in \hat{I}_n(\alpha)) \geq 1-\alpha.$$

Proof. That $X_n \rightsquigarrow \mathcal{N}(0,1)$ means that $P(X_n \in [-\Phi^-(1-\alpha/2), \Phi^-(1-\alpha/2)]) \to \Phi(\Phi^-(1-\alpha/2)) - \Phi(\Phi^-(\alpha/2)) = 1-\alpha$ where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$. □

3.2 Hypothesis testing

We start by recalling some definitions and some vocabulary related to statistical testing. Then we consider no effect tests on the covariates of a regression. These tests play an important role in practice because can be used to quantify the importance of some variables and ultimately to remove some variables. This is described the forward variable selection method.

3.2.1 Definitions

Statistical testing aims at answering whether or not an hypothesis $\mathcal{H}_0$ is likely. It is usually performed by constructing a test statistic $\hat{T}_n$ and deciding to reject, or not, whenever $\hat{T}_n$ is in $\mathcal{R}$, or not. The region $\mathcal{R}$ is called the reject region. As soon as $\hat{T}_n$ and $\mathcal{R}$ are specified, the process is quite simple:

$$\text{Reject whenever } \hat{T}_n \in \mathcal{R}$$
$$\text{Do not reject whenever } \hat{T}_n \notin \mathcal{R}.$$

The terminology "to not reject" rather than "to accept" comes from the fact that $\mathcal{H}_0$ is often too much thin and unlikely to be "accepted", e.g., a simple hypothesis $\boldsymbol{\theta}_1^\star = \pi$. There are basically 2 kinds of errors that we wish to control:

Type-1: to reject whereas $\mathcal{H}_0$ is true

Type-2: not to reject whereas $\mathcal{H}_0$ is not true.

The proportion of Type-1 errors is called the level of the test. One minus the proportion of Type-2 errors is called the power of the test. The consistency of level α imposes that asymptotically, the level is smaller than α while the power is one.

Definition 5. *A statistical test $(\hat{T}_n, \mathcal{R})$ is said to be consistent with level $1-\alpha$ whenever*

$$\limsup_{n\to\infty} P_{\mathcal{H}_0}(\hat{T}_n \in \mathcal{R}) \leq \alpha$$
$$\lim_{n\to\infty} P_{\mathcal{H}_1}(\hat{T}_n \in \mathcal{R}) = 1.$$

3.2.2 Test of no effect

In a linear regression model, a covariate has no effect if and only if its associated regression coefficient is null. A test of no effect of a covariate, say the $k-th$, then consists in testing the nullity of its regression coefficient $\boldsymbol{\theta}_k^\star$:

$$\mathcal{H}_0 : \boldsymbol{\theta}_k^\star = 0.$$

Proposition* 6. *Under the random design model, if* $\mathbb{E}[X_1 X_1^T]$ *and* $\mathbb{E}[Y_1^2]$ *exist and* $\mathbb{E}[X_1 X_1^T]$ *is invertible, the statistic and reject region*

$$\hat{T}_{n,k} = \left(\frac{n^{1/2}}{\hat{s}_{n,k}\hat{\sigma}_n} \right) |\hat{\boldsymbol{\theta}}_{n,k}|,$$
$$\mathcal{R} = (\Phi^-(1-\alpha/2), +\infty),$$

produce a consistent test with level $1-\alpha$.

Proof. For the level, it is very similar to confidence interval. For the power, suppose that $\boldsymbol{\theta}_k^\star \neq 0$. Let $Z_n = (n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)(\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_k^\star)$ and $q = \Phi^-(1-\alpha/2)$. Then $\hat{T}_{n,k} \in \mathcal{R}$ if and only if

$$Z_n + (n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)\boldsymbol{\theta}_k^\star < -q \quad \text{or} \quad Z_n + (n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)\boldsymbol{\theta}_k^\star > q.$$

If $\boldsymbol{\theta}_k^\star$ is positive (resp. negative) one can show that the event in the right (resp. left) has probability going to 1. We consider only the case $\boldsymbol{\theta}_k^\star > 0$. It has been shown in the proof of Proposition 3 that $\hat{s}_{n,k}\hat{\sigma}_n$ converges in probability to a finite value. We can work on the event that $\hat{s}_{n,k}\hat{\sigma}_n < M$. Let $K > 0$. For n large enough $q - (n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)\boldsymbol{\theta}_k^\star < -K$. Hence

$$P(Z_n + (n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)\boldsymbol{\theta}_k^\star > q) \geq P(Z_n > -K).$$

Hence

$$\liminf_{n\to\infty} P(Z_n + (n^{1/2}/\hat{s}_{n,k}\hat{\sigma}_n)\boldsymbol{\theta}_k^\star > q) \geq 1 - \Phi(-K).$$

But K is arbitrary and the result follows. □

Remark 4. *In practice, the statistic* $\hat{T}_{n,k}$ *is scale invariant: if* D *is a positive diagonal matrix, then the statistic* $\hat{T}_{n,k}$ *constructed from the sample* X *is the same as the statistic* $\hat{T}_{n,k}$ *constructed from the sample* XD.

Remark 5. *In the Gaussian case, the test statistic and the reject region are given by*

$$\hat{T}_{n,k} = \left(\frac{n^{1/2}}{\hat{s}_{n,k}\hat{\sigma}_n} \right) |\hat{\boldsymbol{\theta}}_{n,k}|,$$
$$\mathcal{R} = (Q_{n-p-1}(1-\alpha/2), \infty).$$

Such a test has a level exactly equal to $1-\alpha$. *To derive that the power goes to* 1, *one can assume that for all* $n \geq 1$, $\hat{s}_{n,k}\hat{\sigma}_n$ *is bounded.*

	age	sex	bmi	bp	Serum measurements						output
patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93	38	4	4.9	87	151
2	48	1	21.6	87	183	103	70	3	3.9	69	75
...	...										...
...	...										...
441	36	1	30.0	95	201	125	42	5	5.1	85	220
442	36	1	19.6	71	250	133	97	3	4.6	92	57

Table 1: The dataset is composed of $n = 442$ patients, $p = 10$ variables "baseline" body mass index, bmi), average blood pressure (bp), etc... The output is a score corresponding to the desease evolution. Each covariate has been standardized Efron et al. (2004).

Remark 6 (test and confidence intervals)**.** *Making no effect tests consists in rejecting whenever* 0 *(or more generally any tested values) is not lying inside the confidence interval. For instance, in the random design model, to reject is equivalent to*

$$\frac{n^{1/2}}{\hat{s}_{n,k}\hat{\sigma}_n}|\hat{\boldsymbol{\theta}}_{n,k}| \in (\Phi^-(1-\alpha/2), +\infty),$$

which is equivalent to

$$0 \notin \hat{\boldsymbol{\theta}}_{n,k} + \left[-\left(\frac{\hat{s}_{n,k}\hat{\sigma}_n}{n^{1/2}}\right)\Phi^-(1-\alpha/2)\,,\,\left(\frac{\hat{s}_{n,k}\hat{\sigma}_n}{n^{1/2}}\right)\Phi^-(1-\alpha/2)\right].$$

3.3 Forward variable selection

In this section present a simple forward variable selection method that directly comes from the tests of no effect presenting before. To this aim we consider the "diabetes" dataset of sklearn presented in Table 1.

The method of forward selection is a stepwise procedure that starts with no covariate and add a new one at each step. This kind of methods is refereed to as greedy methods. The criterion used to select the best covariate is based on the test statistic defined before: $|\hat{\boldsymbol{\theta}}_{n,k}|/\hat{s}_{n,k}$. This criterion has an interpretation in terms of p-values. When the test is described by $(\hat{T}_n, \mathcal{R}_\alpha)$, the p-value is the smallest value of α for which we still reject. For instance, in the Random design model,

$$\inf\{\alpha \in [0,1] \,:\, \hat{T}_{n,k} > \Phi^-(1-\alpha/2)\} = 2(1-\Phi(\hat{T}_{n,k}))$$

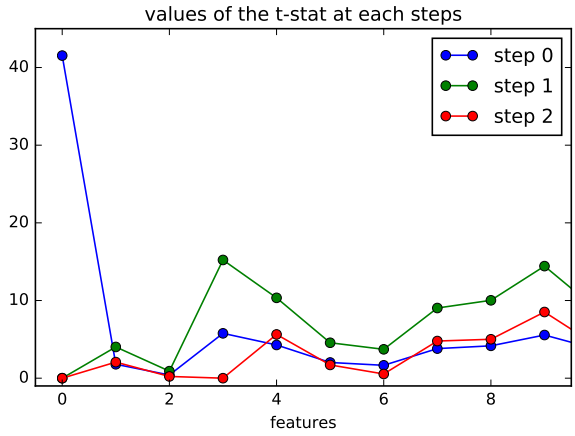

Figure 1: The statistics of each selected variable is 0 in the next step. The intercept is the first selected variable, then X_3, etc...

Hence taking the largest $\hat{T}_{n,k} = n^{1/2}|\hat{\boldsymbol{\theta}}_{n,k}|/(\hat{s}_{n,k}\hat{\sigma}_n)$ is equivalent to take the smallest p-value for the underlying test of no effect. A stopping rule can be based on the p-value: stop as soon as none of the p-value is smaller that 0.05.

Algorithm 1 (forward variable selection)**.**
Inputs: *(Y, X) a threshold p_{stop}. Start with $r = Y$, $\mathcal{S} = \emptyset \subset \mathcal{A} = \{0, \ldots, p\}$.*

(i) For each $k \in \mathcal{A}\backslash\mathcal{S}$, in the model $r \simeq X_k\boldsymbol{\theta}_k$ (where $\boldsymbol{\theta}_k \in \mathbb{R}$), compute $\hat{\boldsymbol{\theta}}_{n,k}$, $\hat{s}_{n,k}$ and $\hat{\sigma}_{n,k}$ (it depends on k here as the underlying model).

(ii) Stop if no p-values are smaller than p_{stop}. Else compare each $|\hat{\boldsymbol{\theta}}_{n,k}|/(\hat{s}_{n,k}\hat{\sigma}_{n,k})$ and keep the larger, say $\tilde{k}$.

(iii) Update the residuals $r := r - X_{\tilde{k}}\hat{\boldsymbol{\theta}}_{n,\tilde{k}}$ and the set $\mathcal{S} := \mathcal{S} \cup \{\tilde{k}\}$.

Figure 3.3 illustrates the procedure described by Algorithm 1.

Remark 7. *Different stopping rules might be considered. For instance, in Zhang (2009), it is advocate to consider the residuals sum of squares and to stop as soon as $\|r\|^2 < \epsilon$.*

4 Regularization

4.1 Singular value decomposition

We first give two results dealing with the spectral decomposition of matrices. The first one is a classic called the eigendecomposition of a symmetric matrix. The second one is called the singular-value-decomposition (SVD) and applies to any matrices, not necessary squared.

Proposition 5. *Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Then there exist $\lambda_1 \geq \ldots \geq \lambda_d$, called eigenvalues, and an orthogonal matrix $U \in \mathbb{R}^{d\times d}$ of eigenvectors, such that $A = UDU^T$, where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$.*

Proposition* 7. *Let $X \in \mathbb{R}^{n\times p}$. Then there exist two orthogonal matrices : $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{n\times n}$ of singular vectors; and $s_1 \geq \ldots \geq s_{\min(n,p)} \geq 0$, called singular values, such that*

$$X = VSU^T,$$

where $S \in \mathbb{R}^{n\times p}$ contains 0 everywhere except on the diagonal formed by $(s_1, \ldots, s_{\min(n,p)})$.

Proof. Without loss of generality, we suppose that $p \leq n$. Otherwise we apply the result to the X^T. Applying Proposition 5 to X^TX, there exists $U \in \mathbb{R}^{p\times p}$ such that $U^T(X^TX)U$ is diagonal with r positive coefficients. Hence $U_1^T(X^TX)U_1 = D \in \mathbb{R}^{r\times r}$ and $XU_2 = 0$. Take $V_1^T = D^{-1/2}U_1^TX^T$ (an orthogonal set of r vectors : $V_1^TV_1 = I_r$) to find that $V_1^TXU_1 = D^{1/2}$. Consequently, $V_1^TX(U_1, U_2) = (D^{1/2}, 0)$. Remarking that v orthogonal to V_1 means that $v^TXU_1 = 0$ implying that $v^TX(U_1, U_2) = 0$ leading to $v^TX = 0$. Now taking V_2 such that $V = (V_1, V_2) \in \mathbb{R}^{n\times p}$ is orthogonal, we obtain the claimed decomposition with $S^2 = \mathrm{diag}(d_1, \ldots, d_p)$. □

We have the following reduced SVD formula, if $r \geq 1$ stands for the dimension of $\mathrm{span}(X)$,

$$X = \tilde{V}_r\tilde{S}_r\tilde{U}_r^T,$$

where $\tilde{U}_r = (U_1, \ldots, U_r)$, $\tilde{V}_r = (V_1, \ldots, V_r)$, and $\tilde{S}_r \in \mathbb{R}^{r\times r}$ contains only the positive singular-values.

Proposition* 8. *Let $X \in \mathbb{R}^{n\times p}$. For any projector $P \in \mathbb{R}^{p\times p}$ with rank smaller than k, it holds that*

$$\|X - XP_k\|_F \leq \|X - XP\|_F,$$

where $P_k = \sum_{i\leq k} U_iU_i^T$.

Proof. Suppose that $1 \leq k < r$. By Pythagorean identity, $\|X - XP\|_F^2 = \|X\|_F^2 - \|XP\|_F^2$. Hence one just has to show that $\|XP_k\|_F^2 \geq \|XP\|_F^2$. Considering the reduced SVD $X = U_r S_r V_r^T$, we have

$$\begin{aligned}\|XP\|_F^2 &= \operatorname{tr}\left((PU_r)S^2(PU_r)^T\right) \\ &= \operatorname{tr}\left(\sum_{i\leq r} s_i^2 W_i W_i\right) \\ &= \sum_{i\leq r} s_i^2 \|W_i\|_2^2,\end{aligned}$$

with $W_i = PU_i$ and the constraints that $\|W_i\|_2^2 \leq 1$ and $\sum_{i\leq r}\|W_i\|_2^2 \leq k$. Note that this correspond to the optimization problem

$$\max_{m_1,\ldots,m_{r'}} \sum_{i\leq r'} s_i^2 m_i \qquad \text{u.c. } m_i \in (0,k_i),\ \sum_{i\leq r'} m_i \leq k,$$

in which we suppose that $s_1 < \ldots < s_{r'}$ with $r' \leq r$ and $k_i \geq 1$ stands for the multiplicity. We derive the maximum. Note first that necessarily $\sum_{i\leq r'} m_i = k$. Then if i is the first index such that $0 < m_i < k_i$, the function cannot achieved its maximum. Then we get that the maximizer is achieved when m_i is either 0 or 1. Clearly the maximum is $\sum_{i\leq k} s_i^2$ which is achieved when $P = \sum_{i\leq k} U_i U_i^T$. □

Definition 6. *Let $X \in \mathbb{R}^{n\times p}$ and define $X_c = X - 1_n\overline{X}^T$. The PCA of X of degree k is given by the k first elements of the SVD of X_c, i.e., the singular values $(s_1,\ldots,s_k)$, the principal components $U_1,\ldots U_k$ and the principal axes $V_1,\ldots,V_k$.*

Introduce the estimated covariance matrix

$$\hat{\Sigma}_n = n^{-1}X_c^T X_c$$

Proposition* 9. *The principal components $U = U_1,\ldots U_k$ forms a set of orthonormal vectors along which the empirical variance is maximal, i.e.,*

$$\sum_{i\leq k} U_i^T\hat{\Sigma}_n U_i \geq \sum_{i\leq k}\tilde{U}_i^T\hat{\Sigma}_n\tilde{U}_i,$$

for any $(\tilde{U}_1,\ldots,\tilde{U}_k)$ orthonormal vectors. The principal components U can be obtained by an eigendecomposition of $\hat{\Sigma}_n$.

Proof. Take $\tilde{U}$ and U as define in the statement. Define $\tilde{P} = \tilde{U}\tilde{U}^T$ and $P = UU^T$, the associated projectors of rank k. Write

$$\sum_{i \leq k} U_i^T \hat{\Sigma}_n U_i = \operatorname{tr}(\hat{\Sigma}_n P) = n^{-1} \operatorname{tr}(X_c^T X_c P) = n^{-1} \|X_c P\|_F^2.$$

Using Proposition 8 and the Pythagorean identity, we get that $\|X_c P\|^2 \geq \|X_c \tilde{P}\|_F^2$. □

Remark 8. *The PCA of X depends on the scale of each covariate and one may prefer in practice to rescale the matrix X so that each covariate has the same empirical variance. This is done by considering $XD^{-1/2}$ rather than X, taking D equal to the diagonal matrix whose elements are $e_k^T \hat{\Sigma}_n e_k$.*

Remark 9. *In practice, one may run PCA on X before OLS. The aim would be to reduce the number of covariates to avoid large matrix inversion. Be careful that after running the PCA, the prediction must be operated with respect to the centered covariate and output and therefore without intercept. This is in virtue of for all $\theta \in \mathbb{R}^p$,*

$$\|Y - 1_n\theta_0 - Z\theta\|^2 \geq \|(Y - \overline{Y}) - (Z - \bar{Z}^T)\theta\|^2$$

with $X = (1_n, Z)$. As the PCA is independent of the output, doing such a process might result in some loss of information.

Remark 10. *The formulas of OLS can be expressed with the SVD of $X = V^T S U$. For instance, $\hat{\boldsymbol{\theta}}_n = \sum_{k=1}^r s_k^{-1} u_k v_k^T y = X^+ Y$ and its variance is $\sigma^2 \sum_{k=1}^r s_i^{-2} u_k u_k^T$.*

4.2 Ridge estimator

The ridge estimator is introduced to overcome the issues caused by a poorly conditioned matrix $\hat{G}_n$, i.e., when some of its eigenvalues are to small. As the singular value indicates $\hat{\theta}_n = \sum_{k=1}^r s_i v_i u_i^T y$ the estimate is not stable when some s_i are too close to 0. As we can see looking at the variance of the OLS or at Proposition 1, the smallest eigenvalues of $\hat{G}_n$ have a bad influence on the behaviour of the estimate. This implies also some difficulties when applying the forward selection method. The ridge estimator is a solution to control such bad effects due to conditioning.

4.2.1 Definition

The ridge estimator is defined as a solution of the following minimization problem

$$\|Y - X\boldsymbol{\theta}\|^2 + n\lambda\|\boldsymbol{\theta}\|^2, \tag{4}$$

where $\lambda > 0$ is called the regularization parameter and is fixed by the analyst. Before dealing with the choice of λ we need to describe the property of the ridge estimate. First of all, let us briefly state some remark:

- As the expression in (4) is a Lagrangian with constraints $\|\boldsymbol{\theta}\|^2 \leq c$ the Ridge is an OLS under the constraint. The link between c and λ is not explicit.
- Doing ridge is adding a regularization term to the square loss of OLS, aiming to penalize for large coefficients in $\boldsymbol{\theta}$. Other norms might be used such as $\sum_k |\boldsymbol{\theta}_k|$.
- Intuitively, when $\lambda \to 0$, we obtain OLS. When $\lambda \to +\infty$, we obtain 0.
- In order that the ridge estimate is scale invariant one might replace X by $XD^{-1/2}$ where D is the diagonal matrix with entries $e_k^T \hat{G}_n e_k$. The ridge estimate is classically defined without intercept. Hence one may first centre Y and X so that the intercept of the OLS is automatically 0. We work under this framework in the following i.e., $X \in \mathbb{R}^{n \times p}$.

Proposition 6. *The minimizer of (4) exists and is unique. It is given by* $\hat{\boldsymbol{\theta}}_n^{(rdg)} = (X^T X + n\lambda I_p)^{-1} X^T Y$.

Proof. Let f denote the objective function of (4). There exists A such that whenever $\|\boldsymbol{\theta}\| > A$, $f(\boldsymbol{\theta}) > f(0)$. But the set $\|\boldsymbol{\theta}\| \leq A$ is compact and so a minimum exists and is achieved. Note that for any $\boldsymbol{\theta}$,

$$\begin{aligned} f(\boldsymbol{\theta}) - f(0) &= -2 < Y, X\boldsymbol{\theta} > + \|X\boldsymbol{\theta}\|^2 + n\lambda\|\boldsymbol{\theta}\|^2 \\ &= -2 < Y, X\boldsymbol{\theta} > + \|A\boldsymbol{\theta}\|^2, \end{aligned}$$

with $A = ((X^T X) + n\lambda I_p)^{1/2}$ a positive matrix. For any u and v we have

$$\|tu + (1-t)v\|^2 = t\|u\|^2 + (1-t)\|v\|^2 - t(1-t)\|u - v\|^2.$$

Suppose that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two minimizers with $f^* = f(\boldsymbol{\theta}_1) = f(\boldsymbol{\theta}_2)$, we have

$$\begin{aligned} &f(t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2) \\ &= tf(\boldsymbol{\theta}_1) + (1-t)f(\boldsymbol{\theta}_2) - t(1-t)\|A(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|^2 < f^*. \end{aligned}$$

Hence $\hat{\boldsymbol{\theta}}_n^{(rdg)}$ is unique. It is a local minimum given by the first order equation:

$$((X^T X) + n\lambda I_p)\boldsymbol{\theta} = X^T Y.$$

□

4.2.2 Bias and variance

We have seen that, similarly to the OLS, the ridge estimator is the solution of a linear system of equations. In the ridge system of equation the matrix that was previously X^TX in the OLS is now replaced by $X^TX + n\lambda I_p$. As λ is chosen by the user, it allows us to control the smallest eigenvalue of the underlying matrix. Such a change of course influence the bias and the variance of the estimate. To express these quantities, we consider the fixed design model.

The bias of the ridge is

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{(rdg)}] - \boldsymbol{\theta}^\star = -\lambda n(X^TX + n\lambda I_p)^{-1}\boldsymbol{\theta}^\star$$

The variance of the ridge estimator expresses as

$$\mathrm{var}(\hat{\boldsymbol{\theta}}_n^{(rdg)}) = \sigma^2(X^TX + n\lambda I_p)^{-1}X^TX(X^TX + n\lambda I_p)^{-1}.$$

Proposition 7. *We have that*

$$\mathrm{var}(\hat{\boldsymbol{\theta}}_n^{(rdg)}) < \mathrm{var}(\hat{\boldsymbol{\theta}}_n).$$

Proof. We use the SVD of X to write that

$$\mathrm{var}(\hat{\boldsymbol{\theta}}_n^{(rdg)}) = \sigma^2 \sum_{k=1}^{p} \frac{s_k^2}{(s_k^2 + n\lambda)^2} u_i u_i^T.$$

In terms of eigenvalues $\hat{\lambda}_k$ associated to $\hat{G}_n$, we have

$$\mathrm{var}(\hat{\boldsymbol{\theta}}_n^{(rdg)}) = \sigma^2 \sum_{k=1}^{p} \frac{\lambda_k}{(\lambda_k + \lambda)^2} u_k u_k^T.$$

Doing the same for $\hat{\boldsymbol{\theta}}_n$ and using that $\lambda_k/(\lambda_k + \lambda)^2 \leq 1/\lambda_k$, we obtain the result. □

4.2.3 Choice of the regularization parameter

As we have seen before, ridge regression reduces the variance of the OLS while it introduced some bias. Actually this is the parameter λ that decides whether we reduce the bias, $\lambda \to 0$, or the variance, $\lambda \to \infty$. As it cannot be accomplished simultaneously, we are facing a trade-off commonly known under the name of bias-variance trade-off. In the next few lines, we promote the use of cross validation.

Divide the data (Y, X) according to the lines into K-folds of approximately equal size $\lfloor K/n \rfloor$. Let $(Y_{(k)}, X_{(k)})$ (resp. $(Y_{-(k)}, X_{-(k)})$ denote the observation in the k-th fold (resp. all the observation outside the k-th fold). Proceed as follows:

(i) Compute $\hat{\boldsymbol{\theta}}_{n,k}^{(rdg)}$ based on each sample $(Y_{-(k)}, X_{-(k)})$.

(ii) Compute the (unnormalized) prediction error over each fold $Y_{(k)} - X_{(k)}\hat{\boldsymbol{\theta}}_{n,k}^{(rdg)}$. The risk is given by

$$\hat{R}(\lambda) = \sum_{k=1}^{K} \|Y_{(k)} - X_{(k)}\hat{\boldsymbol{\theta}}_{n,k}^{(rdg)}\|^2.$$

The quantity $\hat{R}(\lambda)$ reflects the prediction risk associated to λ. It is then natural to minimize $\hat{R}$ over $\lambda \in (0, \infty)$. In practice, this is usually done by taking a finite grid.

Remark 11. *A computational advantage of using the SVD is that even if considering many values of λ the SVD could be done once for each fold.*

References

Efron, B., T. Hastie, I. M. Johnstone, and R. Tibshirani (2004). Least angle regression. *32*(2), 407–499. With discussion, and a rejoinder by the authors.

Zhang, T. (2009). Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems*, pp. 1921–1928.