Lecture notes on ordinary least squares[1]

François Portier

November 28, 2023

[1]This document is a first version. Please let me know if you find typos or mistakes (`francois.portier@gmail.com`). The author is grateful to Joseph Salmon (`http://josephsalmon.eu/`) for some help on the writing of this course and for sharing some materials.

Contents

Notations

- $\langle \cdot, \cdot \rangle$ 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 $e_0, \ldots, e_{d-1}$. Additionally, the ℓ_q-norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|_q^q = \sum_{k=1}^d x_k^q$.
- 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. The vector $1_n \in \mathbb{R}^n$ contains n ones.
- For any sequence $z_1, z_2, \ldots$, the empirical mean over the n first elements is denoted by $\overline{z}^n = \sum_{i=1}^n z_i / n$
- 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. $X_n \xrightarrow{p} X$).

Chapter 1

Definition of ordinary least-squares and first properties

1.1 Definition

The general goal of regression analysis is to learn some relationship between a variable to predict $y \in \mathbb{R}$ and some covariates $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$, with $p \geq 1$. This is done by *learning* a *link function* that maps the input x to the output y. Linear regression is interested in modeling y using a linear link function of x, i.e., the variable y is modeled by $\theta_0 + \theta_1 x_1 + \ldots + \theta_p x_p$ where $(\theta_0, \dots, \theta_p)$ are the parameters of the linear link function. To learn the values of these parameters, $(\theta_0, \dots, \theta_p)$, we observe $n \geq 1$ pairs (x_i, y_i) that are supposed to come from the same generating mechanism. In what follows we introduce the ordinary least squares (OLS) approach which basically consists in minimizing the sum of squares of the distance between the observed values y_i and the predicted values at x_i under the linear model.

We focus on a regression problem with $n \geq 1$ observations and $p \geq 1$ covariates. For notational convenience, for $i = 1, \dots, n$, we consider $y_i \in \mathbb{R}$ and $x_i = (x_{i,0}, \dots, x_{i,p})^T \in \mathbb{R}^{p+1}$ with $x_{i,0} = 1$. This is only to include the intercept in the same way as the other coefficients. The OLS estimator is any coefficient vector $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\theta}}_{n,0}, \dots, \hat{\boldsymbol{\theta}}_{n,p})^T \in \mathbb{R}^{p+1}$ such that

$$\hat{\boldsymbol{\theta}}_n \in \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (y_i - x_i^T \boldsymbol{\theta})^2. \tag{1.1}$$

It is useful to introduce the notations

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

The matrix X which contains the covariates is called the *design matrix*. With the previous notation, (1.1) 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.

1.2 Existence and uniqueness

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

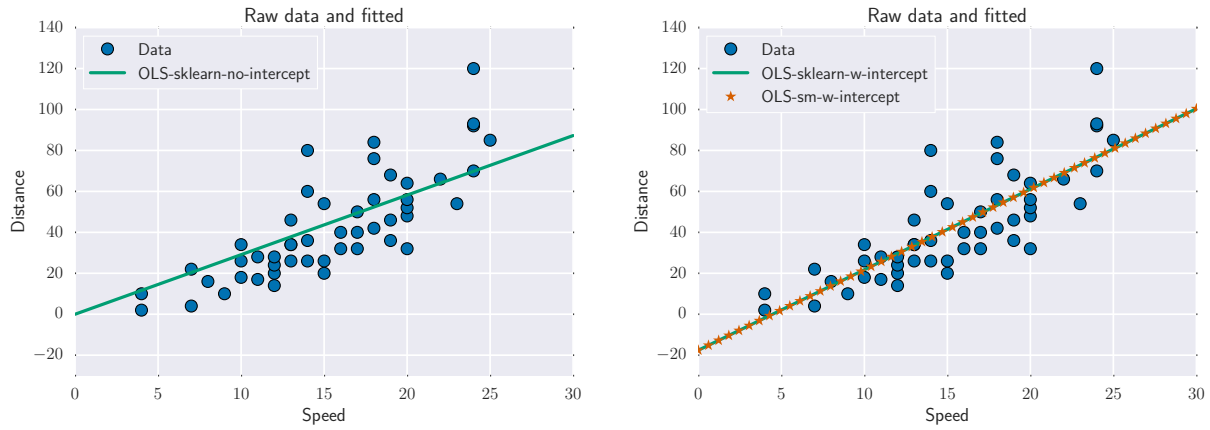

Figure 1.1: The dataset is the cars dataset from the `R` software. We use `sklearn` to compute OLS. The graph on the left represents the OLS line without intercept and the graph on the right is the OLS line computed with intercept.

columns of A). Using the Hilbert projection theorem ($\mathbb{R}^n$ is a Hilbert space, $\text{span}(X)$ is a (closed) linear subspace of $\mathbb{R}^n$), $\hat{Y}$ is unique and is characterized by the fact that the vector $Y - \hat{Y}$ is orthogonal to $\text{span}(X)$. This property is equivalent to the so-called normal equation:

$$X^\top (Y - \hat{Y}) = 0.$$

Since $\hat{Y} = X\hat{\boldsymbol{\theta}}_n$, we obtain that the vector $\hat{\boldsymbol{\theta}}_n$ must verify

$$X^T X\, \hat{\boldsymbol{\theta}}_n = X^T Y. \tag{1.2}$$

Note that in contrast with $\hat{Y}$ (which is always unique), the vector $\hat{\boldsymbol{\theta}}_n$ is not uniquely defined without further assumptions on the data. For instance, take $u \in \ker(X)$ then $\hat{\boldsymbol{\theta}}_n + u$ verifies (1.2) as well as $\hat{\boldsymbol{\theta}}_n$. The uniqueness of the OLS is actually determined by the kernel of X which is related to the invertibility of the so called Gram matrix introduce below (see Exercise 1).

Definition 1. *The matrix $\hat{G}_n = X^T X/n$ is called the Gram matrix. Denote by $\hat{H}_{n,X} \in \mathbb{R}^{n\times n}$ the orthogonal projector*[1] *on* $\text{span}(X)$.

When the Gram matrix is invertible, the OLS is uniquely defined. When it is not the case, (1.1) has an infinite number of solutions.

Proposition 1. *The OLS estimator always exists and the associated prediction is given by $\hat{Y} = \hat{H}_{n,X} Y$. 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 solutions. This happens if and only if $\ker(X) \neq \{0\}$. *In this case, the set of solution writes $\hat{\boldsymbol{\theta}}_n + \ker(X)$ where $\hat{\boldsymbol{\theta}}_n$ is a particular solution.*

[1] Recall that P is the orthogonal projector on E, a subspace of $\mathbb{R}^n$, if and only if $P^2 = P$, $P^T = P$ and $\ker(P) = E^\perp$.

Proof. The existence has already been shown using the Hilbert projection theorem. The linear system (1.2) has therefore a unique solution or an infinite number of solutions depending on whether the Gram matrix is invertible or not. Hence it remains to show that $\ker(X) = \ker(X^T X)$ which follows easily from the identity $\|Xu\|^2 = u^T X^T X u$. □

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 demi-definite positive symmetric matrix with eigenvectors u_i and corresponding eigenvalues $\lambda_i \geq 0$, the Moore–Penrose inverse is given by $\sum_i \lambda_i^{-1} u_i u_i^T 1_{\{\lambda_i > 0\}}$.

Corollary 1. *The set of solution of OLS* (1.1) *is given by* $\{(X^T X)^+ X^T Y + u \,:\, u \in \ker(X)\}$.

Proof. Let $u \in \ker(X)$. Verify that $(X^T X)^+ X^T Y + u$ is a solution (see exercise 7). Then assuming that v is a solution, note that $v - (X^T X)^+ X^T Y$ belongs to $\ker(X)$. □

1.3 To centre the data or not to centre the data

We now state the equivalence between this 2 procedures : doing OLS, with the intercept, on (Y, X) (as defined before) and doing OLS, without the intercept, on the centred variables. The later estimation procedure consists in the following. Let $X = (1_n, \tilde{X})$, $Y_c = Y - 1_n(1_n^T Y)/n$ and $\tilde{X}_c = \tilde{X} - 1_n(1_n^T \tilde{X})/n$. Hence the quantities Y_c and $\tilde{X}_c$ are just centred version of Y and $\tilde{X}$, respectively. Define

$$\hat{\boldsymbol{\theta}}_{n,c} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \|Y_c - \tilde{X}_c \boldsymbol{\theta}\|.$$

Proposition 2. *It holds that*

$$\min_{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^p} \|Y_c - \tilde{X}_c \tilde{\boldsymbol{\theta}}\| = \min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \|Y - X\boldsymbol{\theta}\|.$$

and, assuming that X has full rank, we have the following relationship between the traditional OLS and the OLS based on centred data,

$$(\hat{\boldsymbol{\theta}}_{n,1}, \ldots, \hat{\boldsymbol{\theta}}_{n,p}) = \hat{\boldsymbol{\theta}}_{n,c}^T.$$

Consequently, the 2 *methods gives the same predictor.*

Proof. See exercise 9. □

1.4 The determination coefficient

To avoid trivial cases, we suppose in the following that $\sum_{i=1}^n (y_i - \overline{y}^n)^2 > 0$, i.e., that the sequence y_i is not constant. The determination coefficient, denoted by R^2, is defined as the quotient between the explained sum of squares and the total sum of squares. It is given by

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

Because of the orthogonality between $\hat{Y} - Y$ and $\hat{Y}$ and between $\hat{Y} - Y$ and $\overline{y}^n 1_n$, we have that

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \overline{y}^n)^2} = 1 - \frac{\|\hat{Y} - Y\|^2}{\|Y - \overline{y}^n 1_n\|^2}. \tag{1.3}$$

The last expression involves a new quantity, called the residual sum of squares, which is small as soon as the OLS procedure went well, i.e., as soon as the predicted values are close to the observed values. Hence the closer to 1 the R^2 the better. The following statement justifies the use of the R^2 as a score supporting the quality of the OLS estimation :

- $R^2 = 1$ if and only if $Y = \hat{Y}$.
- $R^2 = 0$ if and only if $\hat{Y} = \hat{H}_{1_n} Y$ implying that $\hat{\boldsymbol{\theta}}_n = (\overline{y}^n, 0, \dots, 0)$ is one OLS estimator.

Exercises

Exercise 1. *Show that* $\ker(X^T X) = \ker(X)$ *and that* $\mathrm{span}(X^T) = \mathrm{span}(X^T X)$ *(for the latter, one might first note that* $\ker(X) = \mathrm{span}(X^T)^\perp$*). Deduce that the normal equations always have at least one solution.*

Exercise 2. *Give* $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}$ *and* $\hat{Y} \in \mathbb{R}^n$ *in the case where* $X = 1_n$ *and* $Y \in \mathbb{R}^n$*.*

Exercise 3. *Show that any invertible transformation on the covariate, i.e.* X *is replaced by* XA *with* A *invertible, does not change the prediction* $\hat{Y}$*.*

Exercise 4. *Show that* $\sum_{i=1}^n \hat{\epsilon}_i = 0$*, where* $\hat{\epsilon} = Y - \hat{Y} = (I - \hat{H}_{n,X})Y$*.*

Exercise 5. *Aim is to express the uniqueness condition of the OLS in terms of the empirical covariance matrix* $\hat{\Sigma}_n = n^{-1} \sum_{i=1}^n (x_i - \overline{x}^n)(x_i - \overline{x}^n)^T$ *.*

(a) Show that $\ker(X) = \ker(X^T X)$*.*

(b) Prove that $X^T X = \sum_{i=1}^n x_i x_i^T$*.*

(c) Verify that $ker(X) = 0$ *if and only if the empirical covariance matrix* $\hat{\Sigma}_n$ *is invertible (hint : one might work on the condition that* $\hat{\Sigma}_n$ *is non-invertible, i.e., there exists* $u \in \mathbb{R}^d \backslash \{0\}$ *such that* $\tilde{X}_c u = 0$*).*

Exercise 6. *Aim is to obtain the formula* $\hat{H}_{n,X} = X(X^T X)^+ X^T$*.*

(a) Verify that for any non-negative symmetric matrix $A \in \mathbb{R}^{p \times p}$*, show that* $A^+ A = A^+$*.*

(b) Show that $X(X^T X)^+ X^T$ *is idempotent and symmetric (making it an orthogonal projector).*

(c) Using that $X(X^T X)^+ X^T$ *writes as* UU^T *for some matrix* U *that we shall specify, obtain that* $\ker(\hat{H}_{n,X}) = \ker(X^T)$*.*

(d) Conclude showing that $\mathrm{span}(\hat{H}_{n,X}) = \mathrm{span}(X)$*.*

Exercise 7. *Show that* $\hat{\boldsymbol{\theta}}_n = (X^T X)^+ X^T Y$ *is a solution of the OLS problem.*

Exercise 8. *Show (1.3).*

Exercise 9. *Aim is to prove Proposition 2.*

(a) Start by obtaining that the inequality $\geq$ *holds true.*

(b) Then show that for any sequence (z_i)*, and for all* $z \in \mathbb{R}$*, it holds that* $\|Z - z1_n\| \geq \|Z - \overline{z}^n 1_n\|$ *, where* $Z = (z_1, \dots, z_n)$ *and* $\overline{z}^n = n^{-1} \sum_{i=1}^n z_i$*.*

(c) Find $\hat{a}_n$ *such that, for any* $\theta_0 \in \mathbb{R}$ *and* $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^p$*,* $\|Y - \theta_0 1_n - \tilde{X}\tilde{\boldsymbol{\theta}}\| \geq \|Y - \hat{a}_n(\tilde{\boldsymbol{\theta}})1_n - \tilde{X}\tilde{\boldsymbol{\theta}}\|$ *where* $\tilde{X} \in \mathbb{R}^{n \times p}$ *is the same as* X *without the first column.*

(d) Conclude that $\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|Y_c - \tilde{X}_c \boldsymbol{\theta}\| = \min_{\boldsymbol{\theta} \in \mathbb{R}^p,\, \theta_0 \in \mathbb{R}} \|Y - X(\theta_0, \boldsymbol{\theta}^T)^T\|$

(e) Use the Lebesgue projection theorem to conclude that whenever $\ker(X) = \{0\}$*,* $(\hat{\boldsymbol{\theta}}_{n,1}, \dots, \hat{\boldsymbol{\theta}}_{n,p}) = \hat{\boldsymbol{\theta}}_{n,c}^T$*.*

Exercise 10 (on-line ols and cross-validation)**.** *The goal of this exercise is to show that the OLS estimator $\hat{\boldsymbol{\theta}}_n$ associated with design matrix $X_{(n)} \in \mathbb{R}^{n\times(p+1)}$ and output $\boldsymbol{y}_{(n)} \in \mathbb{R}^n$ can be easily updated when a new pair of observation $(\boldsymbol{x}_{n+1}^T, y_{n+1}) \in \mathbb{R}^{(p+1)} \times \mathbb{R}$ is given. We apply the result to cross validation procedure in the end.*

To clarify the notation:

$$X_{(n+1)} = \begin{pmatrix} X_{(n)} \\ \boldsymbol{x}_{n+1}^T \end{pmatrix} \in \mathbb{R}^{(n+1)\times(p+1)}, \quad \text{and} \quad \boldsymbol{y}_{(n+1)} = \begin{pmatrix} \boldsymbol{y}_{(n)} \\ y_{n+1} \end{pmatrix} \in \mathbb{R}^{n+1}$$

We assume from now on that $X_{(n)}$ and $X_{(n+1)}$ are full column rank (i.e., the columns of each matrix are independent vectors).

NB : Some of the questions require some computation (in particular obtaining (1.4) and (1.6)). Even if you could not prove it, it can be use later.

(a) Let A, B, C, D be matrices with respective sizes (d, d), (d, k), (k, k), (k, d). Show that if A and C are invertible, then

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}. \tag{1.4}$$

(b) Obtain that

$$(X_{(n+1)}^T X_{(n+1)})^{-1} = (X_{(n)}^T X_{(n)})^{-1} - \frac{\boldsymbol{\zeta}_{n+1}\boldsymbol{\zeta}_{n+1}^T}{1 + b_{n+1}} \tag{1.5}$$

where $\boldsymbol{\zeta}_{n+1} = (X_{(n)}^T X_{(n)})^{-1}\boldsymbol{x}_{n+1}$ and $b_{n+1} = \boldsymbol{x}_{n+1}^T(X_{(n)}^T X_{(n)})^{-1}\boldsymbol{x}_{n+1}$.

(c) Express $X_{(n+1)}^T \boldsymbol{y}_{(n+1)}$ with respect to $X_{(n)}^T \boldsymbol{y}_{(n)}$ and $y_{n+1}\boldsymbol{x}_{n+1}$.

(d) Show that the OLS estimator $\hat{\boldsymbol{\theta}}_{n+1}$ associated with design matrix $X_{(n+1)}$ and output $\boldsymbol{y}_{(n+1)}$ can be obtained as follows:

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n + \frac{u_{n+1}}{1 + b_{n+1}}\boldsymbol{\zeta}_{n+1}, \tag{1.6}$$

where $u_{n+1} = y_{n+1} - \boldsymbol{x}_{n+1}^T\hat{\boldsymbol{\theta}}_n$.

(e) Keeping in memory $(X_{(n)}^T X_{(n)})^{-1}$ and $\hat{\boldsymbol{\theta}}_n$, explain how to update $\hat{\boldsymbol{\theta}}_{n+1}$ using a minimal number of operations of the kind : matrix $(p + 1, p + 1)$ times vector $(p + 1, 1)$. How many such operation are needed?

(f) Using Equation (1.5) above, show that

$$1 + b_{n+1} = \frac{1}{1 - h_{n+1}}$$

where $h_{n+1} = \boldsymbol{x}_{n+1}^T(X_{(n+1)}^T X_{(n+1)})^{-1}\boldsymbol{x}_{n+1}$.

(g) The prediction of y_{n+1} given by the model is $\hat{y}_{n+1} := \boldsymbol{x}_{n+1}^T\hat{\boldsymbol{\theta}}_{n+1}$. With the following formula

$$\hat{y}_{n+1} = \boldsymbol{x}_{n+1}^T\hat{\boldsymbol{\theta}}_n + \frac{u_{n+1}b_{n+1}}{1 + b_{n+1}}.$$

prove that

$$y_{n+1} - \hat{y}_{n+1} = u_{n+1}(1 - h_{n+1}).$$

(h) Given some data $(\boldsymbol{y}, X)$, leave-one-out cross-validation consists in computing the risk

$$R_{cv} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}_{(-i)})^2$$

where $\hat{\boldsymbol{\theta}}_{(-i)}$ is the OLS estimator based on $(\boldsymbol{y}_{(-i)}, X_{(-i)})$, i.e., the data $(\boldsymbol{y}, X)$ without the i-th line. Applying what have been done so far, show that

$$R_{cv} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(1-\hat{h}_i)^2,$$

with $\hat{h}_i = \boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i$ and $\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}_n$, $\hat{\boldsymbol{\theta}}_n$ being the OLS estimator of $(\boldsymbol{y}, X)$.

Chapter 2

Statistical model

In the previous section, we have defined the OLS estimator based on the observed data without any assumption on the generating process associated to the data. When assuming that the observations 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 probabilistic models : fixed design model, random design model and the Gaussian noise model.

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 sequence of deterministic points in $\mathbb{R}^{p+1}$ and (ϵ_i) is a 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 centred random variables 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 sequence (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 determined 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 that we present 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 $\text{span}(X)$ equals $p+1$, implying that $\ker(X) = \{0\}$ and that $\hat{\boldsymbol{\theta}}_n$ is unique. We therefore implicitly assume that $n \geq p+1$. We can now state a useful decomposition: provided that $\ker(X) = \{0\}$, it holds that

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star = (X^T X)^{-1} X^T \epsilon. \tag{2.1}$$

2.1.1 Bias, variance and risk

The bias, the variance and the risk are important quantities because they are measures of the estimation quality. For instance, an estimator is accurate when the bias is 0 and the variance is small. The following notion of bias is related to the whole statistical model (for all $\theta^\star$, not for a particular one).

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 $\mathbb{E}[\boldsymbol{\theta}(X,Y)]=\boldsymbol{\theta}^\star$.*

The risk measures the average error associated to an estimation procedure. Different notions of risk can be defined: the quadratic risk is defined on the regression coefficients β, the prediction risk takes care of the prediction error, i.e., the error when predicting y. Formal definitions are given below.

Definition 3. *The quadratic risk associated to $\hat{\boldsymbol{\theta}}_n$ estimating $\boldsymbol{\theta}^\star$ is*

$$R_{quad}(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^\star)=\mathbb{E}[\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^\star\|^2].$$

The prediction risk is

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

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

Proposition 3. *When $\ker(X)=\{0\}$, the following holds:*

(i) the OLS estimator is unbiased i.e., it holds that $\mathbb{E}[\hat{\boldsymbol{\theta}}_n]=\boldsymbol{\theta}^\star$.

(ii) Its variance is given by $\operatorname{var}(\hat{\boldsymbol{\theta}}_n)=(X^TX)^{-1}\sigma^2$.

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

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

Hence whenever the smallest eigenvalue of $\hat{G}_n$ is larger than b (independently 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.

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 matrices. 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 $u^TV_1u\leq u^TV_2u$ 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 any linear combination of $\hat{\beta}_1$ has a smaller variance than the same linear combination of $\hat{\beta}_2$.

Definition 4. *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 4 (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.,*

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

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

Proof. First note that AY is unbiased if and only if $(A-(X^TX)^{-1}X^T)X\boldsymbol{\theta}^\star=0$ for all $\boldsymbol{\theta}^\star$, equivalently, $BX=0$ with $B=(A-(X^TX)^{-1}X^T)$. Consequently, using that $E[\epsilon\epsilon^T]=\sigma^2I_n$, $\operatorname{cov}(BY,\hat{\boldsymbol{\theta}}_n)=0$. Then, just write

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

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

2.1.3 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 one knew the residuals (ϵ_i), one would take the empirical variance of $\epsilon_1, \dots, \epsilon_n$, 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 (1.2), we have $\sum_{i=1}^n (Y_i - \hat{Y}_i) = 0$. Consequently, $\tilde{\sigma}_n^2$ is simply the empirical variance estimate of the residual vector $Y_i - \hat{Y}_i$. Noting that $\tilde{\sigma}_n^2 = n^{-1}\|(I_n - \hat{H}_{n,X})\epsilon\|^2$ 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

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

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 $i = 1, \dots, n$.

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:

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

where (x_i) is non-random sequence of vector in $\mathbb{R}^{p+1}$. We keep assuming that $\ker(X) = \{0\}$ in the following.

Lemma 1. *Let $\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top Y$. Then,*

$$\hat{\theta} = \boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \epsilon$$

and

$$\hat{\theta} \sim \mathcal{N}(\boldsymbol{\theta}^*, \sigma^2 (X^\top X)^{-1}).$$

Proof. Recall that $\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top Y$, therefore, using the definition of Y we can write this as

$$\begin{aligned} \hat{\theta} &= (X^\top X)^{-1} X^\top Y \\ &= (X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^* + \epsilon) \\ &= (X^\top X)^{-1} X^\top X \boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \epsilon \\ &= \boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \epsilon. \end{aligned} \tag{2.2}$$

This proves the first claim.

Since ϵ follows a multi-variate normal distribution, $\boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \epsilon$ is also normally distributed. Taking expectations we get

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \mathbb{E}(\boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \epsilon) = \boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \mathbb{E}(\epsilon) = \boldsymbol{\theta}^*,$$

since $\mathbb{E}(\epsilon) = 0$.

For the covariance,

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\theta}}) &= \mathrm{Cov}(\boldsymbol{\theta}^* + (X^\top X)^{-1}X^\top \epsilon) \\ &= \mathrm{Cov}((X^\top X)^{-1}X^\top \epsilon) \\ &= (X^\top X)^{-1}X^\top \mathrm{Cov}(\epsilon)((X^\top X)^{-1}X^\top)^\top \\ &= (X^\top X)^{-1}X^\top \sigma^2 I X(X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1}X^\top X(X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1}.\end{aligned} \tag{2.3}$$

□

2.2.1 The hat matrix

In this and the following sections, we will use the Hat matrix and various properties.

$$H = X(X^\top X)^{-1}X^\top.$$

Lemma 2. *The hat matrix H has the following properties:*

- *H is symmetric, i.e., $H^\top = H$.*
- *H is idempotent, i.e., $H^2 = H$.*
- $(I-H)^\top = I - H$
- $(I-H)^2 = I - H$
- $HX = X$
- $(I-H)X = 0$

Proof.

$$\begin{aligned}&H^\top = (X(X^\top X)^{-1}X^\top)^\top = (X^\top)^\top((X^\top X)^{-1})^\top X^\top = X(X^\top X)^{-1}X^\top = H, \\ &H^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top = H. \\ &(I-H)^\top = I^\top - H^\top = I - H \\ &(I-H)^2 = (I-H)(I-H) = I^2 - HI - IH + H^2 = I - H - H + H = I - H.\end{aligned} \tag{2.4}$$

□

2.2.2 The Cochran lemma

Our main tool in this and the following section will be a simplified version of Cochran's theorem:

Lemma 3. *The following statements are true:*

$$\begin{aligned}&H\varepsilon \text{ and } (I-H)\varepsilon \text{ are independent,} \\ &\frac{1}{\sigma^2}\varepsilon^\top H\varepsilon \sim \chi^2_{p+1}) \\ &\frac{1}{\sigma^2}\varepsilon^\top (I-H)\varepsilon \sim \chi^2_{n-p-1}\end{aligned} \tag{2.5}$$

Proof. We start by some definitions and remarks. Let A be an $n \times n$ matrix. A scalar λ is called an eigenvalue of A if there exists a non-zero vector v such that

$$Av = \lambda v.$$

The vector v is called an eigenvector corresponding to the eigenvalue λ. For a diagonal matrix D, the eigenvalues are precisely the entries on the diagonal. The eigenvalues of an idempotent matrix ($A^2 = A$) are either 0 or 1, and the number of eigenvalues equal to 1 is then $tr(A)$.

Let A be a symmetric matrix. The two matrices A and B are similar if there exists an orthogonal matrix P such that $B = P^{-1}AP$. If A and B are similar matrices, they share the same eigenvalues. The Jacobi eigenvalue algorithm is a numerical method for the computation of the eigenvalues of a symmetric matrix.

We now proof the first result. Since H is symmetric ($H^T = H$), we can diagonalize H (using, for example, the Jacobi method). Therefore, there is an orthogonal matrix U such that $D := UHU^\top$ is diagonal, and the diagonal elements of D are the eigenvalues of H. Since H is idempotent, these diagonal elements can only be 0 or 1. Also, since U is orthogonal, we have $U^\top U = I$, and thus

$$U^\top DU = U^\top UHU^\top U = H.$$

The same matrix U also diagonalizes $I - H$, since $U(I - H)U^\top = UU^\top - UHU^\top = I - D$. Exactly one of the diagonal elements D_{ii} and $(I - D)_{ii}$ is 1 and the other one is 0 for every i.

Since $\varepsilon \sim N(0, \sigma^2 I)$, we find that $\eta := U\varepsilon$ is normally distributed with mean $U0 = 0$ and covariance matrix $\sigma^2 UIU^\top = \sigma^2 UU^\top = \sigma^2 I$. Thus, η has the same distribution as ε does: $\eta \sim N(0, \sigma^2 I)$, and the components η_i are independent of each other. We have

$$H\varepsilon = U^\top DU\varepsilon = U^\top D\eta.$$

and

$$(I - H)\varepsilon = U^\top (I - D)U\varepsilon = U^\top (I - D)\eta.$$

Since $(D\eta)_i = 0$ if $D_{ii} = 0$ and $((I - D)\eta)_i = 0$ otherwise, each component of η contributes to exactly one of the two vectors $D\eta$ and $(I - D)\eta$. Thus, $D\eta$ and $(I - D)\eta$ are independent, and thus $H\varepsilon$ and $(I - H)\varepsilon$ are also independent. This proves the first statement of the theorem.

For the second statement, we note that

$$\varepsilon^\top H\varepsilon = \varepsilon^\top U^\top DU\varepsilon = \eta^\top D\eta = \sum_{i=1, D_{ii}=1}^{n} \eta_i^2.$$

Since X has full rank, then $\text{rank}(X) = \min(n, p + 1) = p + 1$, and we know $(X^\top X)^{-1}$ exists. By commutativity of the trace operator, we have

$$\text{tr}(H) := \text{tr}\left(X(X^\top X)^{-1}X^\top\right) = \text{tr}\left(X^\top X(X^\top X)^{-1}\right) = \text{tr}[I_{p+1}] = p + 1$$

Since $X^\top X$ is invertible, $rank(H) = p + 1$, there are $p + 1$ terms contributing to the sum. Thus,

$$\frac{1}{\sigma^2}\varepsilon^\top H\varepsilon = \sum_{i=1}^{n}(\eta_i/\sigma)^2$$

is the sum of the squares of $p + 1$ independent standard normals, and thus is χ^2_{p+1} distributed. □

Finally, the third statement follows in much the same way as the first one, except that H is replaced with $I - H$ and the sum is over the $n - p - 1$ indices i where $D_{ii} = 0$. This completes the proof.

A direct application of the previous proposition gives us the following equality, which is informative on the estimation error, for any $k = 0, \ldots, p$,

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

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

2.2.3 Estimating the Error Variance

We are now ready to establish the result for the residual variance. Recall that we define the residuals as $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

Theorem 1. *The unbiased estimator for the variance of the residuals , i.e., $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$, is given by*

$$\hat{\sigma}^2 := \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

Proof. We first note that

$$\begin{aligned}
(n-p-1)\hat{\sigma}^2 &= (Y-\hat{Y})^\top (Y-\hat{Y}) \\
&= (Y-HY)^\top (Y-HY) \\
&= Y^\top (I-H)^\top (I-H)Y \\
&= Y^\top (I-H)Y \\
&= (X\theta^* + \varepsilon)^\top (I-H)(X\theta^* + \varepsilon) \\
&= \theta^{*\top} X^\top (I-H)X\theta^* + 2\varepsilon^\top (I-H)X\theta^* + \varepsilon^\top (I-H)\varepsilon \quad (\textit{Lemma 2}) \\
&= \varepsilon^\top (I-H)\varepsilon.
\end{aligned} \tag{2.6}$$

Now we can apply Cochran's lemma. This shows that

$$\frac{1}{\sigma^2}(n-p-1)\hat{\sigma}^2 = \frac{1}{\sigma^2}\varepsilon^\top (I-H)\varepsilon \sim \chi^2_{n-p-1}.$$

Since the expectation of a χ^2_k distribution equals k, we find

$$\frac{1}{\sigma^2}(n-p-1)\mathbb{E}(\hat{\sigma}^2) = n-p-1$$

and thus

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2.$$

□

There are two remarkable corollaries to these results.

Proposition 5. *The random vector of coefficients $\hat{\boldsymbol{\theta}}$ and the random number $\hat{\sigma}^2$ are independent of each other.*

Proposition 6. *The quantity*

$$T := \frac{\hat{\theta}_i - \theta_i}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{ii}}}$$

follows a T-student distribution of $n-p-1$ degrees of freedom (provided X is full rank).

2.2.4 A concentration inequality

We now provide an additional guarantee for the OLS estimator under the Gaussian model. It consists of a concentration inequality : an upper bound on the probability that the estimation error exceeds any given $t > 0$. The upper bound unsurprisingly depends on p, n, t, and the smallest eigenvalue of $\hat{G}_n$.

Proposition 7. *Suppose that the Gaussian model is valid. Denote by $\hat{\lambda}_n$ the smallest eigenvalue of $\hat{G}_n$ and suppose that $\hat{\lambda}_n > 0$ for all $n \geq 1$. Then, for any $k \in \{0, \ldots, p\}$, $n \geq 1$ and $\delta > 0$, it holds with probability $1 - \delta$,*

$$\left|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k\right| \leq \sqrt{\frac{2\sigma^2 \hat{s}^2_{n,k} \log(2/\delta)}{n}}.$$

where $\hat{s}^2_{n,k} = e_k^T \hat{G}_n^{-1} e_k$. Moreover,

$$\max_{k=0,\ldots,p} \left|\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}^\star_k\right| \leq \sqrt{\frac{2\sigma^2 \log(2(p+1)/\delta)}{n\hat{\lambda}_n}}.$$

Proof. Let $\tilde{X}_i = (X^T X)^{-1} X_i$. Apply Lemma 7 of Appendix C to the sequence $\sum_{i=1}^n (u^T \tilde{X}_i)\epsilon_i$ to obtain that

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

Choosing $u = e_k$, we have $\sum_{i=1}^n (u^T \tilde{X}_i)^2 = n^{-1}\hat{s}^2_{n,k}$, and using (2.1), we obtain that

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

Choose t appropriately to obtain 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 7 is important as it shows that each coordinate might not behave similarly depending on the associated diagonal element of $\hat{G}_n$. For instance, for the intercept, the bound just becomes $\sqrt{2\sigma^2 \log(2/\delta)/n}$. The quantity $\hat{s}_{n,k}$ will play an important role in practice when building confidence intervals (see section 3).*

Remark 2. *Proposition 7 suggests that the value of the smallest eigenvalue $\hat{\lambda}_n$ of $\hat{G}_n$ plays a certain role on the accuracy of the estimation. The smaller $\hat{\lambda}_n$ the worst the estimation accuracy.*

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. Recall that $X_1 \in \mathbb{R}^{p+1}$ and $X_{1,0} = 1$ almost surely.

Proposition 8. *Suppose that for all $k = 0, \ldots, p$, $\mathbb{E}[X^2_{1,k}] < \infty$ 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. □

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. Note that $\boldsymbol{\theta}^*$ is not unique whenever one variable is a combination of the others. In this case one might consider any of the solution, e.g., $\boldsymbol{\theta}^* = \mathbb{E}[X_1 X_1^T]^+ \mathbb{E}[X_1 Y_1]$. 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 9. *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. For more details, see Exercise 11.

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

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

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

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

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

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

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

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 : $f^*(X_1) = \mathbb{E}[Y_1 | X_1]$, almost surely.

Exercises

Exercise 11 (Asymptotics for the OLS in Random design)**.** *Let $(X_1, Y_1), (X_2, Y_2), \dots$ be an i.i.d. sequence of random vectors. Each pair (X_i, Y_i) is valued in $\mathbb{R}^p \times \mathbb{R}$. Denote by $X_i^T = (X_i^{(1)}, \dots, X_i^{(p)})$. Suppose that for all $(k,l) \in \{1,\dots,p\}^2$, $\mathbb{E}[|X_1^{(k)} X_1^{(l)}|] < \infty$ and $G = \mathbb{E}[X_1 X_1^T]$ is invertible. The goal is to show that*

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

where $\boldsymbol{\theta}^*$ is defined in Proposition 8. Recall that for each $n \in \mathbb{N}_+^*$, the OLS is given by

$$\hat{\boldsymbol{\theta}}_n = \hat{G}_n^+ \left(n^{-1} \sum_{i=1}^n X_i Y_i \right), \tag{2.7}$$

with $\hat{G}_n = n^{-1} \sum_{i=1}^n X_i X_i^T$.

1. *Let $\mathcal{S}_p(\mathbb{R})$ be the space of symmetric matrices with real coefficients. Let $T : \mathcal{S}_p(\mathbb{R}) \to \mathcal{S}_p(\mathbb{R})$ be such that $T(A) = A^+$. Show the continuity of T at each point A such that $\det(A) \neq 0$. We recall that for any A such that $\det(A) \neq 0$, $A^{-1} = (\det(A))^{-1} Com(A)^T$ where $Com : \mathcal{S}_p(\mathbb{R}) \to \mathcal{S}_p(\mathbb{R})$ is continuous (it is called the comatrix).*
2. *What is the limit of $\hat{G}_n^+$? In which sense?*
3. *Show that $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* = \hat{G}_n^+ \hat{\mu}_n + (\hat{G}_n^+ \hat{G}_n - I_p)\boldsymbol{\theta}^*$ with $\hat{\mu}_n = n^{-1} \sum_{i=1}^n X_i (Y_i - X_i^T \boldsymbol{\theta}^*)$.*
4. *Obtain that $\sqrt{n} \hat{G}_n^+ \hat{\mu}_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$ where Σ needs to be determined.*
5. *Prove that $\sqrt{n}(\hat{G}_n^+ \hat{G}_n - I)\beta_0 \xrightarrow{\mathbb{P}} 0$. One can consider the event $\det(G_n) \neq 0$.*
6. *Show that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$.*
7. *Let $k \in \{1,\dots,p\}$, find $\hat{s}_{n,k}$, depending only on σ^2 and $\hat{G}_n$, such that $\sqrt{n}(\frac{\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_k^*}{\hat{s}_{n,k}}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.*
8. *Deduce a $(1-\alpha)$-confidence interval for $\boldsymbol{\theta}_k^*$. Verify it has level $1-\alpha$.*

Chapter 3

Confidence intervals and hypothesis testing

3.1 Confidence intervals

From a practical perspective, building confidence intervals is often an inevitable step as it permits to evaluate the quality of the estimation. The construction of confidence intervals follows the estimation step. Intuitively, a confidence interval is simply a region (based on the observed data) in which the parameter of interest is most likely to lie. The accuracy/quality of the estimation is then naturally measured by the size of the underlying confidence interval. As we shall see, the construction of a confidence interval is based on the estimation of the variance.

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 (where $k =\in \{0, \dots, p\}$).

Definition 5. *A confidence interval of level* $1-\alpha$ *is an interval* $\hat{I}_n(Y, X) \subset \mathbb{R}$ *satisfying, for all* $n \geq 1$*,*

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

3.1.1 Gaussian model

Confidence intervals for the regression coefficients

Confidence intervals 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 **??**. Recall that, when it exists,

$$\hat{s}_{n,k}^2 = e_k^T \hat{G}_n^{-1} e_k.$$

Proposition 10. *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 the distribution* $\mathcal{T}_{n-p-1}$*, is a confidence interval of level* $1-\alpha$*.*

Confidence intervals for the predicted values

We are now interested in building confidence intervals for the predicted value under the true model at a single given point $x = (1, x_1, \dots, x_p) \in \mathbb{R}^p$. The predicted value at x under the true model is defined as

$y^* = x^T\boldsymbol{\theta}^*$. In the Gaussian model, using preservation properties of the Student's distribution, we find the following confidence interval $\mathrm{CI}(x)$ of level $1-\alpha$. With probability equal to $1-\alpha$,

$$y^* \in \mathrm{CI}(x),$$

where

$$\mathrm{CI}(x) = x^T\hat{\boldsymbol{\theta}}_n \pm Q_{n-p-1}(1-\alpha/2)\hat{\sigma}\sqrt{x^T(X^TX)^{-1}x},$$

and $\hat{\sigma}_n^2 = \sum_{i=1}^n \big(Y_i - x_i^T\hat{\boldsymbol{\theta}}_n\big)^2/(n-p-1)$ (it has been introduced in Chapter 2). A related question is to build a confidence interval on the value of y (not y^*) under the true model. This can be done in a similar manner as before but one needs to pay a particular attention to the additive noise in the model. Indeed, we have that $y = y^* + \epsilon$ where $\epsilon \sim \mathcal{N}(0,\sigma^2)$. It follows that

$$y \in \mathrm{PI}(x),$$

with

$$\mathrm{PI}(x) = x^T\hat{\boldsymbol{\theta}}_n \pm Q_{n-p-1}(1-\alpha/2)\hat{\sigma}\sqrt{1 + x^T(X^TX)^{-1}x}.$$

For more details on the derivation of those confidence intervals, see Exercise 12.

3.1.2 Nongaussian case

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 intervals:

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

We start by deriving 2 confidence intervals based, respectively, on two concentration inequalities : the Markov and the Hoeffding inequalities.

Proposition 11. *In the fixed design model, suppose that (for clarity) $X^TX = nI_n$ and that (ϵ_i) is an identically distributed sequence of centered random variables with variance σ^2, then for each $k \in \{0,1,\dots,p\}$, 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, $|\epsilon_i| \le c$ for all $i = 1,\dots,n$, then for each $k \in \{0,1,\dots,p\}$, the interval

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

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

Proof. We have using (2.1), $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star = X^T\epsilon/n$. Applying the Markov inequality

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

leading to the first confidence interval.

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

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

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

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

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 12. *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 the distribution* $\mathcal{N}(0,1)$*, is, asymptotically, a confidence interval of level* $1-\alpha$*, i.e.,*

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

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 as they might quantify the importance of each covariate in the regression. As an application, we consider the forward variable selection method in Section 3.3.

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 "not to 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 $\boldsymbol{\theta}_1^\star = 3.14159$. There are basically 2 kinds of error that we wish to control:

$$\begin{array}{ll} \text{Type-1:} & \text{to reject whereas } \mathcal{H}_0 \text{ is true} \\ \text{Type-2:} & \text{not to reject whereas } \mathcal{H}_0 \text{ is not true.} \end{array}$$

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 imposes that, for any level $1-\alpha$, asymptotically, the level is smaller than α while the power is one. To achieve consistency, it is natural to let the reject region depend on α.

Definition 6. *A statistical test $(\hat{T}_n, \mathcal{R}_\alpha)$ is said to be (asymptotically) consistent whenever for all level $1-\alpha \in (0,1)$*

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

Remark 4. *In practice, a standard choice is $\alpha = 0.05$. Of course when the sample size is too small one cannot be too demanding and larger values of α might be more reasonable.*

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 13. *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, respectively given by*

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

produce a consistent test.

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 $\hat{T}_{n,k} \in \mathcal{R}_\alpha$ 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 on 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 9 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 5. *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 6. *In the Gaussian case, the test statistic and the reject region are given by*

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

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 3.1: 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 disease evolution. Each covariate has been standardized **?**.

Remark 7 (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

The method of forward selection is a stepwise procedure that aims at selecting the most "important" variables. The method starts with no covariate and add a new one at each step. This kind of methods is sometimes referred to as *greedy methods*. The criterion used to select the best covariate follows from the test statistic for the test of no effect: $n^{1/2}|\hat{\boldsymbol{\theta}}_{n,k}|/(\hat{s}_{n,k}\hat{\sigma}_n)$. Intuitively, the larger the statistic, the more important the effect of the k-th variable.

More formally, let $X = (1_n, \tilde{X}_1, \ldots, \tilde{X}_p)$. Each (non-constant) covariate $\tilde{X}_k$ is competing against the others via 1-dimensional regression submodels $Y \simeq \boldsymbol{\theta}_0 + X_k\boldsymbol{\theta}_k$. For any $Y \in \mathbb{R}^n$ and $\tilde{X}_k \in \mathbb{R}^n$, define the OLS

$$\hat{\boldsymbol{\theta}}_n(Y, \tilde{X}_k) = \text{argmin}_{(\theta_0,\theta_1)\in\mathbb{R}^2}\|Y - \theta_0 1_n - \theta_1\tilde{X}_k\|^2.$$

Within each submodel, the Gram matrix and the noise level estimate are given by

$$\hat{G}_n(\tilde{X}_k) = n^{-1}(1_n, \tilde{X}_k)^T(1_n, \tilde{X}_k),$$
$$\hat{\sigma}_n(Y, \tilde{X}_k)^2 = (n-2)^{-1}\|Y - (1_n, \tilde{X}_k)\hat{\boldsymbol{\theta}}_n(Y, \tilde{X}_k)\|^2.$$

Another quantity of interest is $\hat{s}_n(\tilde{X}_k)^2 = e_1^T\hat{G}_n(\tilde{X}_k)^{-1}e_1$. The criterion used to compare the importance of each variable is the test statistic of the test of no effect, computed within each submodel:

$$\hat{T}_n(Y, \tilde{X}_k) = \frac{\hat{\boldsymbol{\theta}}_n(Y, \tilde{X}_k)}{\hat{s}_n(\tilde{X}_k)\hat{\sigma}_n(Y, \tilde{X}_k)}.$$

For each covariate, such a quantity is compared and the largest value is selected. This criterion has an interpretation in terms of p-values. When the test is described by $(\hat{T}_n(Y, \tilde{X}_k), \mathcal{R}_\alpha)$, the p-value is the smallest value of α for which we still reject. For instance, in the *random design model*,

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

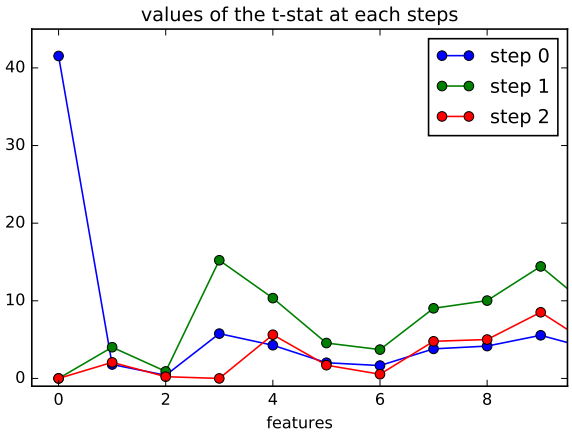

Figure 3.1: The statistics of each selected variable is 0 in the next step. The intercept is the first selected variable, then X_3, etc...

Hence taking the largest $\hat{T}_n(Y, \tilde{X}_k)$ 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 than 0.05. As soon as one variable, say $\tilde{X}_k$, is selected, one needs to account for the predictive information it has brought in the modeling of Y. This is to prevent from selecting 2 identical covariates. This is done by replacing the output Y by the residual $Y - (1_n, \tilde{X}_k)\hat{\boldsymbol{\theta}}_n(Y, \tilde{X}_k)$.

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 $k \in \mathcal{A} \backslash \mathcal{S}$, compute $\hat{T}_n(r, \tilde{X}_k)$.

(ii) Stop if no p-values are smaller than p_{stop}.
Else compute $k^* \in \operatorname{argmax} \hat{T}_n(r, \tilde{X}_k)$.
And update $\mathcal{S} = \mathcal{S} \cup \{k^*\}$ and $r = r - (1_n, \tilde{X}_{k^*})\hat{\boldsymbol{\theta}}_n(Y, \tilde{X}_{k^*})$.

Figure 3.3 illustrates the procedure described by Algorithm 1 applied to the "diabetes" dataset of sklearn presented in Table 3.1.

Remark 8. *Different stopping rules might be considered. For instance, in* **?***, the authors recommend to consider the residuals sum of squares and to stop as soon as $\|r\|^2 < \epsilon$.*

Exercises

Exercise 12 (explicit formulas when $p = 1$ for prediction intervals)**.** *Let us consider the following fixed-design one-dimensional ($p = 1$) linear regression model:*

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i\,, \quad \epsilon_i \sim \mathcal{N}(0, \sigma) \quad i.i.d.\,, \quad i = 1, ..., n\,.$$

Being a particular but simply interpretable case it facilitates intuitive understanding and enables easy two-dimensional visualization. Let $\overline{x}^n = n^{-1}\sum_{i=1}^n x_i$ and $\overline{Y}^n = n^{-1}\sum_{i=1}^n Y_i$. We further assume that x_i is not constant, i.e., that $\sum_{i=1}^n (x_i - \overline{x}^n)^2 \neq 0$.

1. *Show that the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are*

$$\hat{\beta}_0 = \overline{Y}^n - \hat{\beta}_1 \overline{x}^n \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x}^n)(Y_i - \overline{Y}^n)}{\sum_{i=1}^n (x_i - \overline{x}^n)^2}$$

2. *Show that*

$$e_0^T(X^TX)^{-1}e_0 = \Big(\frac{1}{n} + \frac{\overline{x}^{n2}}{\sum_{i=1}^n (x_i - \overline{x}^n)^2}\Big) \quad \textit{and} \quad e_1^T(X^TX)^{-1}e_1 = \frac{1}{\sum_{i=1}^n (x_i - \overline{x}^n)^2},$$

3. *Give the distribution of* $\mathbb{V}[\hat{\beta}_0]^{-1/2}(\hat{\beta}_0 - \beta_0)$ *and* $\mathbb{V}[\hat{\beta}_1]^{-1/2}(\hat{\beta}_1 - \beta_1)$

$$\mathbb{V}[\hat{\beta}_0] = \hat{\sigma}^2\Big(\frac{1}{n} + \frac{\overline{x}^{n2}}{\sum_{i=1}^n (x_i - \overline{x}^n)^2}\Big) \quad \textit{and} \quad \mathbb{V}[\hat{\beta}_1] = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \overline{x}^{n2})^2},$$

where $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \big(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\big)^2$.

4. *Give the reject region for the test* $\mathcal{H}_0 : \beta_j = 0$.

5. *For a new pair* (Y, x) *observed from the Gaussian model above, the value* $\hat{\beta}_0 + \hat{\beta}_1 x$ *is called the point prediction. Show that*

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x) - (\beta_0 + \beta_1 x)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\overline{x}^n)^2}{\sum_{i=1}^n (x_i - \overline{x}^n)^2}}} \sim t(n-2) \quad \textit{and} \quad \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x)}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x}^n)^2}{\sum_{i=1}^n (x_i - \overline{x}^n)^2}}} \sim t(n-2).$$

6. *Build confidence intervals for* $(\beta_0 + \beta_1 x)$ *and* Y. *Note that these intervals correspond, respectively, to CI and PI given in section 3.1.1. The last one is often called prediction interval.*

Chapter 4

Ridge regularization

In this chapter we use the singular-value decomposition (SVD), a matrix decomposition presented in Appendix B. The SVD of X shall provide useful expression for quantities related to the OLS estimate. The SVD is also important to understand principal component analysis (PCA), a method that compresses the data without loosing too much information, also presented in Appendix B.

The ridge estimator is introduced to overcome the issues caused by poorly conditioned Gram matrix $\hat{G}_n$, i.e., when some of the eigenvalues are too small. As indicated by the singular-value decomposition of $X = \sum_{k=1}^{r} s_i u_i v_i^T$, where r stands for the rank of X and s_i (resp. u_i and v_i) are the singular-values (resp. singular-vector) of X, we have that $\hat{\theta}_n = \sum_{k=1}^{r} s_i v_i u_i^T y$. Consequently, the estimate is numerically unstable as soon as some of the s_i are close to 0. As we can see looking at the variance of the OLS or at Proposition 7, the smallest eigenvalues of $\hat{G}_n$ have a bad influence on the statistical behaviour of the OLS. The ridge estimator is a solution to control these bad effects due to poor conditioning. Before going through the definition of the ridge estimator, we discuss another method which consists in doing PCA before running OLS.

For ease of notation (to avoid working with $\tilde{X}_c$ as before), in the rest of the chapter, we consider the formulation of OLS without intercept with centered variable $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$. As established in Proposition 2, this is equivalent to include an intercept in the OLS with non-centered variables.

4.1 PCA before OLS

Many practitioners are familiar with the method of combining PCA and OLS. In addition to visualize and explore the centered covariates X, aim is to reduce the number of covariates to avoid inverting a possibly too large matrix $X^T X$. Be careful that after running PCA, due to its definition (see Definition 8 in Appendix B), the prediction must be operated with respect to the centered covariate X and centered output Y. Hence the intercept is no longer necessary as explained in Proposition 2. The algorithm is as follows.

Algorithm 2 (PCA before OLS)**.**
***Inputs**: (Y, X), centered variables, and an integer k (the number of components to keep). **output**: prediction of Y at $x \in \mathbb{R}^p$.*

(i) Do a PCA on X and keep the k first components $U_1, \dots, U_k$.

(ii) Let $P_k = \sum_{i=1}^{k} U_i U_i^T$. Compute $\hat{\boldsymbol{\theta}}_{n,k}$, the OLS associated with (Y, XP_k).

(iii) Return the prediction $x^T P_k \hat{\boldsymbol{\theta}}_{n,k}$.

Trying to legitimize the approach, one can write

$$\|Y - X\hat{\boldsymbol{\theta}}_n\| \leq \|Y - XP_k\hat{\boldsymbol{\theta}}_{n,k}\| \leq \|Y - X\hat{\boldsymbol{\theta}}_n\| + \|X(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n,k})\| + \|(XP_k - X)\hat{\boldsymbol{\theta}}_{n,k}\|$$

in which, by Proposition 26, the last term should be small. However, the second term in the right hand side might be large. The lack of guarantee for this approach is due to the fact that the PCA used from the beginning is independent of the output Y. Doing such a process might result in some loss in accuracy.

4.2 Definition of the Ridge estimator

For centered variables (Y, X), the ridge estimator is defined as a solution of the following minimization problem

$$\|Y - X\boldsymbol{\theta}\|^2 + n\lambda\|\boldsymbol{\theta}\|^2, \tag{4.1}$$

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

- Intuitively, when $\lambda \to 0$, we obtain the OLS. When $\lambda \to +\infty$, we estimate 0.
- 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 $\sum_{k=1}^{p} |\boldsymbol{\theta}_k|$ (see the next chapter about the LASSO).
- As the expression in (4.1) is a Lagrangian with constraint $\|\boldsymbol{\theta}\|^2 \leq c$ the Ridge is an OLS under constraints. The link between c and λ is not explicit.
- To make the ridge estimate scale invariant, one might replace X by $XD^{-1/2}$ where D is the diagonal matrix with entries $e_k^T X^T X e_k$. Actually, this normalization permits to justify having 1 single parameter λ to control the influence of the penalty. The ridge estimate is classically defined without intercept (to prevent from penalizing the intercept). Hence one needs to first center Y and X so that the intercept of the OLS is automatically 0.

Proposition 14. *The minimizer of (4.1) 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 (4.1). Considering the behaviour of f at the limit of the domain, 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 uniqueness, note that, for any u and v, we have

$$\|tu + (1-t)v\|^2 = t\|u\|^2 + (1-t)\|v\|^2 - t(1-t)\|u - v\|^2. \tag{4.2}$$

Then suppose that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two distinct minimizers with $f^* = f(\boldsymbol{\theta}_1) = f(\boldsymbol{\theta}_2)$. We have, from (4.2),

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

Hence $\hat{\boldsymbol{\theta}}_n^{(rdg)}$ is unique. The first order equation is

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

□

4.3 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 equations 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 Gram 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.

Proposition 15. *In the fixed-design model :*

(i) The bias of the ridge is $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{(rdg)}] - \boldsymbol{\theta}^\star = -\lambda n(X^TX + n\lambda I_p)^{-1}\boldsymbol{\theta}^\star$

(ii) The variance of the ridge estimator expresses as $\mathrm{var}(\hat{\boldsymbol{\theta}}_n^{(rdg)}) = \sigma^2(X^TX + n\lambda I_p)^{-1}X^TX(X^TX + n\lambda I_p)^{-1}$.

(iii) We have that $\mathrm{var}(\hat{\boldsymbol{\theta}}_n^{(rdg)}) < \mathrm{var}(\hat{\boldsymbol{\theta}}_n)$*, where* $\hat{\boldsymbol{\theta}}_n$ *is the OLS solution.*

Proof. For the last point, 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

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

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

4.4 Choice of the regularization parameter

As we have seen before, ridge regression reduces the variance of the OLS but introduces 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 to select the parameter λ. This technique of cross validation works in more general context and is of common use as soon as one needs to choose a parameter to run a method. Examples include the choice of the bandwidth in kernel smoothing methods, the choice of the scale parameter in RKHS as well as the choice of the cut-off parameter in Huber regression.

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 9. *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.*

Exercises

Exercise 13. *Recall the SVD of* $X = VSU^T = \sum_{k=1}^{r} s_i v_k u_k^T$ *where* $r = \mathrm{rank}(X)$.

1. *Show that* $\hat{\boldsymbol{\theta}}_n = \sum_{k=1}^{r} s_k^{-1} u_k v_k^T y = X^+ Y$ *and its variance is* $\mathrm{var}(\hat{\boldsymbol{\theta}}_n) = \sigma^2 \sum_{k=1}^{r} s_i^{-2} u_k u_k^T$.
2. *Show that*

$$(X^T X + n\lambda I_p)^{-1} X^T = X^T (X X^T + n\lambda I_n)^{-1}$$

   (hint : one might prefer to use the complete SVD rather than its reduced form)

3. *If* $n \ll p$, *give an efficient method that would compute the Ridge estimator and would cost less than the formula of Proposition 14. Compare the number of operations required.*

Chapter 5

The LASSO

The LASSO (least absolute shrinkage and selection operator), introduced in **?** is a regression technique that consists in minimizing the usual least-squares loss with an ℓ_1-norm regularization. As another regularization method, it is similar to the Ridge method, presented in the previous chapter, which uses the ℓ_2-norm to regularize. In contrast with the Ridge, some of the LASSO coefficients are usually equal to 0, meaning that the corresponding variables are no longer included in the predictive model. The LASSO thus achieves in the mean time estimation and variable selection.

5.1 Definition

As for the Ridge estimator, the LASSO is usually defined with centered variables so that we can skip estimating the intercept. We consider the following framework: $X \in \mathbb{R}^{n\times p}$ denote the covariates vector is such that $1_n^T X = 0$ and $Y \in \mathbb{R}^n$ is the output and satisfies $1_n^T Y = 0$. In other words, X and Y are supposed to have (empirical) mean 0. The LASSO estimate is defined by

$$\hat{\boldsymbol{\theta}}_{\text{LASSO}} \in \text{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|Y - X\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}, \tag{5.1}$$

where $\|\cdot\|_q$ stands for the ℓ_q-norm. Some remarks are of order:

- Note that the solution of (5.1) can be recovered by working with non centered variables X and Y and adding an intercept.
- As for the Ridge, it is standard to let the LASSO estimate be scale invariant. This is usually done by an additional standardization step that consists in using $XD^{-1/2}$ in place of X, where D is the diagonal matrix with entries $e_k^T X^T X e_k$, $k = 1, \ldots, p$ (e_k being the k-th element of the canonical basis of the space $\mathbb{R}^p$).
- The LASSO is not unique. Some conditions for uniqueness are given and discussed in **?**.
- In contrast with the Ridge approach, the ℓ_1-penalty of the LASSO objective function allows to shrink to 0 the coefficients in $\hat{\boldsymbol{\theta}}_{\text{LASSO}}$ associated to the variables that are useless to predict Y.

5.2 Theoretical properties

From a theoretical perspective, the LASSO takes advantage of *sparse* regression models. A regression model is sparse whenever many of the coefficients of the parameter vector $\boldsymbol{\theta}$ are equal to zero, i.e., many of the covariates are useless to predict Y. We consider the Gaussian regression model

$$Y = X\boldsymbol{\theta}^* + \epsilon, \qquad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{5.2}$$

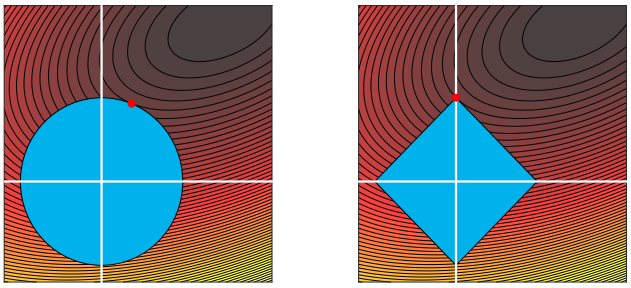

Figure 5.1: Graphical representation of the Ridge and LASSO penalties with the level set of the quadratic loss.

and we define the *active set* $S^* \subset \{1, \dots, p\}$ as

$$S^* = \{j = 1, \dots, p \,:\, \boldsymbol{\theta}^*_j \neq 0\}.$$

The number of elements in S^*, that we denote by s, quantifies the level of sparsity associated to the regression model. We will see that the generalization bounds for the LASSO improve whenever s becomes small. We follow the approach presented in **?**, in which the theoretical analysis of the LASSO is carried out using the *restricted eigenvalue condition*. This condition is basically dealing with the smallest eigenvalue of the matrix $X^T X$. It is called "restricted" because it is concerned only with particular eigenvectors that are "away" from the *not active* directions. More precisely, it only considers vectors living in certain cone that leaves away the direction S^{*c}. The collection of cones of interest are now defined. For $\alpha > 0$ and $S \subset \{1, \dots, p\}$, we set

$$C(\alpha, S) = \{u \in \mathbb{R}^p \,:\, \|u_{S^c}\|_1 \leq \alpha \|u_S\|_1\}.$$

The *restricted eigenvalue condition* (RE) for (γ, α, S) is satisfied whenever

$$n^{-1}\|Xu\|_2^2 \geq \gamma \|u\|_2^2, \qquad \forall u \in C(\alpha, S). \tag{5.3}$$

The following lemma is crucial to understand the role played by the cone $C(3, S^*)$ in the analysis of the LASSO.

Lemma 4. *Whenever* $\lambda \geq 2\|X^T \epsilon\|_\infty$, *then*

$$0 \leq \|X\hat{u}\|_2^2 \leq \lambda(3\|\hat{u}\|_1 - \|\hat{u}_{S^{*c}}\|_1). \tag{5.4}$$

In particular, $(\hat{\boldsymbol{\theta}}_{LASSO} - \boldsymbol{\theta}^*) \in C(3, S^*)$.

Proof. Define

$$G(u) = \|Y - X(\boldsymbol{\theta}^* + u)\|^2/2 + \lambda\|\boldsymbol{\theta}^* + u\|_1 = \|\epsilon - Xu\|^2/2 + \lambda\|\boldsymbol{\theta}^* + u\|_1.$$

Let $\hat{u} = \hat{\boldsymbol{\theta}}_{\text{LASSO}} - \boldsymbol{\theta}^*$. Because $G(\hat{\boldsymbol{\theta}}_{\text{LASSO}}) \leq G(0)$, we have

$$\|X\hat{u}\|_2^2/2 \leq \langle \epsilon, X\hat{u}\rangle + \lambda(\|\boldsymbol{\theta}^*\|_1 - \|\boldsymbol{\theta}^* + \hat{u}\|_1).$$

From the triangle inequality, $\|(\boldsymbol{\theta}^* - (-\hat{u}))_{S^*}\|_1 \geq |\|\boldsymbol{\theta}^*_{S^*}\|_1 - \|\hat{u}_{S^*}\|_1| \geq \|\boldsymbol{\theta}^*_{S^*}\|_1 - \|\hat{u}_{S^*}\|_1$, implying that

$$\begin{aligned}
\|\boldsymbol{\theta}^*\|_1 - \|\boldsymbol{\theta}^* + \hat{u}\|_1 &= \|\boldsymbol{\theta}^*\|_1 - \|(\boldsymbol{\theta}^* + \hat{u})_{S^*}\|_1 - \|(\boldsymbol{\theta}^* + \hat{u})_{S^{*c}}\|_1 \\
&\leq \|\boldsymbol{\theta}^*\|_1 - \|\boldsymbol{\theta}^*_{S^*}\|_1 + \|\hat{u}_{S^*}\|_1 - \|(\boldsymbol{\theta}^* + \hat{u})_{S^{*c}}\|_1 \\
&= \|\hat{u}_{S^*}\|_1 - \|\hat{u}_{S^{*c}}\|_1.
\end{aligned}$$

From Holder inequality, we get $\langle \epsilon, X\hat{u} \rangle \leq \|X^T\epsilon\|_\infty \|\hat{u}\|_1$, which leads to

$$\|X\hat{u}\|_2^2/2 \leq \|X^T\epsilon\|_\infty \|\hat{u}\|_1 + \lambda(\|\hat{u}_{S^*}\|_1 - \|\hat{u}_{S^{*c}}\|_1).$$

Consequently, because $2\|X^T\epsilon\|_\infty \leq \lambda$, we obtain that

$$0 \leq \|X\hat{u}\|_2^2/2 \leq \lambda(\|\hat{u}\|_1/2 + \|\hat{u}_{S^*}\|_1 - \|\hat{u}_{S^{*c}}\|_1),$$

and the conclusion follows.

□

Now we can state the main result dealing with the analysis of the LASSO error.

Theorem 2. *Under the Gaussian model*(5.2)*, assume that* $\forall k \in \{1, \ldots, p\}$, $(X^TX)_{k,k} \leq n$ *and that RE for* $(\gamma, 3, S^*)$ *is satisfied. Then provided that* $\lambda = 2\sqrt{2n\sigma^2 \log(2p)}$*, we have with probability* $1 - \delta$ *that*

$$\|X(\hat{\boldsymbol{\theta}}_{LASSO} - \boldsymbol{\theta}^*)\|_2^2 \leq \frac{64s\sigma^2 \log(2p)}{\gamma}.$$

In addition, we have with probability $1 - \delta$ *that*

$$\|\hat{\boldsymbol{\theta}}_{LASSO} - \boldsymbol{\theta}^*\|_2 \leq \frac{6}{\gamma}\sqrt{\frac{2\sigma^2 s \log(2p)}{n}}.$$

Proof. Let $\hat{u} = \hat{\boldsymbol{\theta}}_{\text{LASSO}} - \boldsymbol{\theta}^*$. Suppose for now that $\lambda \geq 2\|X^T\epsilon\|_\infty$ (this will be shown to hold with probability $1 - \delta$ at the end of the proof). We have, from (5.4) and Jensen inequality, that

$$\|X\hat{u}\|_2^2 \leq 3\lambda\|\hat{u}_{S^*}\|_1 \leq 3\lambda\sqrt{s}\|\hat{u}_{S^*}\|_2 \leq 3\lambda\sqrt{s}\|\hat{u}\|_2. \tag{5.5}$$

By Lemma 4 and the RE condition, it holds that

$$\|\hat{u}\|_2^2 \leq (\gamma n)^{-1}\|X\hat{u}\|_2^2.$$

Injecting this in (5.5), we obtain the first and second statement. It remains to show that with probability $1 - \delta$, $\lambda \geq 2\|X^T\epsilon\|_\infty$. This follows from the use of Lemma 7, a Gaussian concentration result stated in Appendix C. Let $k \in \{1, \ldots, p\}$, note that $(X^T\epsilon)_k$ is distributed as $\mathcal{N}(0, (X^TX)_{k,k}\sigma^2)$. Applying Lemma 7 gives that

$$\mathbb{P}(|(X^T\epsilon)_k| > t) \leq 2\exp(-t^2/(2(X^TX)_{k,k}\sigma^2)).$$

From the union bound, it follows that

$$\mathbb{P}(\|X^T\epsilon\|_\infty > t) \leq (2p)\exp(-t^2/(2n\sigma^2)),$$

or equivalently, that with probability $1 - \delta$,

$$\|X^T\epsilon\|_\infty \leq \sqrt{2n\sigma^2 \log(2p/\delta)} = \lambda/2.$$

□

Some other (asymptotic) properties of the LASSO are derived in **?**. The authors assume the following

$$n^{-1}X^TX \to C, \text{ a positive definite matrix,} \tag{5.6}$$

and

$$(Y_i - \theta_0^\star - X_i^T\boldsymbol{\theta}^\star)_i \text{ is an iid sequence with mean 0 and variance } \sigma^2. \tag{5.7}$$

Theorem 3 (?). *Suppose that (5.6) and (5.7) hold and that* $\lambda/n \to 0$*, then* $\hat{\boldsymbol{\theta}}_{LASSO} \to \boldsymbol{\theta}^\star$*, in probability. If moreover* $\lambda/\sqrt{n} \to \lambda_0 \geq 0$*, then* $\sqrt{n}(\hat{\boldsymbol{\theta}}_{LASSO} - \boldsymbol{\theta}^\star)$ *converges weakly.*

5.3 Computation

In contrast with the OLS or the Ridge, we have no closed formula for the LASSO solutions. This is due to the lack of smoothness of the ℓ_1-norm. In particular, the traditional first order conditions are derived using subgradients (rather than gradients). In Appendix D, some basic definitions and properties are given concerning subgradients and subdifferentials.

The following remark is helpful to characterize the set of LASSO solutions : if $f : \mathbb{R}^p \to \mathbb{R}$ is a convex function, the point $x^\star$ is a minimum if and only if $0 \in \partial f(x^\star)$, where ∂f is the subdifferential of f (see Appendix D for details). This is often refereed to as the Fermat's rule.

Proposition 16. *Denote by X_k the k-th column of X. The LASSO solution satisfies*

$$\forall k = \{1, \ldots, p\}, \quad \langle X_k, Y - X\hat{\boldsymbol{\theta}}_{LASSO} \rangle \in \begin{cases} \{sign(\hat{\boldsymbol{\theta}}_{LASSO,k})\} & \text{if } \hat{\boldsymbol{\theta}}_{LASSO,k} \neq 0 \\ [-1, 1] & \text{if } \hat{\boldsymbol{\theta}}_{LASSO,k} = 0 \end{cases}$$

Actually, the previous set of equations has no explicit solutions. The LASSO problem becomes much simpler when we fix all coordinates except one, and try to minimize with respect to this coordinate. For this reason, the LASSO is usually computed using a coordinate descent, i.e., by iteratively solving the first order conditions (involving subgradients) for each coordinate. For any $\lambda \geq 0$, define the function $\eta_\lambda : \mathbb{R} \to \mathbb{R}$ given by

$$\eta_\lambda(z) = \begin{cases} z + \lambda & \text{if } z < -\lambda \\ 0 & \text{if } z \in [-\lambda, \lambda] \\ z - \lambda & \text{if } z > \lambda \end{cases} \tag{5.8}$$

The function η_λ intervenes in solving least-squares with an absolute penalty (see Exercise 14). Note that $\eta_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+$. As shown in the following development, the function η is useful to update the LASSO in the coordinate descent algorithm. To avoid trivial cases, we suppose in the following that $\|X_k\|^2 \neq 0$. In practice, one just need to remove the constant variables. Let

$$z_k = Y_k - \sum_{j \neq k} X_j \boldsymbol{\theta}_j.$$

Minimizing (5.1) with respect to the k-th coordinates is the same as minimizing

$$\begin{aligned} &\left\{ \frac{1}{2}(\|z_k - X_k\boldsymbol{\theta}_k\|_2^2 - \|z_k\|_2^2) + \lambda|\boldsymbol{\theta}_k| \right\} \\ &= \left\{ \frac{1}{2}(-2\langle z_k, X_k\rangle\boldsymbol{\theta}_k + \boldsymbol{\theta}_k^2\|X_k\|_2^2) + \lambda|\boldsymbol{\theta}_k| \right\}, \end{aligned}$$

which is the same as minimizing

$$\left\{ \frac{1}{2}\left(\left\langle z_k, \frac{X_k}{\|X_k\|_2^2} \right\rangle - \boldsymbol{\theta}_k \right)^2 + \frac{\lambda}{\|X_k\|_2^2}|\boldsymbol{\theta}_k| \right\}.$$

Consequently, the update is given by

$$\hat{\boldsymbol{\theta}}_k = \eta_{\lambda/\|X_k\|_2^2}\left(\left\langle z_k, \frac{X_k}{\|X_k\|_2^2} \right\rangle \right).$$

As it is standard, one can start with a Ridge solution and then update each coordinates with the previous formula.

5.4 Extensions

Among the extension of the LASSO, we have the LSLASSO (Least-Square LASSO) which consists in (i) running the LASSO to find the support and (ii) applying OLS on the non-zero coefficients. Another extension of LASSO is Elastic Net, introduced in **?**, which computes the regression coefficient

$$\hat{\boldsymbol{\theta}}_{\text{E-NET}} \in \operatorname{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|Y - X\boldsymbol{\theta}\|_2^2 + \lambda\{\alpha\|\boldsymbol{\theta}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{\theta}\|_2^2\} \right\}. \tag{5.9}$$

The previous estimate is unique. Finally, a notable extension is adaptive LASSO, introduced in **?**. The adaptive LASSO is an iterative strategy that attributes weights to each coefficient $|\boldsymbol{\theta}_k|$ in the penalty term. The weights might take the form $1/\hat{\boldsymbol{\theta}}_{OLS,k}^{1/2}$ penalizing mostly the small OLS coefficients. The resulting estimate recover the support with probability going to 1.

Exercises

Exercise 14. *Show that* $\eta_\lambda(z) = \operatorname{argmin}_{x\in\mathbb{R}}\{(z-x)^2/2 + \lambda|x|\}$ *where* η_λ *is defined in (5.8).*

Exercise 15. *Following the approach given for the LASSO, find the update for the Elastic Net defined in (5.9).*

Appendix A

Elementary results from linear algebra

The vector space $\mathbb{R}^d$ is endowed with the usual inner product

$$\forall (u,v) \in \mathbb{R}^d \times \mathbb{R}^d, \quad \langle u, v \rangle = u^T v = \sum_{k=1}^{d} u_k v_k,$$

where u^T stands for the transpose of u. If $\langle u, v \rangle = 0$ we say that u and v are orthogonal and we write $u \perp v$. If E is a set of vector in $\mathbb{R}^d$, we define its orthogonal complement as

$$E^\perp = \{u \in \mathbb{R}^d \, : \, x^T u = 0, \quad \forall x \in E\}.$$

Proposition 17. *If E is a linear subspace of $\mathbb{R}^d$, then $(E^\perp)^\perp = E$.*

For any matrix $A \in \mathbb{R}^{p \times d}$, define

$$\begin{aligned} \text{span}(A) &= \{Ax \, : \, x \in \mathbb{R}^d\}, \\ \ker(A) &= \{x \in \mathbb{R}^d \, : \, Ax = 0\}. \end{aligned}$$

The set $\text{span}(A)$ is called the image of the matrix A. It is the linear space generated by the columns of A. The set $\ker(A)$ is called the kernel of A. Both sets are linked by the following property.

Proposition 18. *Let $A \in \mathbb{R}^{p \times d}$. Then $\ker(A) = \text{span}(A^T)^\perp$.*

Proposition 19. *Let $A \in \mathbb{R}^{p \times d}$. Then $\ker(A) = \{0\}$ if and only if $\text{span}(A^T) = \mathbb{R}^d$. Consequently, if $p < d$ then $\ker(A) \neq \{0\}$.*

Let $A \in \mathbb{R}^{p \times d}$, $b \in \mathbb{R}^d$. Let S be the set of solutions of the linear system $Ax = b$.

Proposition 20. *We have only three possible configurations:*

1. *S contains only one element,*
2. *$S = \emptyset$,*
3. *the number of elements in S is infinite.*

Note that S is empty if and only if $b \notin \text{span}(A)$.

Proposition 21. *Suppose that $b \in \text{span}(A)$ and let $x_0 \in S$, then*

$$S = x_0 + \ker(A).$$

We now recall a classical result called the spectral decomposition of symmetric matrices or the eigen decomposition of symmetric matrices.

Proposition 22. *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 orthonormal matrix $U \in \mathbb{R}^{d\times d}$ (i.e., $U^TU = I_d$) of eigenvectors, such that $A = UDU^T$, where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$.*

Definition 7. *A linear transformation $P \in \mathbb{R}^{d\times d}$ is called orthogonal projector if $P^2 = P$ and $P^T = P$.*

The next proposition says that an orthogonal projector is characterized by its span and, therefore, by its kernel from Proposition 17 and 18.

Proposition 23. *The eigenvalues of an orthogonal projector are either 1 or 0. Hence any orthogonal projector can be written as UU^T where $U \in \mathbb{R}^{p\times r}$ forms a basis of $\mathrm{span}(P)$.*

Proposition 24. *The trace of an orthogonal projector is equal to the dimension of its span.*

Appendix B

Singular value decomposition and principal component analysis

Before we present the method of principal component analysis (PCA), it is appropriate to recall some matrix decomposition results and more particularly the singular value decomposition (SVD).

B.1 Matrix decomposition

The usual eigen decomposition of symmetric matrices can be extended to arbitrary matrices (even not squared matrix). The price to pay is that the left and right eigenvectors are different. This is called the SVD.

Proposition 25. *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 22 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 $S^2 = \text{diag}(d_1, \ldots, d_p)$. □

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

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

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.

An attractive property of the SVD is that it defines subspaces on which one can project the data X without loosing too much.

Proposition 26. *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 corresponds 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 achieve 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 $P = \sum_{i\leq k} U_i U_i^T$. □

B.2 Principal component analysis

Definition 8. *Let $X \in \mathbb{R}^{n\times p}$ and define $X_c = X - 1_n \overline{X}^{nT}$. 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 27. *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 26 and the Pythagorean identity, we get that $\|X_c P\|^2 \geq \|X_c \tilde{P}\|_F^2$. □

Remark 10. *As the PCA of X depends on the scale of each covariate, one may prefer in practice to rescale the matrix X before running the PCA algorithm. This can be done by taking $XD^{-1/2}$ rather than X, with D equal to the diagonal matrix whose elements are $e_k^T \hat{\Sigma}_n e_k$, $k = 1,\ldots,n$. Then each covariate of XD has the same empirical variance.*

Appendix C

Concentration inequalities

To derive concentration inequalities for the errors of the estimators, we use the notion of sub-Gaussianity as defined for instance in (**?**, Section 2.3). Recall that the moment generating function of a Gaussian random variable W with mean μ and variance σ^2 is equal to $\lambda \mapsto E[\exp(\lambda W)] = \exp(\mu\lambda + \lambda^2\sigma^2/2)$.

Definition 9. *A centered random variable Y is* sub-Gaussian *with variance factor $\tau^2 > 0$, notation $Y \in \mathcal{G}(\tau^2)$, if $\log E[\exp(\lambda Y)] \leq \lambda^2\tau^2/2$ for all $\lambda \in \mathbb{R}$.*

If $Y \in \mathcal{G}(\tau^2)$, then necessarily $\mathrm{var}(Y) \leq \tau^2$ (**?**, Exercise 2.16). Centered, bounded random variables taking values in an interval $[a, b]$ are sub-Gaussian with variance factor at most $(b-a)^2/4$ (**?**, Lemma 2.2). Chernoff's inequality provides exponential bounds on the tails of sub-Gaussian random variables.

Lemma 5 (Chernoff)**.** *If $Y \in \mathcal{G}(\tau^2)$, then $P(Y > t) \leq \exp(-t^2/(2\tau^2))$.*

Proof. For any $\lambda \in \mathbb{R}$, we have $1_{Y>t} \leq \exp(\lambda(Y-t))$. Hence it holds that $P(Y > t) \leq E[\exp(\lambda(Y-t))] \leq \exp(\lambda^2\tau^2/2 - \lambda t)$. Minimizing the previous bound in λ gives $\lambda = t/\tau^2$, from which we deduce that the stated bound. □

Finally, the sum of independent sub-Gaussian variables is again sub-Gaussian. This is the statement of the following Lemma, which proof is left as an exercise.

Lemma 6. *If $(Y_i)_{i\geq 1}$ is a sequence of independent random variables such that, for all $i \geq 1$, $Y_i \in \mathcal{G}(\tau_i^2)$, then $\sum_{i=1}^n Y_i \in \mathcal{G}(\sum_{i=1}^n \tau_i)$.*

To conclude we state a concentration inequality for the sum of independent sub-Gaussian random variables. This is just a consequence of the two previous Lemma.

Lemma 7. *If $(Y_i)_{i\geq 1}$ is a sequence of independent random variables such that, for all $i \geq 1$, $Y_i \in \mathcal{G}(\tau_i^2)$, then, for all $n \geq 1$, and $t \geq 0$,*

$$P\left(\left|\sum_{i=1}^n Y_i\right| > t\right) \leq 2\exp\left(-t^2/(2\sum_{i=1}^n \tau_i^2)\right)$$

.

Appendix D

Optimization of convex functions

We recall here some definitions and basic properties dealing with the minimization of convex functions.

Definition 10 (convex function)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be convex if*

$$\forall (x, y) \in \mathbb{R}^p \times \mathbb{R}^p, \forall \alpha \in [0, 1], \quad f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

A function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be strictly convex if

$$\forall x \neq y, \forall \alpha \in (0, 1), \quad f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

Definition 11. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex function. A subgradient of f at x is any vector $u \in \mathbb{R}^p$ satisfying*

$$\forall y \in \mathbb{R}^p, \quad f(y) - f(x) \geq u^T(y - x).$$

The subdifferential of f at x, noted $\partial f(x)$, is the set of all subgradients of f at x.

By simply using the definition of the subdifferential, we obtain the following caracterization of minimum points (often refereed to as the Fermat's rule). This is useful in deriving the LASSO first-order conditions.

Proposition 28. *if $f : \mathbb{R}^p \to \mathbb{R}$ is a convex function, the point $x^\star$ is a minimum if and only if $0 \in \partial f(x^\star)$.*

For differentiable (convex) function, the notion of subgradient coincides with the notion of gradient. This is stated in the following.

Definition 12 (differentiable function)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be differentiable at x_0 if there exists a vector $u \in \mathbb{R}^p$ such that*

$$\lim_{h \to 0} \frac{|f(x_0 + h) - f(x_0) - u^T h|}{\|h\|} \text{ exists}$$

As a consequence of the previous definition (when taking $h = te_k, t \to 0$) the partial derivatives (gradient) exists for differentiable functions. The gradient of a differentiable function f at $x \in \mathbb{R}^p$ is denoted by $\nabla f(x)$.

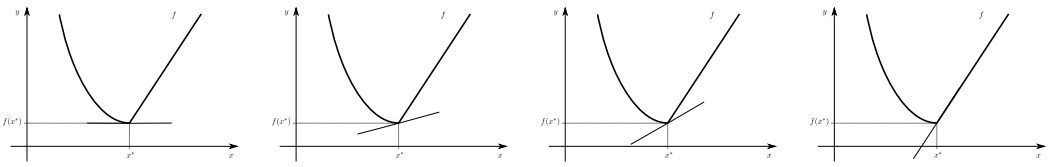

Figure D.1: Draws of subgradients.

Proposition 29 (differentiable function and convexity)**.** *A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is convex if and only if*

$$\forall (x, y) \in \mathbb{R}^p \times \mathbb{R}^p, \quad f(y) - f(x) \geq \nabla f(x)^T (y - x).$$

Moreover, for any $x \in \mathbb{R}^p$, the gradient $\nabla f(x)$ is the only vector satisfying the previous equation for any y. Consequently, for differentiable and convex functions $\partial f(x) = \{\nabla f(x)\}$.