Lecture notes on ordinary least squares[1]

March 4, 2022

[1]

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 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. $X_n \overset{p}{\to} X$).

Chapter 1

Definition of ordinary least-squares and first properties

1.1 Definition

The general goal of regression is to predict with an output variable $y \in \mathbb{R}$, also called the explanatory variable, based on the observation of some input variables $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$, with $p \geq 1$, also called the covariates. The statistical approach consists in *learning* or *estimating* a *regression function* (also called link function) that maps any element x of the input space $\mathbb{R}^p$ to the output space $\mathbb{R}$. Generally, the estimation of the regression function is based on the observation of a sample made of examples. These examples are also called the observations and each example is a pair made of an input variable together with the corresponding output. Let $n \geq 1$ denote the number of observations. Let $(x_i, y_i)_{i=1,\dots,n}$ be the observations such that for each i, $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Many regression model can be estimated on the basis of these observations. In this course, we focus on *linear regression* in which the regression function is defined as a simple linear function, 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 regression function. 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 $z_i = (x_{i,0}, \dots, x_{i,p})^T = (x_{i,0}, x_i^T) \in \mathbb{R}^{p+1}$ 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

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

It is useful to introduce the notations

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}, \qquad Z = \begin{pmatrix} z_1^T \\ \vdots \\ z_n^T \end{pmatrix} = \begin{pmatrix} x_{1,0} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,0} & \dots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}, \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

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

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

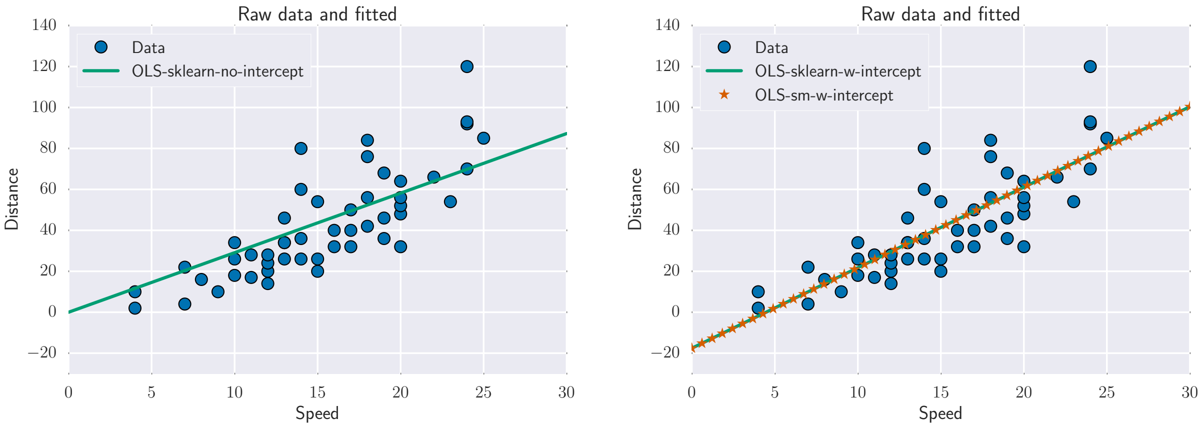

Figure 1.1: The dataset is the cars dataset from the R software. We use sklearn to compute OLS. The graph on the left represents the OLS line without intercept and the graph on the right is the OLS line computed with intercept.

where $\|\cdot\|$ stands for the Euclidean norm. As soon as $\hat{\boldsymbol{\theta}}_n$ is obtained, one can use it to define the estimated regression function

$$\hat{g}(x) = \hat{\boldsymbol{\theta}}_{n,0} + \sum_{k=1}^{p} \hat{\boldsymbol{\theta}}_{n,k} x_k.$$

This function is often called the *predictor* due to its use to predict the output value based on the observation of the input x.

1.2 Existence and uniqueness

With the above formulation, the OLS has a nice geometric interpretation : $\hat{Y} = Z\hat{\boldsymbol{\theta}}_n$ is the closest point to Y in the linear subspace $\operatorname{span}(Z) \subset \mathbb{R}^n$ (where $\operatorname{span}(A)$ stands for the linear subspace generated by the columns of A). Using the Hilbert projection theorem ($\mathbb{R}^n$ is a Hilbert space, $\operatorname{span}(Z)$ 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 $\operatorname{span}(Z)$. This property is equivalent to the so-called normal equation:

$$Z^T(Y - \hat{Y}) = 0.$$

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

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

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 $u \in \ker(Z)$ 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 Z which is related to the invertibility of the so called Gram matrix introduce below (see Exercise 1).

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

[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 $\ker(P) = E^\perp$.

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* $\hat{Y} = \hat{H}_{n,Z} Y$. *It is either*

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

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

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

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 $\ker(Z) = \ker(Z^T Z)$ which follows easily from the identity $\|Zu\|^2 = u^T Z^T Z u$. □

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

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

where $(Z^T Z)^+$ denotes the Moore–Penrose inverse of $Z^T Z$, 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* $\{(Z^T Z)^+ Z^T Y + u \, : \, u \in \ker(Z)\}$.

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

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 described before) and doing OLS, without the intercept, on centered variables. The later estimation procedure consists in the following. Define

$$\overline{Y} = n^{-1} \sum_{i=1}^n Y_i, \quad \text{and} \quad \overline{X} = n^{-1} \sum_{i=1}^n x_i,$$

$Y_c = Y - 1_n \overline{Y}$ and $X_c = X - 1_n \overline{X}^T$. Hence the quantities Y_c and X_c are just centered version of Y and X, respectively. Define

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

The associated predictor is given by

$$\hat{g}_c(x) = \overline{Y} + \hat{\boldsymbol{\theta}}_{n,c}(x - \overline{X}).$$

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 Z *has full rank, we have the following relationship between the traditional OLS and the OLS based on centred data,*

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

Moreover, the 2 *methods give the same predictor, i.e., for all* $x \in \mathbb{R}^p$, $\hat{g}(x) = \hat{g}_c(x)$.

Proof. See exercise 9. □

We conclude with the following proposition which expresses the uniqueness condition given in Proposition 1 in terms of X.

Proposition 3. *The following conditions are equivalent:*

(i) $\ker(Z) = \{0\}$

(ii) $\ker(X_c) = \{0\}$

(iii) The empirical covariance matrix of X, defined as $X_c^T X_c/n$, is invertible

Proof. See Exercise 5. □

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

$$\|Y - \overline{y}^n 1_n\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \overline{y}^n 1_n\|^2$$

It follows that

$$R^2 = 1 - \frac{\|\hat{Y} - Y\|^2}{\|Y - \overline{y}^n 1_n\|^2}. \tag{1.3}$$

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 $\hat{\boldsymbol{\theta}}_n = (\overline{y}^n, 0, \ldots, 0)$ is one OLS estimator.

Exercises

Exercise 1. *Show that* $\ker(Z^T Z) = \ker(Z)$ *and that* $\operatorname{span}(Z^T) = \operatorname{span}(Z^T Z)$ *(for the latter, one might first note that* $\ker(Z) = \operatorname{span}(Z^T)^\perp$*). 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 $Z = 1_n$ and $Y \in \mathbb{R}^n$.*

Exercise 3. *Show that any invertible transformation on the covariate, i.e. Z is replaced by ZA with A invertible, does not change the predicted values $\hat{Y}$ nor the predictor.*

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

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

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

(b) Prove that $Z^T Z = \sum_{i=1}^n z_i z_i^T$.

(c) Verify that $ker(Z) = 0$ *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* $u \in \mathbb{R}^p \backslash \{0\}$ *such that* $X_c u = 0$*).*

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

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

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

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

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

Exercise 7. *Show that* $\hat{\boldsymbol{\theta}}_n = (Z^T Z)^+ Z^T Y$ *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 collection $(w_i)_{i=1,\dots,n}$ *of real numbers, and for all* $w \in \mathbb{R}$*, it holds that* $\|W - w1_n\| \geq \|W - \overline{W}1_n\|$*, where* $W = (z_1, \dots, z_n)$ *and* $\overline{W} = n^{-1}\sum_{i=1}^n z_i$.

(c) Find $\hat{a}_n$ *such that, for any* $\theta_0 \in \mathbb{R}$ *and* $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^p$, $\|Y - \theta_0 1_n - X\tilde{\boldsymbol{\theta}}\| \geq \|Y - \hat{a}_n(\tilde{\boldsymbol{\theta}})1_n - X\tilde{\boldsymbol{\theta}}\|$.

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

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

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:

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

where (z_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

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

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 (z_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 $\operatorname{span}(Z)$ equals $p + 1$, implying that $\ker(Z) = \{0\}$ 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 $\ker(Z) = \{0\}$, it holds that

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

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

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

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

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

The predictive risk is

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

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

Proposition 4. *When $\ker(Z) = \{0\}$, 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 covariance matrix is given by $\mathrm{var}(\hat{\boldsymbol{\theta}}_n) = (Z^T Z)^{-1}\sigma^2$.

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

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

Hence whenever the smallest eigenvalue of $\hat{G}_n$ is larger than b (independently of n), the quadratic error of the OLS decreases with the rate $1/n$, which is the classical estimation rate in statistics, e.g., Z 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 $u^T V_1 u \leq u^T V_2 u$ for every $u \in \mathbb{R}^d$. This partial order is particularly useful to compare the covariance matrices of estimators. Indeed if $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimators with respective covariance V_1 and V_2. Then, $V_1 \leq V_2$ if and only if 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, Z), it writes as AY, where $A \in \mathbb{R}^{(p+1)\times n}$ depends only on Z.*

Proposition 5 (Gauss-Markov)**.** *Under the fixed design model, among all the unbiased linear estimators AY, $\hat{\boldsymbol{\theta}}_n$ is the one with minimal variance, i.e.,*

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

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

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

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

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

2.1.3 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 Z variance of $\epsilon_1, \ldots, \epsilon_n$, but this is not possible. Alternatively, one can take

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

Because of the first normal equations expressed in (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 $\tilde{\sigma}_n^2 = n^{-1}\|(I_n - \hat{H}_{n,Z})\epsilon\|^2$ 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 Z has rank $p+1$, we obtain that $Y_i = \hat{Y}_i$ for all $i = 1, \ldots, n$.

2.2 The Gaussian model

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

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

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

2.2.1 The Cochran lemma

The Student's t-distribution with p degrees of freedom is defined as the distribution of the random variable $W/\sqrt{V/p}$, where W (resp. V) has standard normal distribution (resp. chi-square distribution with p degrees of freedom).

Proposition 6. *Under the Gaussian model, if* $\ker(Z) = \{0\}$ *and* $n > p+1$*, it holds that*

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

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

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

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

$$\mathrm{cov}(Z^T\epsilon,(I-\hat{H}_{n,Z})\epsilon)=\mathbb{E}[Z^T\epsilon\epsilon^T(I-\hat{H}_{n,Z})]=0.$$

Then writing

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

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

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

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

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

Consequently,

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

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

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

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

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

□

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

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

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

2.3 The random design model

In the random design model, we suppose that $(Y_i,Z_i)_{i\geq1}$ is a sequence of independent and identically distributed random vectors defined on the probability space $(\Omega,\mathcal{A},\mathbb{P})$ and each element (Y_i,Z_i) is valued in $\mathbb{R}\times\mathbb{R}^{p+1}$. The aim is to estimate the best linear approximation of Y_1 made up with Z_1 in terms of L_2-risk, i.e., to find $\boldsymbol{\theta}$ that minimizes $\mathbb{E}[(Y_1-Z_1^T\boldsymbol{\theta}^*)^2]$. Such a minimizer can be characterized with the help of the normal equation. Recall that $Z_1\in\mathbb{R}^{p+1}$ and $Z_{1,0}=1$ almost surely.

Proposition 7. *Suppose that for all $k=0,\dots,p$, $\mathbb{E}[X_{1,k}^2]<\infty$ and $\mathbb{E}[Y_1^2]<\infty$, then*

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

if and only if

$$\mathbb{E}[X_1X_1^T]\boldsymbol{\theta}^*=\mathbb{E}[X_1Y_1].$$

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 $Z_{1,0}, \ldots, Z_{1,p}$. 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 $\mathbb{E}[Z_1 Z_1^T]$ 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., $\boldsymbol{\theta}^* = \mathbb{E}[Z_1 Z_1^T]^+ \mathbb{E}[Z_1 Y_1]$. Some asymptotic properties are available. They will be useful to run some statistical tests. We consider the following definition, valid for any $n \geq 1$,

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

Proposition 8. *Suppose that $\mathbb{E}[Z_1 Z_1^T]$ and $\mathbb{E}[Y_1^2]$ exist and that $\mathbb{E}[Z_1 Z_1^T]$ is invertible. Then*

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

where $\sigma^2 = \operatorname{var}(Y_1 - Z_1^T \boldsymbol{\theta}^*)$ and $G = \mathbb{E}[Z_1 Z_1^T]$. 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

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

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 **??**.

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

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

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

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

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

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

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

Remark 1. *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*

$$R(f) = \mathbb{E}[(Y_1 - f(Z_1))^2].$$

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 Z_1 : $f^*(Z_1) = \mathbb{E}[Y_1 | Z_1]$, almost surely.

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 6. Recall that, when it exists,

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

Proposition 9. *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 10.

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 focus on the second approach.

Proposition 10. *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:

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

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:

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

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 $1-\alpha \in (0,1)$*

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

Remark 2. *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 11. *Under the random design model, if $\mathbb{E}[X_1X_1^T]$ and $\mathbb{E}[Y_1^2]$ exist and $\mathbb{E}[X_1X_1^T]$ is invertible, the statistic and reject region, respectively given by*

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

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 8 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 3. *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 4. *In the Gaussian case, the test statistic and the reject region are given by*

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

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.

Remark 5 (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

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

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

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

	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 Efron et al. (2004).

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:

$$\hat{T}_n(Y, \tilde{X}_k) = n^{1/2} \frac{\hat{\boldsymbol{\theta}}_n(Y, \tilde{X}_k)}{\hat{s}_n(\tilde{X}_k)\hat{\sigma}_n(Y, \tilde{X}_k)}.$$

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*,

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

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 $r = Y$, $\mathcal{S} = \emptyset \subset \mathcal{A} = \{0, \dots, p\}$.*

```
(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 6. *Different stopping rules might be considered. For instance, in Zhang (2009), the authors recommend to consider the residuals sum of squares and to stop as soon as $\|r\|^2 < \epsilon$.*

Exercises

Exercise 10 (explicit formulas when $p = 1$ for prediction intervals)**.** *Let us consider the following fixed-design one-dimensional ($p = 1$) linear regression model:*

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

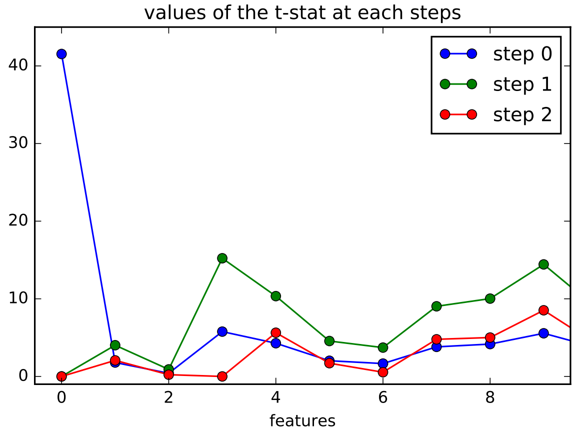

Figure 3.1: The statistics of each selected variable is 0 in the next step. The intercept is the first selected variable, then X_3, etc...

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*

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

2. *Show that*

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

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

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

   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.*

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 12. *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

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

The set $\operatorname{span}(A)$ 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 13. *For any $A \in \mathbb{R}^{p \times d}$, $\ker(A) = \operatorname{span}(A^T)^\perp$.*

Proposition 14. *For any $A \in \mathbb{R}^{p \times d}$, $\ker(A) = \{0\}$ if and only if $\operatorname{span}(A^T) = \mathbb{R}^d$. 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 15. *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 $b \notin \operatorname{span}(A)$.

Proposition 16. *Suppose that $b \in \operatorname{span}(A)$ and let $x_0 \in S$, then*

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

We say that a matrix $A \in \mathbb{R}^{d\times d}$ is invertible if there exists $B \in \mathbb{R}^{d\times d}$ such that $AB = BA = I$.

Proposition 17. *$A \in \mathbb{R}^{d\times d}$. The following are equivalent*

1. *A is invertible*
2. $\ker(A) = \{0\}$

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

Proposition 18. *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., $U^T U = I_d$) of eigenvectors, such that $A = UDU^T$, where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$.*

A symmetric matrix is called positive definite (resp. positive semi-definite) if $u^T Au > 0$ (resp. $u^T Au \geq 0$) for all $u \in \mathbb{R}^d$.

Proposition 19. *Let $A \in \mathbb{R}^{d\times d}$. The following are equivalent*

1. *A is positive definite (resp. positive semi-definite)*
2. *All the eigenvalues of A are positive (resp. nonnegative)*

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 12 and 13.

Proposition 20. *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 $\operatorname{span}(P)$.*

Proposition 21. *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 22. *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 18 to X^TX, there exists $U \in \mathbb{R}^{p\times p}$ such that $U^T(X^TX)U$ is diagonal with r positive coefficients. Hence $U_1^T(X^TX)U_1 = D \in \mathbb{R}^{r\times r}$ and $XU_2 = 0$. Take $V_1^T = D^{-1/2}U_1^TX^T$ (an orthogonal set of r vectors : $V_1^TV_1 = I_r$) to find that $V_1^TXU_1 = D^{1/2}$. Consequently, $V_1^TX(U_1, U_2) = (D^{1/2}, 0)$. Remarking that v orthogonal to V_1 means that $v^TXU_1 = 0$ implying that $v^TX(U_1, U_2) = 0$ leading to $v^TX = 0$. Now taking V_2 such that $V = (V_1, V_2) \in \mathbb{R}^{n\times p}$ is orthogonal, we obtain the claimed decomposition with $S^2 = \mathrm{diag}(d_1, \ldots, d_p)$. □

We have the following reduced SVD formula, if $r \geq 1$ stands for the dimension of 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 23. *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 24. *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 23 and the Pythagorean identity, we get that $\|X_c P\|^2 \geq \|X_c \tilde{P}\|_F^2$. □

Remark 7. *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.*

Bibliography

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

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