Lecture notes on ordinary least squares

François Portier

November 30, 2017

Contents

1 Notation

- $< \cdot, \cdot >$ is the usual inner product in $\mathbb{R}^n$. $\|\cdot\|$ is the Euclidean norm.
- 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.
- 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. $X_n \xrightarrow{p} X$).

2 Definition of the OLS

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

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

It is useful to introduce the notations

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

Then (1) becomes

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

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

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

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

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

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

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

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

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

(i) uniquely defined. This happens if and only if the Gram matrix is invertible, which is equivalent to $\ker(X) = \ker(X^T X) = \{0\}$*.*

(ii) or not unique, with an infinite number of solution for (2). This happens if and only if $\ker(X) \neq \{0\}$*.*

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

When the Gram matrix is invertible, the OLS has the following expression:

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

When not, the OLS is any of the solution of (2). The solution traditionally considered is

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

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

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

3 Statistical model

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

3.1 Fixed-design model

The fixed design model takes the form:

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

where (x_i) is a deterministic sequence of points in $\mathbb{R}^{p+1}$ and (ϵ_i) is a random sequence of identically distributed and independent random variables in $\mathbb{R}$. The probability distribution of ϵ_1, is such that $\mathbb{E}[\epsilon_1] = 0$ and $\sigma^2 = \text{var}(\epsilon_1) > 0$, the level of noise σ reflecting the difficulty of the problem.

The fixed-design model is appropriate when the (x_i) is chosen by the analyst, e.g., in a physics laboratory experiment one can fix some variables such as the temperature, or in a clinical survey one can give to patients a determine quantity of some serum. In contrast, the random design (see Section 3.3) model is appropriate when the covariates are unpredictable as for instance the wind speed observed in the nature or the age of some individuals in a survey.

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

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star = (X^T X)^{-1} X^T \epsilon, \qquad \text{with } \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Bias and variance. The OLS estimator is unbiased i.e., it holds that

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n] = \boldsymbol{\theta}^\star.$$

Its variance is given by

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

Hence whenever $(X^T X)/n \to I_{p+1}$, the variance of the OLS decreases with the rate $1/n$, just as the rate of estimation of an expectation based on iid data.

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

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

Hence whenever the smallest eigenvalue of $(X^T X)/n$ is larger than b, positive and independent of n, the quadratic risk of the OLS decreases with the rate $1/n$.

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

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

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

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

Noise estimation. Providing only an estimate $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}^\star$ is often not enough as it does not give any clue on the accuracy of the estimation. When possible one should also furnish an estimation of the error σ^2. If we knew the error ϵ_i, one would take the empirical estimator of the 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 equation in (2), $\sum_{i=1}^n (Y_i - \hat{Y}_i) = 0$, hence the previous estimate is an empirical variance. Noting that $\tilde{\sigma}^2 = n^{-1} \|(I_n - H_X)\epsilon\|^2$ one can compute the expectation:

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

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

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

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

3.2 Gaussian model

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

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

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

Proposition 2. *Under the Gaussian model, if* $\ker(X) = \{0\}$ *and* $n > p+1$*, it holds that*

- $\hat{\boldsymbol{\theta}}_n$ *and* $\hat{\sigma}_n^2$ *are independent,*
- $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star) \sim \mathcal{N}(0, n\sigma^2(X^T X)^{-1})$ *,*
- $(n-p-1)(\hat{\sigma}_n^2/\sigma^2) \sim \chi^2_{n-p-1}$*,*
- *if* $\hat{s}_{n,k}^2$ *is the k-th term in the diagonal of* $n(X^T X)^{-1}$*, 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 $X^T\epsilon$ and $(I - H_X)\epsilon$ are two independent Gaussian vector:

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

Then writing

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

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

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

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

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

Consequently,

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

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

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

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

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.

□

3.3 Random design model

In the random design model, we observe a sequence (Y_i, X_i) of independent and identically distributed random vectors defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The regression problem might be then formulated in terms of expectation: the regression function f^* is defined as a minimizer of the risk

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

When $\mathbb{E}[Y^2] < \infty$, the minimizer is unique and coincides with the conditional expectation of Y given X : $f^*(X) = \mathbb{E}[Y|X]$.

The aim is to estimate the best linear approximation of Y_1 made up with X_1 in terms of L_2-risk, i.e., to find $\boldsymbol{\theta}$ that minimizes $\mathbb{E}[(Y_1 - X_1^T\boldsymbol{\theta}^*)^2]$. Such a minimizer can be characterized with the help of the normal equation.

Proposition 3. *Suppose that* $\mathbb{E}[X_{1,k}^2] < \infty$ *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 $X_{1,0}, \ldots, X_{1,p}$. As $\mathcal{F}$ is a closed linear subspace (because it has a finite dimension), the minimizer is unique and characterised 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 $\mathbb{E}[X_1X_1^T]$ is invertible, or $\boldsymbol{\theta}^*$ is not uniquely defined, in which case one might take $\boldsymbol{\theta}^* = \mathbb{E}[X_1X_1^T]^+\mathbb{E}[X_1Y_1]$. The case where $\boldsymbol{\theta}^*$ is not unique happens as soon as we add the constant variable or as soon as one variable is a combination of the others. Some asymptotic properties are available. They will be useful to run some statistical tests. We consider the following definition, valid for any $n \geq 1$,

$$\hat{\boldsymbol{\theta}}_n = (X^TX)^+X^TY.$$

Theorem 1. *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*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \rightsquigarrow \mathcal{N}(0, \sigma^2 \mathbb{E}[X_1 X_1^T]^{-1}),$$

where $\sigma^2 = \mathrm{var}(Y_1 - X_1^T \boldsymbol{\theta}^*)$. *Moreover*

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

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.

□