Linear Least Squares Regression

Pierre Gaillard

12 février 2019

In this lesson, we study the simple but still widely used problem of Linear Least Square Regression. The linear regression problem can be traced back to Legendre (1805) and Gauss (1809). The word "regression" is said to have been introduced by Galton in the 19th century. By modeling the size of individuals according to that of their fathers, Galton observed a return (regression) towards average height. Larger-than-average fathers tend to have smaller children and vice versa for smaller fathers.

1 Introduction

Let's start with an example of a practical problem. In order to better optimize its production, a producer is interested in modeling electricity consumption in France as a function of temperature (cf. Figure 1).

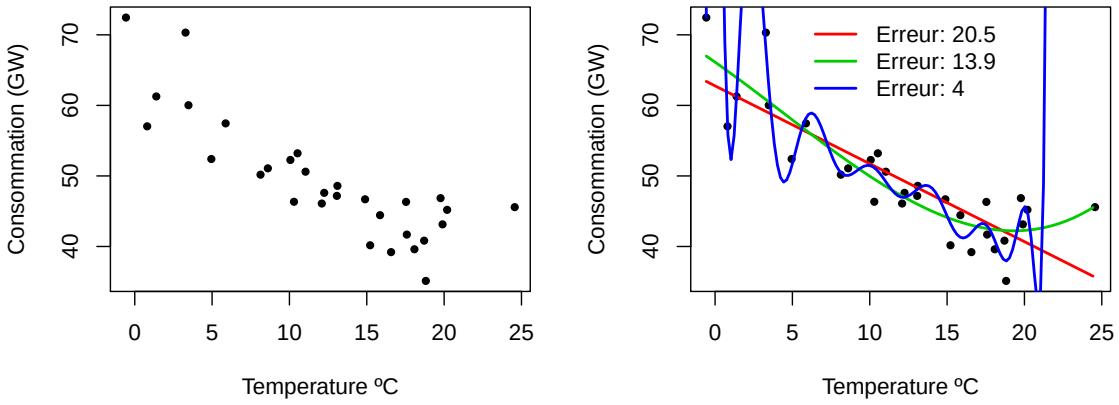

Figure 1: French power consumption (GW) as a function of temperature (°C). To the right are plotted error minimizing functions for polynomial spaces of degrees 1 (red), 3 (green) and 30 (blue).

The objective is to find a function f such that it explains well the power consumption $(y_i)_{1\leqslant i\leqslant n}$ as a function of temperature $(x_i)_{1\leqslant i\leqslant n}$, that is $y_i \approx f(x_i)$. To do this, we can choose a function space $\mathcal{F}$ and solve the empirical risk minimization problem:

$$\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \widehat{\mathcal{R}}_n(f) := \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 . \tag{1}$$

Care must be taken when selecting the function space to avoid overfitting (see Figures 1). Although the empirical mean square error decreases when the $\mathcal{F}$ space becomes larger (larger

polynomial degrees), the $\widehat{f}_n$ estimator loses its predictive power. The question is: will $\widehat{f}_n$ perform well on new data? The linear function space of the form $f : x \mapsto ax + b$ is the simplest. This is the one we will study in this course.

Supervised learning: general setup and notation

Goal. In supervised machine learning, the goal is given some observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ of inputs/ouptus and given a new input $x \in \mathcal{X}$ to predict well the next output $y \in \mathcal{Y}$. The training data set will be denoted $D_n := \{(x_i, y_i), i = 1, \dots, n\}$. We will often make the assumption that the observations (x_i, y_i) are realizations of i.i.d. random variables from a distribution ν.

The distribution ν is unknown to the statistician, it's a matter of learning it from the D_n data. A learning rule $\mathcal{A}$ is a function that associates to training data D_n a prediction function $\widehat{f}_n$ (the hat on f indicates that it is an estimator):

$$\begin{array}{rcl} \mathcal{A} : \cup_{n\in\mathbb{N}} \big(\mathcal{X} \times \mathcal{Y}\big)^n & \to & \mathcal{Y}^{\mathcal{X}} \\ D_n & \mapsto & \widehat{f}_n \end{array} .$$

The estimated function $\widehat{f}_n$ is constructed to predict a new output y from a new x, where (x, y) is a pair of *test data*, i.e. not observed in the training data. The function $\widehat{f}_n$ is an estimator because it depends on the data D_n and not on unobserved parameter (such as ν). If D_n is random, it is a random function.

Risk and empirical risk. The objective is to find an estimator $\widehat{f}_n$ that predicts well new data by minimizing the risk:

$$\mathcal{R}(\widehat{f}_n) := \mathbb{E}\Big[\big(Y' - \widehat{f}_n(X')\big)^2 \,\Big|\, D_n\Big] \qquad \text{where} \qquad (X', Y') \sim \nu\,. \tag{Risk}$$

However, the statistician cannot compute the expectation (and thus the risk) because he does not know ν. A common method in supervised machine learning is therefore to replace the risk with the empirical risk.

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2\,. \tag{Empirical risk}$$

However, one must be careful about overfitting (case where $\widehat{\mathcal{R}}_n(f)$ is much lower than $\mathcal{R}(f)$, see Figure 2). In this class, we will study the performance of the least square estimator in the case of the linear model.

Linear model. When $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, the set of affine functions is natural. To ease the notation, we assume that the first component of the inputs is 1 so that it is sufficient to consider linera functions. The functions $\mathcal{F} = \{x \mapsto \theta^\top x : \theta \in \mathbb{R}^p\}$ can be parametrized by a vector $\theta \in \mathbb{R}^d$ and we thus consider minimizing the empirical risk

$$\widehat{\mathcal{R}}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \big(y_i - \theta^\top x_i\big)^2\,,$$

This expression can be rewritten in matrix notation. Let $Y = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$ be the vector of outputs and $X \in \mathbb{R}^{n\times d}$ the matrix of inputs, which rows are $x_i^\top$. X is called the *design matrix*. The empirical risk is then

$$\widehat{\mathcal{R}}_n(\theta) = \frac{1}{n}\big\|Y - X\theta\big\|_2^2\,. \tag{2}$$

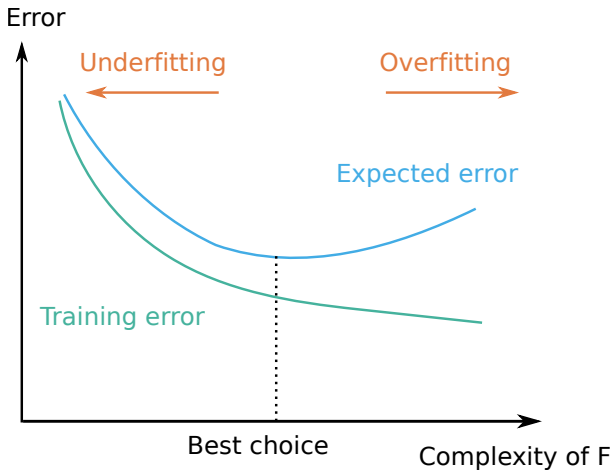

Figure 2: Overfitting and underfitting according to the complexity of $\mathcal{F}$. In blue the risk $\mathcal{R}(f)$ which we want to minimize, in green the empirical risk $\widehat{\mathcal{R}}_n(f)$ that we observe on the training data.

2 Ordinary Least Squares Estimator (OLS)

In the following, we assume that the design matrix X is injective (i.e., the rank of X is d). In particular, $d \leqslant n$.

Definition 2.1. *If X is injective, the minimizer of the empirical risk (i.e., of Equation (2)) is called the Ordinary Least Squares (OLS) estimator.*

Proposition 2.1 (Closed form solution)**.** *If X is injective, the the OLS exists and is unique. It is given by*

$$\widehat{\theta} = \left(X^\top X\right)^{-1} X^\top Y\,.$$

Proof. Since $\widehat{\mathcal{R}}_n$ is coercive and continuous, it admits at least a minimizer. Furthermore, we have

$$\widehat{\mathcal{R}}_n(\theta) = \frac{1}{n}\left\|Y - X\theta\right\|_2^2 = \frac{1}{n}\left(\theta^\top \left(X^\top X\right)\theta - 2\theta^\top X^\top Y + \|Y\|^2\right).$$

Since $\widehat{\mathcal{R}}_n$ is differentiable any minimizer should cancel the gradient:

$$\nabla\widehat{\mathcal{R}}_n(\widehat{\theta}) = \frac{1}{n}\left(\widehat{\theta}^\top (X^\top X) + (X^\top X)\widehat{\theta} - 2X^\top Y\right) = \frac{2}{n}\left((X^\top X)\widehat{\theta} - Y^\top X\right).$$

where the last equality is because $X^\top X \in \mathbb{R}^{p\times p}$ is symmetric. Since X is injective, $X^\top X$ is invertible (Exercise: show that is positive definite). Therefore, a solution of $\nabla\widehat{\mathcal{R}}_n(\widehat{\theta}) = 0$ satisfies

$$\widehat{\theta} = \left(X^\top X\right)^{-1} XY.$$

However, it remains to check that this is indeed a minimum and therefore that the Hessian is defined as positive, which is the case because: $\nabla^2\widehat{\mathcal{R}}_n(\widehat{\theta}) = \frac{2}{n}(X^\top X)$. □

2.1 Geometric interpretation

The linear model seeks to model the output vector $Y \in \mathbb{R}^n$ by a linear combination of the form $X\theta \in \mathbb{R}^n$. The image of X is the solution space, denoted $\mathrm{Im}(X) = \{Z\mathbb{R}^n : \exists \theta \in \mathbb{R}^d \text{ s.t. } Z = X\theta\} \subseteq \mathbb{R}^n$. This is the vector subspace of $\mathbb{R}^n$ generated by the $d < n$ columns of the design matrix. As $\mathrm{rg}(X) = d$, it is of dimension d.

By minimizing $\|Y - X\theta\|$ (cf. Definition 2.1), we thus look for the element of $\mathrm{Im}(X)$ closest to Y. This is the orthogonal projection of Y on $\mathrm{Im}(X)$, denoted $\widehat{Y}$. By definition of the OLS and by the Proposition 2.1, we have:

$$\widehat{Y} \overset{\text{Déf 2.1}}{=} X\widehat{\theta} \overset{\text{Prop. 2.1}}{=} X(X^\top X)^{-1}X^\top Y.$$

In particular, $P_X := X(X^\top X)^{-1}X^\top$ is the projection matrix on $\mathrm{Im}(X)$.

2.2 Numerical resolution

The closed form formula $\widehat{\theta} = \left(X^\top X\right)^{-1}X^\top Y$ from the OLS is useful in analyzing it. However, calculating it naively can be prohibitively expensive. Especially when d is large, one prefers to avoid inverting the design matrix $X^\top X$ which costs $\mathcal{O}(d^3)$ by the Gauss-Jordan method and can be very unstable when the matrix is badly conditioned. The following methods are usually preferred.

QR factorization To improve stability, QR decomposition can be used. Recall that $\widehat{\theta}$ is the solution to the equation:

$$(X^\top X)\widehat{\theta} = X^\top Y\,,$$

We write $X \in \mathbb{R}^{n\times d}$ of the form $X = QR$, where $Q \in \mathbb{R}^{n\times d}$ is an orthogonal matrix (i.e., $QQ^\top = I_n$) and $R \in \mathbb{R}^{d\times d}$ is upper triangular. Upper triangular matrices are very useful for solving linear systems. Substituting in the previous equation, we get:

$$\begin{aligned} R^\top(Q^\top Q)R\widehat{\theta} = R^\top Q^\top Y \quad &\Leftrightarrow \quad R^\top R\widehat{\theta} = R^\top Q^\top Y \\ &\Leftarrow \quad R\widehat{\theta} = Q^\top Y\,. \end{aligned}$$

Then all that remains is to solve a linear system with a triangular upper matrix, which is easy.

Gradient descent We can completely bypass the need of matrix inversion or factorization using gradient descent. It consists in solving the minimization problem step by step by approaching the minimum through gradient steps. For example, we initialize $\widehat{\theta}_0 = 0$, then update:

$$\begin{aligned} \widehat{\theta}_{i+1} &= \widehat{\theta}_i - \eta\nabla\widehat{\mathcal{R}}_n(\widehat{\theta}_i) \\ &= \widehat{\theta}_i - \frac{2\eta}{n}\left((X^\top X)\widehat{\theta}_i - Y^\top X\right), \end{aligned}$$

where $\eta > 0$ is a learning parameter. We see that if the algorithm converges, then it converges to a point canceling the gradient, thus to the OLS solution. To have convergence, the η parameter must be well calibrated, but we will see this in more detail in the class on gradient descent.

If the data set is much too big, $n \gg 1$. It can also be prohibitively expensive to load all the data to make the $\nabla\widehat{\mathcal{R}}_n(\widehat{\theta}_i)$ calculation. The common solution is then to do Stochastic Gradient Descent, where gradient steps are made only on estimates of $\nabla\widehat{\mathcal{R}}_n(\widehat{\theta}_i)$, calculated on a random subset of the data.

2.3 Nonlinear problem: polynomial, spline, kernel regression

The assumption that the observations y_i can be explained as a linear combination of the explanatory variables $x_{i,j}$ may seem strong. However, the previous linear framework can be applied to transformations of the variables $x_{i,j}$. For example, by adding the powers of the variables $x_{i,j}^k$ or their products $x_{i,j}x_{i,j}$, this allows comparison to polynomial spaces. Doing a linear regression on polynomial transformations of variables is like doing a polynomial regression.

Of course other bases (transformations) exist like regression on spline bases (piecewise polynomials with constraints on the edges). This is the model used for example by EDF in operation to predict electricity consumption as a function of variables such as time of day, day of the week, temperature or cloud cover. In general, we can consider transformations $\phi : \mathcal{X} \to \mathbb{R}^d$ and try to explain the outputs y_i with functions of the form $\theta \mapsto \phi(x_i)^\top \theta$. Another form of regression that we will discuss in the following will be kernel regression, which allows to compute efficiently the estimator even when ϕ maps to an infinite dimensional space.

In Figure 1, we have in this way minimized the empirical risk on polynomial spaces of degree 1 (linear model), 3 and 30. We can see that we must be careful not to consider spaces that are too large, at the risk that the model is badly posed (X non injective). Conversely, for the statistical analysis that we will see next to be verified, one must be in the true model $Y = \phi(X)\theta^* +$ centered noise. We must therefore make sure that $\phi(X)$ contains enough descriptors so that the dependency between Y and $\phi(X)$ is indeed linear. Otherwise we pay an additional bias term.

3 Statistical analysis

3.1 Stochastic assumptions and bias/variance decomposition

Any can of guarantees requires assumption about how the data is generated. In this section, we consider a stochastic framework that will allow us to analyze the performance of OLS.

Assumption 1 (Linear model)**.** *We assume that there exists a vector $\theta^* \in \mathbb{R}^d$ such that for all $1 \leqslant i \leqslant n$*

$$Y_i = x_i^\top \theta^* + Z_i \,, \tag{3}$$

where $Z = (Z_1, \dots, Z_n)^\top \in \mathbb{R}^n$ is a vector of errors (or noise). The Z_i are assumed to be centered independent variables $\mathbb{E}[Z_i] = 0$ and with variance $\mathbb{E}[Z_i] = 0$.

We write Y_i and Z_i with capital letters to remind ourselves that (from now on) they are random variables. The noise Z comes from the fact that in practice the observations Y_i never completely fit the linear forecast. They are due to noise or unobserved explanatory variables. As before, we assume that the first vector of the explanatory variables is the constant vector $x_{i,1} = 1$. The Equation (3) can be rewritten in matrix form:

$$Y = X\theta^* + Z$$

where $Y = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$, $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times d}$ and $Z = (Z_1, \dots, Z_n)^\top \in \mathbb{R}^n$.

From here, there are two settings of analysis for least squares:

- *Fixed design.* In this setting, the design matrix X is not random but deterministic and the features $x_1, \dots, x_n$ are fixed. The expectations are thus only with respect to the Z_i and the Y_i and the goal is to minimize the

$$\mathcal{R}_X(\theta) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - x_i^\top\theta\right)^2\right], \tag{4}$$

  for new random observations Y_i (different from the ones observed in the dataset) but on the same inputs.

- *Random design.* Here, both the inputs and the outputs are random. This is the most standard setting of supervised machine learning. The goal is to minimize the risk (sometimes called the generalization error) defined in Equation (Risk).

In this class, we will consider the fixed design setting because it eases the notation and the calculation (we only need simple linear algebra).

Before analyzing the statistical properties of OLS, we state a general result under the linear model which illustrate the tradeoff between estimation and approximation (or bias and variance).

Proposition 3.1 (Risk decomposition)**.** *Under the linear model (Assumption 1) with fixed design, for any $\theta \in \mathbb{R}^d$ it holds*

$$\mathbb{E}\big[\mathcal{R}_X(\theta) - \mathcal{R}_X(\theta^*)\big] = \|\theta - \theta^*\|_\Sigma^2$$

where $\Sigma = \frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$ and $\|\theta\|_\Sigma^2 = \theta^\top\Sigma\theta$. If θ is a random variable (because it depends on a random data set) then

$$\mathbb{E}\big[\mathcal{R}_X(\theta)] - \mathcal{R}(\theta^*) = \underbrace{\big\|\mathbb{E}[\theta] - \theta^*\big\|_\Sigma^2}_{Bias} + \underbrace{\mathbb{E}\Big[\big\|\theta - \mathbb{E}[\theta]\big\|_\Sigma^2\Big]}_{Variance}.$$

Proof. Now, let $\theta \in \mathbb{R}^d$. Then,

$$\begin{aligned}
\mathcal{R}_X(\theta) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - x_i^\top\theta\right)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - x_i^\top\theta^* + x_i^\top\theta^* - x_i^\top\theta\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(Y_i - x_i^\top\theta^*\right)^2\right] + \mathbb{E}\left[(Y_i - x_i^\top\theta^*)(x_i^\top\theta^* - x_i^\top\theta)\right] + \mathbb{E}\left[\left(x_i^\top\theta^* - x_i^\top\theta\right)^2\right] \\
&= \mathcal{R}_X(\theta^*) + 0 + \frac{1}{n}\sum_{i=1}^{n}(\theta^* - \theta)^\top x_i x_i^\top(\theta^* - \theta) \\
&= \mathcal{R}_X(\theta^*) + \big\|\theta - \theta^*\big\|_\Sigma^2.
\end{aligned}$$

If θ is random, we have the following bias-variance decomposition

$$\begin{aligned}
\mathbb{E}\big[\mathcal{R}_X(\theta)] - \mathcal{R}_X(\theta^*) &= \mathbb{E}\Big[\big\|\theta - \mathbb{E}[\theta] + \mathbb{E}[\theta] - \theta^*\big\|_\Sigma^2\Big] \\
&= \mathbb{E}\Big[\big\|\theta - \mathbb{E}[\theta]\big\|_\Sigma^2\Big] + \mathbb{E}\Big[(\theta - \mathbb{E}[\theta])^\top\Sigma(\mathbb{E}[\theta] - \theta^*)\Big] + \mathbb{E}\Big[\big\|\mathbb{E}[\theta] - \theta^*\big\|_\Sigma^2\Big] \\
&= \mathbb{E}\Big[\big\|\theta - \mathbb{E}[\theta]\big\|_\Sigma^2\Big] + \cancel{\mathbb{E}\big[(\theta - \mathbb{E}[\theta])\big]^\top\Sigma(\mathbb{E}[\theta] - \theta^*)} + \big\|\mathbb{E}[\theta] - \theta^*\big\|_\Sigma^2 \\
&= \mathbb{E}\Big[\big\|\theta - \mathbb{E}[\theta]\big\|_\Sigma^2\Big] + \mathbb{E}\Big[\big\|\mathbb{E}[\theta] - \theta^*\big\|_\Sigma^2\Big].
\end{aligned}$$

□

It is worth to note that the optimal risk satisfies

$$\mathcal{R}_X(\theta^*) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left(Y_i - x_i^\top \theta^*\right)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Z_i^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[Z_i^2\big] = \sigma^2\,.$$

3.2 Statistical properties of OLS

We now show some guarantees for the OLS estimator.

Proposition 3.2. *Under the linear model (i.e., Assumption 1) with fixed design, the OLS estimator $\widehat{\theta}$ defined in Definition 2.1 satisfies:*

- *it is unbiased* $\mathbb{E}\big[\,\widehat{\theta}\,\big] = \theta^*$.
- *its variance is* $\mathrm{Var}(\widehat{\theta}) = \frac{\sigma^2}{n}\Sigma^{-1}$.

We can even show that the OLS satisfies the Gauss-Markov property. It is optimal among unbiased estimators of θ, in the sense that it has a minimal variance-covariance matrix.

Proof. Using $\mathbb{E}[Z_i] = 0$ and $Y = X\theta^* + Z$, we have

$$\mathbb{E}[\widehat{\theta}] = \mathbb{E}\big[(X^\top X)^{-1}X^\top Y\big] = \mathbb{E}\big[(X^\top X)^{-1}X^\top X\theta^* + \cancel{(X^\top X)^{-1}X^\top Z}\big] = \theta^*\,.$$

Furthermore, using $\mathrm{Var}(Y) = \mathrm{Var}(Z) = \sigma^2 I_n$, we have

$$\mathrm{Var}(\widehat{\theta}) = \mathrm{Var}\big((X^\top X)^{-1}X^\top Y\big) = (X^\top X)^{-1}X^\top \mathrm{Var}(Y) X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1} = \frac{\sigma^2}{n}\Sigma^{-1}\,.$$

□

Corollary 3.3 (Excess risk of OLS)**.** *Under the linear model with fixed design, the excess risk of the OLS satisfy*

$$\mathbb{E}\big[\mathcal{R}_X(\widehat{\theta})\big] - \mathcal{R}(\theta^*) = \frac{\sigma^2 d}{n}\,.$$

Proof. Using the bias-variance decomposition and the fact that θ^* is unbiased (i.e., $\mathbb{E}[\widehat{\theta}] = \theta^*$), we have

$$\begin{aligned}
\mathbb{E}\big[\mathcal{R}_X(\widehat{\theta})] - \mathcal{R}_X(\theta^*) &= \cancel{\big\|\mathbb{E}[\widehat{\theta}] - \theta^*\big\|_\Sigma^2} + \mathbb{E}\Big[\big\|\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\big\|_\Sigma^2\Big] = \mathbb{E}\Big[\big\|\widehat{\theta} - \theta^*\big\|_\Sigma^2\Big] \\
&= \mathbb{E}\left[(\widehat{\theta} - \theta^*)^\top \Sigma (\widehat{\theta} - \theta^*)\right] \\
&= \frac{1}{n}\mathbb{E}\big[(\widehat{\theta} - \theta^*)^\top X^\top X (\widehat{\theta} - \theta^*)\big] \\
&= \frac{1}{n}\mathbb{E}\big[\,\mathrm{Tr}\left((\widehat{\theta} - \theta^*)^\top X^\top X (\widehat{\theta} - \theta^*)\right)\big] \\
&= \frac{1}{n}\mathbb{E}\big[\,\mathrm{Tr}\left(X(\widehat{\theta} - \theta^*)(\widehat{\theta} - \theta^*)^\top X^\top\right)\big] \qquad \leftarrow \text{because } \mathrm{Tr}(AB) = \mathrm{Tr}(BA) \\
&= \frac{1}{n}\,\mathrm{Tr}\left(X\mathbb{E}\big[(\widehat{\theta} - \theta^*)(\widehat{\theta} - \theta^*)^\top\big]X^\top\right) \qquad \leftarrow \text{because } \mathbb{E} \text{ and Tr are linear operators} \\
&= \frac{1}{n}\,\mathrm{Tr}\left(X\mathrm{Var}(\widehat{\theta})X^\top\right) \\
&= \frac{\sigma^2}{n}\,\mathrm{Tr}\left(X(X^\top X)^{-1}X^\top\right) = \frac{\sigma^2 d}{n}\,,
\end{aligned}$$

where the last equality is because $X(X^\top X)^{-1}X^\top = P_X$ is the orthogonal projection matrix onto the d-dimensional subspace $\mathrm{Im}(X)$. □

Exercise: show that the expected risk $\mathbb{E}\big[\widehat{\mathcal{R}}_X(\widehat{\theta})\big] = \frac{n-d}{n}\sigma^2$. In particular, an unbiased estimator of the noise σ^2 is

$$\widehat{\sigma}^2 = \frac{\|Y - X\widehat{\theta}\|^2}{n-p}\,.$$

Gaussian noise model A very considered special case is Gaussian noise $Z_i \sim \mathcal{N}(0, \sigma^2)$. This choice comes not only from the fact that it allows to compute many additional statistical properties on $\widehat{\theta}$ and to perform tests (confidence intervals, significance of variables, ...). In practice, it is also motivated by the central limit theorem and the fact that noise is often an addition of many phenomena not explained by the linear combination of the explanatory variables.

Proposition 3.4. *In the linear model with Gaussian noise, the maximum likelihood estimators of θ and σ satisfy respectively:*

$$\widehat{\theta}_{MV} = \big(X^\top X\big)^{-1} XY \qquad \textit{and} \qquad \widehat{\sigma}^2_{MV} = \frac{\|Y - X\widehat{\theta}\|^2}{n}\,.$$

We therefore find the least squares estimator obtained by minimizing the empirical risk. The variance estimator is biased. We will see more about maximum likelihood estimators in next lectures.

4 Ridge regression

If X is not injective (i.e., $\mathrm{rg}(X) \neq d$), the matrix $(X^\top X)$ is no longer invertible and the OLS optimization problem admits several solutions. The problem is said to be poorly posed or unidentifiable.

The Proposition 3.2 reminds us that the variance of $\widehat{\theta}$ depends on the conditioning of the matrix $(X^\top X)^{-1}$. The more the columns of the latter are likely to be dependent, the less stable $\widehat{\theta}$ will be. Several solutions allow to deal with the case where $\mathrm{rg}(X) < d$:

- *explicit complexity control* by reducing the $\mathrm{Im}(X)$ solution space. This can be done by removing columns from the X matrix until they become injective (for example, by reducing the degree of polynomials). One can also set identifiability constraints of the form $\theta \in V$ a vector subspace of $\mathbb{R}^d$ such that any element $y \in \mathrm{Im}(X)$ has a unique antecedent $\theta \in V$ with $y = X\theta$. For example, we could choose $V = \mathrm{Ker}(X)^\perp$.
- *implicit complexity control* by regularizing the empirical risk minimization problem. The most common is to regularize by adding $\|\theta\|_2^2$ (Ridge regression, which we see below) or $\|\theta\|_1$ (Lasso regression, see Lecture 7).

Definition 4.1. *For a regularization parameter λ, the Ridge regression estimator is defined as*

$$\widehat{\theta}_\lambda \in \underset{\theta\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2 \right\}\,.$$

The regularization parameter $\lambda > 0$ regulates the trade-off between the variance of $\widehat{\theta}$ and its bias.

Proposition 4.1. *The Ridge regression estimator is unique (even if X is not injective) and satisfies*

$$\widehat{\theta}_\lambda = \left(X^\top X + n\lambda I_n\right)^{-1} X^\top Y\,.$$

The proof is similar to the one of OLS and left as exercise. We can see that there is no longer the problem of inverting $X^\top X$ since the Ridge regression amounts to replacing $\left(X^\top X\right)^{-1}$ by $\left(X^\top X + n\lambda I_n\right)^{-1}$ in the OLS solution.

Proposition 4.2 (Risk of Ridge regression)**.** *Under the linear model (Assumption 1), the Ridge regression estimator satisfies*

$$\mathbb{E}\big[\mathcal{R}_X(\widehat{\theta}_\lambda)\big] - \mathcal{R}_X(\theta^*) = \sum_{j=1}^{d} (\theta_j^*)^2 \frac{\lambda_j}{(1+\lambda_j/\lambda)^2} + \frac{\sigma^2}{n}\sum_{j=1}^{d} \frac{\lambda_j^2}{(\lambda_j+\lambda)^2}\,,$$

where λ_j is the j-th eigenvalue of $\Sigma = \frac{1}{n}X^\top X$. In particular, the choice $\lambda^* = \frac{\sigma\sqrt{\operatorname{Tr}(\Sigma)}}{\|\theta^*\|_2\sqrt{n}}$ yields

$$\mathbb{E}\big[\mathcal{R}_X(\widehat{\theta}_{\lambda^*})\big] - \mathcal{R}_X(\theta^*) \leqslant \frac{\sigma\sqrt{2\operatorname{Tr}(\Sigma)\|\theta^*\|_2}}{\sqrt{n}}\,.$$

The proof, which follows from the bias-variance decomposition (Proposition 3.1) is left as exercise.

Note that as $\lambda \to 0$, its risk converges to the one of OLS. The first term corresponds to the bias of the Ridge estimator. Thus, on the downside the Ridge estimator is biased in contrast to the OLS. But on the positive side, its variance does not involve the inverse of Σ but of $\Sigma + \lambda I_d$ which is better conditioned. It has therefore a lower variance. The parameter λ controls this trade-off.

We can compare the excess risk bound obtained by $\widehat{\theta}_{\lambda^*}$ with the one of OLS which was $\sigma^2 d/n$:

- First, the one of OLS decreases in $O(1/n)$ while this one converges slower in $O(1/\sqrt{n})$ which could seem worse. Yet Ridge has a milder dependence on the noise σ instead of σ^2.
- Furthermore, since $\operatorname{Tr}(\Sigma) \leqslant \max_{1\leqslant i\leqslant n}\|x_i\|^2$, if the input norms are bounded by R, the excess risk of Ridge does not depend on the dimension d, which can even be infinite. It is called a *dimension free* bound.

The calibration of the regularization parameter is essential in practice. It can for example be done analytically as in the proposition (but some quantities are unknown σ^2, $\|\theta^*\|$,...). In practice one resorts to train/validation set or *cross-validation (generalized)*.