Chapter 3

Linear regression

This chapter is mostly based on [7, Chapter 2]. Further reading include [7, Chapter 3,4], [10, 9, 1, 3, 2, 5].

3.1 Introduction

We consider a generative model of the following form $Y_i = f^*(X_i) + \epsilon_i$, $i = 1 \dots, n$, where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T \sim \text{subG}(\sigma^2)$ and $\mathbb{E}[\epsilon] = 0$. The regression function $f^* \colon x \mapsto \mathbb{E}[Y|X = x]$ is assumed to be of the form $f^* \colon x \mapsto x^T\theta^*$ for an unknown $\theta^* \in \mathbb{R}^d$. This generative model is assumed to hold true throughout the chapter.

Design points:

The sample points $X_1, \dots, X_n$ are called *design* points. Depending on the nature of these points one may consider different ways to measure the quality of an estimate.

Random design: The design points are random, given $\mathcal{D}_n$ and a new observation X_{n+1}, one would like to build a predictor $\hat{f}_n$ for Y_{n+1}. In this case $R(\hat{f}_n)$ is a relevant measure.

Fixed design: If the design points are not random, one talks about fixed design and we denote the design points by $x_1, \dots, x_n$. In this situation, there is not much interest in talking about risk or expected prediction errror, since there is no expectation to consider. In this situation, we will consider for any g the mean squared error:

$$\text{MSE}(g) = \frac{1}{n}\sum_{i=1}^{n} \left(g(x_i) - f^*(x_i)\right)^2$$

We denote by $\mathbb{X} \in \mathbb{R}^{n\times d}$ the design matrix for which each row is one of the design points. Our model can then be expressed as follows:

$$Y = \mathbb{X}\theta^* + \epsilon \tag{LM}$$

where $Y = (Y_1, \dots, Y_n)^T$ and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$. In the sequel, we will focus on fixed designs. In this case, the mean squared error is given for any $\theta \in \mathbb{R}^d$, by

$$\text{MSE}(\theta) = \frac{1}{n}\|\mathbb{X}\left(\theta - \theta^*\right)\|_2^2.$$

3.2 Least squares and constrained least squares with fixed design

Least squares estimator

The least squares estimator is given by

$$\hat{\theta}^{LS} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2 \tag{3.1}$$

where we use the Euclidean norm. We start with an algebraic expression for $\hat{\theta}^{LS}$.

Lemma 3.2.1. *We have*

$$\mathbb{X}^T\mathbb{X}\hat{\theta}^{LS} = \mathbb{X}^T Y$$

and one solution is given by $\hat{\theta}^{LS} = (\mathbb{X}^T\mathbb{X})^\dagger \mathbb{X}^T Y$*, where* $\dagger$ *denotes the Moore-Menrose pseusdo inverse.*

Proof. The matrix $\mathbb{X}^T\mathbb{X}$ is positive semidefinite so that the objective in (3.1) is a convex quadratic function of θ. A necessary and sufficient condition for global optimality is that the gradient vanishes. This is the first claim and the second one follows from properties of the pseudoinverse. □

3.2.1 Constrained least squares estimator

Let K denote a closed subset of $\mathbb{R}^d$, the K constrained least squares estimator is given by

$$\hat{\theta}_K^{LS} \in \arg\min_{\theta \in K} \|\mathbb{X}\theta - Y\|_2^2 \tag{3.2}$$

where we use the Euclidean norm. The following lemma will be useful to prove finite sample bounds for $\hat{\theta}_K^{LS}$. The difficulty in bounding mean squared errors comes from the randomness of $\hat{\theta}^{LS}$, here we bound the MSE by a product of the noise and a quantity which can be controled uniformly. The question of how to compute constrained least squares estimates will be the topic of further chapters.

3.3 Finite sample bounds for least squares

We start with a general Lemma for constrained least squares estimators.

Lemma 3.3.1. *Assume that model* (LM) *holds and that* $\theta^* \in K$*, then, almost surely*

$$\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)$$

Proof. Since $\theta^* \in K$ and we have by definition of $\hat{\theta}_K^{LS}$,

$$\|\mathbb{X}\hat{\theta}_K^{LS} - Y\|_2^2 \leq \|\mathbb{X}\theta^* - Y\|_2^2 = \|\epsilon\|_2^2.$$

Furthermore, it holds that

$$\begin{aligned}\|\mathbb{X}\hat{\theta}_K^{LS} - Y\|_2^2 &= \|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^* - \epsilon\|_2^2 \\ &= \|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 - 2\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) + \|\epsilon\|_2^2\end{aligned}$$

So that

$$\begin{aligned}\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 &= \|\mathbb{X}\hat{\theta}_K^{LS} - Y\|_2^2 - \|\epsilon\|_2^2 + 2\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) \\ &\leq 2\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\end{aligned}$$

□

Unconstrained least squares

The following result provides mean squared error estimates for the least squares estimator.

Theorem 3.3.1. *Assume that* (LM) *holds with* $\epsilon \sim \text{subG}(\sigma^2)$*, then*

$$\mathbb{E}\left[\text{MSE}(\hat{\theta}^{LS})\right] \leq 16\sigma^2 \frac{r}{n}$$

where $r = \text{rank}(\mathbb{X}^T\mathbb{X})$*, furthermore, for any* $\delta > 0$*, with probability at least* $1-\delta$*,*

$$\text{MSE}(\hat{\theta}^{LS}) \leq \frac{64\sigma^2\left(2r + \log(1/\delta)\right)}{n}$$

Proof. Denote by $\Phi \in \mathbb{R}^{n\times r}$ a matrix which column constitute an orthonormal basis of the column span of $\mathbb{X}$. One may write $\mathbb{X}\left(\hat{\theta}^{LS} - \theta^*\right) = \Phi\nu$ where $\nu \in \mathbb{R}^r$. We have

$$\frac{\epsilon^T\mathbb{X}(\hat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}(\hat{\theta}^{LS} - \theta^*)\|_2} = \frac{\epsilon^T\Phi\nu}{\Phi\nu} = \left(\epsilon^T\Phi\right)\frac{\nu}{\|\nu\|_2} \leq \|\Phi^T\epsilon\|_2.$$

Applying Lemma 3.3.1, with $K = \mathbb{R}^d$, we have

$$\|\mathbb{X}(\hat{\theta}^{LS} - \theta^*)\|_2^2 \leq 4\left(\frac{\epsilon^T\mathbb{X}(\hat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}(\hat{\theta}^{LS} - \theta^*)\|_2}\right)^2 \leq 4\|\Phi^T\epsilon\|_2^2 = 4\sum_{i=1}^{r}(\Phi_i^T\epsilon)^2,$$

where Φ_i denotes the i-th column of Φ, $i = 1, \ldots, n$. Note that $\Phi_i^T\epsilon \sim \text{subG}(\sigma^2)$ by orthonormality of the Columns of Φ and Theorem 2.1.2 for $i = 1\ldots, r$ and hence using Theorem 2.1.1, we have

$$\mathbb{E}\left[\text{MSE}(\hat{\theta}^{LS})\right] \leq \frac{4}{n}\sum_{i=1}^{r}(\Phi_i^T\epsilon)^2 \leq \frac{16r\sigma^2}{n}.$$

This concludes the bound in expectation. For the bound in probability, we remakr that $\|\Phi^T\epsilon\|_2 = \max_{\|u\|\leq 1} u^T\Phi^T\epsilon$ where $\Phi^T\epsilon \sim \text{subG}(\sigma^2)$. Theorem 2.2.2 and Remark 2.2.2 entails for any $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}\text{MSE}(\hat{\theta}^{LS}) &\leq \frac{4}{n}\left(4\sigma\sqrt{r} + 2\sigma\sqrt{2\log(1/\delta)}\right)^2 \\ &\leq \frac{64\sigma^2\left(2r + \log(1/\delta)\right)}{n}\end{aligned}$$

□

Optimality and high dimensional setting

A natural question arising about Theorem 3.3.1 is "could we do better?". If d is the number of variables an $\mathbb{X}$ has full possible rank, then $r = \min(n, d) = d$ assuming $n \geq d$. We obtain a rate of the order of $\sigma^2 d/n$. In this case, we have

$$\text{MSE}(\hat{\theta}^{LS}) = (\hat{\theta}^{LS} - \theta^*)^T\frac{\mathbb{X}^T\mathbb{X}}{n}(\hat{\theta}^{LS} - \theta^*) \geq \lambda_{\min}\left(\frac{\mathbb{X}^T\mathbb{X}}{n}\right)\|\hat{\theta}^{LS} - \theta^*\|_2^2.$$

It turns out that this rate is optimal in a precise minimax sense.

Theorem 3.3.2. *Suppose that* $Y = \xi + \theta$ *where* $\theta \in \mathbb{R}^d$ *and* $\xi_i \sim \mathcal{N}(0, \sigma^2/n)$*,* $i = 1, \ldots, d$*. Then, for any* $\alpha \in (0, 1/4)$*:*

$$\inf_{\hat{\theta}}\sup_{\theta\in\mathbb{R}^d}\mathbb{P}_\theta\left(\|\hat{\theta} - \theta\|_2^2 \geq \frac{\alpha}{256}\frac{\sigma^2 d}{n}\right) \geq \frac{1}{2} - 2\alpha$$

where the infimum is taken over all measurable functions of Y*.*

The proof of this statement, can be done by reduction to statistical hypothesis testing and use known impossibility results to discriminate between two close hyptheses (See Chapter 4 of Philippe Rigollet's notes). Note that in the specific Gaussian sequence model proposed in the Theorem, the order of decay predicted by Theorem 3.3.1 is precisely $\sigma^2 d/n$. The theorem essentially says that for any estimator, there is a statistical setting for which this rate is attained. This type of result is called *minimax*. The conclusion is that the least squares estimator is optimal among all estimators without any prior knowledge.

In the high dimensional setting, we have $d \geq n$ and in this case, the bound of Theorem 3.3.1 remains bounded away from zero. Since this bound is optimal, it seems that there is no hope to solve high dimensional statistical problems. This is in fact not true, if one has for example prior knowledge that θ^* is in a certain ball of radius δ, then imposing that our estimaor $\hat{\theta}$ is in the same ball allows to estimate θ^* such that $\|\hat{\theta} - \theta^*\|^2 \leq \delta^2$. If δ is small, this may improve over the estimate of Theorem 3.3.1.

How is this compatible with Theorem 3.3.2? In the inf sup expression, the sup is taken over $\mathbb{R}^d$ and considering smaller subsets of $\mathbb{R}^d$ would reduce the right hand side.

ℓ_1 constrained least squares

We let B_1 denote the unit ball of the ℓ_1 norm in $\mathbb{R}^d$,

$$B_1 = \left\{ x \in \mathbb{R}^d, \sum_{i=1}^{d} |x_i| \leq 1 \right\}.$$

This is a polytope with $2d$ vertices given by the elements of the canonical basis and their oposite. The following result shows that under prior knowledge on θ^*, one can hope for better rates.

Theorem 3.3.3. *Let $K = B_1$ and $d \geq 2$. Assume that model* (LM) *holds with $\epsilon \sim \mathrm{subG}(\sigma^2)$ and $\theta^* \in K$. Assume also that the columns of $\mathbb{X}$ are normalized such that $\|\mathbb{X}_j\| \leq \sqrt{n}$, $j =, 1 \ldots, d$. Then, it holds that*

$$\mathbb{E}\left[\mathrm{MSE}(\hat{\theta}_K^{LS})\right] \leq \frac{4\sigma}{\sqrt{n}}\sqrt{2\log(2d)}$$

and for any $\delta > 0$, with probability at least $1 - \delta$, it holds that

$$\mathrm{MSE}(\hat{\theta}_K^{LS}) \leq \sigma\sqrt{\frac{32\log(2d/\delta)}{n}}.$$

Proof. Invoking Lemma 3.3.1, we have

$$\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*).$$

Note that since $\|\hat{\theta}_K^{LS}\|_1 \leq 1$ and $\|\theta^*\|_1 \leq 1$, we have $\|\hat{\theta}_K^{LS} - \theta^*\|_1 \leq 2$ so that

$$\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \leq 2 \sup_{\|v\|_1 \leq 2} \epsilon^T \mathbb{X} v = 4 \sup_{\|v\|_1 \leq 1} \epsilon^T \mathbb{X} v = 4 \sup_{u \in \mathbb{X}K} \epsilon^T u.$$

Now $\mathbb{X}K$ by linearity if v is not an extreme point of K then $\mathbb{X}v$ is not an extreme point of $\mathbb{X}K$. Hence $\mathbb{X}K$ is a polytope with at most $2d$ vertices which are taken among the columns of $\mathbb{X}$. The normalization of the columns of $\mathbb{X}$ ensures that on each of these vertices, $\mathbb{X}_j^T \epsilon \sim \mathrm{subG}(\sigma^2 n)$. Applying Theorem 2.2.3, we have

$$\mathbb{E}\left[\mathrm{MSE}(\hat{\theta}_K^{LS})\right] \leq \frac{4}{n}\sqrt{n}\sigma\sqrt{2\log(2d)} = \frac{4\sigma}{\sqrt{n}}\sqrt{2\log(2d)}.$$

Furthermore, for any $t > 0$, we have

$$\mathbb{P}\left[\mathrm{MSE}(\hat{\theta}_K^{LS}) \geq t\right] \leq \mathbb{P}\left[\sup_{u \in \mathbb{X}K} \epsilon^T u \geq \frac{nt}{4}\right] \leq 2de^{-\frac{nt^2}{32\sigma^2}}.$$

Given any $\delta \geq 0$, one has

$$2de^{-\frac{nt^2}{32\sigma^2}} \leq \delta \quad \Leftrightarrow t^2 \geq \frac{32\sigma^2}{n} \log\left(\frac{2d}{\delta}\right),$$

and the conclusion follows. □

ℓ_0 constrained least squares

We refer to the ℓ_0 norm as the cardinality of the set of non zero coordinates of a vector $\theta \in \mathbb{R}^d$. Note that this is an abuse of notations since this is not a norm. For any $\theta \in \mathbb{R}^d$,

$$\|\theta\|_0 = \sum_{i=1}^{d} \mathbb{I}(\theta_j \neq 0).$$

A vector with small ℓ_0 norm is called sparse. The support of a vector is the set of indices of its nonzero coordinates:

$$\mathrm{supp}(\theta) = \{j \in \{1, \ldots, d\}, \theta_j \neq 0\},$$

so that $\|\theta\|_0 = \mathrm{card}(\mathrm{supp}(\theta))$. By extension, for any $k = 1, \ldots, d$, we denote by $B_0(k)$ the set of k-sparse vectors.

Theorem 3.3.4. *For any $k \in \mathbb{N}^*$, $k \leq d/2$, let $K = B_0(k)$ and assume that model* (LM) *holds with $\epsilon \sim \mathrm{subG}(\sigma^2)$ and $\theta^* \in K$. Then, for any $\delta > 0$, with probability $1 - \delta$, it holds*

$$\mathrm{MSE}(\hat{\theta}_K^{LS}) \leq \frac{32\sigma^2}{n}\left(\log\left(\binom{d}{2k}\right) + 2k\log(6) + \log(1/\delta)\right).$$

Furthermore, we have

$$\mathbb{E}\left[\mathrm{MSE}(\hat{\theta}_K^{LS})\right] \leq \frac{32\sigma^2}{n}\left(1 + \log\left(\binom{d}{2k}\right) + 2k\log(6)\right)$$

Proof. Using Lemma 3.3.1, we have

$$\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \leq 4\frac{\left(\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\right)^2}{\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2}.$$

We have $\|\hat{\theta}_K^{LS} - \theta^*\|_0 \leq 2k$ and we set $\hat{S} = \mathrm{supp}\left(\hat{\theta}_K^{LS} - \theta^*\right)$, we have $|\hat{S}| \leq 2k$. We repeat similar steps as for the unconstrained least squares. For any $S \subset \{1, \ldots, d\}$, denote by $\mathbb{X}_S \in \mathbb{R}^{n \times |S|}$ the matrix composed of the columns of $\mathbb{X}$ indexed by S, by r_S the rank of $\mathbb{X}_S$ and by Φ_S an orthonormal basis of the span of the collumns of $\mathbb{X}$. There exists $\nu \in \mathbb{R}^{r_{\hat{S}}}$, such that

$$\frac{\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)}{\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2} = \frac{\epsilon \Phi_{\hat{S}} \nu}{\|\nu\|} \leq \max_{|S|=2k} \max_{u \in \mathbb{R}^{r_S}, \|u\|_2 \leq 1} u^T \Phi_S^T \epsilon.$$

Using Theorem 2.1.2, for any S, $\Phi_S^T\epsilon \sim \text{subG}(\sigma^2)$. Using a union bound, and Theorem 2.2.2, for any $t > 0$, we have

$$\begin{aligned}
\mathbb{P}\left[\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \geq 4t\right] &\leq \mathbb{P}\left[\max_{|S|=2k} \max_{u\in\mathbb{R}^{r_S}, \|u\|_2\leq 1} \left(u^T\Phi_S^T\epsilon\right)^2 > t\right] \\
&\leq \mathbb{P}\left[\max_{|S|=2k} \max_{u\in\mathbb{R}^{r_S}, \|u\|_2\leq 1} \left|u^T\Phi_S^T\epsilon\right| > \sqrt{t}\right] \\
&\leq \sum_{|S|=2k} \mathbb{P}\left[\max_{u\in\mathbb{R}^{r_S}, \|u\|_2\leq 1} |u^T\Phi_S^T\epsilon| > \sqrt{t}\right] \\
&\leq \sum_{|S|=2k} 6^{|S|} e^{\frac{-t}{8\sigma^2}} \\
&\leq \binom{d}{2k} 6^{2k} e^{\frac{-t}{8\sigma^2}}.
\end{aligned}$$

We deduce that

$$\mathbb{P}\left[\text{MSE}(\hat{\theta}^{LS}) \geq \frac{4t}{n}\right] \leq \binom{d}{2k} 6^{2k} e^{\frac{-t}{8\sigma^2}}$$

and we choose t such that the right hand side is bounded by δ, that is

$$t \geq 8\sigma^2\left(\log\left(\binom{d}{2k}\right) + 2k\log(6) + \log(1/\delta)\right)$$

and the bound in probability follows. The expectation is deduced from the bound in probability. We have, for any $H \geq 0$, using

$$\begin{aligned}
\mathbb{E}\left[\text{MSE}(\hat{\theta}_K^{LS})\right] &= \int_0^{+\infty} \mathbb{P}\left[\text{MSE}(\hat{\theta}_K^{LS}) > u\right] du \\
&\leq H + \int_0^{+\infty} \mathbb{P}\left[\text{MSE}(\hat{\theta}_K^{LS}) \geq (u+H)\right] du \\
&\leq H + \binom{d}{2k} 6^{2k} \int_0^{+\infty} e^{\frac{-n(u+H)}{32\sigma^2}} du \\
&= H + \binom{d}{2k} 6^{2k} e^{\frac{-nH}{32\sigma^2}} \frac{32\sigma^2}{n}.
\end{aligned}$$

Inverting the relation

$$\binom{d}{2k} 6^{2k} e^{\frac{-nH}{32\sigma^2}} = 1,$$

we obtain

$$H = \frac{32\sigma^2}{n}\left(\log\left(\binom{d}{2k}\right) + 2k\log(6)\right)$$

and the result follows. □

Lemma 3.3.2. *For any $1 \leq k \leq n$, it holds*

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

Proof. This is a simple recursion. □

As a consequence, the order of the bounds which we obtain is $\frac{\sigma^2 k}{n}\log\left(\frac{ed}{2k}\right)$. This also turns out to be minimax optimal for sparse estimation.