IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: P Rigollet
Scribe: Alex Gu, Chandler Squires

Lecture 6
Feb. 25, 2020

Goals: In this lecture, we introduce the linear regression model. We present the least squares estimator for this model, and develop results on the finite sample performance of this estimator. Then, we introduce constrained least squares estimation, and develop results for estimator when the regression coefficients are constrained to the ℓ_1-ball.

1. LINEAR REGRESSION SETUP

In this section, we setup the linear regression model. We observe n pairs (X_i, Y_i), such that

$$Y_i = f(X_i) + \varepsilon_i, 1 \le i \le n$$

with $\mathbb{E}\varepsilon_i = 0$. We consider the fixed design model, in which the X_i's are deterministic. In a fixed design setting, we wish to estimate some $\mu = (f(X_1), \cdots, f(X_n))^\top \in \mathbb{R}^n$, given the vector $\mu + \varepsilon$.

In other words, we wish to reconstruct a function $\hat{f}_n$ from our n given samples such that $\hat{f}_n(X_i)$ are as close to the original $f(X_i)$ as possible. We measure the performance of $\hat{f}_n$ using the *mean squared error*, given by

$$\mathsf{MSE}(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(x_i) - f(x_i))^2 = \frac{1}{n}|\hat{\mu_n} - \mu|_2^2$$

where $\hat{\mu_n} = (\hat{f}(X_1), \cdots, \hat{f}(X_n))^\top$.

In linear regression, we consider the class of linear functions f given by $f(x) = x^\top\theta^*$, where θ^* is unknown. This can be rewritten as follows:

$$\begin{bmatrix} | \\ \mu_i \\ | \end{bmatrix} = \begin{bmatrix} | \\ x_i^\top\theta^* \\ | \end{bmatrix} \Rightarrow \mu = \mathbb{X}\theta^*, \qquad \mathbb{X} = \begin{bmatrix} — & x_1^\top & — \\ — & x_2^\top & — \\ & \vdots & \end{bmatrix}$$

We also consider the set of linear candidate estimators $\hat{\mu} = \mathbb{X}\hat{\theta}$. The mean squared error can then be written as

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) = \frac{1}{n}|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 = (\hat{\theta} - \theta^*)^\top\frac{\mathbb{X}^\top\mathbb{X}}{n}(\hat{\theta} - \theta^*)$$

2. THE LEAST SQUARES ESTIMATOR

The least squares estimator is defined as any estimator that minimizes the mean squared error, that is,

$$\hat{\theta}^{\mathrm{LS}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} |Y - \mathbb{X}\theta|_2^2$$

Theorem: Let $\mathbb{X}$, Y be defined as in the previous section. The least-squares estimator is given by

$$\hat{\theta}^{\text{LS}} = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y$$

where $A^\dagger$ is the Moore-Penrose pseudoinverse of a matrix A.

Proof. We have

$$|Y - \mathbb{X}\theta|_2^2 = |Y|^2 + \theta^\top \mathbb{X}^\top \mathbb{X}\theta - 2\theta^\top \mathbb{X}^\top Y$$

Since the MSE function is convex, the estimator that minimizes the MSE must satisfy

$$\nabla_\theta |Y - \mathbb{X}\theta|_2^2 \Big|_{\theta=\hat{\theta}^{\text{LS}}} = 0 \Rightarrow 2\mathbb{X}^\top \mathbb{X}\hat{\theta}^{\text{LS}} - 2\mathbb{X}^\top Y = 0$$

or

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

A solution must exist, since the column space of $\mathbb{X}^\top \mathbb{X}$ is equal to the column space of $\mathbb{X}^\top$. Moreover,

$$\hat{\theta}^{\text{LS}} = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y$$

is a solution since $\mathbb{X}^\top \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top = \mathbb{X}^\top$.

Remark. For a matrix $\mathbb{X}$, if $\mathbb{X}^\top \mathbb{X}$ is not invertible (i.e., the x_i's belong to a subspace of $\mathbb{R}^d$), then $\ker(\mathbb{X}^\top \mathbb{X}) \neq \{0\}$ where $\forall a \in \ker(\mathbb{X}^\top \mathbb{X})$, we have $\mathbb{X}^\top \mathbb{X} a = 0$. Each θ can be divided into two parts, $\theta_1 \in \ker^\perp(\mathbb{X}^\top \mathbb{X})$, and $\theta_2 \in \ker(\mathbb{X}^\top \mathbb{X})$, such that $\theta = \theta_1 + \theta_2$. The Moore-Penrose pseudoinverse picks the solution with $\theta_2 = 0$, which is the solution that minimizes $|\theta|_2$.

If the noise ε is centered and subGaussian, we can bound the expectation of the MSE of the least-squares solution, as shown in the following theorem:

Theorem: Assume the linear regression model $Y = \mathbb{X}\theta^* + \varepsilon_i$, where ε_i are independent and in $\mathsf{subG}(\sigma^2)$. Then,

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}})] \lesssim \frac{\sigma^2 r}{n}$$

where $r = \text{rank}(\mathbb{X}^\top \mathbb{X}) \leq d \wedge n$.

Moreover, for $\delta > 0$, we have with probability $1 - \delta$,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}) \lesssim \frac{\sigma^2}{n} \log(\frac{1}{\delta}) + \frac{\sigma^2 r}{n}$$

Pre-proof. We first provide some intuition around this result. In the proof, we write $\hat{\theta}$ to mean $\hat{\theta}^{\text{LS}}$. First, consider $\hat{\mu} = \mathbb{X}\hat{\theta} = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y$. Let $P = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top$. Observe that P is a projection matrix, because

$$P^2 = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top = P$$

In fact, it is the projection of matrix Y onto the column span of $\mathbb{X}$. Therefore, if $Y = \mathbb{X}\theta^* + \varepsilon$, then $PY = P\mathbb{X}\theta^* + P\varepsilon = \mathbb{X}\theta^* + P\varepsilon$, because we know $\mathbb{X}\theta^*$ is in the column span of $\mathbb{X}$. Therefore, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) = \frac{1}{n}|\mathbb{X}\theta^* - \mathbb{X}\hat{\theta}|_2^2 = \frac{1}{n}|\mathbb{X}\theta^* - PY|_2^2 = \frac{1}{n}|\mathbb{X}\theta^* - \mathbb{X}\theta^* - P\varepsilon|_2^2 = \frac{1}{n}|P\varepsilon|_2^2$$

Observe that if $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then $|P\varepsilon|_2 \sim |\mathcal{N}(0, \sigma^2 I_r)|_2$, which makes $|P\varepsilon|_2^2 \sim \sigma^2 \chi_r^2$. Since $\mathbb{E}[\sigma^2 \chi_r^2] = \sigma^2 \mathbb{E}[\chi_r^2] = \sigma^2 r$, we get that

$$\mathbb{E}[\mathsf{MSE}(X\hat{\theta})] = \frac{\sigma^2 r}{n}$$

Now, we show how to prove the statement for general ε.

Proof. We first use the basic inequality:

$$|Y - \mathbb{X}\hat{\theta}|_2^2 \leq |Y - \mathbb{X}\theta|_2^2, \quad \forall \theta \in \mathbb{R}^d$$

In particular, we can take $\theta = \theta^*$ in the right-hand side, so that

$$|Y - \mathbb{X}\hat{\theta}|_2^2 \leq |Y - \mathbb{X}\theta^*|_2^2$$

Since $Y = \mathbb{X}\theta^* + \varepsilon$, the inequality can be written as

$$|\mathbb{X}\theta^* + \varepsilon - \mathbb{X}\hat{\theta}|_2^2 \leq |\mathbb{X}\theta^* + \varepsilon - \mathbb{X}\theta^*|_2^2$$

Cancelling and expanding both sides:

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 + |\varepsilon|_2^2 - 2\langle \varepsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^* \rangle \leq |\varepsilon|_2^2$$

Then, rearranging:

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 2\langle \varepsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^* \rangle$$

And dividing by a factor of $|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2$,

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2 \leq 2\langle \varepsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2} \rangle$$

Then squaring both sides:

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 4\langle \varepsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2} \rangle^2$$

Normally, we would think of applying Cauchy-Schwarz, which would give us a bound of

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 4\langle \varepsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2} \rangle^2 \leq 4|\varepsilon|^2 \left| \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2} \right|^2 = 4|\varepsilon|^2$$

Looking at what we want to prove, we notice that there is a dependence on the rank of the matrix r. However, after applying Cauchy-Schwarz, we completely got rid of this dependence. Hence, this is a dead end, and we need to try a different approach.

Now, let $\Phi \in \mathbb{R}^{n\times r} = [\phi_1, \phi_2, \cdots, \phi_r]$, where $\phi_1, \cdots, \phi_r$ comprise an orthonormal basis of the column span of $\mathbb{X}$. This means that there exists $\nu \in \mathbb{R}^r$ such that $\mathbb{X}(\hat{\theta} - \theta^*) = \Phi\nu$, or

$$|\mathbb{X}(\hat{\theta} - \theta^*)|_2 = |\Phi\nu|_2 = |\nu|_2$$

Therefore, we can write the MSE as

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 4\langle \varepsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2}\rangle^2 = 4\langle \varepsilon, \frac{\Phi\nu}{|\nu|_2}\rangle^2 = 4\langle \Phi^\top\varepsilon, \frac{\nu}{|\nu|_2}\rangle^2 \leq 4|\tilde{\varepsilon}|_2^2$$

where $\tilde{\varepsilon} = \Phi^\top\varepsilon \in \mathbb{R}^r$ and the last inequality follows by Cauchy-Schwarz. We claim that $\tilde{\varepsilon}^\top u \sim \mathsf{subG}(\sigma^2|u|_2^2)$ for $u \in \mathbb{R}^r$. To see this, notice that

$$\begin{aligned}
\mathbb{E}[\exp\{s\tilde{\varepsilon}^\top u\}] &= \mathbb{E}[\exp\{s\varepsilon^\top \Phi u\}] \\
&= \prod_{i=1}^r \mathbb{E}[\exp\{s\varepsilon_i(\Phi u)_i\}] \\
&\leq \prod_{i=1}^r \exp\left\{\frac{\sigma^2 s^2 (\Phi u)_i^2}{2}\right\} \\
&= \exp\left\{\frac{\sigma^2 s^2 |\Phi u|_2^2}{2}\right\} \\
&= \exp\left\{\frac{\sigma^2 s^2 |u|_2^2}{2}\right\}
\end{aligned}$$

where the inequality comes from the fact that all the ε_i are $\mathsf{subG}(\sigma^2)$. This shows that $\tilde{\varepsilon}^\top u \sim \mathsf{subG}(\sigma^2|u|_2^2)$.

Thus, $\varepsilon_j = \langle \tilde{\varepsilon}, e_j\rangle$ is $\mathsf{subG}(\sigma^2)$, and thus has variance at most σ^2. Therefore,

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \frac{1}{n} \cdot 4\mathbb{E}[|\tilde{\varepsilon}|_2^2] = \frac{4}{n}\sum_{j=1}^r \mathbb{E}[\tilde{\varepsilon}_j^2] \leq \frac{4r\sigma^2}{n}$$

Using the high-probability bounds developed in the last lecture for maximizing over the unit ball, we have

$$\mathbb{P}[|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 > t] \leq \mathbb{P}[4\langle \tilde{\varepsilon}, \frac{\nu}{|\nu|_2}\rangle^2 > t] \leq \mathbb{P}\left[\sup_{u\in\mathcal{B}_2(\mathbb{R}^r)} |\langle \tilde{\varepsilon}, u\rangle| > \frac{\sqrt{t}}{2}\right] \leq \exp\left\{\frac{-t}{32\sigma^2} + r\log 5\right\}$$

Therefore, we get that with probability $1-\delta$, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) \lesssim \frac{\sigma^2}{n}\log(\frac{1}{\delta}) + \frac{\sigma^2 r}{n}$$

3. CONSTRAINED LEAST SQUARES

We may wish to consider cases where $\mu = X\theta^*$ for $\theta^* \in K \subsetneq \mathbb{R}^d$, and constrain our minimization to K, i.e., solve the constrained least squares problem

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in K} |Y - \mathbb{X}\theta|_2^2$$

In this lecture, we consider the case $K = B_1 = \{x \in \mathbb{R}^d : |x|_1 \leq 1\}$. Recall that B_1 has $2d$ vertices at $\pm e_j$ for $j = 1, \ldots, d$. We will establish the following result:

Theorem: Assume the linear regression model $Y_i = X_i^\top \theta^* + \varepsilon_i$, with ε_i independent and $\mathsf{subG}(\sigma^2)$. Further, let $\theta^* \in B_1$, and $\max_j |\mathbb{X}_j|_2 \leq \sqrt{n}$, where $\mathbb{X}_j$ is the j-th column of $\mathbb{X}$. Denote the constrained least square solution $\hat{\theta} \in \operatorname{argmin}_{\theta \in B_1} |Y - \mathbb{X}\theta|_2$. Then

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \lesssim \sigma \sqrt{\frac{\log 2d}{n}}$$

and with probability $1 - \delta$, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) \lesssim \sigma \sqrt{\frac{\log \frac{2d}{\delta}}{n}}$$

Proof. We start with the basic inequality,

$$|Y - \mathbb{X}\hat{\theta}|^2 \leq |Y - \mathbb{X}\theta^*|^2$$

Once again substituting $Y = \mathbb{X}\theta^* + \varepsilon$, cancelling, and re-arranging to obtain

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|^2 \leq 2\langle \varepsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^* \rangle$$

Instead of using the fixed-point argument as in the unconstrained case, we will replace the right-hand side with a worst-case bound over the whole image of B_1 under the linear transformation $\mathbb{X}$

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 4 \sup_{v \in \mathbb{X}B_1} \langle \varepsilon, v \rangle$$

As established in the last lecture, we need only maximize over the vertices of $\mathbb{X}B_1$, i.e. over the columns of $\mathbb{X}$. By our assumptions on ε and $\mathbb{X}_j$, we have $\mathbb{X}_j^\top \varepsilon \sim \mathsf{subG}(\sigma^2 n)$. Thus, we meet the conditions for the maximal inequality from the last lecture and have

$$\mathbb{E}[\max_j |\mathbb{X}_j^\top \varepsilon|] \lesssim 2\sigma\sqrt{n}\sqrt{\log(2d)}$$

Hence,

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \frac{4}{n}\mathbb{E}[\max_j |\mathbb{X}_j^\top \varepsilon|] \lesssim \sigma \sqrt{\frac{\log(2d)}{n}}$$

Similarly, using the high-probability bounds developed in the last lecture for maximizing over a convex polytope, we have

$$\mathbb{P}(\mathsf{MSE}(\mathbb{X}\hat{\theta}) > t) \leq \mathbb{P}(\max_j \mathbb{X}_j^\top \varepsilon > \frac{nt}{4}) \leq 2d \exp\{-\frac{nt^2}{32\sigma^2}\}$$

Thus, picking

$$t = \sigma \sqrt{32 \frac{\log(\frac{2d}{\delta})}{n}}$$

bounds the right-hand side by δ.

□

Summary: In this lecture, we developed finite-sample results for linear regression with a fixed design matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ and $\mathsf{subG}(\sigma^2)$ noise, showing that the expectation of the mean squared error is bounded up to constant factor by $\frac{\sigma^2 r}{n}$, where $r = \text{rank}(\mathbb{X}^\top \mathbb{X})$.

We then considered constrained linear regression over the ℓ_1 ball, with additional constraints on the columns of $\mathbb{X}$.