# Lecture 7: Linear Regression

Isabella Zhu

27 February 2025

## §1 Linear Regression Model With Fixed Design

Pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ where $X_i$ is deterministic. Our model is

$$Y_i = f(X_i) + \epsilon_i, \quad \forall 1 \leq i \leq n$$

where $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[Y_i] = f(X_i)$. Typically we assume $\epsilon_i$ is subgaussian. $f$ is our *regression function* and our parameter of interest.

Our estimator of $f$ is $\hat{f}_n$. Usually we just use MSE for evaluation.

> **Remark 1.1.** *Fixed design* means that we only care about performance over our sample points, instead of over all $\mathbb{R}^d$.

For linear regression, we assume that our function is of the form

$$f(x) = x^T\theta^*, \quad \theta^* \in \mathbb{R}^d$$

We have our design matrix

$$\mathbb{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_n^T \end{bmatrix} \theta^*$$

### §1.1 Least Squares Estimator

Least squares estimator is

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

We are basically projecting $Y$ onto the column span of $\mathbb{X}$.

**Theorem 1.2**

We can choose

$$\hat{\theta}^{LS} = (\mathbb{X}^T\mathbb{X})^+\mathbb{X}^TY$$

where $A^+$ is the pseudo-inverse of $A$.

*Proof.* To minimize MSE, we set

$$\nabla_\theta |Y - \mathbb{X}\theta|_2^2 \big|_{\theta=\hat{\theta}} = 0$$

We solve this to get

$$-2\mathbb{X}^T Y + 2\mathbb{X}^T \mathbb{X}\theta = 0$$

so we want to satisfy the solution

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

In the case that there are multiple solutions, we can take the Moore-Penrose pseudo inverse, which is defined as

$$\hat{\theta} = \underset{\mathbb{X}^T\mathbb{X}\theta = \mathbb{X}^T Y}{\operatorname{argmin}} |\theta|_2$$

## §1.2 Bounds on the LSE

**Theorem 1.3**

Assume $Y = \mathbb{X}\theta^* + \epsilon$, where $\epsilon_i \in \text{subG}(\sigma^2)$ independent. Then, we have

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

where $r = rank(\mathbb{X}^T\mathbb{X})$. Furthermore,

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

*Pre-proof.* For the Gaussian case. Let $\epsilon \sim N(0, \sigma^2 I_n)$. Then, we have

$$\text{MSE} = \frac{1}{n}|\hat{\mu} - \mu|^2 = \frac{1}{n}|P\mu + P\epsilon - \mu|^2 = \frac{|P\epsilon|^2}{n},$$

where $P$ is the projection matrix $\mathbb{X}^T\mathbb{X}$.

A Gaussian projected is a Gaussian in lower dimension. We can rewrite

$$P\epsilon = U\Lambda U^T\epsilon = U\begin{bmatrix}\epsilon_1\\ \cdots\\ \epsilon_r\\ 0\\ \cdots\end{bmatrix}$$

So, this implies that $\mathbb{E}[\text{MSE}] = \frac{\sigma^2 r}{n}$. Furthermore, we get that $\text{Var}(\text{MSE}) = \frac{\sigma^4 r}{n}$ using variance of $\chi^2$ distribution.

*Proof.* We know that

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

for every $\theta \in \mathbb{R}^d$ because it's LSE. We will choose $\theta = \theta^*$. Define $\epsilon = Y - \mathbb{X}\theta^*$. Then, we have

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

which implies that

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

We can rewrite this as

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*| \leq 2\langle \epsilon, \frac{\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|} \rangle \leq 2 \sup_{u \in B_2} \langle \epsilon, u \rangle = 2|\epsilon|$$

Therefore,

$$\mathbb{E}[|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|^2] = 4\mathbb{E}[|\epsilon|^2] \leq 4n\sigma^2$$

so $\mathbb{E}[\text{MSE}] \leq 4\sigma^2$... wow that's a terrible bound.

Why is our bound bad? The intuition is that applying Cauchy-Schwarz in stats blindly usually results in a suboptimal bound. We didn't take advantage of our knowledge that $u$ is equal to $\mathbb{X}$ times something.

We know that $u \in \text{colspan}(\mathbb{X})$. Let $\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \dots \phi_r \end{bmatrix}$ be the orthonormal basis of $\text{colspan}(\mathbb{X})$. Can rewrite $u = \Phi v$.

Then, we have

$$2\langle \epsilon, u \rangle = 2\langle \epsilon, \Phi v \rangle \leq 2|\Phi^T \epsilon||v|$$

Notice that

$$|v|^2 = v^T v = v^T \Phi^T \Phi \Phi v = u^T u = 1.$$

So, we just need to bound $|\Phi^T \epsilon|$.

**Lemma 1.4**

For every $v \in \mathbb{R}^r$, we have $(\Phi^T \epsilon)^T v \sim \text{subG}(\sigma^2 |v|^2)$.

*Proof.* Plug in directly to definition of subgaussian. Only clever idea is that

$$|\Phi v|^2 = v^T \Phi^T \Phi v = v^T v = |v|^2$$

Then, we use our lemma to bound

$$\mathbb{E}[\text{MSE}(\mathbb{X}\hat{\theta})] \leq \frac{4}{n}\mathbb{E}|\Phi^T \epsilon|^2 \leq \frac{4\sigma^2 r}{n}.$$

Define $\tilde{\epsilon} = \Phi^T \epsilon$. We proved above that

$$\frac{1}{n}|X\hat{\theta} - \mathbb{X}\theta^*|^2 \leq \frac{4}{n}|\tilde{\epsilon}|^2$$

For the high probability bound, we impose an $\tilde{\epsilon}$-net, and we have

$$\mathbb{P}(|X\hat{\theta} - \mathbb{X}\theta^*|^2 > t) \leq \mathbb{P}(|\tilde{\epsilon}|^2 > \sqrt{t}/2)$$

$$\begin{aligned}
&= \mathbb{P}(\sup_{u \in B_2^n} \langle u, \tilde{\epsilon} \rangle > \sqrt{t}/2) \\
&\leq \sum_{z \in N} \mathbb{P}(\langle z, \tilde{\epsilon} \rangle > \sqrt{t}/4) \\
&\leq 5^n \cdot e^{-t/16\sigma^2} = \delta
\end{aligned}$$

and then solve for $t$ in terms of $\delta$.

## §1.3 Constrained Least Squares

Setup is as follows: $\mu = \mathbb{X}\theta^*$, $\theta^* \in K \subset \mathbb{R}^d$. Constrained LS estimation is

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

We will investigate the case when $K = B_1$.

**Theorem 1.5**

Assume $Y = \mathbb{X}\theta^* \epsilon$, $|\mathbb{X}_j|_2 \leq n$, and $\epsilon_i \sim \text{subG}(\sigma^2)$ independent. Then,

$$\mathbb{E}[\text{MSE}(\mathbb{X}\hat{\theta}_K)] \lesssim \min(\sigma\sqrt{\log d/n}, \sigma^2 r/n)$$

*Proof.* The $\sigma^2 r/n$ is the same as before. For the other value, we use similar ideas from the proof above.

$$\begin{aligned}
|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|^2 &\leq 2\langle \epsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^* \rangle \\
&= 2\langle \mathbb{X}^T\epsilon, \hat{\theta} - \theta^* \rangle \\
&\leq 2|\mathbb{X}^T\epsilon|_\infty |\hat{\theta} - \theta^*|_1 \quad \text{(Holder)} \\
&\leq 4|\mathbb{X}^T\epsilon|_\infty
\end{aligned}$$

But this is basically just $4\mathbb{E}[\max|\mathbb{X}_j^T\epsilon|]$, which from maximum of subgaussians, we know

$$\mathbb{E}[\max|\mathbb{X}_j^T\epsilon|] \lesssim \sigma \max|\mathbb{X}_j|_2 \sqrt{\log(2d)} \leq \sigma\sqrt{n}\sqrt{\log(2d)}$$

so as a result,

$$\mathbb{E}[\text{MSE}] \lesssim \frac{\sigma\sqrt{n\log(2d)}}{n}$$

which proves the result.