STAT 512: Statistical Inference **Autumn 2020**

Lecture 6: Estimators

Instructor: Yen-Chi Chen

Reference: Casella and Berger Chapter 7.2.

In statistics, we often encounter a problem where we observe a sequence of random variables (data) $X_1, \cdots, X_n$ that are a random sample from a population and we wish to use them to estimate some characteristics of the population. A simple probabilistic model is that these $X_1, \cdots, X_n$ are IID from an unknown PDF p. In this model, the PDF p describes the population (and the underlying sampling scheme). So if we can infer p, we can infer the underlying population.

A simple approach to this idea is to assume that p belongs to some *parametric family*. Namely, $p(x) = p(x; \theta)$, where θ is the underlying parameter. In this case, we say that we are using a *parametric model* . For instance, if we use the normal distribution as the parametric model, then $\theta = (\mu, \sigma^2)$ consists of the mean and variance parameters.

Often we do not know θ so we have to use the data to *estimate* it. An *estimator* is a statistic $W(X_1, \cdots, X_n)$ such that we use $W(X_1, \cdots, X_n)$ to estimate θ[1]. In this lecture, we will discuss some popular approaches to finding a good estimator.

6.1 Method of moments estimator

The method of moments is a very simple but useful approach to finding an estimator. The idea is as follows. For a parametric model $p(x; \theta)$, its moments are determined by the underlying parameter θ. For instance, the first moment is

$$m_1(\theta) = \mathbb{E}[X] = \int x p(x; \theta) dx$$

and the second moment is

$$m_2(\theta) = \mathbb{E}[X^2] = \int x^2 p(x; \theta) dx$$

The moments can be easily estimated using the data:

$$\hat{m}_j(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

for each $j = 1, 2, 3, \cdots$.

Suppose we have k parameters in the model, i.e., $\theta \in \mathbb{R}^k$, then we can use all the moments upto the k-th moments, i.e.,

$$m_j(\theta) = \int x^j p(x; \theta) dx,$$

[1]The concept of estimator can be generalized to other parameter of interest, not necessarily a parameter in a parametric model. For instance, we may be interested in the median of a distribution but we do not assume the distribution is Gaussian. Later in Lecture 10 we will discuss this in a greater details.

for $j = 1, 2, 3, \cdots, k$. And compare them to the data to obtain a unique solution. Namely, we find θ that solves the following equation

$$\begin{aligned} m_1(\theta) &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ m_2(\theta) &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ &\vdots \\ m_k(\theta) &= \frac{1}{n}\sum_{i=1}^{n} X_i^k. \end{aligned}$$

The resulting quantity $\hat{\theta}_{MoM}$ that solves the above equation is called the **method of moment estimator**.

Example: Normal distribution. Consider $X_1, \cdots, X_n$ IID and we use a normal model. In this case, we have two parameters μ and σ^2. It is known that

$$m_1(\mu, \sigma^2) = \mu, \quad m_2(\mu, \sigma^2) = \mu^2 + \sigma^2.$$

Thus, we immediately have

$$\hat{\mu} = \hat{m}_1(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$\hat{\mu}^2 + \hat{\sigma}^2 = \hat{m}_2(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{n} X_i^2,$$

which leads to

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2.$$

Example: Uniform distribution. Suppose that $X_1, \cdots, X_n \sim \mathsf{Uni}[0, \theta]$. We want to estimate θ. The method of moment estimator will lead to

$$\hat{\theta}/2 = \hat{m}_1(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

so $\hat{\theta} = \frac{2}{n}\sum_{i=1}^{n} X_i$.

Example: Exponential distribution. Consider the case where we use the exponential distribution to model $X_1, \cdots, X_n$. Since $p(x; \lambda) = \lambda e^{-\lambda x} I(x \geq 0)$, we have

$$m_1(\lambda) = \frac{1}{\lambda}.$$

As a result,

$$\frac{1}{\hat{\lambda}} = \hat{m}_1(\hat{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

6.2 Maximum likelihood estimator

Another very popular estimator is the **maximum likelihood estimator (MLE)**. The idea is very simple. Suppose we observe only one observation X from a PDF/PMF $p(x)$. The parametric model assumes that such a PDF/PMF can be written as $p(x) = p(x;\theta)$, where θ is the parameter of the model (θ is often the parameter of interest) inside a parameter space Θ ($\theta \in \Theta$). The idea of MLE is to ask the following question: given the observation X, which θ is the *most likely* parameter that generates X? To answer this question, we can vary θ and examine the value of $p(X;\theta)$.

Because we are treating X as fixed and θ being something that we want to optimize, we can view the problem as finding the best θ such that the **likelihood function** $L(\theta|X) = p(X;\theta)$ is maximized. The MLE uses the θ that maximizes the likelihood value. Namely,

$$\hat{\theta}_{MLE} = \mathsf{argmax}_\theta L(\theta|X).$$

When we have multiple observations $X_1, \cdots, X_n$, the likelihood function can be defined in a similar way – we use the joint PDF/PMF to define the likelihood function. Let $p(x_1, \cdots, x_n;\theta)$ be the joinr PDF/PMF. Then the likelihood function is

$$L_n(\theta) = L(\theta|X_1, \cdots, X_n) = p(X_1, \cdots, X_n;\theta).$$

Note that when we assume IID observations,

$$L_n(\theta) = \prod_{i=1}^n L(\theta|X_i) = \prod_{i=1}^n p(X_i;\theta).$$

In many cases, instead of using the likelihood function, we often work with the **log-likelihood function**

$$\ell_n(\theta) = \log L_n(\theta).$$

Because taking the logarithmic does not change the maximizer of a function, the maximizer of the log-likelihood function is the same as the maximizer of the likelihood function. There are both computational and mathematical advantages of using a log-likelihood function over likelihood function. To see this, we consider the case of IID sample. Computationally, the likelihood function often has a very small value due to the product form of PDF/PMFs. So it is very likely that the number if too small, making the computation very challenging. Mathematically, when we take log of the likelihood function, the product of PDF/PMFs becomes an additive form

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^n \log p(X_i;\theta).$$

Under IID assumption, each $\log p(X_i;\theta)$ is an IID random variable so the central limit theorem and the law of large number can be applied to the average, making it possible to analyze it asymptotic behavior.

Since under the IID assumptions, we have many advantages, we will assume IID from now on. Because MLE finds the maximum of $\ell_n(\theta)$, a common trick to find MLE is to study the gradient of the log-likelihood function, which is also known as the **score function**:

$$S_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta) = \sum_{i=1}^n s(\theta|X_i),$$

where $s(\theta|X_i) = \frac{\partial}{\partial\theta}\ell(\theta|X_i) = \frac{\partial}{\partial\theta}\log p(X_i;\theta)$. Under suitable conditions, the MLE satisfies the *score equation*:

$$S_n(\hat{\theta}_{MLE}) = 0.$$

Note that if there are more than one parameter, say $\theta \in \mathbb{R}^p$, the score equation will be a system of p equations.

Because the MLE is at the maximal point of the likelihood function, the curvature of the likelihood function around the maximal will determine its stability. To measure the curvature, we use the **Fisher's information matrix**:

$$I_n(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\ell_n(\theta)\right] = n \cdot I_1(\theta) = n \cdot -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}p(X_1;\theta)\right].$$

If the data is generated from a PDF/PMF $p(x;\theta_0)$ and some regularity conditions are satisfied,

$$\mathbb{E}(S_n(\theta_0)) = 0, I_1(\theta_0) = \mathbb{E}(S_1(\theta_0)S_1^T(\theta_0)).$$

Moreover,

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta_0\right) \xrightarrow{D} N(0, I_1^{-1}(\theta)).$$

Namely, the MLE is asymptotically normally distributed around the true parameter θ_0 and the covariance is determined by the Fisher's information matrix. Note that the asymptotic normality also implies that $\hat{\theta}_{MLE} - \theta_0 \xrightarrow{P} 0$.

Example 1: Binomial Distribution. Assume that we obtain a single observation $Y \sim \mathsf{Bin}(n,p)$, and we assume that n is known. The goal is to estimate p. The log-likelihood function is

$$\ell(p) = Y\log p + (n-Y)\log(1-p) + C_n(Y),$$

where $C_n(Y) = \log\binom{n}{Y}$ is independent of p. The score function is

$$S(p) = \frac{Y}{p} - \frac{n-Y}{1-p}$$

so solving the score equation gives us $\hat{p}_{MLE} = \frac{Y}{n}$. Moreover, the Fisher's information is

$$I(p) = \mathbb{E}\left\{\frac{\partial}{\partial p}S(p)\right\} = -\frac{\mathbb{E}(Y)}{p^2} - \frac{n-\mathbb{E}(Y)}{(1-p)^2} = \frac{n}{p(1-p)}.$$

Example 2: Poisson Distribution. Suppose we observe two integer RVs X_1, X_2. We assume that they are indepedently from Poisson distribution with unknown parameter λ. What will be the MLE of λ? In this case, the joint PDF is

$$p(x_1, x_2; \lambda) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}e^{-\lambda}.$$

Thus, the log-likelihood function will be

$$\ell(\lambda|X_1, X_2) = (X_1 + X_2)\log\lambda - 2\lambda - \log(X_1!) - \log(X_2!)$$

so the score function is

$$S(\lambda|X_1, X_2) = \frac{X_1 + X_2}{\lambda} - 2.$$

This leads to the MLE:

$$\hat{\lambda} = \frac{1}{2}(X_1 + X_2).$$

Example 3: Uniform Distribution. Consider $X_1, \cdots, X_n \sim \mathsf{Unif}[0,\theta]$. What will be the MLE of θ? Recall that the PDF will be

$$p(x_1, \cdots, x_n) = \prod_{i=1}^{n}\frac{1}{\theta}I(0 \le x_i \le \theta).$$

So the likelihood function is

$$L(\theta|X_1,\cdots,X_n)=\frac{1}{\theta^n}\prod_{i=1}^{n}I(0\le X_i\le\theta).$$

An interesting fact is that

$$\prod_{i=1}^{n}I(0\le X_i\le\theta)=I(0\le X_{\min}\le X_{\max}\le\theta),$$

where $X_{\min}=\min\{X_1,\cdots,X_n\}$ and $X_{\max}=\max\{X_1,\cdots,X_n\}$. So the likelihood function increases when θ decreases. However, it will drop to 0 immediately when $\theta<X_{\max}$. Thus, the MLE of θ will be $\hat{\theta}=X_{\max}$.

6.3 Bayesian estimator

The Bayesian inference is an alternative statistical paradigm to the Frequentist approach. The Bayesian approach interprets the probability in a broader sense that include subjective probability, which allows us to assign probability to almost every quantity in our model (including the parameter of interest and even a statistical model). The Bayesian inference relies on a simple decision theoretic rule – if we are competing two or more choices, we always choose the one with higher probability. This simple rule allows us to design an estimator, construct an interval, and perform hypothesis test.

In the Bayesian analysis, we assign a probability to every parameter in our model. For a parametric model $p(x;\theta)$, the parameter of interest θ is given a **prior distribution** $\pi(\theta)$ that reflects our belief about the value of θ. In a sense, the prior distribution quantifies our subjective belief about the parameter θ. The higher value of $\pi(\theta)$ indicates that we believe that θ is a more likely value of it.

How do we interpret this prior distribution? Here is a decision theoretic way of viewing it. To simplify the problem we assume that $\Theta=\{0,1,2\}$. Even without any data at hand, we can ask ourselves about our belief about each parameter value. Some people may think that 1 is the most likely one; some may think that 2 is the most likely one. To make our belief more precise, we use probability to work on it. Let $\pi(j)$ be the number that reflects our belief about $\theta=j$. We interpret the numerical value of $\pi(j)$ as follows. We are forced to guess the answer of $\theta=j$ versus $\theta\neq j$. If the answer is $\theta=j$ and we indeed guess it correctly, we will be rewarded δ dollar. If the answer is $\theta\neq j$ and we get it correct, we will be rewarded 1 dollar. If we get it wrong, we do not lose anything. Our principle is to maximize our expected reward. Now assume that the true value of θ has equal probability of being j or not j. Then what should we choose? $\theta=j$ or $\theta\neq j$? Now we think about this problem by varying δ from 0 to infinity. When δ is small, unless *we have very strong belief on* $\theta=j$, we will not bid on it. When increasing δ, at certain threshold we will switch our decision from bidding on $\theta\neq j$ to $\theta=j$. Let this threshold be η_j. η_j is a number that reflects our belief about $\theta=j$ and we associate it with our prior

$$\pi(j)=\frac{1}{1+\eta_j}\Leftrightarrow\eta_j=\frac{1-\pi(j)}{\pi(j)}\quad(\text{odds of }\theta=j).$$

Here, you see that we only use one simple decision rule – bidding on the one with a higher expected outcome. This allows us to quantify our belief.

Using the prior distribution, the Bayesian probability model can be written as follows:

$$\begin{aligned}X_1,\cdots,X_n|\theta &\overset{IID}{\sim} p(x|\theta)\\ \theta &\sim\pi.\end{aligned}$$

The Bayesian inference focuses on the distribution of θ after observing $X_1, \cdots, X_n$:

$$\pi(\theta|X_1, \cdots, X_n) = \frac{p(X_1, \cdots, X_n, \theta)}{p(X_1, \cdots, X_n)} \propto \underbrace{p(X_1, \cdots, X_n|\theta)}_{\text{likelihood}} \times \underbrace{\pi(\theta)}_{\text{prior}}.$$

This distribution is also known as the **posterior distribution.**

The posterior distribution informs us about how our prior belief is updated after seeing the data. It is the central quantity in Bayesian inference – all our decisions will be related to it. In Bayesian's point of view, probability models are just mathematical tools for analyzing data. We do not assume that the data is generated from a probability distribution. We just *view* the data as generated from $p(x;\theta)$. Given that we do not assume the probability model to be the *true* model, there is NO true parameter so we cannot talk about conventional statistical errors. However, Bayesian does have another way to expressing the error in our inference – the posterior distribution. The posterior distribution reflects our belief about the parameter after seeing the data, we can use it as a measure of *uncertainty* about θ. If the posterior distribution is more spread out, then the uncertainty in our inference is larger. On the other hand, if the posterior distribution is very concentrated, then there is very little (Bayesian) uncertainty.

There are two common estimator in Bayesian inference: the posterior mean and the maximum a posteriori estimation (MAP).

Posterior mean. Just like we often use the sample mean as an estimator of the population mean, the mean of the posterior distribution is a common quantity that was used as an estimator of θ:

$$\hat{\theta}_\pi = \mathbb{E}(\theta|X_1, \cdots, X_n) = \int \theta \cdot \pi(\theta|X_1, \cdots, X_n) d\theta.$$

It represents the average location of our belief about the parameter after seeing the data.

Maximum a posteriori estimation (MAP). Another common estimator of θ is the MAP; it relies on the similar principle as the MLE – we choose the one that is the most likely. Here the 'likely' is interpreted as our posterior belief about the parameter of interest θ. Formally, MAP is defined as

$$\hat{\theta}_{MAP} = \mathsf{argmax}_\theta \pi(\theta|X_1, \cdots, X_n).$$

Example: Binomial. Assume that we have an observation $Y \sim \mathsf{Bin}(N, \theta)$ where N is known and the parameter of interest is θ:

$$P(Y = y|\theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}.$$

We use a Beta distribution with parameters (α, β) as our prior distribution for θ. Namely,

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function and $\alpha, \beta > 0$. Note that (α, β) are called the hypter-parameters and are known quantities (because we know our belief about the data). For a Beta distribution with parameter α, β, the mean is $\frac{\alpha}{\alpha+\beta}$.

The posterior distribution is

$$\begin{aligned}\pi(\theta|Y) &= \frac{\binom{N}{Y}\theta^Y(1-\theta)^{N-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \binom{N}{Y}\theta^Y(1-\theta)^{N-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &\propto \theta^{Y+\alpha-1}(1-\theta)^{N-Y+\beta-1}\end{aligned}$$

so it is a Beta distribution with parameters $(Y+\alpha, N-Y+\beta)$. Then the posterior mean and MAP are

$$\hat{\theta}_\pi = \frac{Y+\alpha}{N+\alpha+\beta}, \qquad \hat{\theta}_{MAP} = \frac{Y+\alpha-1}{N+\alpha+\beta-2}$$

(these are the mean and the mode of a Beta distribution).

Note that in this problem, the MLE is $\hat{\theta}_{MLE} = \frac{Y}{N}$. Thus, the posterior mean has an interesting decomposition:

$$\begin{aligned}
\hat{\theta}_\pi &= \frac{Y+\alpha}{N+\alpha+\beta} \\
&= \hat{\theta}_\pi = \frac{Y}{N+\alpha+\beta} + \frac{\alpha}{N+\alpha+\beta} \\
&= \frac{Y}{N} \times \frac{N}{N+\alpha+\beta} + \frac{\alpha}{\alpha+\beta} \times \frac{\alpha+\beta}{N+\alpha+\beta} \\
&= \hat{\theta}_{MLE} \times W + \text{Prior mean} \times (1-W),
\end{aligned}$$

where $W = \frac{N}{N+\alpha+\beta}$ is a weight that is tending to 1 when $N \to \infty$. This phenomenon – the posterior mean can be written as the weighted average of the MLE and the prior mean – occurs in several scenarios. Moreover, the fact that the weights $W \to 1$ as the sample size $N \to \infty$ means that when we have more and more data, the prior distribution seems to be irrelevant. Thus, the posterior mean would have a similar asymptotic property as the sample mean. However, this is not a general phenomenon; often only certain combination of prior and likelihood models will have this feature.

Example: Normal Bayes estimator. Suppose we model $X_1, \cdots, X_n$ as IID from a normal distribution with mean μ and variance σ^2. Assume that σ^2 is known and the only unknown quantity is μ. We use a prior distribution of μ such that $\mu \sim N(\theta, \tau^2)$, where θ, τ^2 are pre-specified. Now we derive the posterior distribution of μ given $X_1, \cdots, X_n$ and all the specified parameters.

We know that

$$\pi(\mu|X_1, \cdots, X_n) \propto \exp(-\frac{1}{2\tau^2}(\mu-\theta)^2) \prod_{i=1}^n \exp(-\frac{1}{2\sigma^2}(X_i-\mu)^2).$$

Thus, the $\log \pi(\mu|X_1, \cdots, X_n)$ will be a quadratic function of μ, which implies that $\pi(\mu|X_1, \cdots, X_n)$ will still be a normal distribution. A direct computation shows

$$\log \pi(\mu|X_1, \cdots, X_n) = C_0 - \frac{1}{2\tau^2}(\mu-\theta)^2 - \sum_{i=1}^n -\frac{1}{2\sigma^2}(X_i-\mu)^2,$$

which implies

$$\begin{aligned}
\mathbb{E}[\mu|X_1, \cdots, X_n] &= \frac{\tau^2}{\tau^2+\sigma^2/n}\bar{X}_n + \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\theta \\
\mathsf{Var}(\mu|X_1, \cdots, X_n) &= \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}.
\end{aligned}$$

Again, we see that the posterior mean

$$\begin{aligned}
\hat{\mu}_\pi &= \frac{\tau^2}{\tau^2+\sigma^2/n}\bar{X}_n + \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\theta \\
&= \hat{\mu}_{MLE} \times W + \theta \times (1-W) \\
&= \text{MLE} \times W + \text{Prior mean} \times (1-W),
\end{aligned}$$

where $W = W_n = \frac{\tau^2}{\tau^2 + \sigma^2/n}$ is a proportion that converging to 1 as $n \to \infty$.

Remark.

- *Choice of prior and conjugate prior.* The choice of prior reflects our belief about the parameter before seeing any data. Sometimes people want to choose a prior distribution such that the posterior distribution is in the same family as the prior distribution, just like what we have observed in the above example. If a prior distribution and a likelihood function leads to a posterior that belongs to the same family as the prior, we call this prior **conjugate prior**. There are several conjugate priors know to date, see https://en.wikipedia.org/wiki/Conjugate_prior for an incomplete list of cases.

  Another common choice of prior is called the **Jeffreys prior**[2], which chooses a prior $\pi(\theta) \propto \sqrt{\det(I_1(\theta))}$, where $I_1(\theta)$ is the Fisher's information matrix. One can view the Jeffreys prior as the prior that *we do not have any prior belief about* θ; or more formally, an uninformative prior.

- *Challenge of computing the posterior.* In general, if we do not choose a conjugate prior, the posterior distribution could be difficult to compute. The challenge often comes from the normalization quantity $p(X_1, \cdots, X_n)$ in the denominator of the posterior $\pi(\theta|X_1, \cdots, X_n)$ (the numerator is just the prior times the likelihood). In practice we will use Monte Carlo method to compute the posterior – we generate points from $\pi(\theta|X_1, \cdots, X_n)$ and as we generate enough points, these points should approximate the true posterior distribution well. We will talk more about this later in the lecture of MCMC (Monte Carlo Markov Chain).

- *Consistency.* In pure Bayesian's point of view, statistical consistency is not an important property because probability model is a working model to describe the data and we do not need to assume that there exists an actual parameter that generates the data. Thus, the posterior distribution is the quantity that we really need to make our inference. However, sometimes Bayesian estimators, such as the posterior mean or MAP, does have statistical consistency. Namely, $\hat{\theta}_\pi \xrightarrow{P} \theta_0$ and $\hat{\theta}_{MAP} \xrightarrow{P} \theta_0$, where the data $X_1, \cdots, X_n \overset{IID}{\sim} p(x; \theta_0)$. This is often related to the Bernstein-von Mises theorem[3]. Although statistical consistency was not an important property in Bayesian paradigm (because Bayesian does not assume the data is indeed from a probability model; probability models are just a mathematical model to help us analyze the data), still many researchers would prove consistency when proposing a Bayesian approach.

6.4 Empirical risk minimization (ERM) and M-estimation

Another popular idea of finding a good estimator is the **empirical risk minimziation (ERM)**. It is widely used in machine learning and many modern statistical procedure. In fact, the MLE can be viewed as a special case of ERM.

6.4.1 Motivation: least square estimate

To start with, consider the linear regression problem where we observe $(X_1, Y_1), \cdots, (X_n, Y_n)$ and we want to estimate their linear relationship. Here we assume that each covariate $X_i = (X_{i,1} = 1, \cdots, X_{i,p+1})$ such that the first covariate is a constant 1 – this will include the intercept as part of the covariate and we have p covariate. As we have discussed in the previous lecture, a simple way to investigate the linear relationship

[2] see https://en.wikipedia.org/wiki/Jeffreys_prior for more details.
[3] https://en.wikipedia.org/wiki/Bernstein%E2%80%93von_Mises_theorem

is to apply a linear model. Namely, we model that

$$\mathbb{E}[Y|X] = X^T\beta,$$

where $\beta \in \mathbb{R}^p$.

The linear model attempts to find β by minimizing the MSE, i.e.,

$$\beta^* = \mathsf{argmin}_\beta R(\beta) = \mathsf{argmin}_\beta \mathbb{E}[(Y - X^T\beta)^2].$$

If we do not know the distribution of X, Y, we cannot compute the above estimator since we are unable to compute the expectation. However, we can approximate the MSE using the *empirical mean square errors*, i.e., we approximate $R(\beta)$ by

$$\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2.$$

And the estimator corresponds to the famous *least square estimate (LSE)*:

$$\hat{\beta} = \mathsf{argmin}_\beta \hat{R}(\beta) = \mathsf{argmin}_\beta \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2.$$

Here, we see that our estimator $\hat{\beta}$ is the minimizer of an *empirical risk* $\hat{R}_n(\beta)$. The term 'empirical' here refers to something computable from the data/sample. The ERM is an idea similar to the least square approach but with a more general risk function.

6.4.2 A general ERM approach

Before we formally introduce the ERM, we first define a few terms. In prediction, a loss function $L : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ is a function that measures the quality of prediction or estimation. Note that $\mathcal{Y}$ is the support of Y. You can think of the value of $L(a, b)$ as a measure on how much we lost when the true value of b and we make a prediction or estimate with a.

A popular loss is the square loss, i.e., $L(a, b) = (a - b)^2$ but it can be more general. For instance, we can consider the absolute loss, $L(a, b) = |a - b|$ when both a, b are a single number.

The risk is the expected loss. In the case of mean square prediction, the loss function $L(a, b) = (a - b)^2$ and our prediction is $f_\beta(X) = X^T\beta$. So the loss is $L(Y, f_\beta(X)) = (Y - f_\beta(X))^2 = (Y - X^T\beta)^2$ and the risk is $R = \mathbb{E}[L(Y, f_\beta(X))] = \mathbb{E}[(Y - X^T\beta)^2]$. Since the risk will change when we vary β, so we can write the risk as a risk function $R = R(\beta) = \mathbb{E}[(Y - X^T\beta)^2]$.

In general, with any loss function L, we can always write the risk function as

$$R(\beta) = \mathbb{E}[L(Y, f_\beta(X))].$$

As we have discussed, we may not be able to compute the risk function because we do not know the joint distribution of (X, Y). So instead, we attempt to approximate it with something computable from teh data. The empirical risk is the estimated/sample/computable version of the risk function:

$$\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f_\beta(X_i)).$$

With a data at hand, we can compute the value of the empirical risk easily.

The ERM attempts to construct an estimator $\hat{\beta}$ by minimizing $\hat{R}(\beta)$, namely,

$$\hat{\beta} = \mathsf{argmin}_\beta \hat{R}(\beta).$$

Any estimator that can be written as the above expression is called an ERM estimator. As we have seen previously, the least square estimate is an ERM estimator.

Example: least absolute deviation. Consider the loss function $L(a, b) = |a - b|$. Then the regression estimator

$$\hat{\beta}_{LAD} = \mathsf{argmin}_\beta \hat{R}(\beta) = \mathsf{argmin}_\beta \frac{1}{n} \sum_{i=1}^{n} |Y_i - X_i^T \beta|$$

is called the least absolute deviation (LAD) estimator. It is more robust against outliers in the data compared to the LSE due to use of L_1 norm (absolute value) as the loss function. Note: the LSE is approximating the conditional mean of Y given X by a linear function; the LAD will be approximating the conditional *median* of Y given X by a linear function.

6.4.3 M-estimation

The ERM is actually a special case of a more general procedure called **M-estimation**. The M-estimation finds an estimator by maximizing an empirical objective function, i.e.,

$$\hat{\theta} = \mathsf{argmax}_\theta \frac{1}{n} \sum_{i=1}^{n} \omega(\theta; X_i)$$

for some function ω. If we define the negative risk as the objective function, then it is easy to see that the ERM is M-estimation.

The M-estimation generalizes the concept of optimization to a more general scenario. In fact, you can easily see that if we choose the objective function to be the log-likelihood function

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta | X_i) = \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \theta),$$

then the M-estimator is the MLE.