MTH 511a - 2020: Lecture 14

Instructor: Dootika Vats

We have learned a fair amount about sampling from various distributions and estimating integrals. For the next few weeks we will focus our attention to optimization methods for certain statistical procedures.

One common use of optimization in statistics is when obtaining a maximum likelihood estimator (MLE) for a parameter. Thus, we first introduce MLE below briefly, before going into optimization methods.

1 Maximum Likelihood Estimation

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with density $f(x|\theta)$. The "x given θ" implies that given a particular value of θ, $f(\cdot|\theta)$ defines a density.

The parameter θ can be a vector of parameters. After having obtained *real data*, from F, we want to

1. estimate θ
2. construct confidence intervals around the estimator of θ.

A useful method of estimating θ is the method of *maximum likelihood estimation*. Let $\mathbf{X} = (X_1, \ldots, X_n)$. The idea is that we define a function $L(\theta|\mathbf{X})$ which measures "how likely is a particular value of θ given the data observed" and then find the θ that maximizes this likelihood.

This likelihood is defined as

$$L(\theta|\mathbf{X}) = \prod_{i=1}^{n} f(X_i|\theta)\,.$$

It is important to note that $L(\theta|\mathbf{X})$ is not a distribution over θ. The "most likely" value is the value that maximizes the likelihood

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} L(\theta|\mathbf{X})\,.$$

1.1 Examples

Example 1 (Bernoulli). Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(p)$. Then the likelihood is

$$\begin{aligned} L(p|\mathbf{x}) &= \prod_{i=1}^{n} p(x_i|p) \\ &= \prod_{i=1}^{n} \left[p^{x_i}(1-p)^{1-x_i}\right] \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i}\,. \end{aligned}$$

To obtain the MLE of θ, we will maximize the likelihood. Note that maximizing the likelihood is the same as maximizing the log of the likelihood, but the calculations are easier after taking a log. So we take a log:

$$\begin{aligned} \Rightarrow l(p) := \log L(p|\mathbf{x}) &= \left(\sum_{i}^{n} x_i\right)\log p + \left(n - \sum_{i}^{n} x_i\right)\log(1-p) \\ \frac{dl(p)}{dp} &= \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} \overset{\text{set}}{=} 0 \\ \Rightarrow \hat{p} &= \frac{1}{n}\sum_{t=1}^{n} x_i\,. \end{aligned}$$

Verify for yourself that the second derivative is negative for this $\hat{p}$. Thus,

$$\hat{p}_{\text{MLE}} = \frac{1}{n}\sum_{t=1}^{n} x_i\,.$$

Example 2 (Two parameter exponential). The density of a two parameter exponential distribution is

$$f(x|\mu,\lambda) = \lambda e^{-\lambda(x-\mu)} \quad x \geq \mu, \quad \mu \in \mathbb{R}, \lambda > 0\,.$$

We want to compute the MLEs of both λ and μ. The likelihood is

$$\begin{aligned} L(\lambda,\mu|\mathbf{x}) &= \prod_{t=1}^{n} f(x_i|\mu,\lambda) \\ &= \prod_{t=1}^{n} \lambda e^{-\lambda(x_i-\mu)}\, I(x_i \geq \mu) \end{aligned}$$

$$= \lambda^n \exp\left\{-\lambda\left(\sum_i x_i - n\mu\right)\right\} I(x_i \geq \mu) \quad \forall \mu\,.$$

But if $X_1, \ldots, X_n \geq \mu \Rightarrow \min\{X_i\} \geq \mu$. So

$$L(\lambda, \mu|\mathbf{x}) = \lambda^n \exp\left\{-\lambda\left(\sum_i x_i - n\mu\right)\right\} I\left(\min_i\{x_i\} \geq \mu\right) \quad \forall \mu\,.$$

We will first try to maximize with respect to μ and then with respect to λ. Note that $L(\lambda, \mu)$ is an increasing function of μ within the restriction. So that the MLE of μ is the largest value in the support of μ where $\mu \leq \min\{X_i\}$. So

$$\hat{\mu}_{\text{MLE}} = \min_{1 \leq i \leq n}\{X_i\} = X_{(1)}\,.$$

Next, note that

$$\begin{aligned}
L(X_{(1)}, \lambda|\mathbf{x}) &= \lambda^n \exp\left\{-\lambda\left(\sum_i X_i - nX_{(1)}\right)\right\} \\
\Rightarrow l(X_{(1)}, \lambda) := \log L(X_{(1)}, \lambda|\mathbf{x}) &= n\log\lambda - \lambda\left(\sum X_i - nX_{(1)}\right) \\
\Rightarrow \frac{d\,l}{d\,\lambda} &= \frac{n}{\lambda} - \left(\sum X_i - nX_{(1)}\right) \overset{\text{set}}{=} 0 \quad \text{and} \\
\frac{d^2\,l}{d\,\lambda^2} &= -\frac{n}{\lambda^2} < 0\,.
\end{aligned}$$

So, the function is concave, thus there is a unique maximum. Set

$$\begin{aligned}
\frac{d\,l}{d\,\lambda} &= 0 \\
\Rightarrow \frac{n}{\lambda} &= \sum_{t=1}^{n} X_i - nX_{(1)} \\
\Rightarrow \hat{\lambda}_{\text{MLE}} &= \frac{n}{\sum X_i - nX_{(1)}}\,.
\end{aligned}$$

Why MLE?

One main reason of using MLE is that *often* (not always), the resulting estimators are consistent and asymptotically normal. That is, for a general likelihood $L(\theta|x)$

$$\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta \text{ as } n \to \infty$$

and under some additional conditions, we also have

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} N(0, \Sigma^*)\,,$$

where Σ^* is an estimable matrix called the inverse Fisher information matrix. So if we use MLE estimation (and after verifying certain conditions), we know that we can construct confidence intervals around $\hat{\theta}_{\text{MLE}}$. This is great!

Note: The conditions required for consistency and asymptotic normality are important. But we will not be discussing them in this course. Please look at topics under "Inference" for more information.

2 Regression

We will focus a lot on variants of linear regression. Hence, we focus on that specifically here. The following is the setup in regression.

Let $Y_1, Y_2, \dots, Y_n$ be observations known as the *response*. Let $x_i = (x_{i1}, \dots x_{ip})^T \in \mathbb{R}^p$ be the ith corresponding vector of covariates for the ith observation. Let $\beta \in \mathbb{R}^p$ be the *regression coefficient* so that for $\sigma^2 > 0$,

$$Y_i = x_i^T \beta + \epsilon_i \quad \text{where } \epsilon_i \sim N(0, \sigma^2)\,.$$

Let $X = (x_1^T, x_2^T, \dots, x_n^T)^T$. In vector form we have,

$$Y = X\beta + \epsilon \sim N_n(X\beta, \sigma^2 I_n)\,.$$

The linear regression model is built to estimate β, which measures the linear effect of X on Y. There is much more to linear regression and multiple courses are required to study all aspects of it. However, here we will just focus on the mathematical properties and optimization tools required to study them.

Example 3 (MLE for Linear Regression). In order to understand the linear relationship between X and β, we will need to estimate β. We have

$$\begin{aligned}
L(\beta, \sigma^2|y) &= \prod_{t=1}^{n} f(y_i|X, \beta, \sigma^2) \\
&= \prod_{t=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2}\frac{(Y - X\beta)^T(Y - X\beta)}{\sigma^2}\right\} \\
\Rightarrow l(\beta, \sigma^2) := \log L(\beta, \sigma^2|y) &= -\frac{1}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{(Y - X\beta)^T(Y - X\beta)}{\sigma^2}
\end{aligned}$$

Note that

$$\begin{aligned}
(y - X\beta)^T(y - X\beta) &= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
&= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta\,.
\end{aligned}$$

Using this we have

$$
\begin{aligned}
\frac{dl}{d\beta} &= -\frac{1}{2\sigma^2}\left[-2X^Ty + 2X^TX\beta\right] = \frac{X^Ty - X^TX\beta}{2\sigma^2} \overset{set}{=} 0 \\
\frac{dl}{d\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{(y-X\beta)^T(y-X\beta)}{2\sigma^4} \overset{set}{=} 0\,.
\end{aligned}
$$

The first equation leads to $\hat{\beta}_{\text{MLE}}$ satisfying

$$
X^Ty - X^TX\hat{\beta}_{\text{MLE}} = 0 \Rightarrow \hat{\beta}_{\text{MLE}} = (X^TX)^{-1}X^Ty\,,
$$

if $(X^TX)^{-1}$ exists. And $\hat{\sigma}^2_{MLE}$ is

$$
\hat{\sigma}^2_{MLE} = \frac{(y - X\hat{\beta}_{\text{MLE}})^T(y - X\hat{\beta}_{\text{MLE}})}{n}
$$

Verify: that the second derivative is negative, and this is indeed the maximum.

Note: What if $(X^TX)^{-1}$ does not exist? For example, if $p > n$, then the number of observations is less than the number of parameters, and since X is $n \times p$, (X^TX) is $p \times p$ of rank $n < p$. So X^TX is not full rank and cannot be inverted. In this case, the MLE does not exist and other estimators need to be constructed. This is one of the motivations of *penalized regression*, which we will discuss in detail.

2.1 Penalized Regression

Note that in the Linear regression setup, the MLE for β satisfied:

$$
\hat{\beta}_{\text{MLE}} = \arg\min_{\beta}(y - X\beta)^T(y - X\beta)
$$

Suppose X is such that (X^TX) is not invertible, then the MLE does not exist, and we don't know how to estimate β. In such cases, we may used *penalized likelihood*, that penalizes the coefficients β so that some of the βs are pushed towards zero. The corresponding Xs to those small βs are essentially not important, removing singularity from X^TX. The penalized likelihood is

$$
\tilde{Q}(\beta) = L(\beta|y) + \tilde{P}(\beta)\,.
$$

Here $P(\beta)$ is called the *penalization* function.

Since the optimization of $L(\beta|y)$ only depends on $(y - X\beta)^T(y - X\beta)$ term, a penalized (negative) log-likelihood is used and the final penalized (negative) log-likelihood is

$$
Q(\beta) = -\log L(\beta|y) + P(\beta)
$$

There are *many* ways of penalizing β and each method yields a different estimator. A popular one is the *ridge* penalty.

Example 4 (Ridge Regression). The ridge penalization term is $\lambda\beta^T\beta/2$ for $\lambda > 0$ for

$$Q(\beta) = \frac{(y - X\beta)^T(y - X\beta)}{2} + \frac{\lambda}{2}\beta^T\beta\,.$$

We will minimize $Q(\beta)$ over the space of β and since we are adding a arbitrary term that depends on the size of β, smaller sizes of β will be preferred. Small sizes of β means X are less important, and this will eventually nullify the singularity in X^TX. The larger λ is, the more "penalization" there is for large values of β.

$$\hat{\beta} = \arg\min_{\beta}\left\{\frac{(y - X\beta)^T(y - X\beta)}{2} + \frac{\lambda}{2}\beta^T\beta\right\}\,.$$

To carry out the minimization, we take the derivative:

$$\begin{aligned}\frac{dQ(\beta)}{d\beta} &= \frac{1}{2}(-2X^Ty + 2X^TX\beta) + \lambda\beta \overset{set}{=} 0\\ &\Rightarrow (X^TX + \lambda I_p)\hat{\beta} - X^Ty = 0\\ &\Rightarrow \hat{\beta}_{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty\,.\end{aligned}$$

(verify second derivative is positive for yourself).

Note that $(X^TX + \lambda I_p)$ is always positive definite for $\lambda > 0$ since for any $a \in \mathbb{R}^p \neq 0$

$$a^T(X^TX + \lambda I_p)a = a^TX^TXa + \lambda a^Ta \geq 0$$

Thus, the final ridge solution always exists even if X^TX is not invertible.

Questions to think about

1. Under the normal likelihood, what is the distribution of $\hat{\beta}_{\text{MLE}}$ (when it exists) and $\hat{\beta}_{\text{ridge}}$? Are they unbiased? Which one has a smaller variance (covariance)?
2. What other penalization functions can you think of? Recall that $\beta^T\beta = \|\beta\|_2^2$.