Maximum Likelihood Estimation (MLE) with Linear Regression Example

General Idea of MLE

Maximum Likelihood Estimation (MLE) is a method to estimate unknown parameters of a statistical model. Suppose we observe data $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ generated from a distribution with density $f(x \mid \theta)$ 'independently', where θ is an unknown parameter (or parameter vector). The likelihood function is defined as

$$L(\theta; \mathcal{D}) = P(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

The MLE is obtained by maximizing the likelihood function (or equivalently its log):

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \; \ell(\theta), \quad \ell(\theta) = \log L(\theta; \mathcal{D}).$$

When the samples are not independent, we directly maximize $\log(P(x_1, \ldots, x_n \mid \theta))$ with respect to θ.

Thus, the general idea behind MLE is to select the parameter values that make the observed data most probable, i.e., to find the parameters that best approximate the true data-generating mechanism under the assumption that the observed data are most likely to arise from it. In other words, we approximate the true data-generating mechanism by assuming that the parameters producing the highest likelihood are the ones under which the observed data had the greatest chance of occurring.

Linear Regression Setup

Consider the linear regression model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n,$$

where $y_i \in \mathbb{R}$ is the response, $\mathbf{x}_i \in \mathbb{R}^p$ is the covariate vector, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the parameter vector.

The density of y_i given x_i is

$$f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2} \right).$$

Likelihood Function

The likelihood of $(\boldsymbol{\beta}, \sigma^2)$ given data $(y_1, \ldots, y_n)$ is

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2} \right).$$

Taking the log,

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2.$$

MLE for $\boldsymbol{\beta}$ and σ^2

Maximizing with respect to $\boldsymbol{\beta}$ (holding σ^2 fixed) is equivalent to minimizing

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2,$$

which is the familiar *ordinary least squares* (OLS) problem. Thus,

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

where $\mathbf{X}$ is the $n \times p$ design matrix and $\mathbf{y} = (y_1, \ldots, y_n)$.

For σ^2, plugging $\hat{\boldsymbol{\beta}}$ into the likelihood and maximizing gives

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2.$$

Key Insight

The MLE framework shows that linear regression with Gaussian errors naturally leads to:

- OLS estimator for $\boldsymbol{\beta}$,
- Variance estimator based on residual sum of squares.

Hence, OLS can be viewed as a special case of MLE under the assumption of normally distributed errors.

Note that $\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{I}_n-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$ and $(\mathbf{I}_n-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T(\mathbf{I}_n-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, Thus,

$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n (y_i - x_i^\top\hat{\boldsymbol{\beta}})^2 = \frac{1}{n}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\mathbf{y}^T(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}.$

Using the results of χ^2 distribution, we know $\mathbf{y}^T(\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y} \sim \chi^2(n-p)$ as the rank of $(\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$ is $n-p$. Thus $\mathbb{E}(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2$. Thus with increasing p, the MLE estimate of σ^2 contains more bias.

But, $\mathbb{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$, thus MLE of $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$.