Maximum Likelihood

Tadao Hoshino （星野匡郎）

Econometrics II: ver. 2019 Spring Semester

Introduction

Introduction

- Consider an ordinary linear regression model

$$Y_i = \mathbf{X}_i^\top \beta_0 + \varepsilon_i \quad (i = 1, ..., n)$$

- We know that the regression coefficients β_0 can be estimated by the OLS (Ordinary Least Squares) method. Namely,

$$\begin{aligned} \widehat{\beta}_n^{ols} &= \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^\top \beta)^2 \\ &= \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right]^{-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i Y_i \end{aligned}$$

Introduction

- Under certain conditions, the OLS estimator is consistent:

$$\widehat{\beta}_n^{ols} \xrightarrow{P} \beta_0,$$

  and is unbiased:

$$E[\widehat{\beta}_n^{ols}] = \beta_0.$$

- Consistency and unbiasedness are requirements for a "good" estimator.

Introduction

- There are alternative good (consistent and unbiased) estimators for β_0 other than OLS :
  - Generalized Least Squares
  - Maximum Likelihood Method
  - Generalized Method of Moments
  - etc
- Which estimator should be employed?

Introduction

- When various estimators are available, we should choose one which has a smaller stochastic error (variance).
- The smaller the variance, the higher the probability of obtaining a value close to the true parameter.
- Suppose $\widehat{\beta}_{n,1}$ and $\widehat{\beta}_{n,2}$ to be two consistent (and unbiased) estimators of β_0. We say the estimator $\widehat{\beta}_{n,1}$ is more efficient than $\widehat{\beta}_{n,2}$ if

$$Var(\widehat{\beta}_{n,1}) < Var(\widehat{\beta}_{n,2}).$$

---

* When β_0 is a vector, we say $\widehat{\beta}_{n,1}$ is more efficient than $\widehat{\beta}_{n,2}$ if the difference of the variance-covariance matrices $VCM(\widehat{\beta}_{n,2}) - VCM(\widehat{\beta}_{n,1})$ is positive definite.

- It is known that, in general, the most efficient (least error variance) estimator is the maximum likelihood (ML) estimator.
  $=>$ Using the ML estimator is preferable when it is available.

Gauss-Markov Theorem

Under regularity conditions (such as, ε is independent and identically distributed as normal), the OLS estimator is the most efficient estimator in the class of linear unbiased estimators.

- Thus, what the GM theorem states is that the OLS estimator estimator coincides with the ML estimator if such regularity conditions are satisfied.
  - Generally, OLS is inferior to ML (i.e., $Var(\text{OLS}) \geq Var(\text{ML})$).

Outline of this lecture

- What is maximum likelihood?: Informal discussion
- Some examples of ML estimation:
  - flipping an uneven coin
  - upper bound of a uniform distribution
  - linear regression model
- Formal definition of the ML estimator

What is maximum likelihood?: Informal discussion

Maximum Likelihood

- Maximum Likelihood Method is a way to find the most likely function to explain a set of observed data.
- We assume that

  The data in hand is the one theoretically most likely to occur.

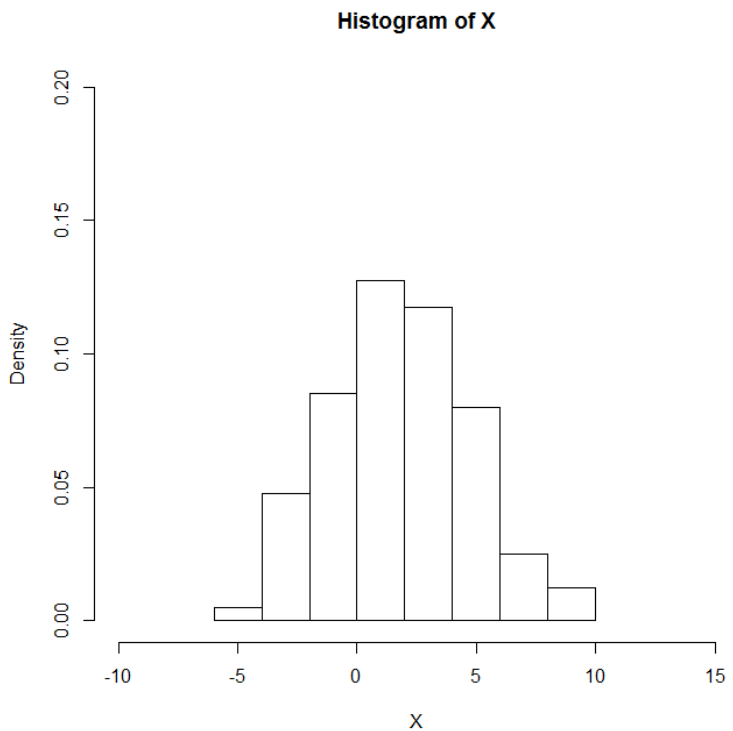
- Here, to what extent it is "likely" (i.e., likelihood) is measured by the probability (or probability density) of the observed data.
- In short, the ML estimation is a method to estimate unknown parameters by maximizing the probability of the data.

OLS = estimate parameters by minimizing the sum of squared errors

ML = estimate parameters by maximizing the probability of the data

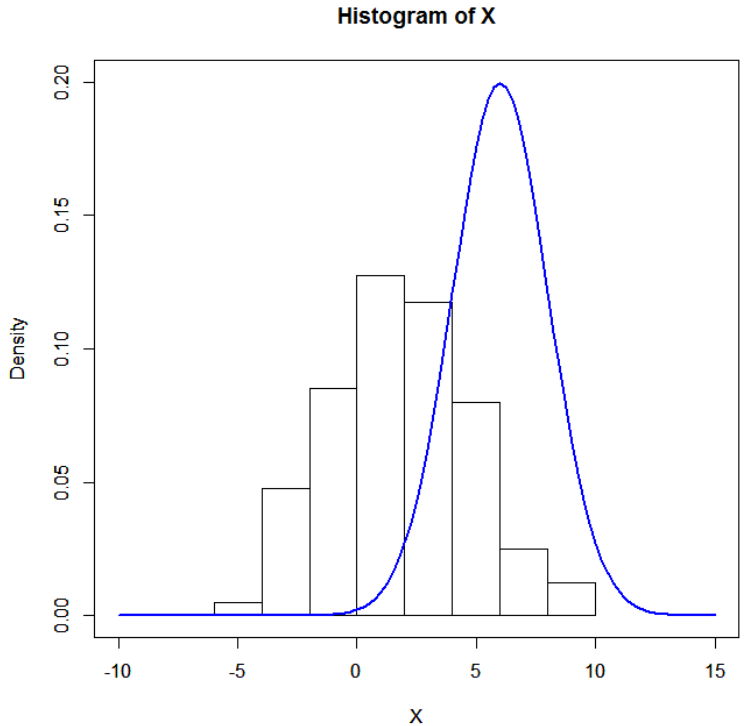
Maximum Likelihood

- For example, suppose that random variable X is distributed as normal; but we do not know its mean $E(X)$ and variance $Var(X)$.
- We have a set of observations $\{X_1, ..., X_n\}$ with sample size n independently drawn from the same distribution as X.
- The histogram of the data is given as follows ($n = 200$):



Maximum Likelihood

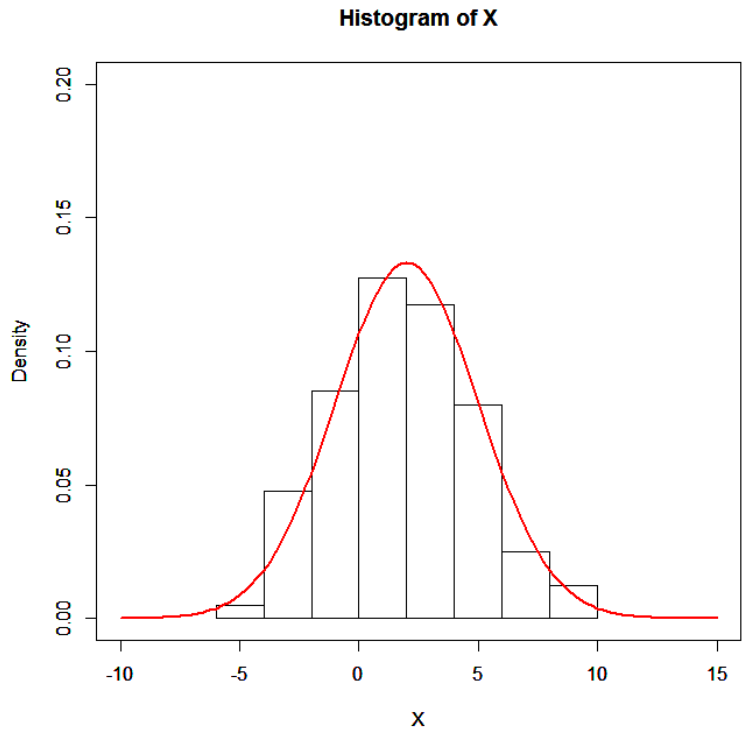
- As a candidate distribution, for example, consider $N(6,2)$ (blue curve):



- Recall that the height of the density function represents the likelihood of realization of the corresponding value.
  - That is, for $N(6,2)$, the values in the neighborhood of 6 are the most likely observable.
  - In other words, it is less likely that our data is generated from $N(6,2)$.

Maximum Likelihood

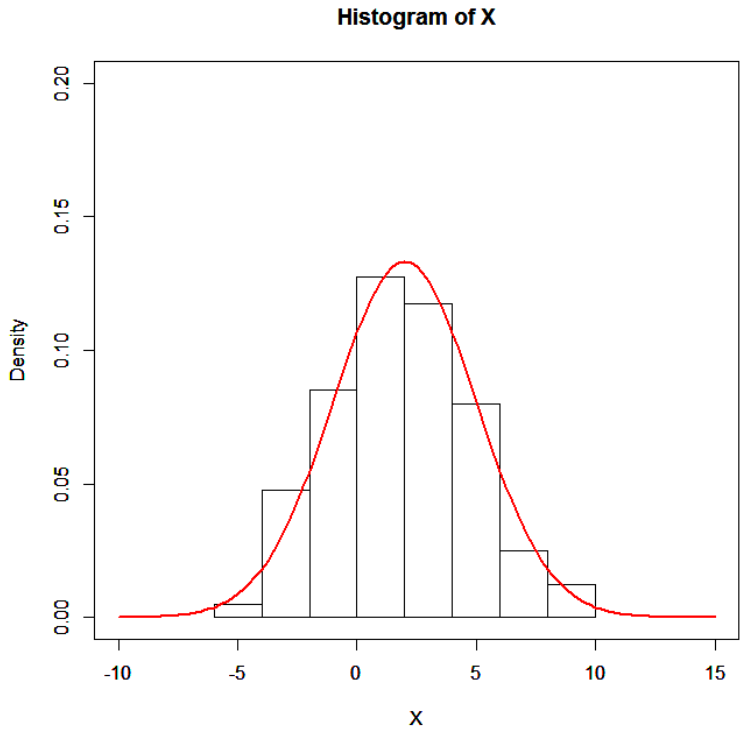
- The distribution that generates the data with the highest likelihood is the one that best fits the histogram:



- The red curve is the normal distribution with its mean and variance being equal to the sample average (1.85) and sample variance (8.55), respectively.

Maximum Likelihood

- The distribution that generates the data with the highest likelihood is the one that best fits the histogram:



- Sample average and sample variance are the maximum likelihood estimator (MLE) of $E(X)$ and $Var(X)$, respectively.
(The true value of $E(X)$ and $Var(X)$ are 2 and 9, respectively.)

Example 1: Flipping an uneven coin

Flipping an uneven coin

- Suppose that we have a possibly uneven coin such that the probability of heads is unknown and may not be equal to 0.5.
- Let X be a dummy variable defined by

$$X = 1 \text{ for "head"},$$
$$X = 0 \text{ for "tail"}.$$

- We use p to denote the coin's true probability of heads: $E(X) = p$.
- We want to estimate p with a set of coin-flipping data $\{X_1, ..., X_n\}$ of n independent trials.
- A natural estimate of p is the sample average, i.e., the sample proportion of heads:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Flipping an uneven coin

- In fact, the sample average $\overline{X}_n$ is the least-squares estimator for p.
- Let

$$\widehat{p}_n^{ls} = \arg\min_p \frac{1}{n}\sum_{i=1}^{n}(X_i - p)^2$$

- The FOC of the minimization problem gives

$$-\frac{2}{n}\sum_{i=1}^{n}(X_i - \widehat{p}_n^{ls}) = 0 \iff \frac{1}{n}\sum_{i=1}^{n} X_i - \widehat{p}_n^{ls} = 0$$
$$\iff \widehat{p}_n^{ls} = \overline{X}_n.$$

Flipping an uneven coin

- As shown below, $\overline{X}_n$ is also an MLE for p.
- The probability distribution of X is

$$\Pr(X = 1) = p, \;\; \Pr(X = 0) = 1 - p$$

  This type of probability distribution is called the Bernoulli Distribution.
- By the independence, the joint probability of $\{X_1, ..., X_n\}$ is given by

$$\Pr(X_1, ..., X_n) = \Pr(X_1) \times \cdots \times \Pr(X_n) = \prod_{i=1}^{n} \Pr(X_i)$$

Flipping an uneven coin

- Note that the probability of realization of each X_i can be written as

$$\Pr(X_i) = p^{X_i}(1-p)^{1-X_i},$$

plugging this into the above yields

$$\Pr(X_1, ..., X_n) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

The function on the right hand side is called the likelihood function.

- Note that the likelihood function is a function of the unknown parameter p.
  - In words, the likelihood function is the joint probability of the data characterized by some unknown parameters.

Flipping an uneven coin

- The maximum likelihood estimator for p is defined as the maximizer of the "log" of the likelihood function (log-likelihood function): that is,

$$\widehat{p}_n^{mle} = \underset{p \in (0,1)}{\operatorname{argmax}} \log \left[\prod_{i=1}^{n} p^{X_i} (1-p)^{1-X_i} \right]$$

- A theoretical reason for maximizing the log-likelihood function, rather than maximizing the likelihood function itself, will be described later.
- An intuitive reason for this is that, without taking the log, the product of the likelihood (probability) vanishes as n increases (since probability is always between 0 and 1).

Flipping an uneven coin

- Observe that

$$\begin{aligned}\log\left[\prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}\right] &= \sum_{i=1}^{n} \log\left[p^{X_i}(1-p)^{1-X_i}\right] \\ &= \sum_{i=1}^{n} \left\{\log[p^{X_i}] + \log[(1-p)^{1-X_i}]\right\} \\ &= \sum_{i=1}^{n} \left\{X_i \log[p] + (1-X_i)\log[1-p]\right\}\end{aligned}$$

- Since $\widehat{p}_n^{mle}$ maximizes the above, we obtain by the FOC that

$$0 = \sum_{i=1}^{n} \left\{\frac{X_i}{\widehat{p}_n^{mle}} - \frac{1-X_i}{1-\widehat{p}_n^{mle}}\right\}.$$

Flipping an uneven coin

- The FOC implies that

$$\sum_{i=1}^{n}\left\{\frac{X_i}{\widehat{p}_n^{mle}}-\frac{1-X_i}{1-\widehat{p}_n^{mle}}\right\}=\sum_{i=1}^{n}\left\{\frac{X_i\cdot(1-\widehat{p}_n^{mle})-(1-X_i)\cdot\widehat{p}_n^{mle}}{\widehat{p}_n^{mle}\cdot(1-\widehat{p}_n^{mle})}\right\}$$
$$=\sum_{i=1}^{n}\left\{\frac{X_i-\widehat{p}_n^{mle}}{\widehat{p}_n^{mle}\cdot(1-\widehat{p}_n^{mle})}\right\}=0$$

- By definition, $\widehat{p}_n^{mle}$ can be neither 0 nor 1.
- Therefore, the maximum likelihood estimator $\widehat{p}_n^{mle}$ must satisfy

$$\sum_{i=1}^{n}(X_i-\widehat{p}_n^{mle})=0 \iff \sum_{i=1}^{n}X_i=n\widehat{p}_n^{mle}$$
$$\iff \widehat{p}_n^{mle}=\overline{X}_n$$

- Thus, in this case, the sample average is the MLE.

Flipping an uneven coin

- We can verify the above result by simulation in **R**.
- Set the true value of p to, for example, 0.7:

```
p0 <- 0.7
```

  You can change it to any number within 0 and 1.
- Draw 1000 random numbers from Uniform$[0,1]$, and judge whether each element is less than `p0` or not:

```
X <- (runif(1000) < p0)
```

Flipping an uneven coin

```
>  p0 <- 0.7
>  X  <- (runif(1000) < p0)
>  head(X)
[1] FALSE FALSE  TRUE  TRUE FALSE FALSE
>  mean(X)
[1] 0.724
```

In **R**, logical TRUE is evaluated as 1, and FALSE is treated as 0.

```
> TRUE + TRUE; TRUE*TRUE; TRUE*FALSE
[1] 2
[1] 1
[1] 0
```

Flipping an uneven coin

- Create the log-likelihood function:

```
LL <- function(p){
    L <- X*log(p) + (1 - X)*log(1 - p)
    return(sum(L))
    }
```

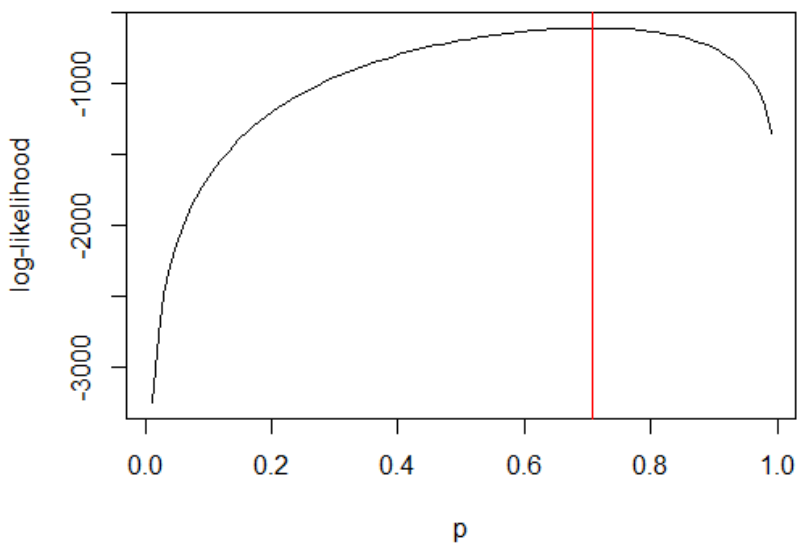
- Create candidate values for p, and evaluate the log-likelihood at each candidate:

```
ps  <- (1:99)/100
LLs <- mapply(LL, ps)
```

The function `mapply(FUN, v)` returns a list of results by applying `FUN` to the elements of the vector `v`. (You can also use for-loop here.)

Flipping an uneven coin

```
plot(ps, LLs, type = "l", xlab = "p", ylab = "log-likelihood")
abline(v = mean(X), col = "red")
```



As the theory suggests, the log-likelihood attains the highest value at the sample average.

Example 2: The Simple Linear Regression

The Simple Linear Regression

- Next, we consider the following simple linear regression model with normally distributed error term:

$$Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i \quad (i = 1, ..., n)$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

- The second line of the above is equivalent to

$$Y_i - \beta_0 - X_i\beta_1 \sim N(0, \sigma^2).$$

- Thus, the density of Y_i conditional on X_i is given by

$$Y_i | X_i \sim N(\beta_0 + X_i\beta_1, \sigma^2).$$

The Simple Linear Regression

- Denote the marginal density of X_i as $f_X(\cdot)$, and let $\phi(\cdot|m, v)$ be the normal density function with mean m and variance v.
- Then, the joint density of (Y_i, X_i) can be written as

$$f_{Y,X}(Y_i, X_i) = \phi(Y_i|\beta_0 + X_i\beta_1, \sigma^2) \cdot f_X(X_i).$$

- Assuming that the data is independently drawn from the population, the likelihood function for our data (joint density of the data) is

$$\text{Joint density of } \{(Y_i, X_i) : 1 \leq i \leq n\} = \prod_{i=1}^{n} \phi(Y_i|\beta_0 + X_i\beta_1, \sigma^2) \cdot f_X(X_i).$$

- Thus, the log-likelihood function is

$$\begin{aligned}
&\log\left[\prod_{i=1}^{n}\phi(Y_i|\beta_0+X_i\beta_1,\sigma^2)\cdot f_X(X_i)\right]\\
&=\sum_{i=1}^{n}\log\left[\phi(Y_i|\beta_0+X_i\beta_1,\sigma^2)\cdot f_X(X_i)\right]\\
&=\sum_{i=1}^{n}\log\phi(Y_i|\beta_0+X_i\beta_1,\sigma^2)+\sum_{i=1}^{n}\log f_X(X_i).
\end{aligned}$$

- Since the second term on the rhs is irrelevant to $(\beta_0,\beta_1,\sigma^2)$, the MLE for $(\beta_0,\beta_1,\sigma^2)$ is defined as the maximizer of the first term:

$$(\widehat{\beta}_{n0}^{mle},\widehat{\beta}_{n1}^{mle},\widehat{\sigma}_n^{2,mle})=\underset{b_0,b_1,s^2}{\operatorname{argmax}}\sum_{i=1}^{n}\log\phi(Y_i|b_0+X_ib_1,s^2).$$

The Simple Linear Regression

- According to the Gauss-Markov theorem, under the normal error assumption, the OLS estimator of (β_0, β_1) is equivalent to the MLE $(\widehat{\beta}_{n0}^{mle}, \widehat{\beta}_{n1}^{mle})$.
- We can verify this by simulation in **R**.
- Set the true value of $(\beta_0, \beta_1, \sigma^2)$ as follows:

```
beta0 <- 1
beta1 <- 1
s2    <- 1
```

The Simple Linear Regression

- The sample size is set to 1000, and draw X and ε from Uniform$[-1, 1]$ and $N(0, s^2)$, respectively:

```
X <- runif(1000, -1, 1)
e <- rnorm(1000, mean = 0, sd = sqrt(s2))
```

Then, create the dependent variable Y

```
Y <- beta0 + X*beta1 + e
```

The Simple Linear Regression

```
> beta0 <- 1
> beta1 <- 1
> s2    <- 1
>
> X <- runif(1000, -1, 1)
> e <- rnorm(1000, mean = 0, sd = sqrt(s2))
>
> Y <- beta0 + X*beta1 + e
```

The Simple Linear Regression

- We first estimate the model by OLS:

```
OLS <- lm(Y ~ X)
```

- The OLS estimate of β_0 and that of β_1 are stored in `OLS$coef`.

```
> OLS <- lm(Y ~ X)
> OLS$coef
(Intercept)           X
  0.9560423   0.9604689
```

The Simple Linear Regression

- We next estimate the same model by maximum likelihood. Create the log-likelihood function as follows:

```
LL <- function(param){
    Ymean <- param[1] + X*param[2]
    Ysd   <- param[3]
    L <- log(dnorm(Y, mean = Ymean, sd = Ysd))
    return(sum(L))
    }
```

- Since `LL` is a function of a vector `param` with 3 elements, it is difficult to visually find a value at which the function has its maximum.

The Simple Linear Regression

- To find the maximizer of the function `LL` with respect to `param`, we can use the `optim()` command.
- The basic syntax of `optim()` is as follows:

$$\texttt{optim(①, ②, } \underbrace{\texttt{control = list(fnscale = -1)}}_{③}\texttt{)}$$

  - ①: Initial candidate value for `param`[1]
  - ②: Objective function to be maximized
  - ③: Since the `optim()` function solves minimization problems by default, we need this option for maximization problems.

[1] By choosing a good (i.e., close to the solution) starting value, you can reduce the computation time.

```
> ML <- optim(c(1,1,1), LL, control=list(fnscale=-1))
> print(ML)
$`par`
[1] 0.9560066 0.9605061 1.0163099
```

- The first two elements of `$par` are the estimated β_0 and β_1, and the third element is the estimate of s. (Recall that the true values of these parameters are all one.)
- The OLS estimates of β_0 and β_1 were about 0.9560 and 0.9605, respectively, which are almost the same as the ML estimates.
  $=$ Gauss-Markov Theorem.

Formal definition of the MLE

The MLE

- Suppose we have a set of observations $\{X_1, ..., X_n\}$ of sample size n, identically independently drawn from the distribution of X. (X may be a scalar or a vector.)
- For continuous (resp. discrete) X, let the population density (resp. probability) function of X be $f(x; \theta_0)$, where θ_0 is a vector of unknown parameters to be estimated.
- The functional form of $f(x; \theta_0)$ must be known up to θ_0; that is, θ_0 is the only unknown component.[2]

[2] If the specification on f is incorrect, then the resulting ML estimator of θ_0 is not consistent. Note that you can use a least-squares approach even when f is totally unknown. In this sense, the maximum likelihood method is more restrictive than the other estimators.

The MLE

Likelihood Function

The likelihood function is defined as

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

Log-likelihood Function

The log-likelihood function is defined as

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

Maximum Likelihood Estimator (MLE)

The maximum likelihood estimator for θ_0 is defined as

$$\widehat{\theta}_n^{mle} = \underset{\theta}{\operatorname{argmax}}\, \ell_n(\theta),$$

or equivalently,

$$\widehat{\theta}_n^{mle} = \underset{\theta}{\operatorname{argmax}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta)}_{=n^{-1}\ell_n(\theta)}.$$

MLE as the minimizer of the KL divergence

Kullback-Leibler (KL) divergence

Let $f(x;\theta_1)$ and $f(x;\theta_2)$ be two density functions. The following quantity is called the Kullback-Leibler divergence between $f(x;\theta_1)$ and $f(x;\theta_2)$:

$$K\left[f(X;\theta_1)\|f(X;\theta_2)\right] = E_{\theta_1}\left[\log\frac{f(X;\theta_1)}{f(X;\theta_2)}\right]$$
$$= \int\left[\log\frac{f(x;\theta_1)}{f(x;\theta_2)}\right] f(x;\theta_1)dx.$$

MLE as the minimizer of the KL divergence

- It is straightforward to see that

$$K\left[f(X;\theta_1)\|f(X;\theta_2)\right] = 0$$

  if $f(x;\theta_1) = f(x;\theta_2)$ (because $\log 1 = 0$).
- Moreover, it holds that for any $f(x;\theta_1) \neq f(x;\theta_2)$,

$$K\left[f(X;\theta_1)\|f(X;\theta_2)\right] > 0$$

  (the proof is omitted).
- Thus, the KL divergence can be interpreted as a measure of the difference between two density functions.

MLE as the minimizer of the KL divergence

- This implies that the true parameter θ_0 can be characterized by

$$\theta_0 = \underset{\theta}{\operatorname{argmin}} K\left[f(X;\theta_0) \| f(X;\theta)\right].$$

- Note that

$$\begin{aligned} K\left[f(X;\theta_0) \| f(X;\theta)\right] &= E\left[\log \frac{f(X;\theta_0)}{f(X;\theta)}\right] \\ &= E[\log f(X;\theta_0)] - E[\log f(X;\theta)]. \end{aligned}$$

- Thus, since the first term on the rhs is independent of θ,

$$\theta_0 = \underset{\theta}{\operatorname{argmax}} E[\log f(X;\theta)],$$

implying that the true parameter θ_0 is the maximizer of the population log-likelihood function.

MLE as the minimizer of the KL divergence

- By the law of large numbers, we can expect that

$$\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta) \xrightarrow{P} E[\log f(X;\theta)].$$

- $\widehat{\theta}_n^{mle} = \text{argmax}_\theta \frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta)$, $\theta_0 = \text{argmax}_\theta E[\log f(X;\theta)]$.
- Therefore, it is also expected that $\widehat{\theta}_n^{mle} \xrightarrow{P} \theta_0$.