Econometrics **August 18, 2026**

Maximum Likelihood Estimation

Instructor: Kalyan Kolukuluri *Content: Lecture Notes*

Contents

1 Probability vs Likelihood

Probability and **Likelihood** are related but distinct concepts in statistics. Probability refers to the chance of observing a particular outcome given a known model and its parameters; it is forward-looking and answers questions like,

> **Probability:** What is the probability of getting heads in a fair coin toss?

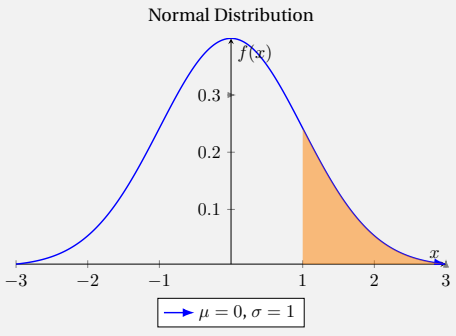
In contrast, likelihood measures how well a given set of parameters explains the observed data; it is backward-looking and answers questions like,

> **Likelihood:** Given that we observed a sequence of coin flips, how likely is it that the coin is fair?

Probability is used in predictive scenarios, while likelihood is fundamental to inference, particularly in Maximum Likelihood Estimation (MLE), where we estimate parameters by maximizing the likelihood of observed data.

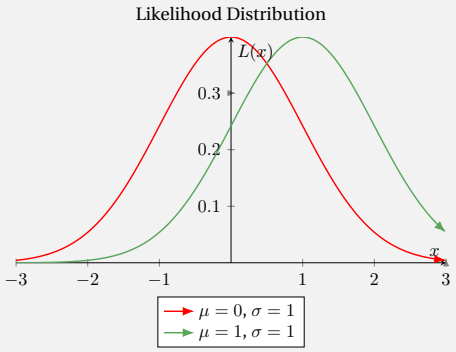
1.1 Probability

- **Probabilities:** Areas under a fixed distribution.
- $P(\text{data}|\text{distribution}) = P(\text{weight} > 1kg|\mu = 0kg; \sigma = 1)$



1.2 Likelihood

- **Likelihood:** Y-axis values with fixed data points and distributions that can be moved.
- $L(\text{distribution}|\text{data}) = L(\mu = 0kg; \sigma = 1|\text{data points})$
- $L(u, \sigma|X_1, X_2, X_3, \ldots, X_N)$
- In likelihood, the measured values (data points) are fixed and the distribution (red and green) is shifted. So, as to encompass as many data points within it as possible.



1.3 Maximum Likelihood (MLE)

Fitting a Distribution onto Observed Data

1. Step 1: Assume a mean (μ) for the distribution and some fixed std. dev (σ). Calculate the Likelihood (L) for all data points.
2. Step 2: If L is low. We iterate the placement of mean until 'L' improved without reversal (maximum likelihood).

1.4 Prob vs Likelihood

For a normal distribution data;

$$\text{Prob. density function (pdf)} \quad \text{Prob}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\text{Likelihood value function (L)} \quad L(u, \sigma \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

2 Mechanics of MLE

If we assume data points to be independent and identically distributed (i.i.d);

$$L(u,\sigma|x_1,x_2,x_3,\ldots,x_n) = L(u,\sigma|x_1) \times L(u,\sigma|x_2) \times \ldots L(u,\sigma|x_n)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/2\sigma^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_2-\mu)^2/2\sigma^2} \times \ldots\ldots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}$$

Taking Logarithm and solving, we get:

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots\ldots - \frac{(x_n-\mu)^2}{2\sigma^2}$$

2.1 Solving for parameters using MLE

$$\boxed{\ln(L) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots\ldots - \frac{(x_n-\mu)^2}{2\sigma^2}}$$

Solving for first-order conditions:

- $\frac{\partial ln(L)}{\partial \mu} = 0 \implies \mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$
- $\frac{\partial ln(L)}{\partial \sigma} = 0 \implies \sigma = \sqrt{\left(\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{n}\right)}$

Caution about MLE variance/S.D estimate

Note that the Maximum likelihood based estimate of S.D $\sigma_{\text{MLE}} = \sqrt{\frac{\Sigma(x_i-\mu)^2}{n}}$ will under-estimate the true variance of the population.

3 ML-Estimation for Linear Regression

Basic model for simple linear regression.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Assume ε to be i.i.d., and following normal distribution $N\left(0, \sigma^2\right)$.

$$\Pr\left(\{Y_i\}_{i=1}^N \mid \{X_i\}_{i=1}^N, \beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma}\right)^2}$$

Take a logarithm on both sides of the equations and calculate the partial derivative to get the following estimates.

$$\begin{aligned}
\hat{\beta}_1^{\text{MLE}} &= \frac{\sum_{i=1}^N \left(X_i - \overline{X}\right) Y_i}{\sum_{i=1}^N \left(X_i - \overline{X}\right)^2} \\
\hat{\beta}_0^{\text{MLE}} &= \overline{Y} - \hat{\beta}_1 \overline{X} \\
\hat{\sigma}^2_{\text{MLE}} &= \frac{\sum_{i=1}^N \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{N}
\end{aligned}$$

Note the estimate of residual variance $\hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^N \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{N}$, if we contrast it with estimate of $\hat{\sigma}^2_{\text{OLS}} = \frac{\sum_{i=1}^N \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{N - K}$, one can see that *MLE underestimates residual variance.*

However, OLS and MLE will generate the same t-statistic, by making the necessary adjustment.

In $Y_i = \beta_0 + \beta_1 X_i + \varepsilon$ with independent errors $\varepsilon_i \sim N\left(0, \sigma^2\right)$, *the MLE of* β_0 *and* β_1 *have a bi-variate normal distribution* with

$$\mathrm{E}\left(\hat{\beta}_0^{\text{MLE}}\right) = \beta_0$$

$$\mathrm{E}\left(\hat{\beta}_1^{\text{MLE}}\right) = \beta_1$$

$$\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^N \left(X_i - \overline{X}\right)^2}$$

$$\text{Var}\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{N} X_i^2}{N \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}$$

$$\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\overline{X}\sigma^2}{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}$$

4 MLE Estimation in R

```r
```{r mle-ab-initio}
data(cars)  # Load the dataset
head(cars)  # View first few rows

# Define the generic negative log-likelihood function for a linear model
neg_log_lik <- function(params, x, y) {
  beta0 <- params[1]  # Intercept
  beta1 <- params[2]  # Slope
  sigma <- params[3]  # Standard deviation

  Y_pred <- beta0 + beta1 * x  # generic Linear model
  log_lik <- dnorm(y, mean = Y_pred, sd = sigma, log = TRUE)  # Normal likelihood

  return(-sum(log_lik))  # Negative log-likelihood (to minimize)
}

# Initial parameter guesses: Intercept, Slope, and Standard Deviation
init_params <- c(0, 1, sd(cars$dist))

# Optimize the negative log-likelihood function
mle_result <- optim(init_params, neg_log_lik,
                    x = cars$speed, y = cars$dist,
                    method = "BFGS")

# Print estimated parameters
mle_result$par


#### OLS Estimates

# Fit using Ordinary Least Squares (OLS)
ols_model <- lm(dist ~ speed, data = cars)

# Compare results
summary(ols_model)
names(cars)
```
```

Appendix :A MLE Estimation of Linear Regression Parameters

$$Y = X^T\beta + \varepsilon$$

Assume ε to follow the normal distribution of $N(0,\sigma^2)$. Thus,

$$Y \backsim X^T\beta \sim N(0,\sigma^2)$$

$$\Pr\left(Y_i \mid X_i,\beta,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - X_i^T\beta}{\sigma^2}\right)}$$

We assume ε to be i.i.d., and thus we can write total likelihood as follows.

$$\Pr\left(\{Y_i\}_{i=1}^N \mid \{X_i\}_{i=1}^N,\beta,\sigma^2\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - X_i^T\beta}{\sigma}\right)^2}$$

In matrix notation, the above can be written as:

$$\begin{aligned} Pr(\{Y_i\}_{i=1}^N|\{X_i\}_{i=1}^N,\beta,\sigma^2) &= N(Y|X\beta,\sigma^2 I) \\ &= (2\sigma^2\pi)^{-n/2} e^{(-\frac{1}{2\sigma^2}(X\beta-Y)^T(X\beta-Y))} \end{aligned}$$

Taking logarithm and writing we get:

$$\log\Pr(\{Y_i\}_{i=1}^N|\{X_i\}_{i=1}^N,\beta,\sigma^2) = -\frac{N}{2}log(2\sigma^2\pi) - \frac{1}{2\sigma^2}(X\beta-Y)^T(X\beta-Y))$$

Now, the idea of maximum likelihood implies

$$\beta^{MLE} \propto \arg\text{MAX}_\beta\{-\frac{1}{2\sigma^2}(X\beta-Y)^T(X\beta-Y)\} \implies \arg\text{MIN}_\beta\{(X\beta-Y)^T(X\beta-Y)\}$$

Using MLE to estimate coefficient $\hat{\beta}$ Here, we assume σ^2 to be what it is, for now, we do not try to find the optimal value for it:

$$\beta^{MLE} \propto \frac{\partial}{\partial\beta}(X\beta-Y)^T(X\beta-Y) = 0$$

$$
\begin{aligned}
-(X\beta - Y)^T X \breve{} (X\beta \breve{} Y)^T X &= 0 \\
2(X\beta - Y)^T X &= 0 \\
(X\beta - Y)^T X &= 0 \\
(\beta^T X^T - Y^T)X &= 0 \\
\beta^T X^T X &= Y^T X \\
\beta^T &= Y^T X (X^T X)^{-1} \\
\beta &= (X^T X)^{-1} X^T Y \\
\hat{\beta} &= (X^T X)^{-1} X^T Y
\end{aligned}
$$

Using MLE to estimate Variance σ^2 Here we estimate the variance-covariance-matrix is a diagonal matrix (i.e., covariance is zero, given the i.i.d. assumption)

$$\sigma^{MLE} = \arg \text{MAX}_\sigma \{-\frac{N}{2} log(2\sigma^2 \pi) - \frac{1}{2\sigma^2}(X\beta - Y)^T (X\beta - Y)\}$$

We can replace β with $\hat{\beta}$ to get

$$\sigma^{MLE} = \arg \text{MAX}_\sigma \{-\frac{N}{2} log(2\sigma^2 \pi) - \frac{1}{2\sigma^2}(X\hat{\beta} - Y)^T (X\hat{\beta} - Y)\}$$

$$
\begin{aligned}
\frac{\partial}{\partial \sigma^2}\{-\frac{N}{2} log(2\sigma^2 \pi) - \frac{1}{2\sigma^2}(X\hat{\beta} - Y)^T (X\hat{\beta} - Y)\} &= 0 \\
\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(X\hat{\beta} - Y)^T (X\hat{\beta} - Y) &= 0 \\
\sigma^2 &= \frac{1}{N}(X\hat{\beta} - Y)^T (X\hat{\beta} - Y) \\
\hat{\sigma}^2 &= \frac{1}{N}(X\hat{\beta} - Y)^T (X\hat{\beta} - Y)
\end{aligned}
$$

Note: The estimated variance in MLE σ^{MLE} is basically the variance of the residual in the linear regression model. Unlike, an OLS estimate σ^{OLS}, which adjusts to degrees of freedom (N-K).