# Maximum Likelihood Estimation
# (Statistical Modelling I)

Dr Lubna Shaheen

Week 6, Lecture 2

## Outline

1. What is a Maximum Likelihood Estimation
   - Likelihood function
   - Maximum Likelihood function
2. MLE of Binomial Distribution
3. Key points of Maximum Likelihood function
4. MLE of Normal distribution
5. MLE of Normal Simple Linear Regression parameters
6. Exams Style Questions

Queen Mary University of London

# Estimating Parameters

So far in this module we have used **Least Squares estimation** to estimate model parameters $\beta_0$ and $\beta_1$.

You can check back to your week 1 notes for how we did this

1. in summary we minimised the sum of squares of errors
2. this involved differentiation and simultaneous equations

There are other methods for estimating parameters. We will now consider one called **Maximum Likelihood Estimation**.

# What is a MLE?

The maximum likelihood estimator $\hat{\theta}$ for a parameter θ, is the estimate which maximises the probability of obtaining the sample we have actually observed

The maximum likelihood estimator is the parameter estimate that maximises the "likelihood function" which is the joint probability function [discrete distribution] or joint pdf [continuous] of the observed sample

Queen Mary
University of London

# Likelihood function

**Definition**: Let $Y_1, Y_2, Y_3, \cdots, Y_n$ be a random sample from a distribution with a parameter $\theta$. In general, $\theta$ can be a vector,$\theta = (\theta_1, \theta_2, \cdots, \theta_k)$.

Suppose that $y_1, y_2, \cdots, y_n$ are the observed values of $Y_1, Y_2, \cdots, Y_n$. If $Y_i$ 's are discrete random variables, we define the likelihood function as the probability of the observed sample as a function of $\theta$

$$L(y_1, y_2, \cdots, y_n; \theta) = P(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n; \theta) = P_{Y_1 Y_2 \cdots Y_N}(y_1, y_2, \cdots, y_n; \theta).$$

If $Y_i$'s are jointly continuous, then the likelihood function is defined as

$$L(y_1, \cdots, y_n; \theta) =$$

In most of the cases, its easier to work with the log Likelihood function given by

$$\log L(y_1, y_2, \cdots, y_n; \theta)$$

# Likelihood function

**Example**:

1. If $X_i \sim Binomial(3, \theta)$, then

$$P_{X_i}(x;\theta) = \binom{3}{x}\theta^x(1-\theta)^{3-x}$$

Thus,

$$\begin{aligned} L(x_1, x_2, x_3, x_4; \theta) &= P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4; \theta) \\ &= P_{X_1}(x_1;\theta)P_{X_2}(x_2;\theta)P_{X_3}(x_3;\theta)P_{X_4}(x_4;\theta) \\ &= \binom{3}{x_1}\binom{3}{x_2}\binom{3}{x_3}\binom{3}{x_4}\theta^{x_1+x_2+x_3+x_4}(1-\theta)^{12-(x_1+x_2+x_3+x_4)}. \end{aligned}$$

Since we have observed $(x_1, x_2, x_3, x_4) = (1, 3, 2, 2)$, we have

$$\begin{aligned} L(1,3,2,2;\theta) &= \binom{3}{1}\binom{3}{3}\binom{3}{2}\binom{3}{2}\theta^8(1-\theta)^4 \\ &= 27 \qquad \theta^8(1-\theta)^4. \end{aligned}$$

Queen Mary University of London

# Likelihood Estimator

**Example**:

2. If $X_i \sim Exponential(\theta)$, then

$$f_{X_i}(x;\theta) = \theta e^{-\theta x} u(x),$$

where $u(x)$ is the unit step function, i.e., $u(x) = 1$ for $x \geq 0$ and $u(x) = 0$ for $x < 0$. Thus, for $x_i \geq 0$, we can write

$$\begin{aligned} L(x_1, x_2, x_3, x_4; \theta) &= f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4; \theta) \\ &= f_{X_1}(x_1;\theta) f_{X_2}(x_2;\theta) f_{X_3}(x_3;\theta) f_{X_4}(x_4;\theta) \\ &= \theta^4 e^{-(x_1+x_2+x_3+x_4)\theta}. \end{aligned}$$

Since we have observed $(x_1, x_2, x_3, x_4) = (1.23, 3.32, 1.98, 2.12)$, we have

$$L(1.23, 3.32, 1.98, 2.12; \theta) = \theta^4 e^{-8.65\theta}.$$

# Likelihood function

More generally for probability distributions we maximise the joint probability of our observations by maximising the **Likelihood function** $L(\theta, y)$

$$L(\theta, y) = \prod_{i=1}^{n} f(y_i|\theta)$$

for discrete observations this becomes

$$L(\theta, y) = \prod_{i=1}^{n} Pr(Y_i = y_i|\theta)$$

The maximum likelihood estimator $\hat{\theta}$ is the value of $\theta$ which maximises the Likelihood function

# Maximising the Likelihood function

## Process of Maximising the Likelihood function

We differentiate the likelihood function with respect to the parameter(s) and set to zero, solving to find the maximum

- Last time in Least Squares we solved for a minimum

For most probability distributions it is much easier to take the log of the likelihood function and differentiate that

- Because the likelihood is the product of probability terms
- log L($\theta, y$) will be maximised at the same $\theta$ as L($\theta, y$)

# Maximising the Likelihood function

**Exponential Example**

For the following observations, find the maximum likelihood estimator (MLE) of $\theta$.

$$X_i \sim \text{Exponential}(m, \theta) \text{ and we have observed } X_1, X_2, X_3, \cdots, X_n$$

$$L(x_1, x_2, \cdots, x_n; \theta) =$$

# Maximising the Likelihood function

## Binomial Example

$n$ binomial trials where $y_i = 1$ if the $i^{\text{th}}$ trial is a success and $y_i = 0$ otherwise

Let the probability of a success be $p$

- $p$ is unknown
- We seek to estimate $p$ by MLE finding $\hat{p}$

Let $y = \sum_{i=1}^{n} y_i$ the total number of successful trials

We first need to find the likelihood function which is the joint probability function for the $n$ trials

- This is a function of $p$

Queen Mary
University of London

# Maximising the Likelihood function

## Binomial Example

$L(p) = L(y_1 \dots y_n|p) = p^y(1-p)^{n-y}$

As $L(p)$ is a product of functions, it will be easier to differentiate *log*

$$logL(p) = \log(p^y(1-p)^{n-y})$$

$$= y\log(p) + (n-y)\log(1-p)$$

# Maximising the Likelihood function

## Binomial MLE

$$\frac{dlogL(p)}{dp} = y\frac{1}{p} + (n-y)\frac{-1}{1-p}$$

Set to zero and solve for $\hat{p}$

$$y\frac{1}{\hat{p}} + (n-y)\frac{-1}{1-\hat{p}} = 0$$

$$\frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} = 0$$

$$\hat{p} = \frac{y}{n}$$

# Maximising the Likelihood function

## Binomial MLE

$$\hat{p} = \frac{y}{n}$$

So the Binomial MLE is the proportion of successful trials observed

which is a natural estimate

# Key properties of MLEs

The Binomial example highlights the key properties of maximum likelihood estimators

- and hence their advantages / disadvantages

With the Binomial MLE $\hat{p} = \frac{y}{n}$ we would expect the quality of the estimate to improve as $n$ increases

We say that the estimator has strong ***asymptotic*** properties

That is as $n \to \infty$

# Maximising the Likelihood function

## Advantages / Disadvantages of MLEs

| + | − |
|---|---|
| Asymptotically unbiased | At small n may be biased |
| Normally distributed | Asymptotic properties may not apply at all n |
| Smallest possible variance | |

# Maximising the Likelihood function

## MLE in the Normal distribution

We need this to use MLE in the simple linear regression model

For a normal distribution with mean $\mu$ and variance $\sigma^2$ we estimate $\mu$ by MLE

Start with the normal pdf

$$f(y|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$$

Then the Likelihood function is the joint pdf for our *n* observations

# Maximising the Likelihood function

## Normal likelihood function

$$f(y|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$$

Remember we seek a MLE of $\mu$

$$L(\mu, y) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum (y - \mu)^2\right)$$

And taking logs

$$logL = -\log\left(\sigma^n (2\pi)^{\frac{n}{2}}\right) - \frac{1}{2\sigma^2}\sum (y - \mu)^2$$

# Maximising the Likelihood function

## Finding the maximum

$$logL = -\log\left(\sigma^n (2\pi)^{\frac{n}{2}}\right) - \frac{1}{2\sigma^2}\sum (y - \mu)^2$$

Differentiating with respect to the parameter

$$\frac{dlogL}{d\mu} = \frac{1}{\sigma^2}\sum (y - \mu)$$

Which equals zero when $\hat{\mu} = \bar{y}$

So the MLE of the normal mean is the sample mean

# Maximising the Likelihood function

# Maximising the Likelihood function

## Normal simple linear regression model

In the simple linear regression model instead of $Y_i \sim N(\mu, \sigma^2)$

we now have $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

we seek to estimate $\beta_0$ and $\beta_1$ by MLE

the Likelihood function is the same normal one but with $\mu$ replaced by our model $\beta_0 + \beta_1 x_i$

# Simple Linear Regression Likelihood

The likelihood is now a function of our two model parameters

$$L(\beta_0, \beta_1, y_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \Sigma \; (y_i - \; \beta_0 + \beta_1 x_i)^2\right)$$

We could solve this in the usual MLE way

- take logs
- Differentiate log L with respect to $\beta_0$ and $\beta_1$
- set to zero and solve the two simultaneous equations

Queen Mary University of London

## But we don't have to ☺

$$L(\beta_0, \beta_1, y_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \Sigma\ (y_i - \ \beta_0 + \beta_1 x_i\ )^2\right)$$

is maximised wherever

$$-\ \Sigma\ (y_i - \ \beta_0 + \beta_1 x_i)^2$$

is maximised (because *n* and $\sigma$ are fixed here)

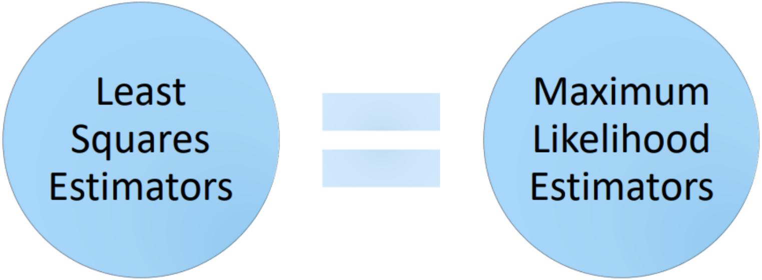
- We already know where this is from Least Squares estimation

# Maximising the Likelihood function

# Maximising the Likelihood function

# Maximising the Likelihood function

## Simple Linear Regression Model




This is not usual in model parameter estimation, we generally have to select one of the methods

# Maximising the Likelihood function

## Exams Style Questions

**Question (2022)**

Let $X_1, X_2, \cdots, X_N$ be random variables from a normal distribution with unknown mean $\mu$ and unknown variance $\sigma^2$. We are interested in finding the maximum likelihood estimates of $\mu$ and $\sigma^2$. Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the maximum likelihood estimates for $\mu$ and $\sigma^2$. The probability density function of $x_i$ is given by

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(x_i-\mu)^2}$$

for $-\infty < \mu < \infty, 0 < \sigma^2 < \infty$ and $i = 1, 2, \cdots n$.

Prove that

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } \widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}$$

# Maximising the Likelihood function

**Solution:**