Linear and Generalized Linear Models (4433LGLM6Y)

Maximum Likelihood Estimation

Meeting 9

Vahe Avagyan

Biometris, Wageningen University and Research

WAGENINGEN UNIVERSITY & RESEARCH

Maximum likelihood (Fox appendix D6.1-D6.4)

- likelihood function, log-likelihood
- maximum likelihood estimator (MLE) and properties
- asymptotic variance of MLE; Fisher information
- likelihood ratio test (LRT)
- inference for single parameter: Wald-test, likelihood ratio test, score test
- inference for several parameters, Fisher information matrix, asymptotic variance-covariance matrix

Maximum-Likelihood Estimation

- Most general estimation principle in statistics.
- Advantages:
  - Broadly applicable, relatively simple to apply.
  - Provides estimators with reasonable intuitive basis
  - Has desirable statistical properties.
  - Theory of MLE provides SE’s, statistical tests, and other results useful for inference.
- Disadvantage: frequently requires strong assumptions about structure of data.

Example: flipping a coin

- Coin is flipped 10 times ($n = 10$) with a probability of getting head π
- Results: H – H – T – H – H – H – T – T – H – H
- Probability function:

$$\Pr(\textbf{data} \mid \textbf{parameter}) = \Pr(\text{HHTHHHTTHH} \mid \pi) =$$

$$= \pi\pi(1-\pi)\pi\pi\pi(1-\pi)(1-\pi)\pi\pi = \pi^7(1-\pi)^3$$

- Likelihood function:

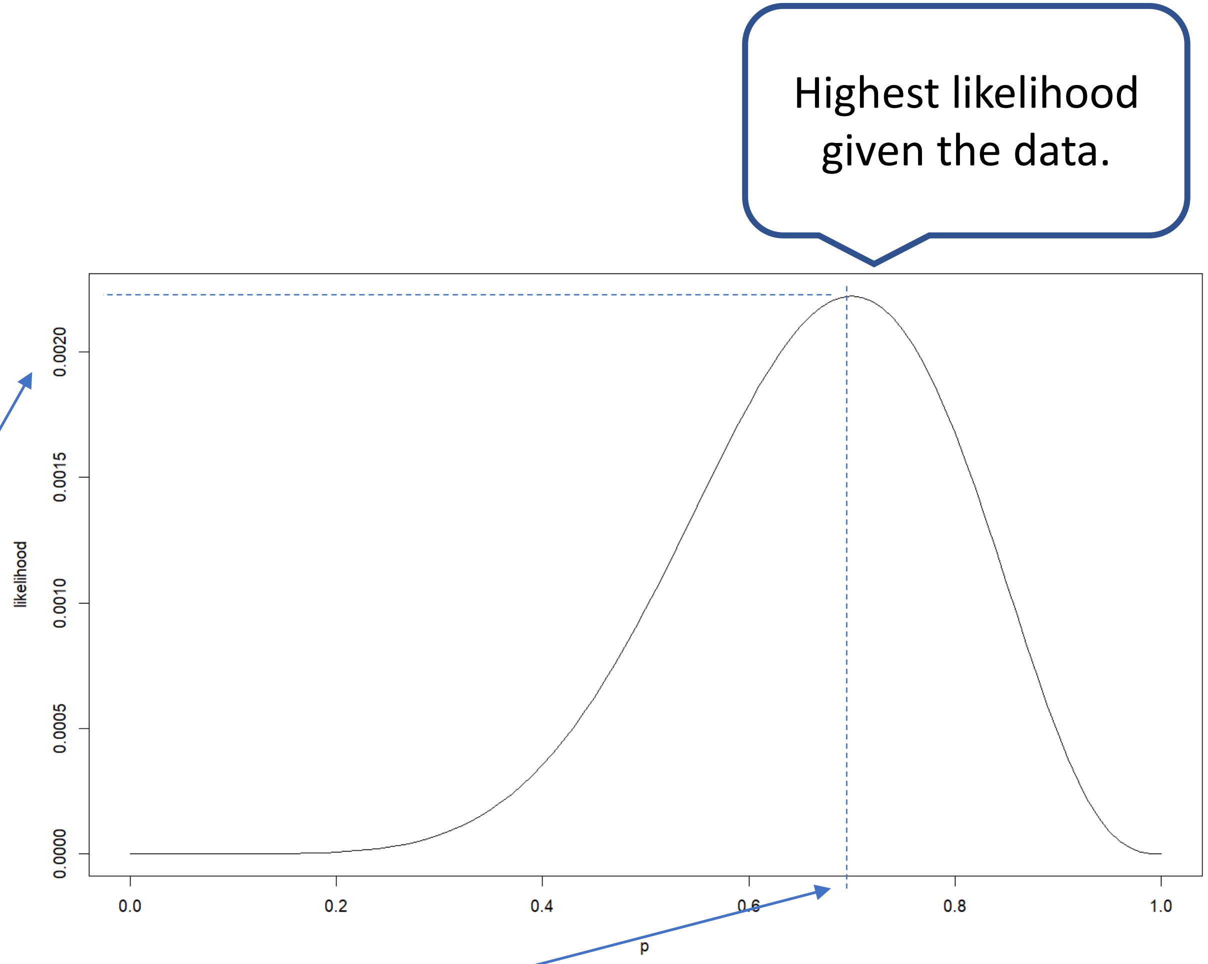
$$L(\textbf{parameter} \mid \textbf{data}) = L(\pi \mid \text{HHTHHHTTHH}) = \pi^7(1-\pi)^3.$$

- NOTE: The first is a function of data, the second is a function of parameter (but they are the same equation).

Example: flipping a coin

$$L(\text{parameter} \mid \text{data}) =$$
$$L(\pi|HHTHHHTTHH) = \pi^7(1 - \pi)^3$$

```
> p <- seq(0, 1, by =0.01)
> likelihood <- p^7 * (1-p)^3
> plot(p, likelihood, type ="l")
```



- MLE selects a parameter value (i.e., $\hat{\pi}$) which is more likely to produce that data at your hand.
- What is the $\hat{\pi}$ here?
- Calculate the probability of obtaining the sample data. Why is it small ?

Generalization of the example

- Consider n independent flips of coin, producing a particular sequence with x heads and $n - x$ tails.

$$L(\pi|\text{data}) \triangleq L(\pi) = \Pr(\text{data}|\,\pi) \;= \pi^{x}\,(1\, - \pi)^{n-x}$$

- Find value of π that maximizes $L(\pi|\text{data})$. Often, it is simpler to maximize the log of the likelihood:

$$\log L(\pi) \;=\; x\log\pi \;+\; (n \;-\; x)\log(1\, - \pi)$$

- Differentiate the log-likelihood w.r.t. π:

$$\frac{d\log L(\pi)}{d\pi} = \frac{x}{\pi} + \;(n \;-\; x)\frac{1}{1-\pi}\;(-1)$$

- Setting it to 0 and solving for π produces the maximum-likelihood estimator (MLE).

- What is the MLE of π ? $\hat{\pi} = X/n$ (i.e., sample average)

Generalization of the example

- Consider n independent flips of coin, producing a particular sequence with x heads and $n - x$ tails.

$$L(\pi|\mathrm{data}) \triangleq L(\pi) = \Pr(\mathrm{data}|\,\pi) \;= \pi^x\,(1\,-\pi)^{n-x}$$

- Find value of π that maximizes $L(\pi|\mathrm{data})$. Often, it is simpler to maximize the log of the likelihood:

$$\log L(\pi) \;=\; x\log\pi \;+\; (n\;-\;x)\log(1\;-\pi)$$

- Differentiate the log-likelihood w.r.t. π:

$$\frac{d\log L(\pi)}{d\pi} = \frac{x}{\pi} + \;(n\;-\;x)\frac{1}{1-\pi}\;(-1)$$

- Setting it to 0 and solving for π produces the maximum-likelihood estimator (MLE).

- What is the MLE of π ? $\hat{\pi} = x/n$ (i.e., sample average)

Likelihood function

- Let $X_1, X_2 \ldots, X_n$ are **iid** random variables with probability functions $P(X_i \mid \theta)$.
- The joint probability function of $X_1, X_2 \ldots, X_n$ is

$$P(X_1, \ldots, X_n | \theta) = \prod_{i=1}^{n} P(X_i \mid \theta).$$

- For an observed sample, the likelihood function is defined as

$$L(\theta | X_1, \ldots, X_n) = P(X_1, \ldots, X_n | \theta)$$

- It's easier to work with log-likelihood function

$$\log L(\theta) = \sum_{i=1}^{n} \log P(X_i, \theta)$$

Properties of Maximum-Likelihood Estimators: Fisher Information

- Asymptotic sampling variance of MLE $\hat{\theta}$ of a single parameter θ

$$V(\hat{\theta}) = \frac{1}{-E\left[\frac{d^2 \log L(\theta)}{d\theta^2}\right]}$$

- The denominator of $V(\hat{\theta})$ is called Fisher Information:

$$I(\theta) = -E\left[\frac{d^2 \log L(\theta)}{d\theta^2}\right]$$

Fisher Information: Example

- Coin tossing example

$$\log L(\pi) = x \log \pi + (n - x) \log(1 - \pi)$$

- Number of successes $X \sim Bin(n, \pi)$, therefore $E(X) = n\pi$

- Calculate $I(\theta)$ for the example above:

$$\frac{d\log L(\pi)}{d\pi} = \frac{x}{\pi} + (n - x)\frac{1}{1 - \pi}(-1)$$

$$\frac{d^2\log L(\pi)}{d\pi} = ?$$

$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n}.$

Properties of Maximum-Likelihood Estimators: Fisher Information

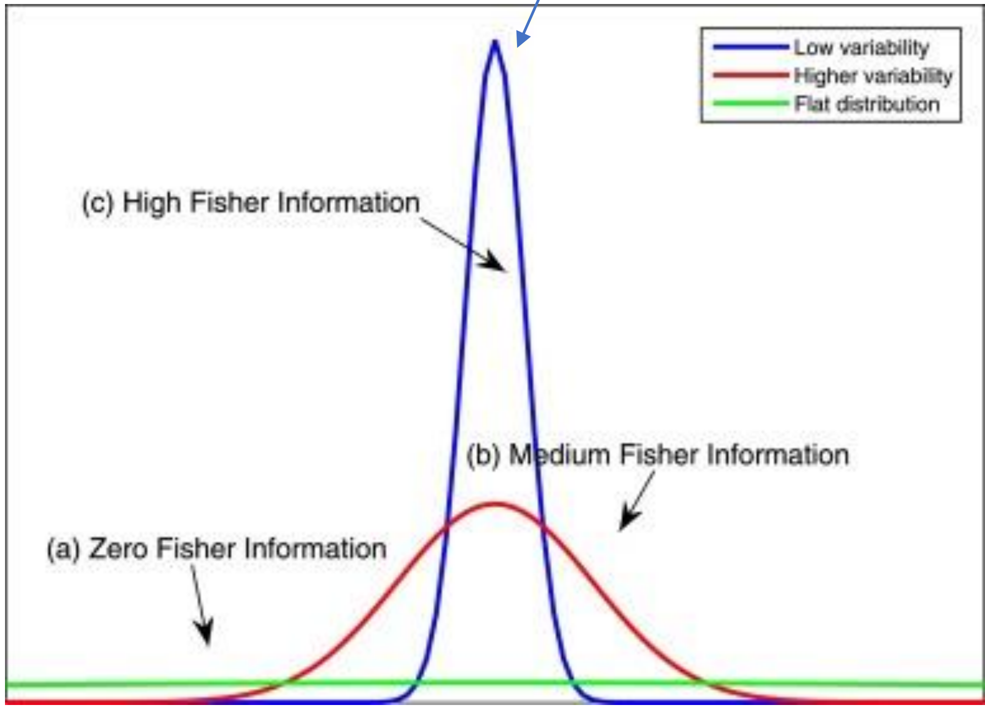
- Fisher Information: $I(\theta) = -E\left[\frac{d^2 \log L(\theta)}{d\theta}\right]$

- Asymptotic sampling variance : $V(\hat{\theta}) = \frac{1}{I(\theta)}$

- Sharp peak ⇒ the second derivative is a small negative number (acceleration) ⇒ lot of "information" in the data ⇒ sampling variance of MLE is small.

- Flat peak ⇒ is little "information"

MLE is clearly differentiated from nearby values



Properties of Maximum-Likelihood Estimators

- **Asymptotically** Consistent

$$\hat{\theta}_{MLE} \to_P \theta,\ \text{when}\ \ n \to \infty,\ \ \text{with probability 1}$$

- **Asymptotically** unbiased (although may be biased finite samples)

$$E\left(\hat{\theta}_{MLE}\right) \to \theta, \text{when}\ \ n \to \infty$$

- **Asymptotically** normally distributed

$$\frac{\hat{\theta}_{MLE}-\theta}{\sqrt{V(\hat{\theta})}} \to N(0,1),\ \text{when}\ \ n \to \infty$$

- **Asymptotically** efficient (Cramér-Rao lower bound) :

$$V(\tilde{\theta}) \geq V(\hat{\theta}_{MLE}),\ \text{for}\ \forall\ \tilde{\theta}\ \text{as. normal estimator}$$

It is the most efficient estimator, among as. normal estimators.

Statistical Inference

- Properties of MLE lead to 3 general procedures for testing a single parameter:

$$H_0 : \theta = \theta_0$$

- Wald test
- Likelihood-ratio test
- Score test

- They are asymptotically equivalent.

Statistical Inference: Wald test

- The test statistic as:

$$Z_0 \equiv \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{V}\,(\hat{\theta})}}$$

which is asymptotically distributed as $N(0, 1)$ under H_0.

- Wald test can be "turned around" to produce confidence intervals.

$$CI(\theta) = \hat{\theta}_{MLE} \pm z_{\alpha/2} \frac{1}{\sqrt{I(\hat{\theta})}}$$

Wald test - Example

- Test the hypothesis

$$H_0\colon \pi = 0.5$$

$$H_0\colon \pi \neq 0.5$$

- The MLE is $\hat{\pi} = 0.7$ with $n = 10$.

Coin toss example, recall $\hat{V}\ (\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{n} = \frac{0.7\times 0.3}{10} = 0.021$.

- $Z_0 \equiv \frac{\hat{\theta}-\theta_0}{\sqrt{\hat{V}\ (\hat{\theta})}} = \frac{0.7-0.5}{\sqrt{0.021}} = 1.38$
- P-value = `2 * pnorm(1.38, lower.tail = FALSE)=0.16` $\Rightarrow$ Fail to reject H_0 .

Statistical Inference: Likelihood-ratio test

- The test statistic is defined as:

$$G_0^2 \equiv -2\log\frac{L(\alpha_0)}{L(\hat{\alpha})} = 2(\log L(\hat{\alpha}) - \log L(\alpha_0))$$

which is asymptotically distributed as χ_1^2 under H_0.

- G_0^2 is twice the difference between the Log-Likelihood at the estimated parameter $\hat{\alpha}$ and at the testing value.

- Is G_0^2 always positive?

Statistical Inference: Score test

- Define the "score" as:

$$S(\alpha) \equiv \frac{d \ \ln L(\alpha)}{d \ \alpha}$$

- What is $S(\hat{\alpha}_{MLE})$?

- The test statistic is defined as

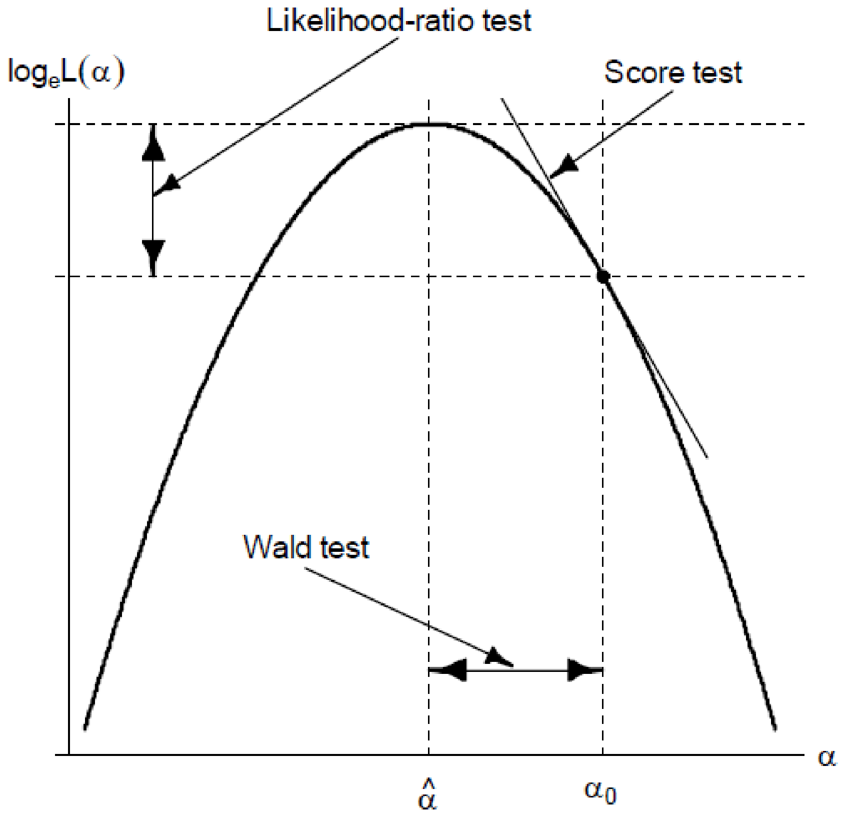
$$S_0 \equiv \frac{S(\alpha_0)}{\sqrt{I(\alpha_0)}}$$

which is asymptotically distributed as $N(0, 1)$ under H_0.

- What is the practical advantage of the score test?

Statistical Inference

- The relationships between 3 test statistics:



Statistical Inference: Several Parameters

- Sample data matrix $\mathbf{X}_{n\times m}$, iid, depending on parameters collected in vector $\boldsymbol{\alpha}_{1\times k}$.
- Joint probability function: $p(\mathbf{X}|\boldsymbol{\alpha}) = p(X_1|\boldsymbol{\alpha}) \times \cdots \times p(X_n|\boldsymbol{\alpha}) = \prod_i P(X_i|\boldsymbol{\alpha})$
- Likelihood function : $L(\boldsymbol{\alpha}) \equiv p(\boldsymbol{\alpha}|\mathbf{X})$
- More convenient with log-transformation: $\log L(\boldsymbol{\alpha})$
- Maximize the likelihood:

$$\frac{\partial \log L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0$$

Statistical Inference: Several Parameters

- Asymptotic variance-covariance $(k \times k)$ matrix of MLE is defined as:

$$V(\hat{\boldsymbol{\alpha}}) = \left\{-E\left[\frac{\partial^2 \log L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'}\right]\right\}^{-1}$$

- Fisher information matrix or expected information matrix is defined as

$$I\,(\boldsymbol{\alpha}) \equiv -E\left[\frac{\partial^2 \log L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'}\right]$$

- MLE is consistent, asymptotically unbiased, asymptotically efficient, asymptotically normal.

Statistical Inference: hypothesis tests

- Generalizations of the tests for $H_0 : \boldsymbol{\alpha} = \boldsymbol{\alpha_0}$ follow directly:

- Wald test uses the following test statistic

$$Z_0^2 \equiv (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha_0})' \hat{V}(\widehat{\boldsymbol{\alpha}})^{-1} (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha_0})$$

which is asymptotically distributed as χ_k^2 under H_0.

Statistical Inference: hypothesis tests

- Likelihood-ratio test uses the following test statistic:

$$G_0^2 \equiv -2 \log \frac{L(\boldsymbol{\alpha}_0)}{L(\widehat{\boldsymbol{\alpha}})}$$

which is asymptotically distributed as χ_k^2 under H_0.

Statistical Inference: hypothesis tests

- Define the score vector

$$S(\boldsymbol{\alpha}) \equiv \frac{\partial \ \ln L(\boldsymbol{\alpha})}{\partial \ \boldsymbol{\alpha}}$$

- Score test statistic is defined as:

$$S_0^2 \equiv S(\boldsymbol{\alpha}_0)' I(\boldsymbol{\alpha}_0)^{-1} S(\boldsymbol{\alpha}_\mathbf{0})$$

which is asymptotically distributed as χ_k^2 under H_0.

- This test can be adapted to more complex hypotheses, e.g., test H_0 that p out of k elements of $\boldsymbol{\alpha}$ are equal to certain values.