STATS 2107
Statistical Modelling and Inference II
Lecture notes
Chapter 5: Likelihood theory

Jono Tuke

School of Mathematical Sciences, University of Adelaide

Semester 2 2017

Maximum likelihood estimation

Joint probability distributions

Consider independent random variables $Y_1,\ Y_2,\ \ldots, Y_n$ and let

$$f_i(y_i;\theta)$$

denote the probability density function if Y_i is continuous and the probability mass function if Y_i is discrete. The joint probability density function or probability mass function is then given by

$$f(\boldsymbol{y};\theta) = \prod_{i=1}^{n} f_i(y_i;\theta).$$

Definitions

- The function
$$L(\theta; \boldsymbol{y}) = f(\boldsymbol{y}; \theta)$$
is called the **likelihood function**.
- The function
$$\ell(\theta; \boldsymbol{y}) = \log L(\theta; \boldsymbol{y})$$
is called the **log-likelihood**.

Examples

- Suppose $y_1, y_2, \ldots, y_n$ are *i.i.d.* $Po(\lambda)$ observations.
- Suppose $y_1, y_2, \ldots, y_n$ are *i.i.d.* $N(\mu, \sigma^2)$ observations with σ^2 known.

Definition

If $y_1, y_2, \ldots, y_n$ are independent observations with log-likelihood $\ell(\theta; \boldsymbol{y})$, then the function

$$S(\theta; \boldsymbol{y}) = \frac{\partial \ell}{\partial \theta}$$

is called the **score function**.

Examples

- Suppose $y_1, y_2, \ldots, y_n$ are *i.i.d.* $Po(\lambda)$
- Suppose $y_1, y_2, \ldots, y_n$ are *i.i.d.* $N(\mu, \sigma^2)$ observations with σ^2 known.

Definition

If $y_1, y_2, \ldots, y_n$ are independent observations with log-likelihood $\ell(\theta; \boldsymbol{y})$, then the maximum likelihood estimate (MLE) $\hat{\theta}$ is the value of θ that maximizes $\ell(\theta; \boldsymbol{y})$.

Remarks

In practice, $\hat{\theta}$ is usually derived by solving the **score equation**

$$S(\theta; \boldsymbol{y}) = 0.$$

We assume $\hat{\theta}$ exists and is unique.

Examples

- Suppose $y_1, y_2, \ldots, y_n$ are *i.i.d.* $Po(\lambda)$ observations.
- Suppose $y_1, y_2, \ldots, y_n$ are *i.i.d.* $N(\mu, \sigma^2)$ observations with σ^2 known.

Fisher information

Cramér-Rao inequality

Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with pdf $f(y;\theta)$. Subject to regularity conditions on $f(y;\theta)$, we have that for any **unbiased** estimator $\tilde{\theta}$ for θ,

$$Var(\tilde{\theta}) \geq I_\theta^{-1}$$

where

$$I_\theta = E\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right].$$

Fisher information

I_θ is known as the Fisher information about θ in the observations.

Alternative form

Under the same regularity conditions as for the Cramér-Rao inequality:

$$I_\theta = -E\left[\frac{\partial^2 \ell}{\partial \theta^2}\right].$$

Proof

Examples

- Suppose $y_1, y_2, \ldots, y_n$ are *i.i.d.* $Po(\lambda)$ observations.
- If $y_1, y_2, \ldots, y_n$ are *i.i.d.* $N(\mu, \sigma^2)$ observations with σ^2 known.

Theorem

Suppose $y_1, y_2, \ldots, y_n$ are independent observations with log-likelihood $\ell(\theta^*; \boldsymbol{y})$, where θ^* denotes the true value of the parameter. Under certain regularity conditions on $f(\boldsymbol{y}; \theta)$,

- $E(S(\theta^*; \boldsymbol{Y})) = 0$.
- $var(S(\theta^*; \boldsymbol{Y})) = I_{\theta^*}$.
- The distribution of
$$\frac{S(\theta^*; \boldsymbol{Y})}{\sqrt{I_{\theta^*}}}$$
converges to $N(0, 1)$ as $n \to \infty$.

Theorem

Suppose $y_1, y_2, \ldots, y_n$ are independent observations with log-likelihood $\ell(\theta^*; \boldsymbol{y})$, where θ^* denotes the true value of the parameter. Under certain regularity conditions on $f(\boldsymbol{y}; \theta)$, then asymptotically

$$\hat{\theta} \sim N(\theta^*, I_{\theta^*}^{-1}),$$

where $\hat{\theta}$ is the MLE for θ.

Examples

- Suppose $y_1, y_2, \ldots, y_n$ are *i.i.d.* $Po(\lambda)$ and recall

$$\hat{\lambda} = \bar{y} \quad \text{and} \quad I_\lambda = \frac{n}{\lambda}.$$

The preceding theory states that

- $\hat{\lambda}$ is "asymptotically unbiased".
- The large-sample standard error is $\sqrt{\lambda/n}$.
- The distribution of

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}}$$

converges to $N(0,1)$ as $n \to \infty$.

Check directly

Confidence intervals

Approximate confidence intervals

Suppose $y_1, y_2, \ldots, y_n$ are independent observations with log-likelihood $\ell(\theta; \boldsymbol{y})$.

An approximate $100(1-\alpha)\%$ confidence interval for θ is given by

$$\left(\hat{\theta} - z_{\alpha/2}\sqrt{I_{\hat{\theta}}^{-1}},\ \hat{\theta} + z_{\alpha/2}\sqrt{I_{\hat{\theta}}^{-1}}\right).$$

Wald test statistic

Suppose $Y_1, \ldots, Y_n$ are i.i.d. observations with log-likelihood $\ell(\theta; \boldsymbol{Y})$.

The Wald test statistic for

$$H_0 : \theta = \theta_0$$

is given by

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{I_{\hat{\theta}}^{-1}}}.$$

If $H_0 : \theta = \theta_0$ is true then, the distribution of Z converges to $N(0, 1)$ as $n \to \infty$.

A test with significance level approximately α is given by the rule:

$$\text{Reject } H_0 \text{ if } |Z| \geq z_{\alpha/2}.$$

Score test statistic

Suppose $Y_1, \ldots, Y_n$ are i.i.d. observations with log-likelihood $\ell(\theta; \boldsymbol{Y})$.

The score test statistic for

$$H_0 : \theta = \theta_0$$

is given by

$$U = \frac{S(\theta_0; \boldsymbol{Y})}{\sqrt{I_{\theta_0}}}.$$

If $H_0 : \theta = \theta_0$ is true then, the distribution of U converges to $N(0,1)$ as $n \to \infty$.

A test with significance level approximately α is given by the rule:

Reject H_0 if $|U| \geq z_{\alpha/2}$.

Log-likelihood ratio test statistic

Suppose $Y_1, \ldots, Y_n$ are i.i.d. observations with log-likelihood $\ell(\theta; \boldsymbol{Y})$.

The log likelihood-ratio test statistic is given by

$$G^2 = -2(\ell(\theta_0; \boldsymbol{Y}) - \ell(\hat{\theta}; \boldsymbol{Y})).$$

If $H_0 : \theta = \theta_0$ is true then, under suitable regularity conditions, the distribution of G^2 converges to χ^2_1 as $n \to \infty$.

A test with significance level approximately α is given by the rule:

$$\text{Reject } H_0 \text{ if } g^2 \geq \chi^2_{1,\alpha}.$$

Example

Suppose $y_1, y_2, \ldots, y_n$ are independent $Po(\lambda)$ observations.

Suppose we wish to test $H_0 : \lambda = \lambda_0$.

Calculate the Wald test statistic, the Score test statistic, and the log-likelihood ratio test statistic.

Transformation of parameters

Setup

Suppose $y_1, y_2, \ldots, y_n$ are independent observations with log-likelihood $\ell(\theta; \boldsymbol{y})$ for a scalar parameter θ and consider an invertible, twice differentiable function, Φ. Taking $\phi = \Phi(\theta)$ we can take ϕ as the parameter of interest rather than θ.

Example Bernoulli

Suppose $y_1, y_2, \ldots, y_n$ are i.i.d. Bernoulli observations with success probability θ.

Consider the log-odds,

$$\Phi(\theta) = \log\left(\frac{\theta}{1-\theta}\right).$$

Relationship between likelihoods

Let the log-likelihoods with respect to θ and ϕ be given respectively by

$$\ell_\theta(\theta; \boldsymbol{y}) \quad \text{and} \quad \ell_\phi(\phi; \boldsymbol{y}).$$

It can be checked that the two likelihoods are related by

$$\ell_\phi(\phi; \boldsymbol{y}) = \ell_\theta(\Phi^{-1}(\phi); \boldsymbol{y})$$

and

$$\ell_\theta(\theta; \boldsymbol{y}) = \ell_\phi(\Phi(\theta); \boldsymbol{y}).$$

Example

Calculate the log-likelihoods for both parameterizations of the Bernoulli.

Theorem

Suppose $\ell_\theta(\theta; \boldsymbol{y})$ and $\ell_\phi(\phi; \boldsymbol{y})$ are equivalent parametrizations of the same problem. Then

$$\hat{\phi} = \Phi(\hat{\theta}).$$

Proof

Invariance of HT

For independent Bernoulli observations $y_1, y_2, \ldots, y_n$ the hypothesis $H_0 : \theta = 0.5$ can be expressed equivalently as $H_0 : \phi = 0$ if

$$\phi = \log\left(\frac{\theta}{1-\theta}\right).$$

However, it can be checked that the Wald test statistics corresponding to the two equivalent formulations of the problem are not equal.

Theorem

Suppose $y_1, y_2, \ldots, y_n$ are independent observations with log-likelihood function

$$\ell_\theta(\theta; \boldsymbol{y}) = \ell_\phi(\phi; \boldsymbol{y})$$

where $\phi = \Phi(\theta)$. Consider the hypothesis

$$H_0 : \theta = \theta_0 \quad \Leftrightarrow \quad H_0 : \phi = \phi_0$$

and let u_θ and u_ϕ be the score statistics defined from the two log-likelihood functions.

If Φ is $1-1$ and onto and twice continuously differentiable with $\Phi'(\theta) \neq 0$ then

$$|u_\phi| = |u_\theta|.$$

“Goodness of fit”

Multinomial distribution

The integer-valued random variables, $Y_1, Y_2, \ldots, Y_k$ are said to follow the multinomial distribution if their joint probability function is given by

$$p(y_1, y_2, \ldots, y_k) = \binom{n}{y_1, y_2, \ldots y_k} \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_k^{y_k}$$

for

$$y_1, \ldots, y_k \geq 0; \quad y_1 + y_2 + \ldots + y_k = n,$$

where

$$\pi_1, \pi_2, \ldots, \pi_k \geq 0$$

with

$$\pi_1 + \pi_2 + \cdots + \pi_k = 1$$

Mean and variance of multinomial

It can be shown for the multinomial-$(n, \boldsymbol{\pi})$ distribution that

$$\begin{aligned} E(Y_i) &= n\pi_i \\ var(Y_i) &= n\pi_i(1 - \pi_i) \\ cov(Y_i, Y_j) &= -n\pi_i\pi_j \text{ if } i \neq j. \end{aligned}$$

Setup

Suppose now that we observe multinomial data $Y_1, Y_2, \ldots, Y_k$ and wish to test the hypothesis

$$H_0: \boldsymbol{\pi} = \boldsymbol{\pi}_0$$

for some specific set of probabilities $\boldsymbol{\pi}_0$.

Definition

The "goodness of fit" test statistic is given by

$$X^2 = \sum_{i=1}^{k} \frac{(Y_i - n\pi_{0i})^2}{n\pi_{0i}}.$$

Theorem

If $H_0: \boldsymbol{\pi} = \boldsymbol{\pi}_0$ is true, then for large n the distribution of X^2 is approximately χ^2_{k-1}.

Reject if $x^2 \geq \chi^2_{k-1,\alpha}$.

Example

A die was rolled 48 times and the following data recorded.

Value	1	2	3	4	5	6	total
Frequency	8	10	9	7	7	7	48

Test the hypothesis that the die is fair.

Example

In an occupational health and safety study, 414 machinists at a particular factory were recorded over 3 months and the number of accidents were recorded.

Accidents	0	1	2	3	4	5	6	7	8	total
Frequency	296	74	26	8	4	3	2	0	1	414

Can these observations be modelled by i.i.d. Poisson observations?

Theorem

If H_0 is true, then for large n, the distribution of X^2 is approximately χ^2_{k-q-1} where q is the number of parameters that have to be estimated to compute $\boldsymbol{\pi}$.