Although the multiple linear regression model provide a versatile framework for modelling, there are many important statistical problems that cannot be expressed within this framework. For this reason, it is useful to consider more general methods for statistical estimation and hypothesis testing. In this module, we will discuss the method of maximum likelihood.

Maximum likelihood estimation

- Other commonly used methods of estimation:
  - Method of moments
  - Method of Bayesian

- Least square estimation minimizes the SSE
- Maximum likelihood estimation maximizes the likelihood function

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

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

Note that this is a function of y. θ is a parameter (treated like a constant).

Likelihood and log-likelihood

- The function

$$L(\theta;\, \boldsymbol{y}) = f(\boldsymbol{y};\theta)$$

For independent Y_i:

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

is called the likelihood function.

Note that the likelihood is a function of θ.

- The function

$$\ell(\theta;\, \boldsymbol{y}) = \log L(\theta;\boldsymbol{y})$$

For independent Y_i:

$$= \log\left[\prod_{i=1}^{n} f_i(y_i;\theta)\right]$$

$$= \sum_{i=1}^{n} \log f_i(y_i;\theta)$$

is called the log-likelihood function.

Example 5.1

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

(a) Give the likelihood function.

(b) Give the log-likelihood function.

$$f(y_i;\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \quad \text{for } y_i = 0, 1, 2, \ldots$$

(a) $$L(\lambda; y) = f(y;\lambda) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = e^{-n\lambda}\,\lambda^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n}\left(\frac{1}{y_i!}\right)$$

(b) $$\begin{aligned} \ell(\lambda; y) &= \log L(\lambda; y) \\ &= \log\left[e^{-n\lambda}\,\lambda^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n}\left(\frac{1}{y_i!}\right)\right] \\ &= -n\lambda + \left(\sum_{i=1}^{n} y_i\right)\log\lambda + \log \prod_{i=1}^{n}\left(\frac{1}{y_i!}\right) \end{aligned}$$

Example 5.2

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

(a) Give the likelihood function.

(b) Give the log-likelihood function.

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

(a) $$L(\mu; y) = f(y;\mu) = \prod_{i=1}^{n} f(y_i;\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

(b) $$\ell(\mu; y) = \log L(\mu; y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Exercise: Suppose σ^2 is also unknown. So $\theta = (\mu, \sigma^2)$. What is $L(\mu, \sigma^2; y)$ and $\ell(\mu, \sigma^2; y)$?

They are the same as (a) and (b) above.

$$L(\mu, \sigma^2; \boldsymbol{y}) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

$$\ell(\mu, \sigma^2; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Score function

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.

If θ has more than one element, then we have a score vector.

e.g. $\theta = (\theta_1, \theta_2)$.

$$S(\theta; y) = \begin{bmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \end{bmatrix}$$

Example 5.3

Find the score function if $y_1, y_2, \ldots, y_n$ are $i.i.d.$ with the following distributions:

(a) $Po(\lambda)$.

(b) $N(\mu, \sigma^2)$ with σ^2 known.

(a) $$S(\lambda; y) = \frac{\partial l(\lambda; y)}{\partial \lambda}$$
$$= \frac{\partial}{\partial \lambda}\left[-n\lambda + \left(\sum_{i=1}^{n} y_i\right)\log\lambda + \log\left(\prod_{i=1}^{n} \frac{1}{y_i!}\right)\right]$$
$$= -n + \frac{1}{\lambda}\left(\sum_{i=1}^{n} y_i\right)$$

(b) $$S(\mu; y) = \frac{\partial l(\mu; y)}{\partial \mu}$$
$$= \frac{\partial}{\partial \mu}\left[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)$$

Maximum likelihood estimation

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 $\ell(\theta; \boldsymbol{y})$.

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta; y)$$

In practice, we obtain $\hat{\theta}$ by solving $\frac{\partial \ell(\theta; y)}{\partial \theta} = 0$ for θ.

equivalently: $S(\theta; y) = 0$

Maximum likelihood estimation

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.

Example 5.4

Suppose $y_1, y_2, \ldots, y_n$ are $i.i.d. Po(\lambda)$ observations. Find the maximum likelihood estimate $\hat{\lambda}$ of λ.

The score function is $S(\lambda; y) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} y_i$.

Solve for λ in $0 = S(\lambda; y)$

$$0 = -n + \frac{1}{\lambda}\sum_{i=1}^{n} y_i$$

$$n = \frac{1}{\lambda}\sum_{i=1}^{n} y_i$$

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

So $\hat{\lambda} = \bar{y}$.

Example 5.5

Suppose $y_1, y_2, \ldots, y_n$ are $i.i.d. N(\mu, \sigma^2)$ observations with σ^2 known. Find the maximum likelihood estimate $\hat{\mu}$ of μ.

Solve for μ in $0 = S(\mu; y)$

$$0 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)$$

$$0 = \sum_{i=1}^{n} (y_i - \mu)$$

$$0 = \sum_{i=1}^{n} y_i - n\mu$$

$$n\mu = \sum_{i=1}^{n} y_i$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

So $\hat{\mu} = \bar{y}$.

In this case, the MLE of μ is maximizing $\ell(\mu; \boldsymbol{y})$, which is equivalent to minimizing $\sum_{i=1}^{n} (y_i - \mu)^2$.

So, for normally distributed data, the MLE of μ is the same as the least squares estimate (LSE) of μ.

Example 5.6

Suppose $y_1, y_2, \ldots, y_n$ are $i.i.d. N(\mu, \sigma^2)$ observations where both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ are unknown. Find the maximum likelihood estimate $\hat{\mu}$ and $\hat{\sigma}^2$ of μ and σ^2, respectively.

$$S(\sigma^2; y) = \frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2; y)$$

$$= \frac{\partial}{\partial \sigma^2} \left[-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \right]$$

$$= -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2$$

The score vector is

$$S(\mu, \sigma^2; y) = \begin{bmatrix} \frac{\partial \ell(\mu, \sigma^2; y)}{\partial \mu} \\ \frac{\partial \ell(\mu, \sigma^2; y)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 \end{bmatrix}.$$

Example 5.6 Solutions

To find $\hat{\mu}$ and $\hat{\sigma}^2$, solve for μ and σ^2 simultaneously in

$$S(\mu, \sigma^2; y) = 0.$$

We have found $\hat{\mu} = \bar{y}$ from Example 5.5.

Solve for σ^2 in $0 = S(\sigma^2; y)$.

$$0 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$\frac{n}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$n\sigma^2 = \sum_{i=1}^{n}(y_i - \mu)^2$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2.$$

So $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2.$