MLE for a gaussian distribution

Susan Thomas
IGIDR, Bombay

August 28, 2008

Recap

- Properties of estimators: unbiasedness, efficiency.
- The Cramer Rao lower bound and the Fisher Information number
- Properties of MLE: consistency, asymptotic efficiency, invariance

MLE of the gaussian distribution

Economic problem: A model for wages

- Wages are observations in the positive real number space.
- Distribution: continuous, positive only – log normal.
- Log(wages): normal distribution where

$$f(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w-\mu)^2}{2\sigma^2}} \qquad \forall \mu \in R, \sigma^2 \in R+$$

- Setting up the model for log(wages), w:
  1. independence
  2. identical distribution
  3. normally distributed

  Parameter space for the estimation of the model: μ, σ^2.
- What is the MLE for μ, σ^2?

The likelihood function for the normal distribution

- Each w has a probability, $f(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w-\mu)^2}{2\sigma^2}}$
- Likelihood, L is $L_{w_1,w_2,\ldots,w_N}(\mu,\sigma^2)$

$$L_{w_1,\ldots,w_N}(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w_1-\mu)^2}{2\sigma^2}} * \ldots * \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w_N-\mu)^2}{2\sigma^2}}$$

- The log likelihood, $l = \ln L$, is

$$l_{w_1,\ldots,w_N}(\mu,\sigma^2) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(w_i-\mu)^2$$

The likelihood function for the normal distribution

- For the MLE, we need to differentiate $l_{w_1,\ldots,w_N}(\mu, \sigma^2)$ by μ, and σ^2.

$$\partial l_{w_1,\ldots,w_N}(\mu, \sigma^2)/\partial\mu$$

$$\partial l_{w_1,\ldots,w_N}(\mu, \sigma^2)/\partial\sigma^2$$

  and set each to zero.

MLE for μ

- The first derivative of $l_{w_1,\ldots,w_N}(\mu, \sigma^2)$ wrt μ is:

$$\begin{aligned} \frac{\partial}{\partial \mu}\left(l_{w_1,\ldots,w_N}(\mu,\sigma^2)\right) &= \frac{\partial}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(w_i-\mu)^2\right) \\ 0 &= -\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{n}(w_i-\mu)^2\right) \end{aligned}$$

- The term $\sum_{i=1}^{n}(w_i-\mu)^2$ is called the sum of squared errors, SSE.
- Maximimising the likelihood function is the same as minimising the SSE.
- It is quadratic in μ. This will have a unique minimum.
- We find: $\mu_{mle} = \sum_{i=1}^{n} w_i/n = \hat{w}$
  The MLE for μ is the sample mean.

Features of the MLE for μ

- Like in the case of the Bernoulli model, the sample mean $\hat{w}$ is the ML solution for the gaussian model as well.
- Since it minimises the SSE, it is also called the least-squares estimator.
  More typically, it is called the *ordinary least-squares* or OLS estimator.
- **HW:** We know the solution is unique for a quadratic. Prove it using the second derivative of $l_{w_1,\ldots,w_N}(\mu, \sigma^2)$.

Features of the MLE for σ^2

- The first derivative of $l_{w_1,\ldots,w_N}(\mu, \sigma^2)$ wrt σ^2 is:

$$\frac{\partial}{\partial \sigma^2} l_{w_1,\ldots,w_N}(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (w_i - \mu)^2$$

- Setting this to zero, we get:

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

- We replace μ with $\mu_{mle} = \hat{w}$ and get:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (w_i - \mu_{mle})^2 = \frac{1}{n} \sum_{i=1}^{n} (w_i - \hat{w})^2$$

- **HW:** Show that the sample variance maximises the likelihood function.

Information matrix for the normal distribution

- We have two parameters, μ, σ^2. Therefore, the second derivative of the log likelihood function has three terms:

$$\partial^2/\partial\mu^2, \quad \partial^2/\partial\mu\sigma^2, \quad \partial^2/\partial\sigma^4$$

- These are:

$$\begin{aligned}
\partial^2/\partial\mu^2 &= \frac{-n}{\sigma^2} \\
\partial^2/\partial\mu\sigma^2 &= \frac{-1}{\sigma^4}\sum_{i=1}^{n}(w_i - \mu) \\
\partial^2/\partial(\sigma^2)^2 &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(w_i - \mu)^2
\end{aligned}$$

Inference for the MLE $\mu_{mle}, \sigma^2_{mle}$

- The Cramer-Rao bound for the MLE estimators is defined as a 2x2 matrix as follows:

$$[\mathrm{I}(\theta)]^{-1} = \begin{bmatrix} \partial^2 l/\partial\mu^2 & \partial^2 l/\partial\mu\sigma^2 \\ \partial^2 l/\partial\sigma^2\mu & \partial^2 l/\partial\sigma^4 \end{bmatrix}$$

- For the MLE of the normal distribution, this is

$$[\mathrm{I}(\theta)]^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}$$

- The sampling distribution of the MLE is set by this bound.

Inference for μ_{mle}

- $E(\mu_{mle}) = E(\sum_{i=1}^{n} w_i/n) = n\mu/n = \mu$
  Thus, the MLE is an unbiased estimator of μ.
- $\mathrm{var}(\mu_{mle}) = \mathrm{var} \sum_{i=1}^{n}(w_i/n) = \sigma^2/n$
- Under the CLT, the sampling distribution for μ_{mle} is:

$$\frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} = \sqrt{n}\left(\frac{\hat{\mu} - \mu}{\sigma}\right) \sim N(0, 1)$$

Inference for μ_{mle}

- We have two estimators for σ^2 – the ML estimator for σ^2 and the sample variance $\hat{\sigma}^2$:

$$\begin{aligned} \sigma^2_{mle} &= \sum_{i=1}^{n} (w_i - \hat{w})^2 / n \\ \hat{\sigma}^2 &= \sum_{i=1}^{n} (w_i - \hat{w})^2 / (n-1) \end{aligned}$$

  where $\hat{\sigma}^2 = n/(n-1)\sigma^2_{mle}$.

Inference for μ_{mle}

- Theoretically, it can be shown that

$$\frac{\hat{\mu} - \mu}{\sqrt{\hat{\sigma}^2 / n}} \sim t[n-1]$$

- The t-distribution can be used to determine the 95% confidence intervals for small samples.
- However, since the t-distribution tends rapidly to the standard normal (for $n > 10$), the 95% confidence interval becomes:

$$\hat{\mu} - 2\hat{\sigma}/\sqrt{n} \leq \mu \leq \hat{\mu} + 2\hat{\sigma}/\sqrt{n}$$

Inference for σ^2_{mle}

- There are two estimators for σ^2 of the normal distribution: σ^2_{mle} and $\hat{\sigma}^2$.
- Are they both unbiased?

$$\sigma^2_{mle} = \frac{(n-1)}{n}\hat{\sigma}^2$$
$$E(\sigma^2_{mle}) < E(\hat{\sigma}^2)$$

  The MLE is biased slightly downward compared to the sample variance, which is an unbiased estimate.
- Are they both efficient?

$$\mathrm{var}(\sigma^2_{mle}) = \left(\frac{n-1}{n}\right)^2 \mathrm{var}(\hat{\sigma}^2)$$
$$\mathrm{var}(\sigma^2_{mle}) < \mathrm{var}(\hat{\sigma}^2)$$

  The MLE variance is lower than the sample variance. MLE is more efficient.
- Software report $\hat{\sigma}^2$ as the estimation variance.