Properties of the score and MLE

- We will briefly look at the basic properties of the score and the MLE: their expected value, variance, and distribution
- The MLE has many desirable properties, such as
  - asymptotically unbiased
  - asymptotically normally distributed
  - asymptotically efficient (i.e. achieves Cramér-Rao lower bound)
  - a consistent estimator
  - invariant under transformation of data and parameters

Theorem 13

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

1. $E[S(\theta^*; \boldsymbol{Y})] = 0$
2. $\operatorname{var}\big(S(\theta^*; \boldsymbol{Y})\big) = I_{\theta^*}$. $= E[S(\theta^*; Y)^2]$
3. The distribution of

$$\frac{S(\theta^*; \boldsymbol{Y})}{\sqrt{I_{\theta^*}}}$$

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

Proof of Theorem 13

① $E[S(\theta;y)] = \int_{-\infty}^{\infty} S(\theta;y)\, L \, dy$

$= \int_{-\infty}^{\infty} \frac{\partial \log L}{\partial \theta} L \, dy$

$= \int_{-\infty}^{\infty} \left(\frac{1}{\cancel{L}} \frac{\partial L}{\partial \theta}\right) \cancel{L} \, dy$ chain rule: $\frac{\partial \log f(x)}{\partial x} = \frac{f'(x)}{f(x)}$

$= \int_{-\infty}^{\infty} \frac{\partial L}{\partial \theta} dy$

$= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} L \, dy$ under regularity conditions

$= \frac{\partial}{\partial \theta} 1$

$= 0$

Proof of Theorem 13

(2) $\mathrm{var}[S(\theta;y)] = E[S(\theta;y)^2] - E[S(\theta;y)]^2$ variance formula

$= E[S(\theta;y)^2] - 0$ from (1) $E[S(\theta;y)] = 0$

$= E[S(\theta;y)^2]$

$= I_\theta$

(3) Write $S(\theta;y) = \sum_{i=1}^{n} S(\theta;y_i)$, where $S(\theta;y_i)$ is the score based on y_i.

We know $E[S(\theta;y_i)] = 0$ and $\mathrm{var}(S(\theta;y_i)) = i_\theta$,

where i_θ is the Fisher information based on y_i. By the Central Limit Theorem,

$$\frac{\overline{S(\theta;y)} - E[\overline{S(\theta;y)}]}{\sqrt{\frac{\mathrm{var}(S(\theta;y_i))}{n}}} \to N(0,1) \quad \text{as } n \to \infty$$

$$\frac{S(\theta;y)}{\sqrt{I_\theta}} \to N(0,1)$$

Proof of Theorem 13

Remarks:

① provides some justification for the principles of maximum likelihood. In MLE, we set $S(\theta; \boldsymbol{y}) = 0$ and then solve for θ. We know from ① that $E[S(\theta\ ; \boldsymbol{y})] = 0$ under regularity conditions. So we are essentially equating $S(\theta; \boldsymbol{y})$ with its expected value.

Theorem 14

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 of $f(\boldsymbol{y}; \theta)$, then asymptotically

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

$\sqrt{I_{\theta^*}}\,(\hat{\theta} - \theta) \to N(0,1)$

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

① $\hat{\theta}$ is an (asymptotically) unbiased estimator of θ

② An 'approximate' standard error for $\hat{\theta}$ is $\frac{1}{\sqrt{I_{\theta^*}}}$

③ $\hat{\theta}$ is (asymptotically) a minimum variance unbiased estimator of θ.
(because it achieves the Cramér-Rao lower bound and is unbiased asymptotically.)

In practice, θ^* is unknown. We approximate it with $\hat{\theta}$.
So an approximation of $\frac{1}{\sqrt{I_{\theta^*}}}$ is $\frac{1}{\sqrt{I_{\hat{\theta}}}}$.

Example 5.10

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

Recall that $\hat{\lambda} = \bar{y}$ and $I_\lambda = \frac{n}{\lambda}$.

Theorem 14 states that

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

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

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

Check directly that the above is true.

Example 5.10 Solution

$Y_i \sim Po(\lambda)$, $\quad E[Y_i] = \lambda$, $\quad var(Y_i) = \lambda$

(a) $E[\hat{\lambda}] = E[\bar{Y}] = \lambda$

In this case, $\hat{\lambda}$ is exactly unbiased for any value of n.

(b) $var(\hat{\lambda}) = var(\bar{Y}) = \frac{\lambda}{n}$

In this case, $\hat{\lambda}$ is exactly $\sqrt{\frac{\lambda}{n}}$ for any n.

(c) From the Central Limit Theorem, we have

$$\hat{\lambda} = \bar{Y} \longrightarrow N\left(\lambda, \frac{\lambda}{n}\right)$$

So
$$\frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\lambda}{n}}} \longrightarrow N(0, 1)$$

In the case of $Y_i \sim N(\mu, \sigma^2)$ with σ^2 known, we have

$$\hat{\mu} = \bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Hence, the asymptotic results hold exactly in this case.

The Poisson and normal examples are both special cases in that the asymptotic properties in Theorem 14 hold exactly for all n or can be obtained directly by other means. The usefulness of Theorem 14 lies in the fact that it gives an approximate distribution for $\hat{\theta}$ when exact calculations are not possible. For example, in many situations there is no formula for the MLE in terms of y, but Theorem 14 can still be used.