“The third variance” and “predictive estimator”

Kunio Takezawa
Statistical Modeling Unit, Division of Informatics and Inventory,
Institute for Agro-Environmental Sciences
Kannondai 3-1-3, Tsukuba, Ibaraki 305-8604, Japan

The maximum likelihood method estimates parameters by maximizing log-likelihood. Log-likelihood is defined as

$$l = \sum_{i=1}^{n} \log \left( f(x_i|\theta) \right),$$

where \{x_1, x_2, \ldots, x_n\} is available data. \( \theta \) is parameters. \( f(x|\theta) \) stands for probability density function. The true value of \( \theta \) is represented as \( \theta_0 \). \( \theta \) which maximizes \( l \) is called the maximum likelihood estimator.

On the other hand, we assume the maximum likelihood method in the light of future data. This method estimates parameters by maximizing the log-likelihood in the light of future data. The log-likelihood in the light of future data is defined as

$$l^* = \sum_{i=1}^{n} \log \left( f(x_i^*|\theta) \right),$$

where \{x_1^*, x_2^*, \ldots, x_n^*\} is future data.

Since the number of future data is infinite, we assume the expectation of \( l^* \) with respect to future data. That is, we calculate the average of \( l^* \) by sampling infinite number of future data. Then, we have

$$E_{\{x_i^*\}}[l^*] = \left[ \sum_{i=1}^{n} \log \left( f(x_i^*|\theta) \right) \right] = n \int \log \left( f(x|\theta) \right) f(x|\theta_0) dx.$$

This is called expected log-likelihood. We cannot estimate the parameters by maximizing this value.

Then, we assume the expectation of \( l^* \) with respect to both of available data and future data. The expectation of \( l^* \) given by sampling infinite number of available data and future data is represented as

$$E_{\{x_i, x_i^*\}}[l^*].$$

The estimator which maximizes the value above is called “predictive estimator”. The concept of estimation by increasing the value of \( E_{\{x_i, x_i^*\}}[l^*] \) is similar to those of AIC (Akaike’s Information Criterion) and GCV (Generalized Cross-Validation). However, while AIC and GCV are used for model selection, the predictive estimator is used for deriving estimates.

Next, the variance of normal distribution is obtained as an example of the predictive estimator. Available data is represented as \{x_1, x_2, \ldots, x_n\}, and future data is represented as \{x_1^*, x_2^*, \ldots, x_n^*\}. Then, the expectation with respect of available data and future data is written as

$$E_{\{x_i, x_i^*\}}[l^*] = E_{\{x_i^*\}} \left[ -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\sigma}^2) - \frac{\sum_{i=1}^{n}(x_i^* - \bar{x})^2}{2\hat{\sigma}^2} \right].$$

where \( \hat{\sigma}^2 \) is the variance which is obtained using available data. Here, \( \hat{\sigma}^2 \) is set as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - \alpha},$$

\( \alpha \) is a constant. \( \bar{x} \) is the average of \{x_i\}.
Then, $l^*$ is written as

$$l^* = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left( \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - \alpha} \right) - \frac{(n - \alpha)}{2} \frac{\sum_{i=1}^{n} (x_i^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \quad (7)$$

Taking the expectation of available data and future data gives

$$E_{\{x_i, x_i^*\}}[l^*] = -\frac{n}{2} \log(2\pi) - \frac{n}{2} E_{\{x_i\}} \left[ \log \left( \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - \alpha} \right) \right] - \frac{n - \alpha}{2} E_{\{x_i, x_i^*\}} \left[ \frac{\sum_{i=1}^{n} (x_i^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]. \quad (8)$$

For calculating the last term of the right hand side of the equation above, we use the equation below.

$$\frac{\sum_{i=1}^{n} (x_i^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\chi^2_{n+1}}{\chi^2_{n-1}} = \frac{n + 1}{n - 1} F_{(n+1, n-1)}, \quad (9)$$

$\chi^2_{n+1}$ is a random variable which obeys chi-squared distribution with $(n+1)$ degrees of freedom. $\chi^2_{n-1}$ is a random variable which obeys chi-squared distribution with $(n-1)$ degrees of freedom. $\chi^2_{n+1}$ and $\chi^2_{n-1}$ are independent each other. $F_{(n+1, n-1)}$ is a random variable which obeys F-distribution with $(n+1, n-1)$ degree of freedom. Therefore, we have

$$E_{\{x_i, x_i^*\}} \left[ \frac{\sum_{i=1}^{n} (x_i^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] = \frac{n + 1}{n - 1} E[F_{(n+1, n-1)}] = \frac{n + 1}{n - 1} \cdot \frac{n - 1}{n - 3} = \frac{n + 1}{n - 3}. \quad (10)$$

Then, $E_{\{x_i, x_i^*\}}[l^*]$ defined by Eq.(8) is represented as

$$E_{\{x_i, x_i^*\}}[l^*] = -\frac{n}{2} \log(2\pi) - \frac{n}{2} E_{\{x_i\}} \left[ \log \left( \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - \alpha} \right) \right] - \frac{(n - \alpha)(n + 1)}{2(n - 3)}. \quad (11)$$

By extracting the essential part from the equation above for estimating optimal value of $\alpha$, we have the equation:

$$g(\alpha) = \frac{n}{2} \log(n - \alpha) + \frac{\alpha(n + 1)}{2(n - 3)}. \quad (12)$$

Differentiation of the equation above with respect to $\alpha$ and setting it to 0, we obtain the optimal value of $\alpha$:

$$\hat{\alpha} = \frac{4n}{n + 1} \approx 4. \quad (13)$$

The value of $\hat{\alpha}$ here is independent of the true parameters. However, the predictive estimators in most cases depend on the true parameters; it may require ingenuity to handle it.

We have entered into a new age of variance estimator because three typical estimators of variance are available now:

- Maximum likelihood variance: $\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$,
- Unbiased variance: $\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$,
- Third variance: $\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 4}$.

If prediction is the main purpose, the third variance is the first choice. However, we can not deny the possibility that other estimators are superior to the third variance in terms of predition.