MAXIMUM LIKELIHOOD ESTIMATION FOR COMPLEX SURVEY DATA

Hee-Choon Shin
Institute for Survey Research, Temple University, 4646 40th St., N.W., Suite 320, Washington, D.C. 20016

KEY WORDS: Maximum Likelihood Estimation, Complex Sample Survey, Weighted Likelihood Function

Introduction

Cochran (1977, p. 150) noted that estimation techniques for sample survey work were "restricted in scope." Cochran's diagnosis on the estimation in sample surveys is still true and valid after a quarter century later. Indeed a large body of theory in theoretical statistics on how to make good estimates from data has not been fully applied to sample survey theory. To author's knowledge, the maximum likelihood estimation (MLE) methods have not been applied to complex survey data.

In this short paper, the current author would like to emphasize the following three research activities in the sample survey community: 1) functional forms, 2) deductive approach in estimation, and 3) Likelihood. Concretely, this paper discusses the specification of likelihood functions, and the method of finding MLE estimate(s) for samples with unequal weights, and MLE for stratified samples. A different interpretation of sampling weights is suggested too.

Independent Sample

As an example, consider an independent sample of size n from a binomial population with parameter θ and x successes in the sample. The likelihood function can be expressed:

$$L(x\mid n,\theta)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x}. \quad (1)$$

As Fisher (1922) described in his seminal paper eighty years ago, the maximum likelihood estimate, $\hat{\theta}$, can be obtained by equating the first derivative of log likelihood to 0:

$$\frac{\partial}{\partial\theta}\log L(x\mid n,\theta)=0. \quad (2)$$

The variance of $\hat{\theta}$ is obtained by

$$\frac{\partial^2}{\partial\theta^2}\log L(x\mid n,\theta)=-\frac{1}{\sigma^2_{\hat{\theta}}}. \quad (3)$$

For a sample from independently and identically distributed population, the estimate is

$$\hat{\theta}=\frac{x}{n}. \quad (4)$$

The variance of the estimate is

$$\sigma^2_{\hat{\theta}}=\frac{\hat{\theta}(1-\hat{\theta})}{n}. \quad (5)$$

Independent Sample with Unequal Weights

Eventually, the lack of MLE approaches in sample survey is due to the fact the most of samples for large-scale surveys are not independent ones. This section deals with MLE for independent samples with unequal selection probabilities or unequal weights. Now consider a sample of n independent observations with unequal weights. The likelihood function is

$$L(x\mid n,\theta)\propto\theta^{\sum_{i=1}^{n}w_iI_i}(1-\theta)^{\sum_{i=1}^{n}w_i(1-I_i)}, \quad (6)$$

where w_i is the sampling weight for the ith observation, and $I_i=1$ for a success and $I_i=0$ otherwise. The MLE estimate is

$$\hat{\theta}=\frac{\sum_{i=1}^{n}w_iI_i}{\sum_{i=1}^{n}w_i}. \quad (7)$$

The variance of the estimate is

$$\sigma^2_{\hat{\theta}} = \frac{\hat{\theta}(1-\hat{\theta})}{\sum_{i=1}^{n} w_i}. \qquad (8)$$

A Different Interpretation of Sampling Weights

Usually the sampling weight is defined as the inverse of the selection probability. Now we can define the sampling weights as relative contribution to the likelihood given the size of n sample elements. The relative contribution of the ith observation is

$$n \cdot \frac{w_i}{\sum_{i=1}^{n} w_i}. \qquad (9)$$

Stratified Sample

For a stratified sample with H strata, the likelihood function can be expressed as

$$L = \prod_{h=1}^{H} \omega_k L_h , \qquad (10)$$

where ω_h is relative hth stratum size, i.e., $\sum_{h=1}^{H} \omega_h = 1$, and L_h is the likelihood function for hth stratum. The estimate, θ_h, for the hth stratum can be obtained by equating the first derivative of the likelihood function (with respect to θ_h) to 0. The variance of the estimate is estimated by equation (3). As an instance, for independent sample with equal weights within hth stratum,

$$\hat{\theta}_h = \frac{x_h}{n_h}, \qquad (11)$$

$$\sigma^2_{\hat{\theta}_h} = \frac{\hat{\theta}_h(1-\hat{\theta}_h)}{n_h}. \qquad (12)$$

Now the population estimate is a linear combination of stratum-specific estimates, i.e.,

$$\hat{\theta} = \sum_{h=1}^{H} \omega_h \hat{\theta}_h . \qquad (13)$$

The variance of the estimate is

$$\sigma^2_{\hat{\theta}} = \sum_{h=1}^{H} \omega_h^2 \frac{\hat{\theta}_h(1-\hat{\theta}_h)}{n_h}. \qquad (14)$$

Concluding Remarks

The author strongly suggests that MLE methods can be applied to “less” complex data. Ultimately, applicability of MLE methods in sample survey depends on one fundamental foundation: specification of likelihood function for related data. At this juncture, there no clear solution for this big question. There might be a good solution for the big question. Not perfect but decent!

REFERENCES

Cochran, W. G. 1977. *Sampling Techniques, 3rd ed.* New York: Wiley.

Fisher, R. A. 1922. “On the mathematical foundations of theoretical statistics.” *Philosophical Transactions of the Royal Society of London, Series A,* Vol. 222 Pp. 309 – 368.