Stat 8112 Lecture Notes

Unbiased Estimating Equations

Charles J. Geyer

April 29, 2012

1 Introduction

In this handout we generalize the notion of maximum likelihood estimation to solution of unbiased estimating equations. We are much less formal in this handout, merely giving a broad overview. Unlike in Geyer (preprint) there is no "no-n" version of these asymptotics (and as far as I can see there cannot be). Thus these asymptotics are based on the law of large numbers (LLN) and the central limit theorem (CLT), and n is sample size.

The mathematics we explicitly show in this handout will be for independent and identically distributed (IID) data. If one has non-IID data, then one must use an LLN or a CLT for such data.

Suppose X_1, X_2, ... are IID and $g(x, \theta)$ is some continuously differentiable function of data and the the parameter that satisfies

$$E_\theta\{g(X_i, \theta)\} = 0, \qquad \text{for all } \theta. \tag{1}$$

$S(\theta)$

Write

$$h_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta).$$

We seek estimators satisfying

$$h_n(\hat{\theta}_n) = 0. \tag{2}$$

If g is a vector-to-vector function, then so is h_n and thus we say (2) are estimating equations (plural), thinking of each component of (2) as one scalar equation. We say the estimating equations are *unbiased* if

$$E_\theta\{h_n(\theta)\} = 0, \qquad \text{for all } \theta, \tag{3}$$

which follows from (1). The terminology is a bit different from the usual applications of unbiasedness. Clearly (3) says $h_n(\theta)$ is an unbiased estimator of zero if θ is the true unknown parameter value, but we usually don't think of random variables containing unknown parameters as "estimators." Nevertheless, this is the accepted terminology for saying that (3) holds.

One application of unbiased estimating equations takes $h_n(\theta) = \nabla l_n(\theta)$, where l_n is the log likelihood. But this isn't a generalization of maximum likelihood; it is maximum likelihood. More applications will arrive in due course.

In this handout we will assume we can expand the estimating equations in a Taylor series with negligible error

$$0 = h_n(\hat{\theta}_n) = h_n(\theta_0) + \big[\nabla h_n(\theta_0)\big](\hat{\theta}_n - \theta_0) + o_p(n^{-1/2}), \tag{4}$$

where θ_0 is the true unknown parameter value. In this handout are vague about how one might establish (4), which is not at all obvious. In Geyer (preprint, Appendix C) a lot of work goes into establishing it from more easily verifiable assumptions.

From the CLT we have

$$n^{1/2} h_n(\theta_0) \xrightarrow{w} \text{Normal}(0, V), \tag{5}$$

where

$$V = \overset{J(\theta)}{\operatorname{var}_{\theta_0}\{g(X_i, \theta_0)\}}.$$

From the LLN we have

$$-\nabla h_n(\theta_0) \xrightarrow{w} U, I \tag{6}$$

where

$$I \quad U = -E_{\theta_0}\{\nabla g(X_i, \theta_0)\}.$$

In the theory of maximum likelihood we have $V = U$ by the second Bartlett identity (Ferguson, 1996, p. 120). Here U need not even be a symmetric matrix and even if it is, there is nothing to make $V = U$ hold and, in general, $U \neq V$.

We do assume that U and V have the same dimensions, so U is square and h_n is a vector-to-vector function between vector spaces of the same dimension, and there are the same number of estimating equations as parameters in (2). This does not assure that solutions of the estimating equations exist, nor does it assure that solutions are unique if they exist, but uniqueness is impossible without as many estimating equations as parameters. We also assume that U and V are both nonsingular.

We can rewrite (4) as

$$n^{1/2}(\hat{\theta}_n - \theta_0) = \big[-\nabla h_n(\theta_0)\big]^{-1}\big[n^{1/2} h_n(\theta_0)\big] + o_p(1),$$

from which by Slutsky's theorem, we get

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{w} U^{-1} Z,$$

where Z is a normal random vector having the distribution on the right side of (5). The distribution of $U^{-1}Z$ is normal with mean zero and variance $U^{-1}V(U^{-1})^T$.

Thus, under our assumptions, (4) plus U and V being nonsingular, we have

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{w} \text{Normal}\bigl(0, U^{-1}V(U^{-1})^T\bigr) \tag{7}$$

(and that is the theory of unbiased estimating equations).

Note that if we are doing maximum likelihood so $U = V$, we have U symmetric and $U^{-1}VU^{-1} = V^{-1}$, and (7) gives the usual asymptotic distribution for the MLE, normal with mean zero and variance inverse Fisher information. But for general estimating equations $U \neq V$ and $U \neq U^T$, so the variance of the normal distribution in (7) does not simplify.

The matrix $U^T V^{-1} U$ is sometimes called the *Godambe information* matrix because its specialization to the theory of maximum likelihood is U or V (because $U = V$ in maximum likelihood), which is the Fisher information matrix, and Godambe initiated the theory of unbiased estimating equations (more than that, Godambe studied *efficient* unbiased estimating equations in which the Godambe information is as large as possible so the asymptotic variance of the estimator is as small as possible). We can say that the variance of the normal distribution in (7) is inverse Godambe information.

Of course, U and V depend on the unknown parameter and hence are themselves unknown and must be estimated. Equation (6) suggests

$$U_n = -\nabla h_n(\hat{\theta}_n) \qquad I_{obs}$$

as a natural estimator of U, but

$$U_n \xrightarrow{w} U \tag{8}$$

does not follow from (6) and (7), so we will just add it as another assumption. The natural estimator of V is the empirical variance with the estimator plugged in

$$V_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i, \hat{\theta}_n) g(X_i, \hat{\theta}_n)^T,$$

but

$$V_n \xrightarrow{w} V \tag{9}$$

does not follow from (5) and (7), so we will just add it as another assumption. Then by Slutsky's theorem we have

$$\bigl[U_n^{-1} V_n (U_n^{-1})^T\bigr]^{-1/2} \bigl[n^{1/2}(\hat{\theta}_n - \theta_0)\bigr] \xrightarrow{w} \text{Normal}(0, I),$$

-1

I_{obs}

which we write sloppily as

$$\hat{\theta}_n \approx \text{Normal}\bigl(\theta_0, n^{-1} U_n^{-1} V_n (U_n^{-1})^T\bigr). \tag{10}$$

Another name for the estimator $U_n^{-1} V_n (U_n^{-1})^T$ of the asymptotic variance is *sandwich* estimator (think of U_n as slices of bread and V_n as ham).

2 Misspecified Maximum Likelihood

One application of the theory of unbiased estimating equations is the theory of misspecified maximum likelihood, that is, maximum likelihood done when the model is wrong and the true unknown distribution of the data is not any distribution in the model.

Let λ denote the Kullback-Leibler information function, defined by

$$\lambda(\theta) = E_f \left(\log \frac{f_\theta(X)}{f(X)} \right), \tag{11}$$

where f is the true unknown density of the data. Suppose λ achieves its maximum over the parameter space at some point θ_0 that is in the interior of the parameter space so

$$\nabla \lambda(\theta_0) = 0.$$

Define

$$g(x, \theta) = \nabla_\theta \log f_\theta(x).$$

Assuming we can move a derivative inside the expectation in (11) we have (1) where

$$E_f\{g(X, \theta_0)\} = 0, \tag{12}$$

and this is enough to get the theory of unbiased estimating equations going. Note that (12) is not quite the same as (1), but we do have (12) for all θ_0 that can arise as described (corresponding to some true unknown distribution f).

The theory of misspecified maximum likelihood is a bit odd in that we are not estimating the true unknown parameter value. There is no true unknown *parameter* value, because the true unknown distribution is not in our parametric model. We are estimating θ_0, which is the parameter value specifying the distribution in the model which is closest to the true unknown distribution in the sense of maximizing Kullback-Leibler information.

We can say that even when the model is misspecified the MLE is a consistent and asymptotically normal estimator of θ_0 and the asymptotic variance is inverse Godambe information (estimated by the sandwich estimator). The only simplification that arises in this situation is that

$$U = E_f\{\nabla^2 l_n(\theta_0)\},$$

so U is a symmetric matrix and we can write $UV^{-1}U$ for the Godambe information matrix and $U_n^{-1}V_nU_n^{-1}$ for the sandwich estimator (omitting transposes), but $U \neq V$ when the model is misspecified so these do not simplify further.

3 Composite Likelihood

It may come as a surprise to those who have no exposure to spatial statistics, but there are statistical models that are easily described but for which it is impossible in practice to evaluate the likelihood. Here is a simple example.

The data for an Ising model are an $r \times c$ matrix of two-valued random variables Y_{ij}. For mathematical convenience, we let the values be -1 and $+1$, and also define $Y_{ij} = 0$ for i and j outside the range of values of indices for this matrix.

Consider two statistics

$$\begin{aligned} t_1(Y) &= \sum_{i=1}^{r}\sum_{j=1}^{c} Y_{ij} \\ t_2(Y) &= \sum_{i=1}^{r}\sum_{j=1}^{c}\bigl(Y_{ij}Y_{i,j+1} + Y_{ij}Y_{i+1,j}\bigr) \end{aligned}$$

We may think of the matrix Y as a black and white image with -1 coding black pixels and $+1$ coding white pixels, in which case $t_1(Y)$ is the total number of white pixels minus the total number of black pixels and $t_2(Y)$ is the total number of concordant (same color) neighbor pairs minus the total number of discordant (different color) neighbor pairs, where pixels are *neighbors* if they are adjacent either horizontally or vertically. The Ising model is the full exponential family of distributions that has these two natural statistics and contains the distribution that makes all 2^{rc} possible data matrices equally likely.

In theory, this model is very simple. Like all exponential families, it has log likelihood

$$l(\theta) = \langle t(Y), \theta \rangle - c(\theta),$$

where

$$c(\theta) = \log \left(\sum_{y \in \mathcal{Y}} e^{\langle t(y), \theta \rangle} \right), \tag{13}$$

and where $\mathcal{Y}$ is the set of all 2^{rc} possible data matrices. The sum with 2^{rc} terms in (13) does not simplify by any known method and hence is undoable by any means other than brute force summation over all 2^{rc} terms, which is completely impractical when rc is more than 100, even if one had all the computers in the world harnessed to the task.

The method of composite likelihood (Lindsay, 1988) is a generalization of the method of pseudo-likelihood (Besag, 1974, 1975), which was designed specifically to tackle problems like this. Varin, Reid, and Firth (2011) review the current state of composite likelihood theory and practice. Okabayashi, Johnson and Geyer (2011) apply composite likelihood to the Ising model.

The general idea of composite likelihood is the following. Suppose we have a statistical model in which the likelihood is difficult to compute, which means we cannot compute the joint density for arbitrary values of the data and parameter. But suppose we can calculate some marginal or conditional densities derived from the joint density. Suppose we can calculate the conditional density of $r_k(Y)$ given $s_k(Y)$ for $k = 1, \ldots, m$. If $r_k(Y)$ and $s_k(Y)$ are stochastically independent, for example, when s_k is a constant function, then the conditional density of $r_k(Y)$ given $s_k(Y)$ is the same as the marginal density of $r_k(Y)$. Thus we can use the same notation for both conditional and marginal densities.

Let $f_{k,\theta}$ denote the conditional density of $r_k(Y)$ given $s_k(Y)$. Considered as a function of the parameter with the observed data plugged in, this is *a* log likelihood; it just isn't *the* log likelihood for the given statistical model. In particular, we have the first Bartlett identity

$$E_\theta \bigl\{ \nabla_\theta \log f_{k,\theta}\bigl(r_k(Y) \mid s_k(Y)\bigr) \bigr\} = 0, \qquad \text{for all } \theta.$$

Since the expectation of a sum is the sum of the expectations, regardless of whether the terms are stochastically dependent or independent, we also have

$$E_\theta \left\{ \sum_{k=1}^{m} \nabla_\theta \log f_{k,\theta}\bigl(r_k(Y) \mid s_k(Y)\bigr) \right\} = 0, \qquad \text{for all } \theta.$$

Thus

$$\sum_{k=1}^{m} \nabla_\theta \log f_{k,\theta}\bigl(r_k(Y) \mid s_k(Y)\bigr) = 0 \tag{14}$$

are unbiased estimating equations to be solved for an estimate $\hat{\theta}$ of the parameter. This leads us to define the function the left side of (14) is the derivative of

$$l(\theta) = \sum_{k=1}^{m} \log f_{k,\theta}\bigl(r_k(Y) \mid s_k(Y)\bigr), \tag{15}$$

which is called a *composite likelihood* for the problem. Any maximizer of (15), necessarily a solution of the estimating equations (14), is called the *maximum composite likelihood estimator* (MCLE).

The method of pseudo-likelihood applied to the Ising model needs double subscripts, so we write $r_{ij}(Y)$ and $s_{ij}(Y)$ rather than $r_k(Y)$ and $s_k(Y)$. We take $r_{ij}(Y) = Y_{ij}$ and we let $s_{ij}(Y)$ be the matrix that is the same as Y except that the i, j element is zero. Thus $r_{ij}(Y)$ is the data for the i, j pixel and $s_{ij}(Y)$ is the data for all other pixels. It is clear that since the log likelihood contains terms only involving single pixels and neighbor pairs of pixels, that the conditional distribution of Y_{ij} given the rest of the data only involves the four pixels that are neighbors of the i, j pixel. Furthermore, since Y_{ij} has only two possible values, normalizing its conditional distribution involves a sum with only two terms. Define

$$X_{ij} = Y_{i,j+1} + Y_{i,j-1} + Y_{i+1,j} + Y_{i-1,j}.$$

Then

$$f_{i,j,\theta}\bigl(Y_{ij} \mid s_{ij}(Y)\bigr) = \frac{\exp(Y_{ij}[\theta_1 + \theta_2 X_{ij}])}{\exp(Y_{ij}[\theta_1 + \theta_2 X_{ij}]) + \exp(-Y_{ij}[\theta_1 + \theta_2 X_{ij}])} \tag{16}$$

The pseudo-likelihood is just the product of (16). Algebraically, it has the form of the log likelihood for a logistic regression. Since

$$\log \frac{f_{i,j,\theta}\bigl(+1 \mid s_{ij}(Y)\bigr)}{f_{i,j,\theta}\bigl(-1 \mid s_{ij}(Y)\bigr)} = 2[\theta_1 + \theta_2 X_{ij}],$$

we can estimate θ by doing a logistic regression with response vector having components $2Y_{ij} - 1$ (the Y matrix recoded to have values zero and one and strung out in a vector) and one non-constant predictor vector having components $2X_{ij}$ and one constant predictor having components 2.

Using composite likelihoods with $r_{ij}(Y)$ involving more than one component of Y is more complicated but doable (Okabayashi, et al., 2011).

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *Statistician*, **24**, 179–195.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. London: Chapman & Hall.

Geyer, C. J. (preprint). Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, **80**, 220–239.

Okabayashi, S., Johnson, L. and Geyer, C. J. (2011). Extending pseudo-likelihood for Potts models. *Statistica Sinica*, **21**, 331–347.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5–42.