CHAPTER 13

Higher order asymptotics

This chapter covers some theory of higher order asymptotics. The results described here show how to adjust empirical likelihood confidence intervals to have good one-sided coverage errors for one-dimensional parameters, and how to obtain similar effects for multidimensional parameters. Bartlett corrections are also presented, as are some large deviations results.

13.1 Bartlett correction

The profile empirical likelihood ratio statistic commonly has an asymptotic chi-squared distribution: $-2\log\mathcal{R}(\theta_0) \to \chi^2_{(p)}$ in distribution as $n \to \infty$, where p is the dimension of θ. A Bartlett correction is made by replacing the threshold $\chi^{2,1-\alpha}_{(p)}$, by a scale multiple $(1+an^{-1})\chi^{2,1-\alpha}_{(p)}$. In parametric settings the value a can often be chosen to make the mean of minus twice the log likelihood ratio more nearly equal to p. Surprisingly, one Bartlett correction a improves the asymptotic error rate for all coverage levels $1-\alpha$.

A Bartlett-corrected empirical likelihood confidence region takes either of the following asymptotically equivalent forms

$$\left\{\theta \mid -2\log\mathcal{R}(\theta) \le \left(1+\frac{a}{n}\right)\chi^{2,1-\alpha}_{(p)}\right\}, \quad \text{or}$$
$$\left\{\theta \mid -2\log\mathcal{R}(\theta) \le \left(1-\frac{a}{n}\right)^{-1}\chi^{2,1-\alpha}_{(p)}\right\}.$$

The appropriate value of a is typically unknown. We will consider $\theta \in \mathbb{R}^p$ obtained as a smooth function $h(\mu)$ of the mean of a random variable $X \in \mathbb{R}^d$, where $d \ge p$. For smooth functions of means, the value a can be expressed in terms of moments of X and derivatives of h. Plugging in sample versions of the moments and evaluating the derivatives at sample moments leads to a $n^{1/2}$ consistent estimator $\hat{a}$. This value can be substituted for a, giving regions such as

$$\left\{\theta \mid -2\log\mathcal{R}(\theta) \le \left(1+\frac{\hat{a}}{n}\right)\chi^{2,1-\alpha}_{(p)}\right\},$$

with the same asymptotic order of coverage accuracy as those based on a.

The proof of Bartlett correctability requires Cramér's condition

$$\lim\sup_{\|t\|\to\infty} \left|E\left(\exp\left(it'X\right)\right)\right| < \infty, \tag{13.1}$$

which rules out some distributions supported on lattices like the integers. It also requires finiteness of some moments of X, and the existence of sufficiently many derivatives of h. These conditions are used to justify Edgeworth expansions, from which the Bartlett correctability follows. Under these conditions, Bartlett-corrected confidence regions have coverage $1-\alpha+O(n^{-2})$, instead of $1-\alpha+O(n^{-1})$, as holds for uncorrected regions.

The Bartlett correction for a univariate mean $\theta = E(X)$ is

$$a = \frac{1}{2}\frac{\mu_4}{\mu_2^2} - \frac{1}{3}\frac{\mu_3^2}{\mu_2^3} = \frac{\kappa+3}{2} - \frac{\gamma^2}{3}$$

where $\mu_k = E((X-E(X))^k)$, for integers $k \geq 2$. The skewness γ and kurtosis κ are defined on page 3. The natural sample estimates use $\hat{\mu}_k = (1/n)\sum_{i=1}^n (X_i - \bar{X})^k$. As a point of reference, $a = 1.5$ for univariate normal distributions. For heavy tailed distributions having $\kappa > 0$, a larger Bartlett correction is applied, while nonzero skewness reduces the Bartlett correction. If κ exists, then $\kappa \geq \gamma^2 - 2$, from which we know that $a \geq 1/2 + \gamma^2/6 \geq 1/2$.

A vector mean, $\theta = E(X) \in \mathbb{R}^d$, has $p = d$. Let $Y = V_0^{-1/2}(X-\mu_0)$, where $\mu_0 = \int x dF_0(x)$, and $V_0 = \int (x-\mu_0)(x-\mu_0)' dF_0(x)$ is assumed to be nonsingular. Introduce component superscripts $Y = (Y^1, \ldots, Y^p)$, and define

$$\mu_{jk\ell} = E\left(Y^j Y^k Y^\ell\right), \quad \text{and} \quad \mu_{jk\ell m} = E\left(Y^j Y^k Y^\ell Y^m\right). \tag{13.2}$$

Then the Bartlett correction is given by

$$a = \frac{5}{3}\sum_{j=1}^p\sum_{k=1}^p\sum_{\ell=1}^p (\mu_{jk\ell})^2 - 2\sum_{j=1}^p\sum_{k=1}^p\sum_{\ell=1}^p \mu_{jj\ell}\mu_{kk\ell} + \frac{1}{2}\sum_{j=1}^p\sum_{k=1}^p \mu_{jjkk}.$$

For a normally distributed vector X, the vector Y has the standard normal distribution with all $\mu_{jkl} = 0$, and $\mu_{jjkk} = 1$, so that $a = p(p-1) + 3p/2 = p^2 + p/2 = d^2 + d/2$.

The Bartlett correction for the mean can be computed in $O(nd^3)$ time. If θ is defined through p estimating equations $E(m(X,\theta)) = 0$, then we may apply the Bartlett correction for the mean of $Z_i = m(X_i, \theta)$ in order to determine whether θ should be in the confidence region.

13.2 Bartlett correction and smooth functions of means

Now consider Bartlett correction for a smooth function of a vector mean. As before, let Y be a standardized version of X, and define the moments μ_{jkl} and μ_{jklm} through (13.2). Suppose that $\theta = h(\mu) = (h_1(\mu), \ldots, h_p(\mu))' \in \mathbb{R}^p$, where $\mu = E(X)$. Now for $i = 1, \ldots, p$, let $\psi^i(\nu) = h_i(V_0^{1/2}\nu)$, and define the partial derivatives

$$\psi^i_{j_1,\ldots,j_r} = \left.\frac{\partial^r \psi^i(\nu)}{\partial \nu^{j_1}\cdots\partial\nu^{j_r}}\right|_{\nu = V_0^{-1/2}\mu_0}.$$

Let Ψ be the $p \times d$ matrix with elements ψ^i_j and define the matrices

$$Q = (\Psi\Psi')^{-1}, \quad M = \Psi' Q \Psi, \quad \text{and} \quad N = \Psi' Q.$$

Then the Bartlett correction is

$$a = \frac{1}{p}\left(\frac{5}{3}a_1 - 2a_2 + \frac{1}{2}a_3 - a_4 + \frac{1}{4}a_5\right),$$

where

$$a_1 = \sum_{j,k,\ell,m,n,o} \mu_{jk\ell}\mu_{mno} M^{jm} M^{kn} M^{\ell o},$$

$$a_2 = \sum_{j,k,\ell,m,n,o} \mu_{jk\ell}\mu_{mno} M^{jk} M^{\ell m} M^{no},$$

$$a_3 = \sum_{j,k,\ell,m} \mu_{jk\ell m} M^{jk} M^{\ell m},$$

$$a_4 = \sum_{j,k,\ell,m,n,u} \mu_{jk\ell} N^{ju} \psi^u_{mn} (I-M)^{mk} (I-M)^{n\ell},$$

$$a_5 = \sum_{\substack{j,k,\ell \\ m,u,v}} Q^{uv} \psi^u_{jk} \psi^v_{\ell m} \left[(I-M)^{jk}(I-M)^{\ell m} + 2(I-M)^{j\ell}(I-M)^{km}\right],$$

indicating the elements of M, N, and Q with superscripts, and summing every index over its entire range.

The moments required by the Bartlett correction for a smooth function of means are dominated by the $\mu_{jk\ell m}$ in the a_3 term. These take $O(nd^4)$ time to compute. The cost of the six-fold summations used in the Bartlett correction does not grow with n, but assuming that $p \le d$, this cost is $O(d^6)$. For d as large as 10 or 20 this cost could be significant. For the vector mean itself, the cost is of the lower order $O(nd^3)$, suggesting that there may be a practical computational savings to Bartlett correcting estimating equations instead of smooth functions of means.

13.3 Pseudo-likelihood theory

Empirical likelihood provides a data-determined shape for confidence regions. For statistics that are smooth functions of means, this shape is asymptotically ellipsoidal. For a fixed sample size and confidence level, the confidence regions are not exactly ellipsoidal, and in some instances are far from ellipsoidal. For a mean, the confidence regions are elongated in the directions of greater skewness.

The coverage errors in empirical likelihood are typically $O(n^{-1})$, and this rate can also be achieved by ellipsoidal confidence regions. Pseudo-likelihood theory shows that the shape of the empirical likelihood confidence regions is informative. The contours of the profile empirical likelihood ratio for θ are close to contours of the probability density function of an approximate pivotal statistic.

The case $p = 1$ is simplest. Let $\theta = h(\mu)$, where h is a smooth function from $\mathbb{R}^d$ to $\mathbb{R}$ and $\mu = \int x dF(x)$. The true value is $\theta_0 = h(\mu_0)$, estimated by $\hat{\theta} = h(\bar{X})$. Let

$$\hat{\eta}_0 = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\hat{\sigma}}, \tag{13.3}$$

where $\hat{\sigma}^2$ is a $n^{1/2}$ consistent estimator of $\sigma^2 = \mathrm{Var}(n^{1/2}(\hat{\theta} - \theta_0))$. Let f_η be the probability density function of $\hat{\eta}_0$. For a properly chosen constant ψ described below, the set

$$C_n^{1-\alpha} = \left\{\theta + \psi\sigma n^{-1} \mid -2\log\mathcal{R}(\theta) \le \chi_{(1)}^{2,1-\alpha}\right\}$$

closely matches the set

$$P_\eta^{1-\alpha} = \left\{\theta \mid f_\eta\left(\frac{\sqrt{n}(\hat{\theta} - \theta)}{\hat{\sigma}}\right) \ge v^{1-\alpha}\right\}$$

where $v^{1-\alpha}$ satisfies

$$\int_{P_\eta^{1-\alpha}} f_\eta(z)dz = 1 - \alpha.$$

Both $C_n^{1-\alpha}$ and $P_\eta^{1-\alpha}$ are asymptotically intervals, and their endpoints differ only by $O(n^{-3/2})$. If one shifts the contours (here endpoints) of the empirical likelihood function by $\psi\sigma n^{-1}$, then to high accuracy they match the contours of the density of $\hat{\eta}_0$.

For $\theta \in \mathbb{R}^p$, with $p \ge 1$, let $\hat{V}$ denote a $n^{1/2}$ consistent estimator of $V = \mathrm{Var}(n^{1/2}(\hat{\theta} - \theta_0))$, and let

$$\hat{\eta}_0 = \sqrt{n}\left(V^{1/2}\hat{V}^{-1}V^{1/2}\right)^{1/2} V^{-1/2}\left(\hat{\theta} - \theta_0\right). \tag{13.4}$$

Definition (13.4) reduces to (13.3) when $p = 1$. Now let

$$C_n^{1-\alpha} = \left\{\theta + \psi V^{1/2} n^{-1} \mid -2\log\mathcal{R}(\theta) \le \chi_{(1)}^{2,1-\alpha}\right\},$$

for a properly chosen vector ψ (see Chapter 13.6), and let

$$P_\eta^{1-\alpha} = \left\{\theta \mid f_\eta\left(\sqrt{n}\left(V^{1/2}\hat{V}^{-1}V^{1/2}\right)^{1/2} V^{-1/2}\left(\hat{\theta} - \theta\right)\right) \ge v^{1-\alpha}\right\},$$

with $v^{1-\alpha}$ chosen as before. Then the boundaries of $C_n^{1-\alpha}$ and $P_\eta^{1-\alpha}$ agree to $O(n^{-3/2})$.

The interpretation is that empirical likelihood contours are very close to those we would construct if we knew the density of $\hat{\eta}_0$. They have the same shape, size, and orientation as the contours of the density of $\hat{\eta}_0$, but they need to be shifted by a bias correction of order $1/n$ to make the match accurate. The term pseudo-likelihood is used because it is the density of an observed statistic that is being studied, as opposed to the probability or density of the data.

For a multivariate mean, $\psi = (\psi_1, \dots, \psi_p)'$ where

$$\psi_j = -\frac{1}{2}\sum_{r=1}^{p} \mu_{jrr},$$

and μ_{jrr} is defined at (13.2). For more general location adjustments, see Chapter 13.6.

Pseudo-likelihood arguments suggest shifting the confidence regions for θ by an amount of order $1/n$. The resulting confidence regions can also be Bartlett corrected, although a different Bartlett constant is required than for unshifted confidence regions.

13.4 Signed root corrections

Here we consider parameters defined as smooth function of means: $X \in \mathbb{R}^d$, $\mu = E(X)$, and $\theta = \theta(\mu) \in \mathbb{R}^p$, for $p \leq d$. The random vector X is assumed to satisfy Cramér's condition, to have sufficiently many finite moments, and to have a variance matrix of full rank d. The function $\theta(\mu)$ is assumed to have sufficiently many derivatives, and to have a gradient of full rank p at the true mean μ_0.

For smooth functions of means, the coverage error is typically $O(n^{-1})$ for empirical likelihood confidence regions. For a scalar-valued parameter θ, this $O(n^{-1})$ coverage error arises as two one-sided coverage errors of size $O(n^{-1/2})$ that nearly cancel each other. Bartlett correction reduces the two-sided error to $O(n^{-2})$, but leaves the one-sided error at $O(n^{-1/2})$. In some applications it is desirable to get smaller one-sided coverage errors, even if this means that some excluded parameter values have higher likelihood than some that are included.

This same phenomenon holds in parametric likelihood. Confidence intervals found by thresholding the likelihood function typically have coverage errors of order $O(n^{-1})$ for two-sided inferences but $O(n^{-1/2})$ for one-sided inferences. Bartlett correction improves the two-sided error rate but not the one-sided one. When $p = 1$, one solution in both parametric and empirical likelihoods is to consider the signed root of the log likelihood ratio statistic. For an empirical likelihood, let

$$R_0 = n^{-1/2}\mathrm{sign}(\hat{\theta} - \theta_0)\sqrt{-2\log \mathcal{R}(\theta_0)},$$

so that the standard $\chi^2_{(1)}$ calibration is based on $n^{1/2}R_0 \dot{\sim} N(0,1)$. Sharper inferences can be obtained by comparing

$$\frac{n^{1/2}(R_0 - \hat{a}/n)}{1 + \hat{b}/(2n)} \qquad (13.5)$$

to $N(0,1)$, where $\hat{a}$ and $\hat{b}$ are mean and variance adjustments to the signed root R_0. The adjusted signed root (13.5) has the $N(0,1)$ distribution to $O(n^{-3/2})$, and so it can be used to construct one-sided confidence intervals with coverage error $O(n^{-3/2})$.

For a univariate mean, the signed root correction uses sample analogs of

$$a = -\frac{\gamma}{6}, \quad \text{and} \quad b = \frac{5}{6}\kappa - \frac{31}{36}\gamma^2 + \frac{3}{2},$$

where γ and κ are the skewness and kurtosis of X. Corrections for more general smooth scalar functions of means are known to exist, but formulas for them have not been developed. The proper formula for b is likely to be awkward. It may be possible to use a form of bootstrap calibration to find $\hat{a}$ and $\hat{b}$.

For general $p \geq 1$, the signed root is constructed as a vector $R_0 \in \mathbb{R}^p$ such that

$$R_0' R_0 = -\frac{2}{n} \log \mathcal{R}(\theta_0).$$

The vector $n^{1/2} R_0$ has the $N(0, I_p)$ distribution to order $O(n^{-1/2})$, but after a mean shift the approximation improves to $O(n^{-1})$. Computing R_0 and the mean correction is awkward. A simpler computation with the same order of accuracy is to take the confidence region

$$\left\{\theta \mid -2 \log \mathcal{R}(\theta) + \widehat{C}(\theta) \leq \chi^{2,1-\alpha}_{(p)}\right\}, \tag{13.6}$$

where $\widehat{C}$ is described below. The region (13.6) has coverage error $O(n^{-1})$, just as ordinary empirical likelihood does, but when specialized to $p = 1$ it has one-sided errors of order $O(n^{-1})$.

Suppose that the i'th data point is X_i with components X_i^j for $j = 1, \ldots, d$, and that $\theta = \theta(\mu)$ has components θ^u for $u = 1, \ldots, p$. The expression for the adjustment quantity $\widehat{C}$ depends on moments

$$\begin{aligned} \bar{X}^j &= \frac{1}{n}\sum_{i=1}^n X_i^j, \quad \widehat{\Omega}^{jk} = \frac{1}{n}\sum_{i=1}^n (X_i^j - \bar{X}^j)(X_i^k - \bar{X}^k), \quad \text{and} \\ \widehat{\alpha}^{jk\ell} &= \frac{1}{n}\sum_{i=1}^n \sum_{r,s,t=1}^d (\widehat{\Omega}^{-1/2})^{jr} (\widehat{\Omega}^{-1/2})^{ks} (\widehat{\Omega}^{-1/2})^{\ell t} \\ &\qquad \times (X_i^r - \bar{X}^r)(X_i^s - \bar{X}^s)(X_i^t - \bar{X}_i^t), \end{aligned}$$

where $\widehat{\Omega}$ is the $d \times d$ matrix of $\widehat{\Omega}^{jk}$ values, on the partial derivatives

$$\widehat{\theta}_j^u = \left.\frac{\partial \theta^u(\mu)}{\partial \mu^j}\right|_{\mu = \bar{X}}, \quad \widehat{\theta}_{jk}^u = \left.\frac{\partial^2 \theta^u(\mu)}{\partial \mu^j \partial \mu^k}\right|_{\mu = \bar{X}}$$

and on the related matrices

$$\begin{aligned} \widehat{\Theta} &= (\widehat{\theta}_j^u), & \widehat{Q} &= (\widehat{\Theta}\widehat{\Omega}\widehat{\Theta}')^{-1}, \\ \widehat{N} &= \widehat{\Omega}^{1/2}\widehat{\Theta}'\widehat{Q}, & \widehat{M} &= \widehat{N}\widehat{\Theta}\widehat{\Omega}^{1/2}. \end{aligned}$$

The dimension of $\widehat{\Theta}$ is $q \times d$. In terms of these quantities, we may write

$$\widehat{C}(\theta) = \sum_{j,k,\ell,u,v} \left(\frac{1}{3}\widehat{\alpha}^{jkl}\widehat{M}^{jk}\widehat{N}^{lu} - \widehat{Q}^{uv}\widehat{\theta}^{v}_{jk}(I - \widehat{M})^{jk} \right) \left(\widehat{\theta} - \theta\right)^{u}. \qquad (13.7)$$

Equation (13.7) requires a summation over five variables. Exercise 13.2 shows that for a vector mean, as one might use in conjunction with estimating equations, that the sum is simpler and only over three variables. For a scalar mean $\hat{C}(\mu_0) = (1/3)(\bar{X} - \mu_0)\hat{\mu}_3/\hat{\mu}_2^2$ for sample second and third central moments $\hat{\mu}_2$ and $\hat{\mu}_3$.

13.5 Large deviations

The power of tests that $\theta = \theta_0 \in \mathbb{R}^p$ is usually investigated for Pitman alternatives of the form $\theta_0 + n^{-1/2}\tau$. This is a critical distance. Alternatives that are $o(n^{-1/2})$ cannot be distinguished from θ_0 in large samples, the power to reject Pitman alternatives usually tends to a limit (depending on τ), and for alternatives that are not $O(n^{-1/2})$ away from the null, the power typically approaches 1 as $n \to \infty$.

The asymptotic power at Pitman alternatives is usually the same for empirical, Euclidean, and the other likelihoods. A sharper distinction between tests can be drawn by considering a fixed alternative $\theta_1 \neq \theta_0$. For such alternatives, tests can be constructed at the level $\alpha_n = \exp(-n\eta)$ with power $1 - \beta_n = 1 - \exp(-n\gamma)$. That is, both type I and type II errors converge to zero exponentially fast in n. Because the errors are so small, they correspond to large deviations of the underlying test statistics.

For multinomial samples $X_1, \dots, X_n \in \{z_1, \dots, z_k\}$ the likelihood ratio test is optimal under the generalized Neyman-Pearson (GNP) criterion. Among tests that satisfy $\limsup_{n\to\infty} n^{-1}\log(\alpha_n) \leq -\eta$ for some $\eta > 0$, it minimizes $\limsup_{n\to\infty} n^{-1}\log(\beta_n)$. This property is referred to as universal optimality, because it holds for very general hypotheses and for very general alternatives to F.

For samples from a possibly continuous distribution on $\mathbb{R}^d$, the original argument establishing the GNP for multinomial likelihood breaks down. But other arguments can be used to show a kind of GNP optimality for empirical likelihood based tests of composite hypotheses under sampling from continuous distributions.

Here we quote the large deviations optimality result for empirical likelihood. This discussion leaves out measure theoretic details. For those, see the reference in Chapter 13.6.

Suppose that $X_1, \dots, X_n \in \mathbb{R}^d$ are independent random vectors with common distribution F. For $\theta \in \Theta \subseteq \mathbb{R}^p$, let $m : \mathbb{R}^d \times \mathbb{R}^p \longrightarrow \mathbb{R}^q$ be an estimating function placing the constraint $E(m(X,\theta)) = \int m(X,\theta)dF(X) = 0$ on F.

For the rest of this section, θ is a freely varying nuisance parameter, and the rest of the hypothesis is specified in m. As a concrete example, to test the hypothesis that $X \in \mathbb{R}$ has variance 7.2, we take

$$m(X,\theta) = \left(\begin{array}{c} X - \theta \\ (X-\theta)^2 - 7.2 \end{array} \right) \in \mathbb{R}^2, \quad \theta \in \Theta \subset \mathbb{R}.$$

Let $\mathcal{F}$ be the set of probability distributions on $\mathbb{R}^d$. Let

$$\mathcal{F}_\theta = \left\{ F \in \mathcal{F} \mid \int m(X,\theta) dF(X) = 0 \right\}, \quad \text{and} \quad \mathcal{H}_0 = \bigcup_{\theta \in \Theta} \mathcal{F}_\theta.$$

The empirical likelihood test of the hypothesis $F \in \mathcal{H}_0$, which specifies that $E(m(X,\theta)) = 0$ for some $\theta \in \Theta$, is based on

$$\mathcal{R}(\mathcal{H}_0) = \sup_{\theta \in \Theta} \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i m(X_i,\theta) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}.$$

The empirical likelihood test rejects $\mathcal{H}_0$ if and only if $\mathcal{R}(\mathcal{H}_0)$ is small. Because this test depends on the data only through the empirical distribution function $\hat{F}$, we may write the rejection region itself as a set $\Lambda = \Lambda_n \subset \mathcal{F}$ of distributions. The empirical likelihood test rejects $\mathcal{H}_0$ if and only if $\hat{F} \in \Lambda_n$.

Let $\Omega = \Omega_n \subset \mathcal{F}$ be the rejection region for some other test. We introduce the concept of a Lévy-ball around a set of distributions, to define a regularity condition on Ω. For distributions F and G, let 1 be a vector of d ones. Then the Lévy distance between F and G is

$$\rho_{F,G} = \inf \left\{ \epsilon > 0 \mid F((-\infty, X - \epsilon 1]) - \epsilon \leq G(X) \leq F((-\infty, X + \epsilon 1]) + \epsilon \right\}.$$

Suppose that $X \sim F$ and that $Y = Y(X)$. For $\rho_{F,G}$ to be small, it must hold that $\|Y(X) - X\|$ is small with probability close to 1 under $X \sim F$. We can keep G within a small Lévy distance of F by making large changes $Y(X) - X$ to a small (under F) fraction of the X values, and/or small changes to a large fraction of the X values.

The rejection region Ω_n is a set of empirical distributions. The Lévy ball of radius $\delta > 0$ around Ω is

$$\Omega^\delta = \bigcup_{F \in \Omega} \left\{ G \mid \rho_{F,G} < \delta \right\}.$$

The test sequence Ω_n is regular for $\mathcal{H}_0$ if

$$\lim_{\delta \to 0} \sup_{F \in \mathcal{H}_0} \limsup_{n \to \infty} \frac{1}{n} \log \Pr(\hat{F} \in \Omega_n^\delta; F) \leq \sup_{F \in \mathcal{H}_0} \limsup_{n \to \infty} \frac{1}{n} \log \Pr(\hat{F} \in \Omega_n; F).$$

For example, the test Ω_n that rejects $\mathcal{H}_0$ if and only if $\mathcal{R}(\mathcal{H}_0)$ is a rational number is not regular.

Theorem 13.1 *Suppose that*

A: *For all* $F \in \mathcal{H}_0$, $\Pr(\sup_{\theta \in \Theta} \|m(X,\theta)\| = \infty; F) = 0$,

B: $m(x,\theta)$ *is continuous in* θ *for all* x,

C: *and let* $\Lambda = \{\hat{F} \mid \frac{1}{n} \log \mathcal{R}(\mathcal{H}_0) \leq -\eta\}$ *for some* $\eta > 0$.

Then

$$\sup_{F \in \mathcal{H}_0} \limsup_{n \to \infty} \frac{1}{n} \log \Pr(\hat{F} \in \Lambda; F) \leq -\eta. \tag{13.8}$$

If (13.8) *holds with* Λ *replaced by any regular (for* $\mathcal{H}_0$*) test* Ω*, then for all* $F \in \mathcal{F}$*,*

$$\limsup_{n \to \infty} \frac{1}{n} \log \Pr(\hat{F} \notin \Omega_n; F) \geq \limsup_{n \to \infty} \frac{1}{n} \log \Pr(\hat{F} \notin \Lambda_n; F). \tag{13.9}$$

Equation (13.9) *also holds for a test* Ω*, not necessarily regular for* $\mathcal{H}_0$*, if that test satisfies* (13.8) *with* Λ *replaced by* Ω^δ *for some* $\delta > 0$.

Proof. Kitamura (2001). □

In this theorem, the empirical likelihood test rejects $\mathcal{H}_0$ if and only if $\mathcal{R}(\mathcal{H}_0) \leq \exp(-n\eta)$. Result (13.8) compares the type I error α_n to $\exp(-n\eta)$. Result (13.9) shows that any regular test for $\mathcal{H}_0$ that also meets condition (13.8) has, asymptotically, a type II error at least as large as the empirical likelihood test has. Result (13.9) is universal, applying to any sampling distribution F, any regular test Ω, and very general hypotheses $\mathcal{H}_0$.

Condition A requires a bound in θ on $m(X, \theta)$ to hold with probability 1. Such a bound may require m to be a bounded function, or Θ to be a bounded domain.

13.6 Bibliographic notes

Adjustments

Most of the results on higher order asymptotics presented here were based on Edgeworth expansions for smooth functions of means. Bhattacharya & Ghosh (1978) established the validity of these expansions assuming Cramér's condition and that certain moments are finite.

The first Bartlett correction for empirical likelihood was published by DiCiccio et al. (1991) for smooth functions of means. The material on Bartlett correction for the mean is based on Hall & La Scala (1990).

Zhang (1996*b*) showed that the Bartlett correction for the univariate mean can be applied for $\theta \in \mathbb{R}$ defined through the estimating function $m(X, \theta) \in \mathbb{R}$. Lazar & Mykland (1999) showed that Bartlett correctability does not hold for empirical profile likelihoods obtained by maximizing over some nuisance parameters. This is the principle way in which empirical likelihood shows different behavior from parametric likelihoods. Mykland (1999) traces the discrepancy to a condition on the fifth moment of the signed square root of the empirical log likelihood.

Chen & Hall (1993) studied Bartlett correction of quantiles, and Chen (1993) established a Bartlett correction for regression with nonrandom predictors. Jing & Wood (1996) show that exponential empirical likelihood (empirical entropy) is not Bartlett correctable. Baggerly (1998) shows that empirical likelihood is the only member of the Cressie-Read family to be Bartlett correctable. Corcoran

(1998) constructed other Bartlett correctable nonparametric likelihoods given by (3.39).

The pseudo-likelihood theory for empirical likelihood was established by Hall (1990). The presentation here also makes use of the account in Hall & La Scala (1990).

Signed root corrections to empirical likelihood were established by DiCiccio & Romano (1989). Signed root corrections for parametric likelihoods are given by DiCiccio (1984) and by Barndorff-Nielsen (1986). The additive correction (13.7) is from DiCiccio & Romano (1989). McCullagh (1984) presents a similar correction for parametric likelihood. DiCiccio & Monti (2001) present adjustments for a scalar parameter based on third and fourth derivatives of the empirical log likelihood.

Large deviations

Dembo & Zeitouni (1998) is a monograph on large deviations. The generalized Neyman-Pearson result for finite multinomials, established by Hoeffding (1965), is of large deviations type. Tusnády (1977) extends Hoeffding's idea to more general distributions, by using finite partitions of the sample space that get finer as n increases. He finds that a likelihood ratio test on the partitions is asymptotically optimal (in Bahadur's sense). No construction is given for the partition. His methods require that the partition have $o(n/\log(n))$ elements.

Kitamura (2001) proves Theorem 13.1. He reports some simulations comparing empirical likelihood and three versions of generalized method of moments: 2-step, 10-step, and continuous updating (Euclidean likelihood). The problem had $n = 200$ observations and 3 parameters. A simulation at the null hypothesis showed that every method gave confidence regions that undercovered the true parameter. After an adjustment to make coverage 95%, the power was compared in simulations that varied the parameters at 8 places along 4 line segments through the null. Of 32 simulations empirical likelihood had the greatest power 22 times, 2-step updating did this 5 times, 10-step updating 7 times, and continuous updating never had the greatest power. There were two cases in which empirical likelihood and 2-step updating tied at power 1.0. As might be expected for a large deviations result like Theorem 13.1, empirical likelihood's power ranking was best at hypotheses farther from the null. Where any of simulated methods achieved power over 80%, empirical likelihood had the greatest power.

13.7 Exercises

Exercise 13.1 The Bartlett constant for a normal distribution is $d^2 + d/2$. An alternative to the scaled F critical value

$$r_F^{1-\alpha} = \frac{d(n-1)}{n-d} F_{d,n-d}^{1-\alpha}$$

is to employ the normal theory Bartlett correction

$$r_{NB,1}^{1-\alpha} = \left[1 + \frac{d^2 + d/2}{n}\right] \chi_{(d)}^{2,1-\alpha}$$

or

$$r_{NB,2}^{1-\alpha} = \left[1 - \frac{d^2 + d/2}{n}\right]^{-1} \chi_{(d)}^{2,1-\alpha}.$$

As n increases, the Bartlett correction becomes smaller and so empirical likelihood with a normal theory Bartlett correction is asymptotically properly calibrated. For small n, the Bartlett correction appropriate to a normal distribution might be competitive with the F calibration. For $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$ and $d \in \{1, 2, , 5, 10, 20, 50, 100, 200, 500\}$ find the values of n for which the scaled F critical value is larger than the first normal theory Bartlett correction. Find those values of $n > d^2 + d/2$ for which the scaled F critical value is larger than the second normal theory Bartlett correction.

Exercise 13.2 Show that for a vector mean $\theta = \mu$, formula (13.7) simplifies to

$$\widehat{C}(\theta) = \frac{1}{3}\widehat{\alpha}^{jjl} B^{lu} \left(\widehat{\theta} - \theta\right)^u, \tag{13.10}$$

where $B = S^{-1/2}$ for $S = (1/n)\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$.