Likelihood Ratio, Wald, and (Rao) Score Tests
Stat 543 Spring 2005

There are common large sample alternatives to likelihood ratio testing. These are the "Wald tests" (the testing version of the confidence ellipsoid material considered earlier) and the (Rao) "score tests." These are discussed in Section 6.3.2 of B&D and, for example, on pages 115-120 of Silvey's *Statistical Inference*.

Consider $\Theta \subset \Re^k$ and l ("genuinely different"/"independent") restrictions on $\boldsymbol{\theta}$

$$g_1(\boldsymbol{\theta}) = g_2(\boldsymbol{\theta}) = \cdots = g_l(\theta) = 0 \ ,$$

(like, for instance, $\theta_1 - \theta_1^0 = \theta_2 - \theta_2^0 = \cdots = \theta_l - \theta_l^0 = 0$). Define

$$\boldsymbol{g}(\boldsymbol{\theta}) \doteq (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), ..., g_l(\boldsymbol{\theta}))' \ .$$

We consider testing H_0:$g_1(\boldsymbol{\theta}) = g_2(\boldsymbol{\theta}) = \cdots = g_l(\boldsymbol{\theta}) = 0$, that is H_0:$\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$. The obvious likelihood ratio test statistic for this hypothesis is

$$\lambda_n = \frac{\sup\limits_{\boldsymbol{\theta}} f(X|\boldsymbol{\theta})}{\sup\limits_{\boldsymbol{\theta} \text{ s.t. } \boldsymbol{g}(\boldsymbol{\theta})=\mathbf{0}} f(X|\boldsymbol{\theta})}$$

and standard theory suggests (at least in iid cases) that

$$2 \ln \lambda_n \overset{\mathcal{L}_{\boldsymbol{\theta}}}{\to} \chi_l^2$$

(which leads to setting critical values for likelihood ratio tests).

Suppose that $\hat{\boldsymbol{\theta}}_n$ is an MLE of $\boldsymbol{\theta}$. Then if H_0:$\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$ is true, $\boldsymbol{g}(\ \hat{\boldsymbol{\theta}}_n)$ ought to be near $\mathbf{0}$, and one can think about rejecting H_0 if it is not. The questions are how to measure "nearness" and how to set a critical value in order to have a test with size approximately α. The Wald approach to doing this is as follows.

We expect (under suitable conditions) that under $P_{\boldsymbol{\theta}}$ the (k-dimensional) estimator $\hat{\boldsymbol{\theta}}_n$ has

$$\sqrt{n}(\ \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \ \overset{\mathcal{L}_{\boldsymbol{\theta}}}{\to} \ \mathrm{N}_k(\mathbf{0}, \boldsymbol{I}_1^{-1}(\boldsymbol{\theta})) \quad .$$

Then if

$$\underset{k \times p}{\boldsymbol{G}}(\boldsymbol{\theta}) \doteq \left(\frac{\partial g_i\,(\boldsymbol{\theta})}{\partial \theta_j} \right) \ ,$$

the delta method suggests that

$$\sqrt{n} \left(\boldsymbol{g}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{g}(\boldsymbol{\theta}) \right) \overset{\mathcal{L}_{\boldsymbol{\theta}}}{\to} \mathrm{N}_l(0, \boldsymbol{G}(\boldsymbol{\theta}) \boldsymbol{I}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{G}(\boldsymbol{\theta})') \ .$$

Abbreviate $\boldsymbol{G}(\boldsymbol{\theta}) \boldsymbol{I}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{G}(\boldsymbol{\theta})'$ as $\boldsymbol{B}(\boldsymbol{\theta})$. Since $\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$ if H_0 is true, the above suggests that

$$n \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)' \boldsymbol{B}(\boldsymbol{\theta})^{-1} \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n) \ \overset{\mathcal{L}}{\to} \ \chi_l^2 \ \text{ under H}_0 \ .$$

Now $n \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)' \boldsymbol{B}(\boldsymbol{\theta})^{-1} \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)$ can not serve as a test statistic, since it involves $\boldsymbol{\theta}$, which is not completely specified by H_0. But it is plausible to consider the statistic

$$W_n \doteq n \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)' \boldsymbol{B}(\hat{\boldsymbol{\theta}}_n)^{-1} \boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)$$

and hope that under suitable conditions

$$W_n \ \overset{\mathcal{L}}{\to} \ \chi_l^2 \text{ under H}_0 \ .$$

If this can be shown to hold up, one may reject for $W_n > (1 - \alpha)$ quantile of the χ_l^2 distribution. This is the use of "expected Fisher information" in defining a Wald statistic. With $\boldsymbol{H}_n\,(\boldsymbol{\theta})$ the matrix of second partials of the log-likelihood evaluated at $\boldsymbol{\theta}$, an "observed Fisher information" version of the above is to let

$$\boldsymbol{B}_n^*(\boldsymbol{\theta}) = \boldsymbol{G}(\boldsymbol{\theta}) \left(-\frac{1}{n} \boldsymbol{H}_n\,(\boldsymbol{\theta}) \right)^{-1} \boldsymbol{G}(\boldsymbol{\theta})'$$

and use the test statistic

$$W_n^* \doteq n\boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)'\boldsymbol{B}_n^*(\hat{\boldsymbol{\theta}}_n)^{-1}\boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)$$

The "(Rao) score test" or "χ^2 test" is an alternative to the LR and Wald tests. The motivation for it is that on occasion it can be easier to maximize the loglikelihood $l_n(\boldsymbol{\theta})$ subject to $\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$ than to simply maximize $l_n(\boldsymbol{\theta})$ without constraints. Let $\tilde{\boldsymbol{\theta}}_n$ be a "restricted" MLE (i.e. a maximizer of $l_n(\boldsymbol{\theta})$ subject to $\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$). One might expect that if H_0:$\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$ is true, then $\tilde{\boldsymbol{\theta}}_n$ ought to be nearly an unrestricted maximizer of $l_n(\boldsymbol{\theta})$ and the partial derivatives of $l_n(\boldsymbol{\theta})$ should be nearly 0 at $\tilde{\boldsymbol{\theta}}_n$. On the other hand, if H_0 is not true, there is little reason to expect the partials to be nearly 0. So (again/still with $l_n(\boldsymbol{\theta}) = \sum_{i=1}^n \ln f(X_i|\boldsymbol{\theta})$) one might consider the statistic

$$R_n \doteq \frac{1}{n}\left(\frac{\partial}{\partial\theta_i} l_n(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_n}\right)' I_1(\tilde{\boldsymbol{\theta}}_n)^{-1}\left(\frac{\partial}{\partial\theta_i} l_n(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_n}\right)$$

or using the observed Fisher information,

$$R_n^* \doteq \left(\frac{\partial}{\partial\theta_i} l_n(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_n}\right)' \left(-\boldsymbol{H}_n\left(\tilde{\boldsymbol{\theta}}_n\right)\right)^{-1}\left(\frac{\partial}{\partial\theta_i} l_n(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_n}\right)$$

What is not obvious is that R_n (or R_n^*) can typically be shown to differ from $2\ln\lambda_n$ by a quantity tending to 0 in P_θ probability. That is, this statistic can be calibrated using χ^2 critical points and can form the basis of a test of H_0 asymptotically equivalent to the likelihood ratio test. *In fact*, all of the test statistics mentioned here are asymptotically equivalent to the LRT (Wald and Rao score alike).

Observe that

1. an LRT requires computation of an MLE, $\hat{\boldsymbol{\theta}}_n$, and a restricted MLE, $\tilde{\boldsymbol{\theta}}_n$,
2. a Wald test requires only computation of the MLE, $\hat{\boldsymbol{\theta}}_n$, and
3. a Rao score test requires only computation of the restricted MLE, $\tilde{\boldsymbol{\theta}}_n$.

Any of the tests discussed above (LRT, Wald, score) can be inverted to find confidence regions for the values of l parametric functions $u_1(\boldsymbol{\theta}), u_2(\boldsymbol{\theta}), \ldots, u_l(\theta)$. That is, for any vector of potential values for these functions $\boldsymbol{c} = (c_1, c_2, \ldots, c_l)'$, one may define $g_{\boldsymbol{c},i}(\boldsymbol{\theta}) = u_i(\boldsymbol{\theta}) - c_i$ and a test of some type above for H_0:$\boldsymbol{g}_{\boldsymbol{c}}(\boldsymbol{\theta}) = \mathbf{0}$. The set of all $\boldsymbol{c}$ for which an approximately α-level test does not reject constitutes an approximately $(1-\alpha)\times 100\%$ confidence set for the vector $(u_1(\boldsymbol{\theta}), u_2(\boldsymbol{\theta}), \ldots, u_l(\boldsymbol{\theta}))'$. When Wald tests are used in this construction, the regions will be ellipsoidal.