

Likelihood 2: Wald Tests¹

STA442/2101 Fall 2017

¹See last slide for copyright information.

Background Reading

Davison Chapter 4, especially Sections 4.3 and 4.4

Vector of MLEs is Asymptotically Normal

That is, Multivariate Normal

This yields

- Confidence intervals
- Z -tests of $H_0 : \theta_j = \theta_0$
- Wald tests
- Score Tests
- Indirectly, the Likelihood Ratio tests

Under Regularity Conditions

(Thank you, Mr. Wald)

- $\hat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}$
- $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{T} \sim N_k(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta})^{-1})$
- So we say that $\hat{\boldsymbol{\theta}}_n$ is asymptotically $N_k(\boldsymbol{\theta}, \frac{1}{n}\mathcal{I}(\boldsymbol{\theta})^{-1})$.
- $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher Information in one observation.
- A $k \times k$ matrix

$$\mathcal{I}(\boldsymbol{\theta}) = \left[E\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f(Y; \boldsymbol{\theta}) \right] \right]$$

- The Fisher Information in the whole sample is $n\mathcal{I}(\boldsymbol{\theta})$

$\hat{\theta}_n$ is asymptotically $N_k(\theta, \frac{1}{n}\mathcal{I}(\theta)^{-1})$

- Asymptotic covariance matrix of $\hat{\theta}_n$ is $\frac{1}{n}\mathcal{I}(\theta)^{-1}$, and of course we don't know θ .
- For tests and confidence intervals, we need a good *approximate* asymptotic covariance matrix,
- Based on a consistent estimate of the Fisher information matrix.
- $\mathcal{I}(\hat{\theta}_n)$ would do.
- But it's inconvenient: Need to compute partial derivatives and expected values in

$$\mathcal{I}(\theta) = \left[E\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f(Y; \theta) \right] \right]$$

and then substitute $\hat{\theta}_n$ for θ .

Another approximation of the asymptotic covariance matrix

Approximate

$$\frac{1}{n} \mathbf{I}(\boldsymbol{\theta})^{-1} = \left[n E \left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(Y; \boldsymbol{\theta}) \right] \right]^{-1}$$

with

$$\widehat{\mathbf{V}}_n = \left(\left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\boldsymbol{\theta}, \mathbf{Y}) \right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \right)^{-1}$$

Details of why it's a good approximation are omitted.

Compare

Hessian and (Estimated) Asymptotic Covariance Matrix

- $\hat{\mathbf{V}}_n = \left(\left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\boldsymbol{\theta}, \mathbf{Y}) \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \right)^{-1}$
- Hessian at MLE is $\mathbf{H} = \left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\boldsymbol{\theta}, \mathbf{Y}) \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}$
- So to estimate the asymptotic covariance matrix of $\boldsymbol{\theta}$, just invert the Hessian.
- The Hessian is usually available as a by-product of numerical search for the MLE.

Connection to Numerical Optimization

- Suppose we are minimizing the minus log likelihood by a direct search.
- We have reached a point where the gradient is close to zero. Is this point a minimum?
- The Hessian is a matrix of mixed partial derivatives. If all its eigenvalues are positive at a point, the function is concave up there.
- Partial derivatives are often approximated by the slopes of secant lines – no need to calculate them symbolically.
- It's *the* multivariable second derivative test.

So to find the estimated asymptotic covariance matrix

- Minimize the minus log likelihood numerically.
- The Hessian at the place where the search stops is usually available.
- Invert it to get $\hat{\mathbf{V}}_n$.
- This is so handy that sometimes we do it even when a closed-form expression for the MLE is available.

Estimated Asymptotic Covariance Matrix $\widehat{\mathbf{V}}_n$ is Useful

- Asymptotic standard error of $\widehat{\theta}_j$ is the square root of the j th diagonal element.
- Denote the asymptotic standard error of $\widehat{\theta}_j$ by $S_{\widehat{\theta}_j}$.
- Thus

$$Z_j = \frac{\widehat{\theta}_j - \theta_j}{S_{\widehat{\theta}_j}}$$

is approximately standard normal.

Confidence Intervals and Z-tests

Have $Z_j = \frac{\hat{\theta}_j - \theta_j}{S_{\hat{\theta}_j}}$ approximately standard normal, yielding

- Confidence intervals: $\hat{\theta}_j \pm S_{\hat{\theta}_j} z_{\alpha/2}$
- Test $H_0 : \theta_j = \theta_0$ using

$$Z = \frac{\hat{\theta}_j - \theta_0}{S_{\hat{\theta}_j}}$$

And Wald Tests

$$W_n = (\mathbf{L}\hat{\boldsymbol{\theta}}_n - \mathbf{h})^\top \left(\mathbf{L}\hat{\mathbf{V}}_n\mathbf{L}^\top \right)^{-1} (\mathbf{L}\hat{\boldsymbol{\theta}}_n - \mathbf{h})$$

A very important special case of the earlier

$$\begin{aligned} W_n &= n(\mathbf{L}\mathbf{T}_n - \mathbf{h})^\top \left(\mathbf{L}\hat{\boldsymbol{\Sigma}}_n\mathbf{L}^\top \right)^{-1} (\mathbf{L}\mathbf{T}_n - \mathbf{h}) \\ &= (\mathbf{L}\mathbf{T}_n - \mathbf{h})^\top \left(\mathbf{L}\frac{1}{n}\hat{\boldsymbol{\Sigma}}_n\mathbf{L}^\top \right)^{-1} (\mathbf{L}\mathbf{T}_n - \mathbf{h}) \end{aligned}$$

Score Tests

Thank you Mr. Rao

- $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, dimension $k \times 1$
- $\hat{\boldsymbol{\theta}}_0$ is the MLE under H_0 , dimension $k \times 1$
- $\mathbf{u}(\boldsymbol{\theta}) = \left(\frac{\partial \ell}{\partial \theta_1}, \dots, \frac{\partial \ell}{\partial \theta_k}\right)^\top$ is the gradient.
- $\mathbf{u}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$.
- If H_0 is true, $\mathbf{u}(\hat{\boldsymbol{\theta}}_0)$ should also be close to zero too.
- Under H_0 for large N , $\mathbf{u}(\hat{\boldsymbol{\theta}}_0) \sim N_k(\mathbf{0}, \frac{1}{n}\mathcal{I}(\boldsymbol{\theta}))$, approximately.
- And,

$$S = \mathbf{u}(\hat{\boldsymbol{\theta}}_0)^\top \frac{1}{n} \mathcal{I}(\hat{\boldsymbol{\theta}}_0)^{-1} \mathbf{u}(\hat{\boldsymbol{\theta}}_0) \sim \chi^2(r)$$

Where r is the number of restrictions imposed by H_0 .

Or use the inverse of the Hessian (under H_0) instead of $\frac{1}{n}\mathcal{I}(\hat{\boldsymbol{\theta}}_0)$.

Three Big Tests

- Score Tests: Fit just the restricted model
- Wald Tests: Fit just the unrestricted model
- Likelihood Ratio Tests: Fit Both

Comparing Likelihood Ratio and Wald tests

- Asymptotically equivalent under H_0 , meaning $(W_n - G_n^2) \xrightarrow{p} 0$
- Under H_1 ,
 - Both have the same approximate distribution (non-central chi-square).
 - Both go to infinity as $n \rightarrow \infty$.
 - But values are not necessarily close.
- Likelihood ratio test tends to get closer to the right Type I error probability for small samples.
- Wald can be more convenient when testing lots of hypotheses, because you only need to fit the model once.
- Wald can be more convenient if it's a lot of work to write the restricted likelihood.

Copyright Information

This slide show was prepared by **Jerry Brunner**, Department of Statistical Sciences, University of Toronto. It is licensed under a **Creative Commons Attribution - ShareAlike 3.0 Unported License**. Use any part of it as you like and share the result freely. The L^AT_EX source code is available from the course website:

<http://www.utstat.toronto.edu/~brunner/oldclass/appliedf17>