Likelihood 2: Wald (and Score) Tests[1]

STA442/2101 Fall 2012

[1] See 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)

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

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

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

$H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$

Suppose $\boldsymbol{\theta} = (\theta_1, \ldots \theta_7)$, and the null hypothesis is

- $\theta_1 = \theta_2$
- $\theta_6 = \theta_7$
- $\frac{1}{3}(\theta_1 + \theta_2 + \theta_3) = \frac{1}{3}(\theta_4 + \theta_5 + \theta_6)$

We can write null hypothesis in matrix form as

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Suppose $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$ is True, and $\widehat{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})}_n \stackrel{p}{\rightarrow} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$

By Slutsky 6a (Continuous mapping),

$$\sqrt{n}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{L}\boldsymbol{\theta}) = \sqrt{n}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \stackrel{d}{\rightarrow} \mathbf{LT} \sim N_k\left(\mathbf{0}, \mathbf{L}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}\mathbf{L}'\right)$$

and

$$\widehat{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})}_n^{-1} \stackrel{p}{\rightarrow} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}.$$

Then by Slutsky's (6c) Stack Theorem,

$$\left(\begin{array}{c} \sqrt{n}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \\ \widehat{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})}_n^{-1} \end{array} \right) \stackrel{d}{\rightarrow} \left(\begin{array}{c} \mathbf{LT} \\ \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1} \end{array} \right).$$

Finally, by Slutsky 6a again,

$$\begin{array}{rcl} W_n & = & n(\mathbf{L}\widehat{\boldsymbol{\theta}} - \mathbf{h})'(\mathbf{L}\widehat{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})}_n^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}} - \mathbf{h}) \\ & \stackrel{d}{\rightarrow} & W = (\mathbf{LT} - \mathbf{0})'(\mathbf{L}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}\mathbf{L}')^{-1}(\mathbf{LT} - \mathbf{0}) \sim \chi^2(r) \end{array}$$

The Wald Test Statistic

$W_n = n(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})'(\mathbf{L}\widehat{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})}_n^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})$

- Again, null hypothesis is $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$
- Matrix $\mathbf{L}$ is $r \times k$, $r \leq k$, rank r
- All we need is a consistent estimator of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$
- $\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}})$ would do
- But it's inconvenient
- Need to compute partial derivatives and expected values in

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

Observed Fisher Information

- To find $\widehat{\boldsymbol{\theta}}_n$, minimize the minus log likelihood.
- Matrix of mixed partial derivatives of the minus log likelihood is

$$\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\boldsymbol{\theta},\mathbf{Y})\right] = \left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{i=1}^{n}\log f(Y_i;\boldsymbol{\theta})\right]$$

- So by the Strong Law of Large Numbers,

$$\begin{aligned}\boldsymbol{\mathcal{J}}_n(\boldsymbol{\theta}) &= \left[\frac{1}{n}\sum_{i=1}^{n}-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y_i;\boldsymbol{\theta})\right] \\ &\stackrel{a.s.}{\rightarrow} \left[E\left(-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y;\boldsymbol{\theta})\right)\right] = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\end{aligned}$$

A Consistent Estimator of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$

Just substitute $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$

$$
\begin{aligned}
\boldsymbol{\mathcal{J}}_n(\widehat{\boldsymbol{\theta}}_n) &= \left[\frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y_i;\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \\
&\stackrel{a.s.}{\rightarrow} \left[E\left(-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y;\boldsymbol{\theta})\right)\right] = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})
\end{aligned}
$$

- Convergence is believable but not trivial.
- Now we have a consistent estimator, more convenient than $\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_n)$: Use $\widehat{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})}_n = \boldsymbol{\mathcal{J}}_n(\widehat{\boldsymbol{\theta}}_n)$

Approximate the Asymptotic Covariance Matrix

- Asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_n$ is $\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}$.

- Approximate it with

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

Compare

Hessian and (Estimated) Asymptotic Covariance Matrix

- $\widehat{\mathbf{V}}_n = \left(\left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\boldsymbol{\theta}, \mathbf{Y}) \right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \right)^{-1}$
- Hessian at MLE is $\mathbf{H} = \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\boldsymbol{\theta}, \mathbf{Y}) \right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$
- 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.
- Its *the* multivariable second derivative test.
- The Hessian at the MLE is exactly the observed Fisher information matrix.
- Partial derivatives are often approximated by the slopes of secant lines – no need to calculate them.

So to find the estimated asymptotic covariance matrix

- Minimize the minus log likelihood numerically.
- The Hessian at the place where the search stops is exactly the observed Fisher information matrix.
- Invert it to get $\widehat{\mathbf{V}}_n$.
- 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 jth 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 $Z_j = \frac{\widehat{\theta}_j - \theta_j}{S_{\widehat{\theta}_j}}$ approximately standard normal, yielding

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

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

And Wald Tests

Recalling $\widehat{\mathbf{V}}_n = \frac{1}{n}\boldsymbol{\mathcal{J}}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}$

$$
\begin{aligned}
W_n &= n(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})'(\mathbf{L}\widehat{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})}_n^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \\
&= n(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})'(\mathbf{L}\boldsymbol{\mathcal{J}}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \\
&= n(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})'\left(\mathbf{L}(n\widehat{\mathbf{V}}_n)\mathbf{L}'\right)^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \\
&= n(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})'\frac{1}{n}\left(\mathbf{L}\widehat{\mathbf{V}}_n\mathbf{L}'\right)^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \\
&= (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})'\left(\mathbf{L}\widehat{\mathbf{V}}_n\mathbf{L}'\right)^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})
\end{aligned}
$$

Score Tests

Thank you Mr. Rao

- $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, size $k \times 1$
- $\widehat{\boldsymbol{\theta}}_0$ is the MLE under H_0, size $k \times 1$
- $\mathbf{u}(\boldsymbol{\theta}) = (\frac{\partial \ell}{\partial \theta_1}, \ldots \frac{\partial \ell}{\partial \theta_k})'$ is the gradient.
- $\mathbf{u}(\widehat{\boldsymbol{\theta}}) = 0$
- If H_0 is true, $\mathbf{u}(\widehat{\boldsymbol{\theta}}_0)$ should also be close to zero.
- Under H_0 for large N, $\mathbf{u}(\widehat{\boldsymbol{\theta}}_0) \sim N_k(\mathbf{0}, \boldsymbol{\mathcal{J}}(\boldsymbol{\theta}))$, approximately.
- And,

$$S = \mathbf{u}(\widehat{\boldsymbol{\theta}}_0)' \boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}_0)^{-1} \mathbf{u}(\widehat{\boldsymbol{\theta}}_0) \sim \chi^2(r)$$

Where r is the number of restrictions imposed by H_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

- Asymptotically equivalent under H_0, meaning $(W_n - G_n) \stackrel{p}{\rightarrow} 0$
- Under H_1,
  - Both have approximately the same 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 rate 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