Likelihood Part Two[1]

STA2101 Fall 2019

[1] See last slide for copyright information.

Background Reading

Appendix A, Section 6

Vector of MLEs is Asymptotically Normal

That is, Multivariate Normal

This yields

- Confidence intervals for the parameters.
- 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)

- $\widehat{\boldsymbol{\theta}}_n \stackrel{a.s.}{\rightarrow} \boldsymbol{\theta}$
- $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \stackrel{d}{\rightarrow} \mathbf{T} \sim N_k\left(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}\right)$
- 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 $n\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$

$\widehat{\boldsymbol{\theta}}_n$ is asymptotically $N_k\left(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1}\right)$

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

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

and then substitute $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$.

Another approximation of the asymptotic covariance matrix

Approximate

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

with

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

$\widehat{\mathbf{V}}_n^{-1}$ is called the "observed Fisher information."

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.
- 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 $\widehat{\mathbf{V}}_n$.
- 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 for $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$

Based on $(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2(p)$

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

$\widehat{\boldsymbol{\theta}}_n \stackrel{.}{\sim} N_p(\boldsymbol{\theta}, \mathbf{V_n})$ so if H_0 is true, $\mathbf{L}\widehat{\boldsymbol{\theta}}_n \stackrel{.}{\sim} N_r(\mathbf{h}, \mathbf{L}\mathbf{V}_n\mathbf{L}^\top)$.
Thus $(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})^\top \left(\mathbf{L}\mathbf{V}_n\mathbf{L}^\top\right)^{-1} (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \stackrel{.}{\sim} \chi^2(r)$.
And substitute $\widehat{\mathbf{V}}_n$ for $\mathbf{V}_n$.

Slutsky arguments omitted.

Score Tests

Thank you Mr. Rao

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

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

Where r is the number of restrictions imposed by H_0.
Or use the inverse of the Hessian (under H_0) instead of $\frac{1}{n}\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_0)$.

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 $(W_n - G_n^2) \stackrel{p}{\rightarrow} 0$
- 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 LaTeX source code is available from the course website:

http://www.utstat.toronto.edu/~brunner/oldclass/2101f1