Additional Topics Related to Likelihood

Information Criteria

Akaike's Information criterion is given by

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}) + 2k,$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the maximized log likelihood and k is the dimension of the model parameter space.

- AIC $= -2\ell(\hat{\boldsymbol{\theta}}) + 2k$ can be used to determine which of multiple models is "best" for a given data set.
- Small values of AIC are preferred.
- The $+2k$ portion of AIC can be viewed as a penalty for model complexity.

Schwarz's Bayesian Information Criterion is given by

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + k\log(n).$$

BIC is the same as AIC except the penalty for model complexity is greater for BIC (when $n \geq 8$) and grows with n.

- AIC and BIC can each be used to compare models even if they are not nested (i.e., even if one is not a special case of the other as in our reduced vs. full model comparison discussed previously).

- However, if REML likelihoods are used, compared models must have the same model for the response mean.

- Different models for the mean would yield different error contrasts and different datasets for computation of maximized REML likelihoods.

Large n Theory for MLEs

- Suppose $\boldsymbol{\theta}$ is a $k \times 1$ parameter vector.
- Let $\ell(\boldsymbol{\theta})$ denote the log likelihood function.
- Under regularity conditions discussed in, e.g., Shao, J.(2003) *Mathematical Statistics*, 2^{nd} Ed. Springer, New York; we have the following.

1. There is an estimator $\hat{\boldsymbol{\theta}}$ that solves the score equations $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ and is a (weakly) consistent estimator of $\boldsymbol{\theta}$.

   This means that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}$, i.e.,

$$\lim_{n \to \infty} Pr[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \varepsilon] = 0 \text{ for any } \varepsilon > 0.$$

2. For sufficiently large n,

$$\hat{\boldsymbol{\theta}} \dot{\sim} N(\boldsymbol{\theta}, \boldsymbol{I}^{-1}(\boldsymbol{\theta})), \text{ where}$$

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\theta}) &= E\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right] \\ &= -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \\ &= \left[-E\left\{\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right\}\right]_{i,j \in \{1,\ldots,k\}}. \end{aligned}$$

- $I(\boldsymbol{\theta})$ is known as the *Fisher Information* matrix.
- $I(\boldsymbol{\theta})$ can be approximated by the *observed Fisher Information* matrix, which is given by

$$\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) \equiv \frac{-\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} .$$

- $\boldsymbol{I}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})$ may depend on unknown nuisance parameters. In such cases, nuisance parameters are replaced by consistent estimators.

A Simple Example

- Suppose $y_1, \ldots, y_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$.
- If we are interested in inference for μ, we can take $\theta = \mu$ and treat σ^2 as a nuisance parameter.
- It is straightforward to show that $\bar{y}_{\cdot}$ is the unique solution to the likelihood equation.
- Furthermore, it is straightforward to show that

$$I(\theta) = \hat{I}(\hat{\theta}) = n/\sigma^2.$$

A Simple Example (continued)

Thus, we have

$$\hat{\theta} = \bar{y}. \sim N(\theta = \mu, I^{-1}(\theta) = \sigma^2/n)$$

and

$$\bar{y}. \dot{\sim} N(\mu, s^2/n)$$

for sufficiently large n, where

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}.)^2}{n-1}.$$

Wald Tests and Confidence Intervals

Suppose for large n that

$$\hat{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} N(\mathbf{0}, \boldsymbol{I})$$

and

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} \chi^2_k.$$

For example, suppose $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})$ is the observed information matrix. Then under regularity conditions, we have

$$[\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})]^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} N(\mathbf{0}, \boldsymbol{I})$$

and

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} \chi^2_k$$

for sufficiently large n.

An approximate $100(1-\alpha)\%$ confidence interval for θ_i is

$$\hat{\theta}_i \pm z_{1-\alpha/2}\sqrt{\hat{\boldsymbol{\Sigma}}_{ii}},$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the $N(0,1)$ distribution and $\hat{\boldsymbol{\Sigma}}_{ii}$ is element (i,i) of $\hat{\boldsymbol{\Sigma}}$.

An approximate p-value for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is

$$Pr[\chi_k^2 \geq (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)],$$

where χ_k^2 is a central χ^2 random variable with k degrees of freedom.

Multivariate Delta Method

- Suppose $\boldsymbol{g}$ is a function from $\mathbb{R}^k$ to $\mathbb{R}^m$, i.e.,

$$\text{for } \boldsymbol{\theta} \in \mathbb{R}^k, \boldsymbol{g}(\boldsymbol{\theta}) = \begin{bmatrix} g_1(\boldsymbol{\theta}) \\ g_2(\boldsymbol{\theta}) \\ \vdots \\ g_m(\boldsymbol{\theta}) \end{bmatrix}$$

  for some functions $g_1, \ldots, g_m$.
- Suppose $\boldsymbol{g}$ is differentiable with derivative matrix

$$\boldsymbol{D} \equiv \begin{bmatrix} \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial g_m(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_k} & \cdots & \frac{\partial g_m(\boldsymbol{\theta})}{\partial \theta_k} \end{bmatrix}.$$

Now suppose $\hat{\theta}$ has mean θ and variance Σ. Then Taylor's Theorem implies

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{D}'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

which implies

$$E[\boldsymbol{g}(\hat{\boldsymbol{\theta}})] \approx \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{D}'E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = g(\boldsymbol{\theta})$$

and

$$\mathrm{Var}[\boldsymbol{g}(\hat{\boldsymbol{\theta}})] \approx \mathrm{Var}[\boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{D}'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] = \boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D}.$$

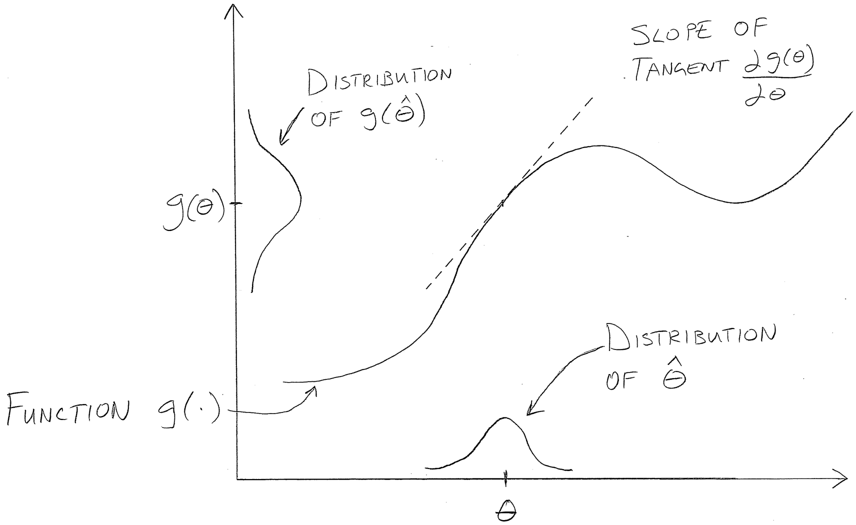
- If $\hat{\boldsymbol{\theta}} \dot{\sim} N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, it follows that

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \dot{\sim} N(\boldsymbol{g}(\boldsymbol{\theta}), \boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D}).$$

- In practice, we often need to estimate $\boldsymbol{D}$ by replacing $\boldsymbol{\theta}$ in $\boldsymbol{D}$ with $\hat{\boldsymbol{\theta}}$ to obtain $\hat{\boldsymbol{D}}$.

- Similarly, we often need to replace $\boldsymbol{\Sigma}$ with an estimate $\hat{\boldsymbol{\Sigma}}$.

THE DELTA METHOD



Likelihood Ratio Based Inference

Suppose we wish to test the null hypothesis that a reduced model provides an adequate fit to a dataset relative to a more general full model that includes the reduced model as a special case.

- Define Λ as

$$\frac{\text{Reduced Model Maximized Likelihood}}{\text{Full Model Maximized Likelihood}}.$$

- Λ is known as the *likelihood ratio*.
- $-2\log\Lambda$ is known as the *likelihood ratio test statistic*.
- Tests based on $-2\log\Lambda$ are called *likelihood ratio tests*.

- Under the regularity conditions in Shao (2003) mentioned previously, the likelihood ratio test statistic $-2\log\Lambda$ is approximately distributed as central $\chi^2_{k_f-k_r}$ under the null hypothesis, where k_f and k_r are the dimensions of the parameter space under the full and reduced models, respectively.

- This approximation can be reasonable if n is "sufficiently large."

Likelihood Ratio Tests and Confidence Regions for a Subvector of the Full Model Parameter Vector $\boldsymbol{\theta}$

- Suppose $\boldsymbol{\theta}$ is $k \times 1$ vector and is partitioned into vectors $\boldsymbol{\theta}_1$ $k_1 \times 1$ and $\boldsymbol{\theta}_2$ $k_2 \times 1$, where $k = k_1 + k_2$ and $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}$.

- Consider a test of $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$.

- Suppose $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$ maximizes $\ell\left(\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}\right)$ over $\boldsymbol{\theta}_2$ for any fixed value of $\boldsymbol{\theta}_1$.

- Then $2\left[\ell(\hat{\boldsymbol{\theta}}) - \ell\left(\begin{bmatrix} \boldsymbol{\theta}_{10} \\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_{10}) \end{bmatrix}\right)\right]$ is approximately $\chi^2_{k_1}$ under the null hypothesis by our previous result when n is "sufficiently large."

Also,

$$Pr\left\{2\left[\ell(\hat{\boldsymbol{\theta}}) - \ell\left(\left[\begin{array}{c}\boldsymbol{\theta}_1\\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\end{array}\right]\right)\right] \leq \chi^2_{k_1,1-\alpha}\right\} \approx 1-\alpha$$

which implies

$$Pr\left\{\ell\left(\left[\begin{array}{c}\boldsymbol{\theta}_1\\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\end{array}\right]\right) \geq \ell(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\chi^2_{k_1,1-\alpha}\right\} \approx 1-\alpha.$$

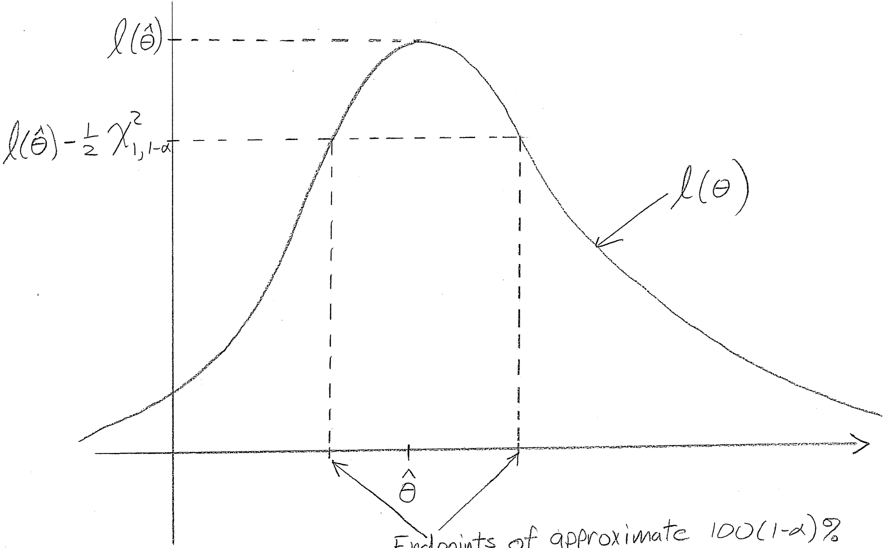
- Thus, the set of values of $\boldsymbol{\theta}_1$ that, when maximizing over $\boldsymbol{\theta}_2$, yield a maximized likelihood within $\frac{1}{2}\chi^2_{k_1,1-\alpha}$ of the likelihood maximized over all $\boldsymbol{\theta}$, form a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}_1$.

- Such a confidence region is known as a *profile likelihood confidence region* because

$$\ell\left(\begin{bmatrix} \boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1) \end{bmatrix}\right)$$

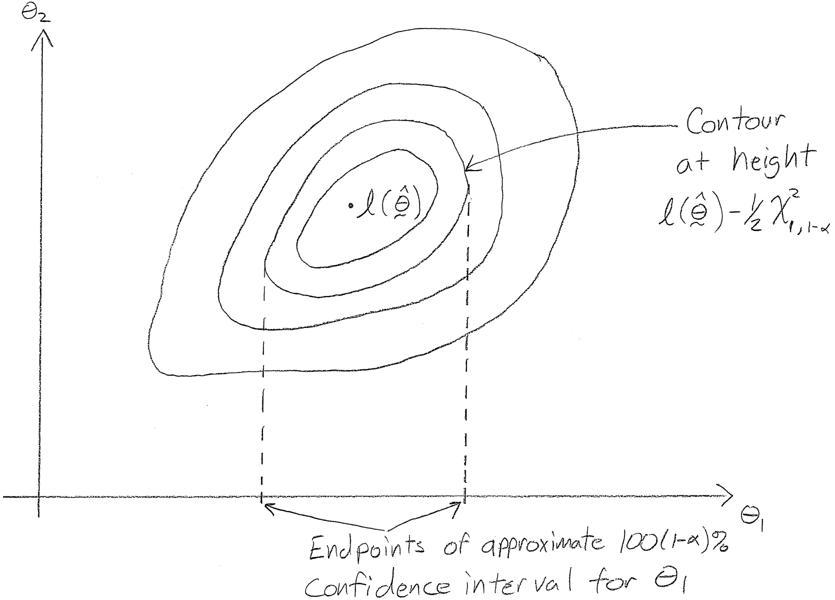
is the *profile log likelihood* for $\boldsymbol{\theta}_1$.

Sketch for the case $k=1$:

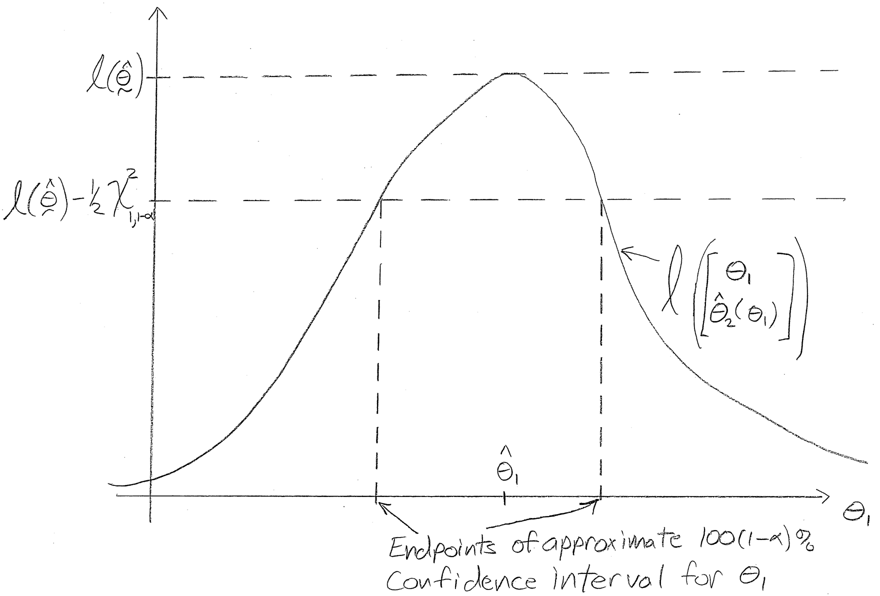


Endpoints of approximate $100(1-\alpha)\%$ Confidence interval for θ.

Sketch for the case $k=2$:



Sketch for the case $K_1 = 1$, K_2 arbitrary



Warnings

- The normal and χ^2 approximations mentioned in these notes may be crude if sample sizes are not sufficiently large.
- The regularity conditions mentioned in these notes do not hold if the true parameter falls on the boundary of the parameter space. Thus, as an example, testing $H_0 : \sigma_u^2 = 0$ is not covered by the methods presented here.