PLSC 504 – Fall 2017
Maximum Likelihood

August 24, 2017

A Model

$$Y \sim N(\mu, \sigma^2)$$

$$\begin{aligned} \mathrm{E}(Y) &= \mu \\ \mathrm{Var}(Y) &= \sigma^2 \end{aligned}$$

Probabilities, Marginal and Joint

$$\Pr(Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\text{exp}\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right]$$

More generally:

$$\begin{aligned}
\Pr(Y_i = y_i \; \forall \; i) \;&\equiv\; L(Y|\mu, \sigma^2) \\
&=\; \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}}\text{exp}\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]
\end{aligned}$$

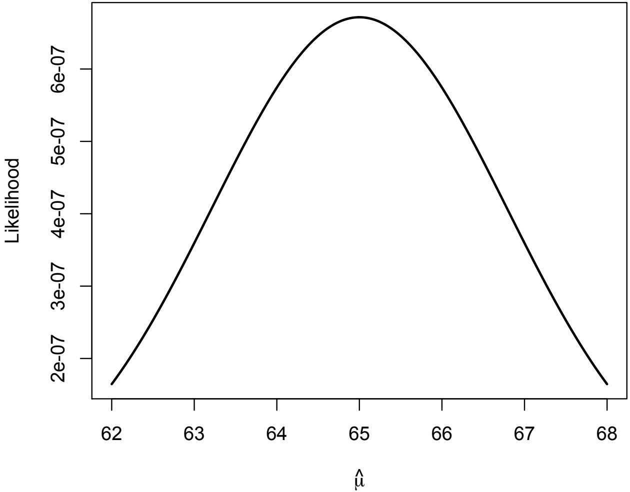
Likelihood

$$L(\hat{\mu}, \hat{\sigma}^2 | Y) \propto \Pr(Y | \hat{\mu}, \hat{\sigma}^2)$$

For (e.g.) $\hat{\mu} = 68$, $\hat{\sigma} = 4$:

$$\begin{aligned} L &= \frac{1}{\sqrt{2\pi 16}} \exp\left[-\frac{(64-68)^2}{32}\right] \times \\ &\quad \frac{1}{\sqrt{2\pi 16}} \exp\left[-\frac{(63-68)^2}{32}\right] \times \\ &\quad \frac{1}{\sqrt{2\pi 16}} \exp\left[-\frac{(59-68)^2}{32}\right] \times \ldots \\ &= \text{some reeeeeally small number...} \end{aligned}$$

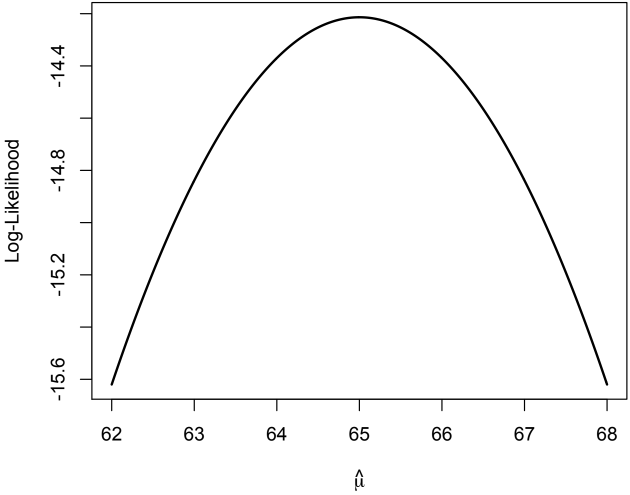
What a Likelihood Looks Like



Log-Likelihood

$$
\begin{aligned}
\ln L(\hat{\mu}, \hat{\sigma}^2 | Y) &= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right] \\
&= \sum_{i=1}^{N} \ln \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right] \right\} \\
&= -\frac{N}{2} \ln(2\pi) - \left[\sum_{i=1}^{N} \frac{1}{2} \ln\ \sigma^2 + \frac{1}{2\sigma^2} (Y_i - \mu)^2 \right]
\end{aligned}
$$

What a Log-Likelihood Looks Like



The "Maximum" Part

For $L = f(Y, \theta)$,

- Calculate $\frac{\partial \ln L}{\partial \theta}$,
- Set $\frac{\partial \ln L}{\partial \theta} = 0$, solve for $\hat{\theta}$,
- Calculate $\frac{\partial^2 \ln L}{\partial \theta^2}$,
- Verify $\frac{\partial^2 \ln L}{\partial \theta^2} < 0$.

MLE in General

$$\Pr(Y) = f(\mathbf{X}, \theta)$$

$$L = \prod_{i=1}^{N} f(Y_i | \mathbf{X}_i, \theta)$$

$$\ln L = \sum_{i=1}^{N} \ln f(Y_i | \mathbf{X}_i, \theta)$$

$$\ln L(\hat{\theta} | Y, \mathbf{X}) = \max_{\theta} \{\ln L(\theta | Y, \mathbf{X})\}$$

The Gradient

$$\mathbf{g}(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta})}{\partial \hat{\theta}}$$

Taylor series:

$$\frac{\partial \ln L}{\partial \hat{\theta}} \approx \frac{\partial \ln L}{\partial \theta} + \frac{\partial^2 \ln L}{\partial \theta^2}(\hat{\theta} - \theta)$$

$$\begin{aligned} \hat{\theta} - \theta &= \left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1} \frac{\partial \ln L}{\partial \theta} \\ &= -\mathbf{H}(\theta)^{-1}\mathbf{g}(\theta) \end{aligned}$$

Summary

MLEs:

- Maximize $L(\theta|Y, \mathbf{X})$
- Are consistent in N
- Are asymptotically efficient
- Are asymptotically Normal
- Are invariant to (injective) transformations and varying sampling methods

Optimization

The Basic Problem

Find

$$\max_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^k} \ln L(\hat{\boldsymbol{\beta}} | Y, \mathbf{X})$$

Unconstrained optimization problem...

The Intuition

- Start with $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$
- Adjust:

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = \hat{\boldsymbol{\beta}}_{\mathbf{0}} + \mathbf{A_0}$$

- Repeat.

More Specifically...

$$\hat{\boldsymbol{\beta}}_\ell = \hat{\boldsymbol{\beta}}_{\ell-\mathbf{1}} + \mathbf{A}_{\ell-\mathbf{1}}$$

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\ell \ni \hat{\boldsymbol{\beta}}_\ell - \hat{\boldsymbol{\beta}}_{\ell-1} (\equiv \mathbf{A}_\ell) < \tau$$

What's $\mathbf{A}$?

$$\mathbf{A} = f[\mathbf{g}(\hat{\beta})]$$

- $\mathbf{g}(\hat{\beta})$ = "directionality" of change
  - $\mathbf{g}(\hat{\beta}_k) < 0 \rightarrow A_k < 0$
  - $\mathbf{g}(\hat{\beta}_k) > 0 \rightarrow A_k > 0$

“Steepest Ascent”

$$\mathbf{A}_\ell = \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_\ell}$$

$$\hat{\boldsymbol{\beta}}_\ell = \hat{\boldsymbol{\beta}}_{\ell-1} + \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}}$$

“Step Size”

$$\hat{\boldsymbol{\beta}}_\ell = \hat{\boldsymbol{\beta}}_{\ell-\mathbf{1}} + \lambda_{\ell-\mathbf{1}} \boldsymbol{\Delta}_{\ell-\mathbf{1}}$$

- $\boldsymbol{\Delta} \rightarrow$ *direction*
- $\lambda \rightarrow$ *amount* (“step size”)

Newton-Raphson

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_\ell &= \hat{\boldsymbol{\beta}}_{\ell-1} - \left(\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}^2} \right)^{-1} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} \\
&= \hat{\boldsymbol{\beta}}_{\ell-1} - \mathbf{H}(\hat{\boldsymbol{\beta}}_{\ell-1})^{-1} \mathbf{g}(\hat{\boldsymbol{\beta}}_{\ell-1})
\end{aligned}
$$

Other Approaches: "Method of Scoring"

Uses:

$$\hat{\boldsymbol{\beta}}_\ell = \hat{\boldsymbol{\beta}}_{\ell-1} - \left[\mathsf{E}\left(\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}^2_{\ell-1}} \right)^{-1} \right] \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} \tag{1}$$

$$= \hat{\boldsymbol{\beta}}_{\ell-1} - \{\mathsf{E}[\mathbf{H}(\hat{\boldsymbol{\beta}}_{\ell-1})]\}^{-1} \mathbf{g}(\hat{\boldsymbol{\beta}}_{\ell-1})$$

- Due to Fisher
- Advantages:
  - $\approx$ Newton-Raphson
  - Can be faster/simpler

Berndt, Hall^2, and Hausman ("BHHH")

Uses:

$$\hat{\boldsymbol{\beta}}_\ell = \hat{\boldsymbol{\beta}}_{\ell-1} - \left(\sum_{i=1}^{N} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}}' \right)^{-1} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}}$$

Advantages:

- (Relatively) very easy to compute
- Reasonably accurate...

Other "Newton Jr.s"

- Davidson-Fletcher-Powell ("DFP")
- Broyden et al. ("BFGS")
- They are:
    - Very fast/efficient
    - Pretty bad at getting $-\left(\mathbf{H}(\hat{\boldsymbol{\beta}})\right)^{-1}$

Summary

Method	"Step size" (∂^2) matrix	Variance-Covariance Estimate
Newton	Inverse of the observed second derivative (Hessian)	Inverse of the negative Hessian
Scoring	Inverse of the expected value of the Hessian (information matrix)	Inverse of the negative information matrix
BHHH	Outer product approximation of the information matrix	Inverse of the outer product approximation

Software Issues: R

Lots of optimizers:

- `maxLik` package: options for Newton-Raphson, BHHH, BFGS, others
- `optim` (in `stats`) – quasi-Newton, plus others
- `nlm` (in `stats`) – nonlinear minimization "using a Newton-type algorithm"
- `newton` (in `Bhat`) – Newton-Raphson solver
- `solveLP` (in `linprog`) – linear programming optimizer

R : Using `maxLik`

- *Must* provide log-likelihood function
- Can provide $\mathbf{H}(\hat{\beta})$, $\mathbf{g}(\hat{\beta})$, both, or neither
- Choose optimizer (Newton, BHHH, BFGS, etc.)
- Returns an object of class `maxLik`

Practical Optimization...

- Potential Problems
- Likely Causes
- Tips

Problems

Enemy # 1: Noninvertable $\mathbf{H}(\hat{\boldsymbol{\beta}})$

- "Non-concavity," "non-invertability," etc.
- (Some part of) the likelihood is "flat"
- Why? (Bob Dole...)

Problems

Identification

- Possible due to functional form alone...
- "Fragile"
- Manifestation: parameter instability

Poor Conditioning

- Numerical issues
- Potentially:
    - Collinearity
    - Other weirdnesses (nonlinearities)

Potential Causes

- Misspecification. SAD!
- Missing data
- Variable scaling
- Typical $\Pr(Y)$

Hints

- T-h-i-n-k!
- Know thy data
- Keep an eye on your iteration logs...
- Don't overreach

Inference, In General

1. Pick some $\mathbf{H}_A : \boldsymbol{\Theta} = \boldsymbol{\Theta}_A$
2. Estimate $\hat{\boldsymbol{\Theta}}$
3. Determine distribution of $\hat{\boldsymbol{\Theta}}$ under $\mathbf{H}_A$
4. Use (2) and (3) $\rightarrow \hat{\mathbf{S}} \sim h(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}})$ (*test statistic*)
5. Assess $\Pr(\hat{\mathbf{S}}|\mathbf{H}_A)$

MLEs and Inference

$$\hat{\mathbf{\Theta}} \overset{a}{\sim} \mathbf{N}[\Theta, \mathbf{I}(\hat{\mathbf{\Theta}})]$$

Means that

$$\frac{\hat{\theta}_k - \theta_k}{\sqrt{\hat{\sigma}_k^2}} \sim N(0, 1)$$

Tests, Conceptually (C. Franklin remix)

- The LR asks, "Did the likelihood change much under the null hypotheses versus the alternative?"

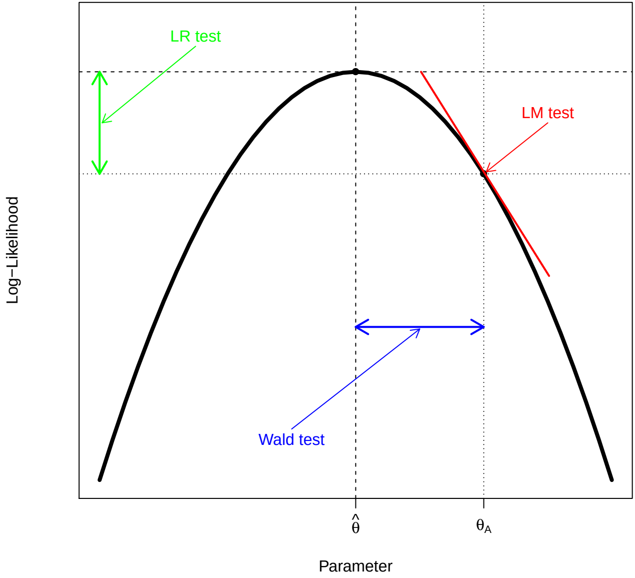
- The Wald test asks, "Are the estimated parameters very far away from what they would be under the null hypothesis?"

- The LM test asks, "If I had a less restrictive likelihood function, would its derivative be close to zero here at the restricted ML estimate?"

Tests, Conceptually (h.t.: Buse 1982)

- LR test $\approx$ manic mountaineer
- Wald test $\approx$ tired mountaineer
- LM test $\approx$ lazy mountaineer

Tests, Conceptually (A Picture)



Tests, Practically

- All are asymptotically identical...
- Require different estimates, but similar information
- Generally, LR $>$ Wald $>$ LM