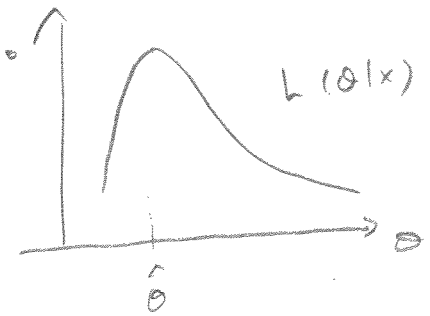
So far; • likelihood function $L(\theta|x)$

- Shape of L captures uncertainty about θ in the sample data x
- To compare $L(\theta|x)$ for different values of θ in a data-scale independent fashion we use the likelihood ratio $LR = \frac{L(\theta_1|x)}{L(\theta_2|x)}$



⇒ summarize the likelihood function numerically?

(location of maximum $\hat{\theta}_{mle} = \underset{\theta}{\operatorname{argmax}}\{S(\theta|x) = 0\}$

where $S(\theta|x) = \frac{\partial}{\partial\theta} \log L(\theta|x)$

if we can approximate $\log L(\theta|x)$ by a quadratic function we only need one more value; curvature

@ $\hat{\theta}$, $I(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2} \log L(\theta)\Big|_{\hat{\theta}}$

⇒ Quadratic approximation (around $\hat{\theta}$)

$$\log L(\theta) \approx \log L(\hat{\theta}) + \underbrace{S(\hat{\theta})}_{=0}(\theta - \hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

$$\Rightarrow \log \frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

normalized log-likelihood

} (can plot and compare LH & RH to see if quad approx ok.

$\Rightarrow$ follows that

$$S(\theta) \approx -I(\hat{\theta})(\theta - \hat{\theta})$$

or

$$I(\hat{\theta})^{-1/2} S(\theta) \approx I(\hat{\theta})^{1/2} (\hat{\theta} - \theta)$$

Can show (Wed) that

$E_\theta(S(\theta)) = 0$

$V_\theta(S(\theta)) = I(\hat{\theta})$

and approx $\sim N(0,1)$ $\Rightarrow$ $\sim N(0,1)$ $\Rightarrow$ $\hat{\theta} \sim N(\theta, I(\theta)^{-1/2})$

want distribution properties of this (bias, variance etc.)

$\Downarrow$

testing

confidence intervals etc

Note $W = 2\log\frac{L(\hat\theta)}{L(\theta)} \sim \chi^2_1$ under null $H_0: \theta$

- came from know the sampling distribution of $\bar{x}$

⇒ Interval estimate

$$\{\theta : \frac{L(\theta)}{L(\hat\theta)} > c\} = \{\theta : 2\log\frac{L(\hat\theta)}{L(\theta)} < -2\log c\}$$

$\sim \chi^2_1$

→ pick $-2\log c = \chi^2_1(1-\alpha)$ for $(1-\alpha)$ Conf. interval

or $c = e^{-\frac{1}{2}\chi^2_1(1-\alpha)}$

⇒ Can also use for hypothesis testing

$H_0: \theta = \theta_0$

Reject if $2\log\frac{L(\hat\theta)}{L(\theta_0)} > \chi^2_1(1-\alpha)$

Note If the quadratic approximation is poor

- no easy way to pick c.

For some special cases (x's and models pol) We can work out a probabilistic statement for $\frac{L(\theta)}{L(\hat\theta)}$ (see 7.6 in book)

Likelihood interval vs Confidence interval

$$\{\theta : \frac{L(\theta)}{L(\hat\theta)} > c\} \quad (*)$$

Likelihood interval

If plug in the quadratic approximation

$$\log \frac{L(\theta)}{L(\hat\theta)} \approx -\frac{I(\hat\theta)}{2}(\theta-\hat\theta)^2 \quad \text{in } (*)$$

$$\Rightarrow \{\theta : -\frac{I(\hat\theta)}{2}(\theta-\hat\theta)^2 > \log c\}$$

$$= \{\theta : (\theta-\hat\theta)^2 \le I(\hat\theta)^{-1}(-2\log c)\}$$

$$= \{\hat\theta \pm \sqrt{-2\log c}\, I(\hat\theta)^{-1/2}\} \quad \text{(Wald interval)}$$

Remember, by quad approx $\hat\theta \sim N(\theta, \underbrace{I(\hat\theta)^{-1}}_{V(\hat\theta)})$

so pick $\sqrt{-2\log c} = 1.96$ for a 95% confidence interval

Pro/Con w. Wilks vs Wald

Both are based on a quadratic approximation but Wilks (likelihood interval) allows for asymmetric intervals and tend to be better. Wald relies on approx $\frac{\hat\theta-\theta}{I(\hat\theta)^{1/2}} \sim N(0,1)$ whereas for Wilks we only need there to be some (unknown) $g(\theta)$ s.t. $g(\hat\theta)-g(\theta)/se(g(\hat\theta)) \sim N(0,1) \Rightarrow 2(\log \frac{L(\hat\theta)}{L(\theta)}) \sim \chi^2_1$

Quick recap @ multi-parameter case...

- In scalar case — can plot $L(\theta)$
- In multi-parameter case — difficult to graph directly
  - issues to consider — can likelihood be factorized? how related are the components of θ?

Notation & terminology $\underset{\sim}{\theta} = \{\theta_1, \ldots, \theta_p\}$

- $\underset{p\times 1}{S(\underset{\sim}{\theta})} = \nabla_\theta \log L(\underset{\sim}{\theta})$ = vector of 1st order derivatives
  $= \frac{\partial}{\partial \theta_j} \log L(\underset{\sim}{\theta})$

- $\underset{p\times 1}{\hat{\underset{\sim}{\theta}}_{MLE}} = \{\theta : S(\underset{\sim}{\theta}) = \underset{\sim}{0}\}$
- (observed) Fisher information matrix $I_{ij}(\hat{\theta}) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta)\Big|_{\hat{\underset{\sim}{\theta}}}$

because now need to care about curvature in high-dim space.

(can still try to approximate $L(\underset{\sim}{\theta})$ by a quadratic shape (2nd order approx)

$$\underset{1\times 1}{\log L(\underset{\sim}{\theta})} \approx \underset{1\times 1}{\log L(\hat{\underset{\sim}{\theta}})} + \underset{1\times p}{S(\hat{\underset{\sim}{\theta}})'} \underset{p\times 1}{(\underset{\sim}{\theta} - \hat{\underset{\sim}{\theta}})} - \frac{1}{2} \underset{1\times p}{(\underset{\sim}{\theta} - \hat{\underset{\sim}{\theta}})'} \underset{p\times p}{I(\hat{\underset{\sim}{\theta}})} \underset{p\times 1}{(\underset{\sim}{\theta} - \hat{\underset{\sim}{\theta}})}$$

$\underset{1\times p}{S(\hat{\underset{\sim}{\theta}})'} = \underset{\sim}{0}$

Would be nice if $I(\hat{\theta})$ was diagonal

- seperate uncertainty about components.

$\Rightarrow$ more later

Working w. multiparameter models in practice

$\Rightarrow$ In complex data situations, not so easy to solve score equation $S(\underset{\sim}{\theta}) = 0$ for all components of data simultaneously

- parameters of interest + 'nuisance' parameters
  - Ex μ, σ^2
  - (more later)
- complex dependencies between components of θ or restrictions
  - $\rightarrow$ constrained optimization, unstable solution
  - Ex. Mixture model
    - $X \sim \sum_k \pi_k N(z'\theta_k, \Sigma_k)$
    - some restricted to 0

One way to approach this is to use profile likelihood

Definition

θ - parameter of interest

η - other parameters (nuisance)

$L(\theta, \eta)$ (full) joint likelihood

$L(\theta) \equiv \max_{\eta} L(\theta, \eta)$ is called the profile likelihood for θ

max is taken for a fixed value of θ (unknown pa general)

$\Rightarrow$ usually $\Rightarrow$ MLE of η for fixed θ is a function of θ

so $\max_{\eta} (L(\theta, \eta) = \eta(\theta)$

$\Rightarrow L(\theta) = L(\theta, \eta(\theta))$ $\Rightarrow$ now maximize wrt θ

$\neq$ Integrating out η in joint distribution is a Bayesian sense

$$P(x)\pi(\theta) = \int_{\eta} P(x|\theta, \eta)\pi(\theta, \eta)d\eta$$

Example normal-mean model

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right\}$$

For fixed $\mu \Rightarrow \arg\max_{\sigma^2} L(\mu, \sigma^2) = \hat{\sigma}^2_\mu = \frac{1}{n} \sum (x_i - \mu)^2$

(μ: fixed unknown)

$$\Rightarrow L(\mu) = (\text{constant}) \times \left(\hat{\sigma}^2_\mu\right)^{-n/2}$$

1) Note: not the same as $\max_{\mu, \sigma^2} L(\mu, \sigma^2) = L(\hat{\mu}, \hat{\sigma}^2)$

and plugging in $\hat{\sigma}^2$ into $L(\mu, \sigma^2)$

2) Note: $L(\sigma^2) = \max_{\mu} L(\mu, \sigma^2)$ easy $\Rightarrow$ same $\hat{\mu}_{\sigma^2} = \hat{\mu}_{MLE} = \bar{x}$

$$= (\text{constant}) \cdot (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum (x_i - \bar{x})^2\right\}$$

$n\hat{\sigma}^2$

- Curse of dimensionality — hard to do inference in high-dim case.
- $L(\theta)$ can have complex shape and small values of the normalized likelihood not so unusual.
- So — picking reasonable cutoffs for $\frac{L(\theta)}{L(\hat{\theta})}$ ⇒ interval estimates etc, requires that we take dimensionality of the problem into account.

Recap scalar case:

Wilks stat (quad approx) large sample

$$W = 2\log\frac{L(\hat{\theta})}{L(\theta)} \sim \chi^2_1$$

⇒ What is a reasonable cut off c for $\frac{L(\theta)}{L(\hat{\theta})} > c$?

(so $L(\theta)$ not too small cmp $L(\hat{\theta})$)

$$\Rightarrow P\left(\frac{L(\theta)}{L(\hat{\theta})} > c\right) = P\left(\underbrace{2\log\frac{L(\hat{\theta})}{L(\theta)}}_{\chi^2_1} < -2\log c\right) = 1-\alpha$$

if we pick $-2\log c = \chi^2_1(1-\alpha)$

$$\left\{\theta : \frac{L(\theta)}{L(\hat{\theta})} > e^{-\frac{1}{2}\chi^2_1(1-\alpha)}\right\}$$

Testing: $H_0: \theta = \theta_0$ reject H_0 if $2\log\frac{L(\hat{\theta})}{L(\theta_0)} > \chi^2_{1(\text{thresh})}$

Multiparameter case

$$W = 2\log\frac{L(\hat{\theta})}{L(\theta)} \sim \chi^2_p \quad \text{(approx)}$$

$$\Rightarrow \left\{\theta : \frac{L(\theta)}{L(\hat{\theta})} > \exp\left\{-\frac{1}{2}\chi^2_p(1-\alpha)\right\}\right\}$$

Usually used for testing

$H_0: \theta = \theta_0 \qquad W = 2\log\frac{L(\hat{\theta})}{L(\theta_0)} \sim \chi^2_p$ under null

$\Rightarrow$ Reject H_0 if $W > \chi^2_p(1-\alpha)$

pvalue $P\left(\chi^2_p > -2\log\frac{L(\theta_0)}{L(\hat{\theta})}\right)$ (**)

observed LR

Notice how cutoff or critical value of test $\chi^2_p(1-\alpha)$ depends on p. So a LR of 25% means something different for $p=2$ or 10.

That is, we cannot used a fixed LR comparison but have to take dimensionality of problem into account

In fact, for fixed $LR = \frac{L(\theta_0)}{L(\hat{\theta})}$ one can show that the pvalue (**) $\to 1$ as $p \to \infty$ so with a fixed LR cutoff we would reject too often.

Heuristic derivation for AIC (model selection)

Want a good model, but as simple as possible

Example, null model $\theta_0 = 0$ (parameter = 0 for some subset of) parameters (like in regression modeling)

'Best' model to match data is the MLE $\hat{\theta}$, but that can be an overfit with many spurious effects estimated (all components of $\hat{\theta}$ are nonzero)

- Let's say $L(\underset{\sim}{\theta}) \approx L(\theta_1) L(\theta_2) \cdots L(\theta_p)$ (factorizes parameter components)
- In scalar case we view $\frac{L(\theta_0)}{L(\hat{\theta})} \leq c$ as an indication that θ_0 is a poor fit (for some value c, e.g. 15%)
- Let's now say we want a similar support for all components of $\underset{\sim}{\theta}$ to prefer model $\underset{\sim}{\theta}_0$

$$\frac{L(\theta_{01})}{L(\hat{\theta}_1)} > c, \quad \frac{L(\theta_{02})}{L(\hat{\theta}_2)} > c \quad \text{etc}$$

- Or $\frac{L(\underset{\sim}{\theta}_0)}{L(\underset{\sim}{\theta})} > c^p$ to "accept" $\underset{\sim}{\theta}_0$ model or

$\frac{L(\underset{\sim}{\theta}_0)}{L(\underset{\sim}{\theta})} < c^p$ to reject it (rejecting all!)

$$\frac{L(\underset{\sim}{\theta}_0)}{L(\hat{\underset{\sim}{\theta}})} \leq c^p \Rightarrow 2\log\frac{L(\hat{\underset{\sim}{\theta}})}{L(\underset{\sim}{\theta}_0)} > 2p\cdot\log(c)$$

$\Rightarrow$ Reject $H_0: \underset{\sim}{\theta} = \underset{\sim}{\theta}_0$ if

$$\underbrace{-2\log L(\hat{\theta}) + 2p}_{\log(c)} < \underbrace{-2\log L(\theta_0)}_{\log(c)}$$

Special case, $\log(c) = 1 \quad c = e^1 = 0.37$

(37% reduction in $-2\log L$ -> rejection for scalar case)

$\Rightarrow$ Reject $H_0: \underset{\sim}{\theta} = \underset{\sim}{\theta}_0$ if

$$-2\log L(\hat{\theta}) + 2p < -2\log L(\theta_0)$$

(residual) deviance — penalty — null deviance

(residual) deviance: All of model $\hat{\theta}$

null deviance: All of null model

To prefer model $\hat{\theta}$ to null, the deviance of $\hat{\theta}$ has to beat the null deviance by at least $2p$.

That is how much ($2p$) we think $-2\log L(\hat{\theta})$ can be reduced compared to fixed $-2\log L(\hat{\theta})$ just by chance

Strategy : Minimize $-2 \times$ log likelihood + penalty $2p$

$\Rightarrow$ Winning model finds a good balance between fit (-2log like) and complexity (# of parameters)

The penalty $2p$ is more aggressive than $\chi^2_p(1-\alpha)$ cutoff in Wilks so AIC tends to pick smaller models (favor the null) more often than Wilks