

## Model Comparison

SS - PSS 201 about 100

STATISTICAL INFERENCE

We've been bumping into this topic (but skirting it) since the beginning of class (discussing Stats philosophy), and in lab (with AIC comparison). But now we'll face it head on (especially interesting as we have interesting models to compare now).

- When we maximized our likelihood functions and decided that the value(s) of the parameters yielding the highest likelihood were the "best" estimate(s) of statistical params,
  - We implicitly made a decision based on differences in likelihood (or comparing likelihoods).

In the same way that we can compare likelihoods of par. values w/in a model to reach a conclusion, we can compare likelihoods (specif. the max like's) across models to gain a sense of how likely (or "probable") one model or another is relative to the data that we have.

— Essential: can only compare likelihoods of models explaining the same Y data.

We also have a sense of parsimony

- Complex models should have ↑ likelihoods due to ↑ complexity alone, but we want a sense of whether an increase in complexity increases likelihood more than due to parameters, indicating it actually has some explanatory traction.

- First we'll discuss this kind of comparison / competition between models, then investigate quantifying uncertainty w/in models
- Comparison of uncertainty

# Akaike information Criteria, Hirotsugu Akaike, 1974

See EMD book pgs 209-212

$$AIC = 2k - 2 \ln(L)$$

$$= 2k - 2 \bar{L}$$

# of pars

$\bar{L}$  how likely the model is given the data investigated

We can arrive @ different AIC values for each model investigated, and compare differences.

## Rules of Thumb (Vague)

### AIC

0-2

Support for model w/ lower AIC

Relatively equivalent

3-7

Positively clearly distinguishable

8-10

Strange

>10

Very strange definitely different

Note: AIC doesn't directly provide a test for a classic Null Hypothesis (how well a model fits data in absolute sense, w/ p-values, etc)

→ the "best" model is only judged to be the best of those examined to begin with. If all models are crap, the best of crap is still crap.

↳ simpler "Null models" can be included in the model

Set However

① AIC doesn't account for small sample sizes (only valid asymptotically)

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \quad (\text{think abt } n \rightarrow \infty)$$

Many others exist, including

$$BIC = \ln(n)k - 2\bar{L}$$

(more conservative than AIC when  $n > 8$  obsvs. or

$$\text{So } \rightarrow \ln(8) \approx 2$$

## Model Weights

$$w_i = \frac{e^{-\Delta AIC_i/2}}{\sum_{j=1}^n e^{-\Delta AIC_j/2}}$$

Useful for model averaging.

$$\text{where } \Delta AIC_i = AIC_i - AIC_{\text{Best Model}}$$

As  $\Delta AIC_i \rightarrow \infty$ ,  $e^{-\Delta AIC_i/2} \rightarrow 0$ , so  $w_i \rightarrow 0$

## Model Averaging

- What is it?

Compete  $\theta \sim \text{model 1}$  or  $\theta \sim \text{model 2}$

↳ If both are somewhat equivalent such that

We can't justify choosing 1 or 2 over the other,

We don't have to give up / pretend we know nothing

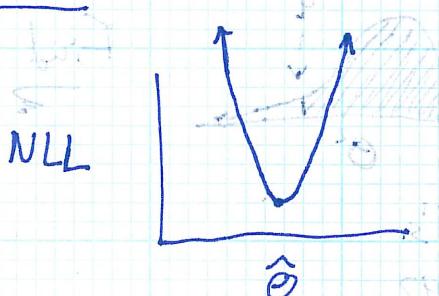
- Can make predictions by model averaging given the AIC weights

- Important, pragmatic solution allowing action in the face of scientific uncertainty.

$$\theta \sim w_1 \cdot \text{Model 1} + w_2 \cdot \text{Model 2} \dots \text{and so on.}$$

The model comparison techniques we've just investigated allow us to compare a version of likelihoods across models. Now let's return to considering uncertainty within a model in our estimate of parameters.

## Motivation!

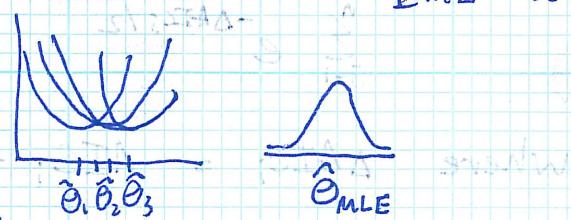


Curvature of NLL function around  $\hat{\theta}$  gives us information about how certain we can be in our estimate of the point value  $\hat{\theta}$ .

It's known that the distribution of MLE estimates  $\hat{\theta}$  (say across repeated, duplicate data sets) follows a normal distribution for large enough sample sizes

BMD book pg 191

- Assuming that  $-L(\theta | x)$  is quadratic around  $\hat{\theta}_{MLE}$

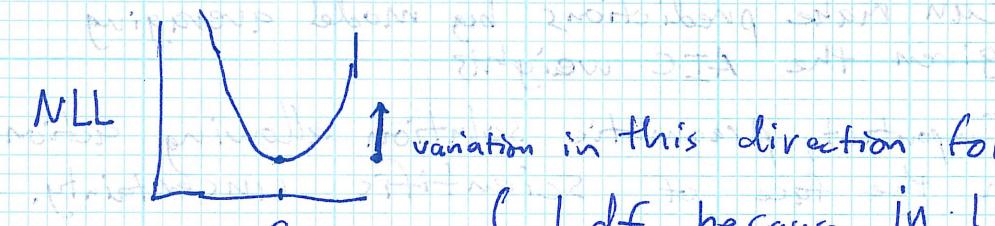


then changes in NLL resulting from variation in  $\hat{\theta}_{MLE}$  will follow

$$\sum_{i=1}^r \hat{\theta}_i^2 \quad (\text{for } r \# \text{ of parameters being estimated})$$

When  $\hat{\theta}$  is normally distributed,  $\sum \hat{\theta}_i^2 \sim \chi^2(r)$

The punch line is that if we look at NLL w/ variation in some particular  $\hat{\theta}_i$ :



variation in this direction follows  $\chi^2(1)$

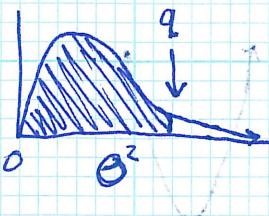
(1 df because in 1 case we fix  $\hat{\theta}_i = \hat{\theta}$ , and otherwise are considering variation in  $\hat{\theta}_i$ )

This means that we can use the  $\chi^2$  distribution to tell us in some probabilistic sense when changing  $\hat{\theta}_i$  results in a substantial enough shift in NLL as to be significantly different from our  $\hat{\theta}$ :

↳ i.e., we can estimate CI's for  $\hat{\theta}_i$ :

$$\int_0^q \chi^2(1) d\theta^2 = 0.95$$

α. level  
for our CI's.



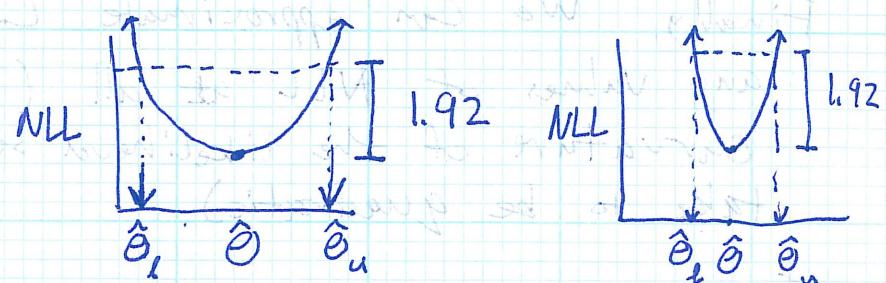
Find q

For  $\chi^2(1)$  and  $\alpha = 0.95$ ,  $q = 1.92$

So, if we can find

95% CI for  $\hat{\theta}$

is  $(\hat{\theta}_L, \hat{\theta}_U)$



(Smaller CI's)  
= Steeper NLL

For models with multiple parameters,  $\hat{\theta} = \{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n\}$   
We can look @ the likelihood surface in 2 different ways:

### ① Likelihood Slices

Plot  $-\mathbb{L}(\theta_1, \hat{\theta}_2, \hat{\theta}_3, \dots, \hat{\theta}_n | x)$

ie, fix all  $\theta_i$  at their  $\hat{\theta}_i$  except  $\theta_1$  and  
calculate the  $-\mathbb{L}$  value,  
→ determine CI's as above.

### ② Likelihood Profiles

Plot  $-\mathbb{L}(\theta_1, \hat{\theta}_2^*, \hat{\theta}_3^*, \dots, \hat{\theta}_n^*)$  (fancy  $\theta$ 's)

where  $\hat{\theta}_j^*$  is a new estimate of  $\hat{\theta}_j$  given the  
particular value of  $\theta_1$  being queried

↳ essentially re-optimize over all other  $\theta_j$  for  
each diff't value of  $\theta_1$  investigated.

Accounts for covariation btwn parameters

(changing  $\hat{\theta}_1$  makes some other choice of  $\theta_j$  better than the  
current  $\hat{\theta}_j$ ).

→ determine CI's as above.

\* Less biased, much more computation

Finally, we can approximate CI's without calculating any values of NLL at all (based only on the local curvature of the likelihood surface, which we again take to be quadratic).

## Fisher Information CI's

EMD Book pg 196 - 200

Remember that  $\frac{\partial^2 f}{\partial \theta^2}$  tells us about the curvature of function  $f$  w/ respect to changes in dimension  $\theta$ .

Applied to a likelihood surface,

$$\frac{df(\theta)}{d\theta} = \frac{f(\theta + \Delta\theta) - f(\theta)}{\Delta\theta}$$

$$\left. \frac{d^2 f(\theta)}{d\theta^2} \right|_{\theta=\hat{\theta}} = \frac{f(\hat{\theta} + \Delta\theta) - 2f(\hat{\theta}) + f(\hat{\theta} - \Delta\theta)}{(\Delta\theta)^2}$$

$\left. \frac{d^2 L}{d\theta^2} \right|_{\theta=\hat{\theta}}$  tells us about the curvature of the likelihood of  $\theta$  given data  $X$  around our MLE estimate of  $\hat{\theta}$

When we have a multiparameter  $\theta = \{\theta_1, \theta_2, \dots, \theta_n\}$

We'll actually have a matrix of partial 2nd derivatives called a Hessian

$$\text{Hessian} = \begin{pmatrix} \frac{\partial^2 L}{\partial \theta_1^2} & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_n} \\ \frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_2^2} & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L}{\partial \theta_n \partial \theta_1} & \dots & \dots & \frac{\partial^2 L}{\partial \theta_n^2} \end{pmatrix}$$

$H$  ~~E~~ ~~N~~ ~~N~~ = Fisher Information

$-E(H)$   $|_{\theta_{MLE}}$  = Observed information

Inverse of the Hessian is the estimate of the Variance-Covariance of the matrix of parameters. Don't want to invert these by hand

$\text{VCov}()$  helps us understand CI's for parameters in a manner similar to profiles

→ when assumptions are satisfied.

width of the CI's should follow

- diagonals of  $\text{VCov}$  give parameter variance
- assumed normality
- $\text{sqrt}$  gives Std. dev / Std. error values per parameter

$$\text{lower} = \hat{\theta} - 1.96 \cdot \text{s.e.}$$

$$\text{upper} = \hat{\theta} + 1.96 \cdot \text{s.e.}$$

these are what you see in mle outputs, when it can estimate the Hessian

Likelihood Ratio Test - what is nested model

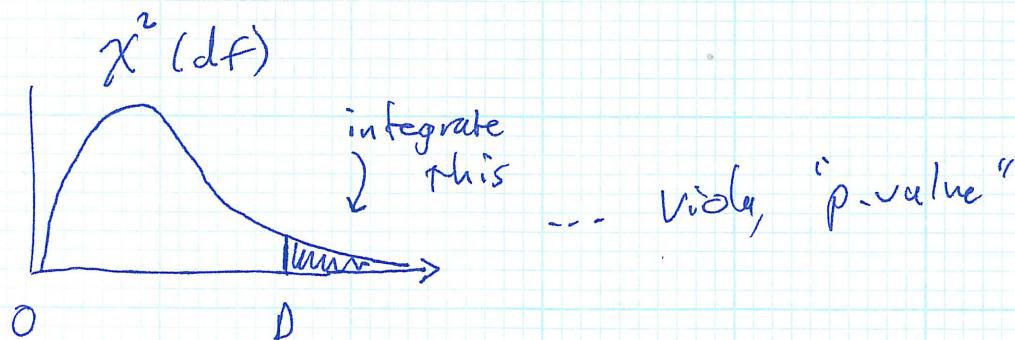
$$\text{Likelihood ratio, } D = -2 \ln \left( \frac{\text{likelihood of reduced model}}{\text{likelihood of complex model}} \right)$$

$$D = -2 \ln(\text{like reduced}) + 2 \ln(\text{alternative})$$

$D$  is  $\chi^2$  distributed test statistic

w/ df equal to dif. btwn alternative + reduced

So we can do the same thing as w/ slices + profiles



- examples of nested models, context of LTER  
fert + disturb.

Wish to find the left to right

symmetric distribution with mean  $\mu$  (normal)  $\rightarrow$   
standard deviation  $\sigma$

symmetric very often from the left to the right

$$\mu + 1.96 \sigma = 2.5 \rightarrow \text{mean}$$

$$\mu + 1.96 \sigma = 2.5 \rightarrow \text{upper}$$

standard normal  $\rightarrow$  normal with mean 0 and standard deviation 1

standard normal  $\rightarrow$  standard  $\rightarrow$  the standard distribution

(standardization to standard)

standard  $\rightarrow$  standard  $\rightarrow$  standard  $\rightarrow$  standard

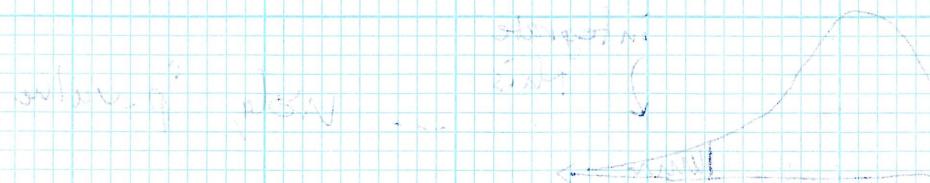
(standardization)  $\rightarrow$   $S + (\text{standard deviation}) \cdot Z = 0$

standard  $\rightarrow$  standard  $\rightarrow$  standard  $\rightarrow$  standard

standard  $\rightarrow$  standard  $\rightarrow$  standard  $\rightarrow$  standard  $\rightarrow$  standard

standard  $\rightarrow$  standard  $\rightarrow$  standard  $\rightarrow$  standard  $\rightarrow$  standard

(4b) x



standard  $\rightarrow$  standard  $\rightarrow$  standard  $\rightarrow$  standard  $\rightarrow$  standard

standard  $\rightarrow$  standard