Lecture 7 – Model-fitting - Maximum Likelihood and AIC

Announcements:

Today: lecture, paper discussion & start PS3

Concepts:

Maximum likelihood & AIC

Recall that least squares estimates of parameters are "most likely" values given the data:

$$\mathcal{L}(\beta|Y) \qquad \text{sometimes} \quad \mathcal{L}(\vec{\beta}|\vec{Y})$$

$\beta =$ (vector of) parameters
$y =$ (vector of) data

Likelihood of a particular parameter value, β, given a data point y_i is proportional to the probability of observing y_i given that β is true.

$$\mathcal{L}(\beta|y_i) \propto P(y_i|\beta)$$

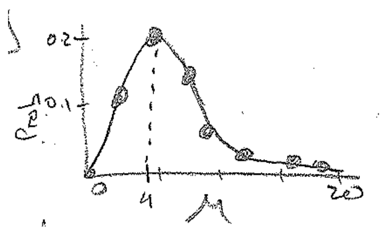
Thus, if the data Y are described by a particular distribution (e.g., Binomial, Poisson, Normal), we can quantify the likelihood using the probability density functional of that distribution.

Example: Poisson whales
Poisson describes freq. of rare events with a single parameter, μ.
(e.g., encountering a whale on an ocean transect)
Say we see 4 whales in one transect... What is the likelihood of a given value of μ?

$$\mathcal{L}(\mu|4) \propto P(4|\mu) = \frac{e^{-\mu}\mu^4}{4!}$$

$\mu =$ "encounter rate"

Evaluate over all possible values of μ.
The value that *maximizes* $P(4|\mu)$ is the MLE of $\mu \Rightarrow$ MLE of $\mu = max.\mathcal{L}(\mu|y_i)$
Show R plot



...shows that MLE of encounter rate = 4 per transect
(not surprising given 4 whales encountered in 1 transect)

Perform 2nd transect, observe 6 whales. But $P(y_i = 6|\mu = 4) = 0.1$ only (low probability).

Therefore: *Joint probability*!

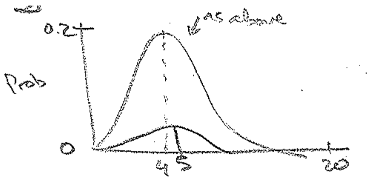
Joint probability of two independent events is the product of their probabilities.

$$P(A \cap B) = P(A) \cdot P(B)$$

Therefore:

$$\mathcal{L}(\mu|[4,6]) = \mathcal{L}(\mu|4) \cdot \mathcal{L}(\mu|6)$$

Again, evaluate over all possible μ values...
Show R plot



...shows that MLE of encounter rate = 5 per transect

But notice that joint probability declines with each additional observation!

$$\mathcal{L}(\mu|y_i) \propto \prod P(y_i|\mu)$$

Therefore take log...

$$\text{Log(small number) = - normal-sized number}$$

Therefore take negative log... That's why we use *Negative Log Likelihood* (NLL)

$$NLL(\mu|y_i) \propto \sum_i^n -log(P(y_i|\mu))$$

Because we've taken the negative $\Rightarrow$ Value that *minimizes* NLL is the MLE.

How to find MLE analytically?
Class Q: How does one find the min or max of a function?
A: Take derivative, set to zero, solve!

Back to Popn Growth data
Assume process-error only.
Process model:

$$N_{t+1} = F(N_t)$$

Assume $log\mathcal{N}$ residual error distribution, thus...

$$\ln\left(\frac{N_{t+1}}{N_t}\right) = \ln\left(\frac{F(N_t)}{N_t}\right) + \epsilon_t$$
$$\epsilon_t \sim \mathcal{N}(\mu, \sigma^2)$$

For Normal distribution:

$$-\ln \mathcal{L}(\beta \mid Y) = \frac{n}{2}\ln(2\pi\sigma_y^2) + \frac{1}{2\sigma_y^2}SSE$$

(see Morris & Doak eqn. 4.5) where

$$\sigma_y^2 = \frac{1}{n-1}\sum_i^n (y_i - \bar{y})^2.$$

In our context

$$-\ln \mathcal{L}(\beta \mid Y) = \frac{n}{2}(2\pi) - \frac{n}{2}\ln(\sigma_y^2) + \frac{1}{2\sigma_y^2}\sum_i^n (obs.growth_i - pred.growth_i)^2$$

where

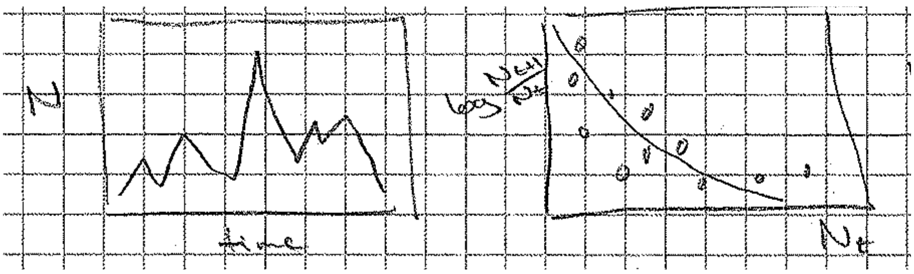
$$y_i = \ln\left(\frac{N_{t+1}}{N_t}\right)$$

such that σ_y^2 is the variance of the observed growth rates.

Note: The MLE of $\epsilon_t \sim \mathcal{N}(\mu, \sigma^2)$ = least squares estimate.
In R we can thus use: *lm* (linear least squares) or *nls* nonlinear least squares.

Model comparison

R-exercise Great tit dataset (setup for models used in PS3)

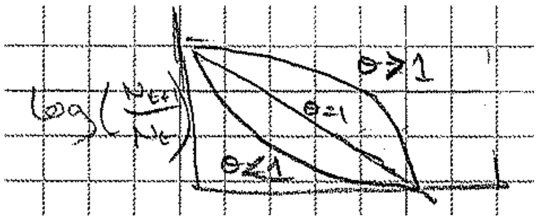


Note curvature!

Three hypothesized models:

	N_{t+1}	$\ln\left(\frac{N_{t+1}}{N_t}\right)$
Density-independent	$N_t e^r$	r
Ricker (linear DD)	$N_t e^{r(1-N/K)}$	$r\left(1-\frac{N}{K}\right)$
Theta-logistic (nonlinear DD)	$N_t e^{r(1-N/K)^\theta}$	$r\left(1-\frac{N}{K}\right)^\theta$

Note: Implicitly using $e^{r\cdot 1}$ since $\Delta t = 1$



Aside: Advise against using Theta-logistic. Has serious problems. Use in PS3 only for illustrative purposes.

For each model, plug in predicted values for each time step into NLL eqn.

	NLL
Density-independent	22.526
Ricker (linear DD)	14.299
Theta-logistic (nonlinear DD)	14.058

$\Rightarrow$ Theta-logistic fits best!

So is Theta-logistic the best model?

"Best fit", but "best-performing"??? Remember polynomial from first class!

$\Rightarrow$ Akaike Information Criterion (AIC)

Penalize models for number of parameters (p)

$$\mathrm{AIC} = 2p - 2\cdot log(\mathcal{L}_{\mathrm{MLE}}) = 2\cdot \mathrm{NLL}_{\mathrm{MLE}} + 2p$$

Small sample size correction:

$$\mathrm{AIC}_c = 2\cdot \mathrm{NLL}_{\mathrm{MLE}} + 2p\left(\frac{n}{n-p-1}\right)$$

where n is number of data points.

Model with lowest AIC is the "best-performing" model.
Typically given using ΔAIC of ith model:

$$\Delta\mathrm{AIC}_i = \mathrm{AIC}_i - min(\mathrm{AIC})$$

Relative likelihood of models - Akaike weights:

$$w_i = \frac{e^{-\frac{1}{2}\Delta\mathrm{AIC}_i}}{\sum_k e^{-\frac{1}{2}\Delta\mathrm{AIC}_k}}$$

	NLL	p	AIC	AICc	ΔAICc	w
Density-independent	22.526	1	47.05	47.2	14.14	0.006
Ricker (linear DD)	14.299	2	32.60	33.06	0	0.73
Theta-logistic (nonlinear DD)	14.058	3	34.12	35.08	2.02	0.27