Lecture 7 – Model-fitting - Maximum Likelihood and AIC

Announcements: Today: lecture & paper discussion
Next time: Bring laptops! (Playing with chaos)
Concepts: Maximum likelihood & AIC

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

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

β = (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 of that distribution (or probability density function).

Example: Poisson whales
Poisson describes freq. of rare events with a single parameter, λ.
(e.g., encountering a whale on an ocean transect)

$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, ... \qquad \text{Probability distribution}$$

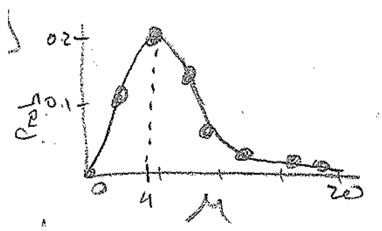
$$\mathcal{L}(\lambda|x) = \prod_{i=1}^{n} P(x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \qquad \text{Likelihood function}$$

Say we see 4 whales in one transect... What is the likelihood of a given value of λ?

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

λ = "encounter rate"

Evaluate over all possible values of λ.
The value that *maximizes* $P(4|\lambda)$ is the MLE of $\lambda \Rightarrow$ MLE of $\lambda = max.\mathcal{L}(\lambda|y_i)$
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 $P(y_i = 6|\lambda = 4) = 0.1$ only (low probability).

Therefore: *Joint probability*!

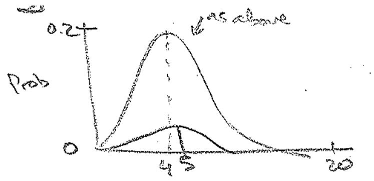
Joint probability of two independent events is the product of their probabilities.

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

Therefore:

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

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



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

But notice that joint probability declines with each additional observation!

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

Therefore take log...

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

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

$$-\ln \mathcal{L}(\lambda|y_i) \propto \sum_{i}^{n} -\ln(P(y_i|\lambda))$$

Because we've taken the negative ⇒ Value that *minimizes* NLL is the MLE.

How to find MLE of parameter analytically?
Class Q: How does one find the min or max of a function?
A: Take derivative with respect to focal parameter, 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:

$$f(y \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{-(y-\mu)^2}{2\sigma^2} \qquad \text{Probability density function}$$
$$-\ln \mathcal{L}(\mu, \sigma \mid Y) = \frac{n}{2}\ln\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

In our context of model-fitting

$$-\ln \mathcal{L}(\beta \mid Y) = \frac{n}{2}\ln\left(2\pi\sigma_y^2\right) + \frac{1}{2\sigma_y^2}\sum_{t=1}^{n}(obs.growth_t - pred.growth_t)^2$$

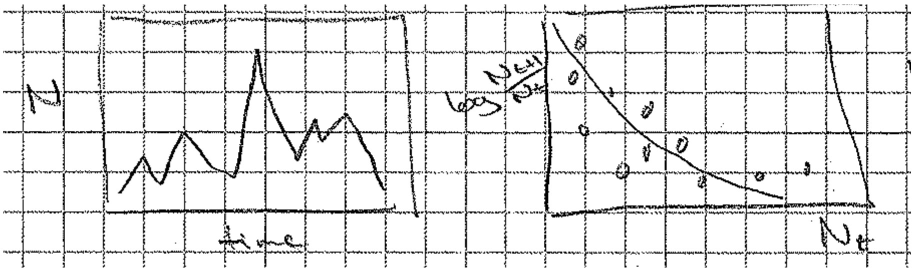
where

$$y_t = \ln\left(\frac{N_{t+1}}{N_t}\right)$$
$$\sigma_y^2 = \frac{1}{n-1}\sum_{t}^{n}(y_t - \bar{y})^2 \quad \text{=Variance of observed growth rates}$$

Remember: The MLE of $\epsilon_t \sim \mathcal{N}(\mu, \sigma^2)$ = least squares estimate.
Thus 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)

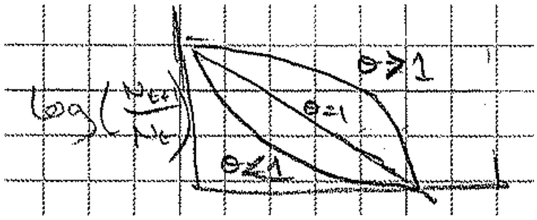


Note curvature!

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 conceptual problems. Use in PS3 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) [Note: don't forget σ for normal!]

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

Small sample size correction:

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

where n is number of data points.

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

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

Relative likelihood of models - Akaike weights:
"Probability of model given the data"

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

Paper discussion

Extra: Continuous probability functions vs. discrete probability distributions
Q: Why do I sometimes write

$$\mathcal{L}(\beta \mid Y) \propto P(Y \mid \beta),$$

and other times

$$\mathcal{L}(\beta \mid Y) = P(Y \mid \beta)$$

It turns out that both are in some ways correct! There are two relevant distinctions:

Distinction 1: $\mathcal{L}(\beta \mid Y) \propto P(Y \mid \beta)$ is correct for *continuous* distributions while $\mathcal{L}(\beta \mid Y) = P(Y \mid \beta)$ is correct for *discrete* distributions. The reason is that, unlike for a discrete distribution, the probability of any specific value on a continuous distribution is zero! It's only over some interval of values that we can speak of a continuous distribution having some probability.

Distinction 2: However, the equality is still correct when the right hand side is not a *probability distribution*, $P(Y \mid \beta)$, but rather a *probability density function*, $f(Y \mid \beta)$. These differ even though $\int f(Y \mid \beta) d\beta = \sum P(Y \mid \beta) = 1$ (i.e the total area under both equals 1). Most people loosely (but technically incorrectly) use these two terms interchangeably.