Maximum Likelihood Estimation Part I

October 23rd 2014

Goals

- Introduce you to the concept of likelihood (I.e. relative probabilities).
- Evaluate a simple model with the concept of likelihood.
- Familiarize you with the basic computational tools in R to evaluate likelihoods.

Resources

- Bolker,B. EMD book 2008. Pp 170-176, 182-185, 187-194, 204-208, 298-303.
- Burnham and Anderson: sections 1.2.1& 1.2.2

I have also posted a useful review article.

Resources

- Lynch, M & Walsh, B. 1997. Genetics and the Analysis of Quantitative traits. Appendix #4
- Bruce Walsh's Course notes as well.

http://nitro.biosci.arizona.edu/courses/EEB519A-2007/pdfs/MLE.pdf

These provide a short but detailed algebraic description of MLE and LRT.

Relationship between Estimation methods

MCMC

Bayesian Estimation

$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)*p(\theta)$

Resampling methods

Maximum Likelihood

$\mathrm{Max}(L(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta))$

Monte Carlo Simulation

Least Squares

$$Min(\sum \varepsilon_i^2)$$

$\varepsilon \sim N(0, \varepsilon^2/\mathrm{n})$

Least Squares Estimates of Parameters for Linear Models

- Minimize the residual error (unexplained variation) of the model.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Criterion for optimization (objective function): Minimize $\Sigma(\varepsilon_i)^2$

- Parameter estimation is non-parametric.
- Most inferences on LSE require assumptions that errors are normally distributed (from sampling theory).

Maximum Likelihood estimation

- What is the probability of observing the data (X) , given a set of parameters & a model (θ)

  $P(X|\theta)$
- Reframed as the likelihood of θ given X, $l(\theta|X)$
- Criterion for estimation: Finding the value(s) of θ that maximizes the probability of observing the data, X.

  θ^* where $\max(p(X)) = p(X|\theta^*)^{\S}$

Likelihood

- Gives same estimates of coefficients as Least Squares when errors are independent and normally distributed.

- Much more general - easy to deal with non-normal errors.

- Still requires the specification of a probability distribution for inference (parametric) but does not need to be normal.

- Developed by R.A. Fisher (1926, 1934, 1950)
- Much more intuitive

- Often involves an iterative approach - computer intensive

- Forms the basis for Bayesian inference

Likelihood

- P(data|parameters)
  - Estimate parameters (e.g., μ, σ)
  - Then look at probability of observing data given estimates or null hypothesis parameters
    - $P(\text{data}|H0{:}\mu_0=0)$
- Likelihood:
  - $l(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})$

Sometimes people refer to “hypothesis” instead of “parameters” in this context. I will use them interchangeably.

Likelihood defined

The likelihood l ($\theta|x$) of the parameters (θ) given the data (x), is proportional to p (x|θ) with the constant of proportionality being arbitrary.

Edwards 1992

$$l\,(\theta|x) \propto p\,(x|\theta)$$

General Notation

- l $(\theta|x, g)$
- Is read as the likelihood of a parameter (θ), given the data, x, and the model, g.

Sometimes just l $(\theta|x)$

Log likelihood

- For computational and algebraic simplification we use log transformed values of the probabilities.

- $L(\theta|x) = \ln(l(\theta|x))$

- $L(\theta|x)$ is referred to as the **support function**.

Algebraically

$$l(\theta) \propto p(X;\theta) = p(X \mid \theta) = \prod_{i=1}^{N} p(x_i;\theta)$$

$$L(\theta \mid X) \propto \sum_{i=1}^{N} \ln(p(x_i;\theta))$$

The likelihood of theta equals the joint probability density function, which is (given independent observations) equal to the product of the individual probabilities (or the sum of the log transformed probabilities).

Salient point….

- For each particular set of estimated values for the parameters, you will get a single number out (the likelihood).
- Eg..

$\boldsymbol{\theta}_1 = (\mu_1, \sigma_1)$ ..... $L(\boldsymbol{\theta}_1)$

$\boldsymbol{\theta}_2 = (\mu_2, \sigma_2)$ ..... $L(\boldsymbol{\theta}_2)$

$\boldsymbol{\theta}_3 = (\mu_3, \sigma_3)$ ..... $L(\boldsymbol{\theta}_3)$

What we just did in R

- L(mean| data)

$$l(\theta) = \prod_{i=1}^{N} 2\pi^{-\frac{1}{2}} e^{\frac{-(x_i - \mu)^2}{2}}$$

```
sample.prob <- dnorm(x=femur.sample, mean=0.55, sd=1)

prod(sample.prob)
```

What we want to find is the parameter estimates that maximize the likelihood (biggest single number)

Sometimes we are interested in the negative log likelihood, in which case we are looking for a minimum.

Principle of Maximum Likelihood

Find an estimate for θ such that it maximizes the likelihood of observing the data that were actually observed. In other words, given a sample of observations **x** for the random variable **X**, find the solution for θ that maximizes the joint probability function $p(\mathbf{x}|\theta)$.

Eliason 1993

How do we find the MLE?

Basic calculus reminds us that we can find minima and maxima by taking the derivative of our likelihood function with respect to the parameter(s) of interest.

For a model with a single parameter

$$S(\theta) = \frac{dL(\theta)}{d\theta}$$

S is called the **Score function**. Evaluating S(θ) = 0 will provide the MLE for θ.

More Generally

$$\mathbf{S}(\theta) = \frac{\partial L(\theta)}{\partial \theta} = \begin{matrix} \dfrac{\partial L(\theta)}{\partial \theta_1} \\ \dfrac{\partial L(\theta)}{\partial \theta_2} \\ \dfrac{\partial L(\theta)}{\partial \theta_N} \end{matrix}$$

i.e. partial derivative of the likelihood with respect to the mean.

i.e. partial derivative of the likelihood with respect to the variance.