Maximum Likelihood

FW8051 Statistics for Ecologists

Department of Fisheries, Wildlife and Conservation Biology

Learning Objectives

Understand how to use Maximum Likelihood to estimate parameters in statistical models

Understand how to create confidence intervals for parameters estimated using Maximum Likelihood

Estimation

We've covered a number of statistical distributions, described by a small set of parameters.

- How do we determine appropriate values of the parameters?
- How do we incorporate the effects of covariates?

Methods of estimation:

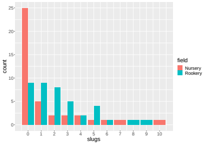
- Least squares
- Maximum likelihood
- Bayesian methods

Example from Crawley 2002. Statistical Computing and also his The R Book (2007).

- Counted slugs in 2 fields (rookery, nursery)
- 40 observations in each

Barplot

```
ggplot(slugs, aes(slugs, fill=field))+
  geom_bar(position=position_dodge())+
  theme(text = element_text(size=20))+
   scale_colour_colorblind()+
   scale_x_continuous(breaks=seq(0,11,1))
```



Hypothesis test

What if we want to use a t-test to test $H_0 : \mu_{rookery} = \mu_{nursery}$? What do we have to assume?

- The data are normally distributed or sample size is "large enough" for the CLT to apply

Are these assumptions reasonable in light of:

- We have counts (discrete data)
- There were 40 tiles in each of the 2 grasslands (field, nursery)

. . . maybe (n = 30 is a common rule for CLT to apply)

Also, side note, is this an interesting hypothesis to test?

Alternative Statistical Distributions

Are there other, more appropriate statistical distributions we could use (instead of Gaussian)?

Given that we have count data, we might consider a Poisson or Negative Binomial distribution for the data

We could assume:

- Nursery $\sim$ $Poisson(\lambda_1)$
- Rookery $\sim$ $Poisson(\lambda_2)$

Test whether $\lambda_1 = \lambda_2$

How would we estimate the parameters?

Lets start with the simper case of $Y_i \sim Poisson(\lambda)$ (ignoring field type)

Maximum Likelihood

Start by writing down a probability statement regarding the data.

Consider the first data point from the Nursery (3 slugs):

$P(X = 3) = \frac{\exp(-\lambda)(\lambda)^3}{3!}$ if the counts are Poisson distributed

or. . .

$P(X = 3) = \binom{3+\theta-1}{3}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{3}$ if NegBinomial

What about the other observations?

Constructing the Likelihood

Assume the data come from a random sample, and that the points are independent. Then:

$P(X_1 = 3 \;\&\; X_2 = 0 \;\&\; \cdots X_{40} = 4) =$

$$P(X_1 = 3)P(X_2 = 0)\cdots P(X_{40} = 4)$$

$$= \frac{\exp(-\lambda)(\lambda)^3}{3!}\frac{\exp(-\lambda)(\lambda)^0}{0!}\cdots\frac{\exp(-\lambda)(\lambda)^4}{4!}$$

More Generally: Construction of the Likelihood

We obtain a random sample of n observations from some statistical distribution.

Write down the probability of obtaining the data:

$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_n = x_n)$
$= \prod_{i=1}^{n} P(X_i = x_i)$

For the Poisson distribution:

$$L(\lambda; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i}\exp(-\lambda)}{x_i!}$$

$$= \frac{\exp(-\lambda)(\lambda)^{x_1}}{X_1!}\frac{\exp(-\lambda)(\lambda)^{x_2}}{x_2!}\cdots\frac{\exp(-\lambda)(\lambda)^{x_n}}{x_n!}$$

$$= \frac{\exp(-n\lambda)(\lambda)^{\sum_{i=1}^{n} x_i}}{x_1!x_2!\cdots x_n!}$$

This gives us the Likelihood of the data!

Likelihood

For discrete distributions, the likelihood gives us the probability of obtaining the observed data for a particular set of parameters (in this case, λ).

$$\text{P(data; }\lambda) = \frac{\exp(-n\lambda)(\lambda)^{\sum_{i=1}^{n} x_i}}{x_1!x_2!\cdots x_n!} = L(\lambda;\text{ data})$$

P(data; parameter):

- Views the data as random, the parameter as fixed.

Likelihood(parameters; data):

- Conditions on the data and considers probability as a function of the parameter.

The maximum likelihood estimate is the value of the parameter, λ, that maximizes the likelihood (*makes the the observed data most likely*)

Maximum Likelihood

The maximum likelihood estimate is the value of the parameter, λ, that *makes the the observed data most likely* (i.e., maximizes the likelihood)

$$L(\lambda; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i}\exp(-\lambda)}{x_i!}$$

How can we find the value of λ that maximizes $L(\lambda; x_1, x_2, \ldots, x_n)$?

Calculus (take derivatives with respect to λ and set = 0).

Log-likelihood

For practical and theoretical reasons, we usually work with the log-likelihood (maximizing the log-likelihood is equivalent to maximizing the likelihood)

$$logL(\lambda; x_1, x_2, \ldots, x_n) = log(L(\lambda; x_1, x_2, \ldots, x_n))$$
$$= \log(\prod_{i=1}^n P(X_i = x_i))$$
$$= \sum_{i=1}^n log(P(X_i = x_i))$$

For the Poisson model:

$$logL(\lambda; x_1, x_2, \ldots, x_n) = -n\lambda + \log(\lambda)\sum_{i=1}^n X_i - \sum_{i=1}^n \log(x_i!)$$

To maximize, take derivatives and set the expression = 0, giving:

$$-n + \frac{\sum_{i=1}^n X_i}{\lambda} = 0$$
$$\Rightarrow \hat{\lambda} = \sum_{i=1}^n \frac{X_i}{n}$$

Some notes

To verify that $\hat{\lambda}$ maximizes (rather than minimizes) $logL(\lambda|x)$:

- Verify that the $\frac{\partial^2 logL(\lambda;x)}{\partial\lambda^2}$ evaluated at $\lambda = \hat{\lambda} = \bar{x} < 0$

Note: If $X \sim$ Poisson(λ), $E[X] = \lambda$

- It makes sense to estimate λ by the sample mean!

Some constants, e.g., $\sum_{i=1}^n \log(x_i!)$ do not matter when maximizing the likelihood

- Statistical software may drop/ignore these
- Can matter when comparing models for different probability distributions using AIC

Finding the “best” value of λ

What if we do not remember calculus? How can we find the value of λ that maximizes:

$$L(\lambda; x_1, x_2, \ldots, x_n) = \prod_{i=1}^n \frac{\lambda^{x_i}\exp(-\lambda)}{x_i!}$$

Graph this expression for different values of λ

[Excel in-class exercise]

Finding the “best” values for λ_1 and λ_2

What if we had a function of more than 1 parameter? How could we numerically find the value of λ that maximizes:

$$L(\lambda_1, \lambda_2; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n_{field}} \frac{\lambda_1^{x_i}\exp(-\lambda_1)}{x_i!} \prod_{j=1}^{n_{rookery}} \frac{\lambda_2^{x_j}\exp(-\lambda_2)}{x_j!}$$

Use *solver* in Excel or `optim` (or `glm`) in R

[In-class exercise R]

Optim? When would you use something like this?

Bolker, B.M. 2008. Ecological Models and Data in R. Princeton University Press, Oxford, UK.

Tadpole predation: Example 6.3.1.1 starting on p. 182

$$p = \frac{a}{1+ahN}$$
$$k \sim \text{Binomial}(p, N)$$

- N = number of tadpoles in a tank
- k = number eaten by predators

We will come back to this or a similar example (in this section & later after introducing Bayesian methods).

Properties of Maximum Likelihood Estimators

$\hat{\theta}$ = maximum likelihood estimate of θ.

For large n (asymptotically):

- Maximum likelihood estimators are unbiased (not always true for small n):
  - $\sigma^2_{MLE} = \sum(x_i - \mu)^2/n$ (biased by a factor of $n/(n-1)$)
- Have minimum variance among estimators
- Will be normally distributed: $\hat{\theta} \sim N(\theta, I^{-1}(\theta))$

$I(\theta)$ is called the Information matrix

Information Matrix

Observed information matrix, observed $I(\theta) = -\frac{\partial^2 logL(\theta)}{\partial\theta^2}$ evaluated at $\theta = \hat{\theta}$

Estimated information matrix, expected $I(\theta) = E\left(-\frac{\partial^2 logL(\theta)}{\partial\theta^2}\right)$ evaluated at $\theta = \hat{\theta}$

The matrix of second derivatives of logL with respect to θ is called the Hessian:

Hessian$(\theta) = \left[\frac{\partial^2 logL(\theta)}{\partial\theta^2}\right]$

Observed Information Matrix

- Used to numerically maximize functions (get for "free") (note: typically we *minimize* -logL rather than maximize logL, so the minus sign is already included)
- Inverse of observed information matrix is usually what is reported as $var(\hat{\theta})$ by statistical software

Hessian

The Hessian(θ) = $\left[\frac{\partial^2 logL(\theta)}{\partial \theta^2}\right]$ describes curvature in the log-likelihood curve (surface)

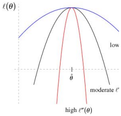


Fig. 1 Curvature and information

If $\left[\frac{\partial^2 logL(\theta)}{\partial \theta^2}\right]$ is close to 0

- The likelihood surface is flat
- LogL is similar across a range of parameter values

Leads to larger confidence intervals since $\widehat{var}(\hat{\theta}) = I^{-1}(\theta) =$ Hessian$^{-1}(\theta)$

Curvature	Information	$\mathrm{Var}(\hat{\theta})$	Confidence interval for θ
high	high	low	narrow
low	low	high	wide

Likelihood Ratio Test

A likelihood ratio test can be used to test nested models with:

- The same probability generating mechanism (i.e., same statistical distribution)
- All of the same parameters, except that in one model some parameters are set to specific values (typically 0)

Slug data example:

Full model:

- $Y_i|$ Nursery $\sim Poisson(\lambda_1)$
- $Y_i|$ Rookery $\sim Poisson(\lambda_2)$

Reduced model:

- $Y_i \sim Poisson(\lambda)$ (i.e., $\lambda_2 = \lambda_1$)

Likelihood Ratio Test

Test statistic:

$$LR = 2log\left[\frac{L(\lambda_1, \lambda_2|Y)}{L(\lambda|Y)}\right] = 2[logL(\lambda_1, \lambda_2|Y) - logL(\lambda|Y)]$$

Null distribution (appropriate when n is large):

$$LR \sim \chi^2_1$$

…and more generally χ^2_p, where p is the difference in the number of parameters in the two models.

[See Section 10.9 in book]

Profile Likelihood Confidence Intervals

Can “invert” the LR test to get profile likelihood-based confidence intervals. Consider generating a CI for λ under the common λ model.

We could use the likelihood ratio test to evaluate $H_0 : \lambda = \lambda_0$ vs. $H_A : \lambda \neq \lambda_0$:

$$LR = 2log\left[\frac{L(\hat{\lambda}|Y)}{L(\lambda_0|Y)}\right] \sim \chi^2_1$$

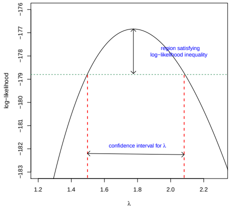
where $\hat{\lambda}$ is the MLE of λ.

- Reject $\lambda = \lambda_0$ at $\alpha = 0.05$ if LR $> \chi^2_1(0.95)$, where $\chi^2_1(0.95)$ is the 95% of the χ^2_1 distribution.
- Fail to reject if LR $< \chi^2_1(0.95)$ (these values are plausible, given the data)

CI for λ: include all values for which we do not reject the null hypothesis

Profile Likelihood Intervals

So, include in our CI all values of λ that lie within $\chi^2_1(0.95)$ = `qchisq(alpha, df=1)/2` = 1.92 units of the maximum.



Profile Likelihood Intervals

- Can extend to multi-parameter models
- Typically more accurate than normal-based CIs (Wald intervals) when n is small.

See Chapter 6 in Bolker's book (listed in **Readings** section on Canvas).

Least Squares and Maximum Likelihood

For Normally distributed data:

$L(\mu, \sigma^2; y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)$

With linear regression, we assume $Y_i \sim N(\beta_0 + x_i\beta_1, \sigma^2)$, so...

$$L(\beta_0, \beta_1, \sigma; x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i-\beta_0+x_i\beta_1)^2}{2\sigma^2}\right)$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^{n} \frac{(y_i-\beta_0+x_i\beta_1)^2}{2\sigma^2}\right)$$

$$\Rightarrow logL = -nlog(\sigma) - \frac{n}{2}log(2\pi) - \sum_{i=1}^{n} \frac{(y_i-\beta_0+x_i\beta_1)^2}{2\sigma^2}$$

$$\Rightarrow \text{maximizing logL} \Rightarrow \text{minimizing } \sum_{i=1}^{n} \frac{(y_i-\beta_0+x_i\beta_1)^2}{2\sigma^2}$$

$$\text{or, equivalently } \sum_{i=1}^{n}(y_i - \beta_0 - x_i\beta_1)^2$$