Chapter 7
Point Estimation
Maximum Likelihood

Outline

Introduction

Procedure

- Bernoulli Trials
- Normal Random Variables
- Uniform Random Variables
- Mark and Recapture
- Linear Regression

Introduction

We begin with observations $X = (X_1, \ldots, X_n)$ of random variables chosen according to one of a family of probabilities P_θ indexed by the parameter space, Θ. In addition,

$$\mathbf{f}(\mathbf{x}|\theta), \quad \mathbf{x} = (x_1, \ldots, x_n)$$

will be used to denote the joint density function when θ is the true state of nature.

Definition. The likelihood function is the density function regarded as a function of θ.

$$\mathbf{L}(\theta|\mathbf{x}) = \mathbf{f}(\mathbf{x}|\theta), \;\; \theta \in \Theta.$$

The maximum likelihood estimate (MLE),

$$\hat{\theta}(\mathbf{x}) = \arg\max_{\theta \in \Theta} \mathbf{L}(\theta|\mathbf{x}).$$

Thus, we are presuming that a *unique* global maximum exists.

Introduction

This class of estimators has two important properties.
If $\hat{\theta}(\mathbf{x})$ is a maximum likelihood estimate for θ,

- then $g(\hat{\theta}(\mathbf{x}))$ is a maximum likelihood estimate for $g(\theta)$.
    - If $\hat{\theta}$ is the maximum likelihood estimate for the variance, then $\sqrt{\hat{\theta}}$ is the maximum likelihood estimator for the standard deviation.
- and if $T(\mathbf{x})$ is a minimal sufficient statistic, then $\hat{\theta}$ is a function of $T(\mathbf{x})$
    - Form the Neyman-Fisher Factorizaton Theorem

$$\mathbf{L}(\theta|\mathbf{x}) = \mathbf{f}(\mathbf{x}|\theta) = h(\mathbf{x})g(\theta, T(\mathbf{x})).$$

and the argurment for θ in the maximization step depend only on $T(\mathbf{x})$

Introduction

For independent observations, the likelihood

$$\mathbf{L}(\theta|\mathbf{x}) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta).$$

is the product of density functions. Using the properties of the logarithm of a product,

$$\ln \mathbf{L}(\theta|\mathbf{x}) = \ln f(x_1|\theta) + \ln f(x_2|\theta) + \cdots + \ln f(x_n|\theta).$$

Finding zeroes of the score function, $\partial \ln \mathbf{L}(\theta|\mathbf{x})/\partial\theta$, the derivative of the logarithm of the likelihood, will be easier.

Bernoulli Trials

If the experiment consists of n Bernoulli trials with success probability p, then

$$\mathbf{L}(p|\mathbf{x}) = p^{x_1}(1-p)^{(1-x_1)}\cdots p^{x_n}(1-p)^{(1-x_n)} = p^{(x_1+\cdots+x_n)}(1-p)^{n-(x_1+\cdots+x_n)}.$$

$$\ln \mathbf{L}(p|\mathbf{x}) = \ln p(\sum_{i=1}^{n} x_i) + \ln(1-p)(n - \sum_{i=1}^{n} x_i) = n(\bar{x}\ln p + (1-\bar{x})\ln(1-p)).$$

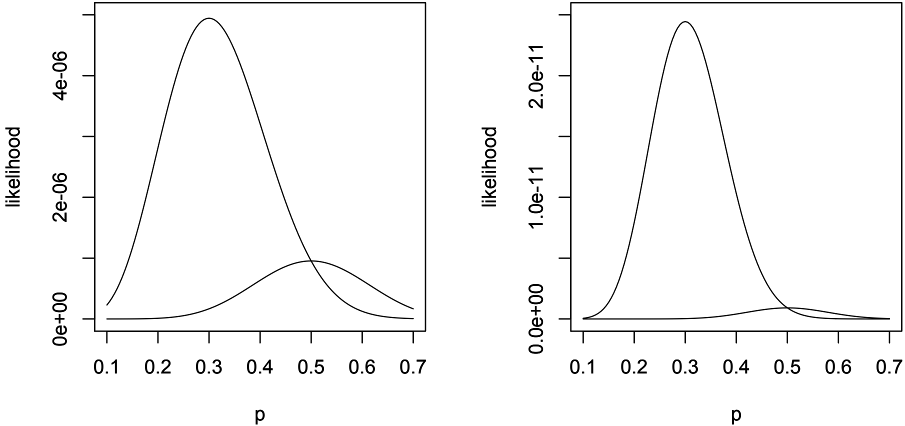
$$\frac{\partial}{\partial p}\ln \mathbf{L}(p|\mathbf{x}) = n\left(\frac{\bar{x}}{p} - \frac{1-\bar{x}}{1-p}\right) = n\frac{\bar{x}-p}{p(1-p)}$$

This equals zero when $p = \bar{x}$, the minimal sufficient statistic.

Exercise. Check that this is a maximum.
Check values both above and below $p = \bar{x}$ and use the first derivative test.

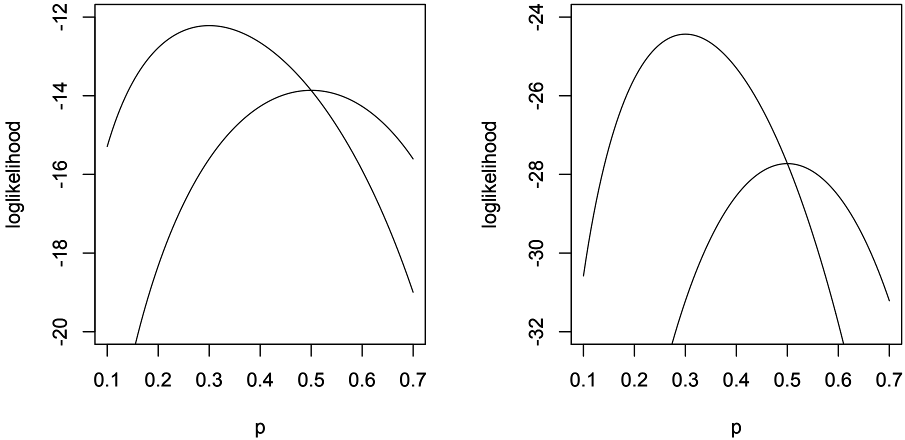
In this case, the maximum likelihood estimator is also unbiased.

Bernoulli Trials



Graph of $\mathbf{L}(p|\mathbf{x})$ with (left) 6 and 10 successes in 20 trials and (right) 12 and 20 successes in 40 trials.

Bernoulli Trials



Graph of $\ln \mathbf{L}(p|\mathbf{x})$ with (left) 6 and 10 successes in 20 trials and (right) 12 and 20 successes in 40 trials.

Bernoulli Trials

Notice

- Both $\mathbf{L}(p|\mathbf{x})$ and $\ln \mathbf{L}(p|\mathbf{x})$ have their maximum at $p = \bar{x}$.
- The maxima when $\bar{x} = 0.3$ is greater than the corresponding maxima when $\bar{x} = 0.5$. However, for the case $n = 20$ there is a factor of

$$\binom{20}{10} \Big/ \binom{20}{6} = \frac{143}{30}$$

  that produce 10 successes than produce 6.
- The maxima are more peaked with larger values of n.
  - We will soon learn that the variance in the estimator is closely tied to the curvature of the log likelihood function at the maximum likelihood estimate.

Normal Random Variables

For a simple random sample of n normal random variables, we can use the properties of the exponential function to simplify the likelihood function.

$$\begin{aligned} \mathbf{L}(\mu, \sigma^2|\mathbf{x}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_1-\mu)^2}{2\sigma^2}\right) \cdots \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_n-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2. \end{aligned}$$

The log-likelihood $\ln \mathbf{L}(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2$.

The score function is now a vector $\left(\frac{\partial}{\partial \mu} \ln \mathbf{L}(\mu, \sigma^2|\mathbf{x}), \frac{\partial}{\partial \sigma^2} \ln \mathbf{L}(\mu, \sigma^2|\mathbf{x})\right)$. Next we find the zeros to determine the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$.

Normal Random Variables

$$\ln \mathbf{L}(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$0 = \frac{\partial}{\partial \mu} \ln \mathbf{L}(\hat{\mu}, \hat{\sigma}^2|\mathbf{x}) = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = \frac{1}{\hat{\sigma}^2} n(\bar{x} - \hat{\mu}).$$

Because the second partial derivative with respect to μ is negative, $\hat{\mu}(\mathbf{x}) = \bar{x}$ is the maximum likelihood estimator. For the derivative with respect to σ^2,

$$0 = \frac{\partial}{\partial \sigma^2} \ln \mathbf{L}(\hat{\mu}, \hat{\sigma}^2|\mathbf{x}) = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = -\frac{n}{2(\hat{\sigma}^2)^2}\left(\hat{\sigma}^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right).$$

Recalling that $\hat{\mu}(\mathbf{x}) = \bar{x}$, we obtain a *biased estimator*,

$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Uniform Random Variables

If our data $X = (X_1, \ldots, X_n)$ are a simple random sample drawn from uniformly distributed random variable whose maximum value θ is unknown, then each random variable has density

$$f(x|\theta) = \begin{cases} 1/\theta & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the joint density or the likelihood

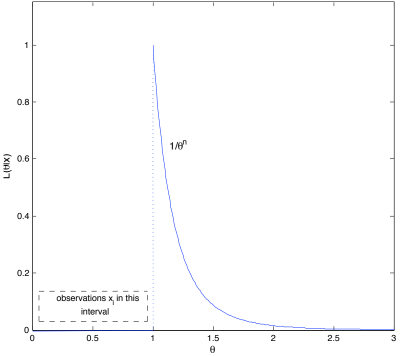
$$\mathbf{f}(x|\theta) = \mathbf{L}(\theta|\mathbf{x}) = \begin{cases} 1/\theta^n & \text{if } 0 \le x_i \le \theta \text{ for all } i, \\ 0 & \text{otherwise.} \end{cases}$$

The joint density is zero whenever *any* of the $x_i > \theta$. Consequently, any value of θ less than any of the x_i has likelihood 0. Symbolically,

$$\mathbf{L}(\theta|\mathbf{x}) = \begin{cases} 0 & \text{for } \theta < \max_i x_i = x_{(n)}, \\ 1/\theta^n & \text{for } \theta \ge \max_i x_i = x_{(n)}. \end{cases}$$

Uniform Random Variables

As promised, $\hat{\theta}$ is a function of $T(\mathbf{x}) = \max_i x_i$ the minimal sufficient statistic.



Likelihood function for uniform random variables on the interval $[0, \theta]$. The likelihood is 0 up to $T(\mathbf{x}) = \max_{1 \le i \le n} x_i$ and $1/\theta^n$ afterwards. Thus, $\hat{\theta}(\mathbf{x}) = T(\mathbf{x})$

Uniform Random Variables

We have seen that the density

$$f_{T(X)}(t|\theta) = \frac{nt^{n-1}}{\theta^n}, 0 < t \leq \theta.$$

Thus,

$$E_\theta T(X) = \int_0^\theta t f_{T(X)}(t|\theta)\ dt = \int_0^\theta \frac{nt^n}{\theta^n}\ dt = \frac{n}{n+1}\frac{t^{n+1}}{\theta^n}\Big|_0^\theta = \frac{n}{n+1}\theta < \theta.$$

Consequently, $T(X)$ is biased downward and

$$\frac{n+1}{n}T(X)$$

is unbiased.

Mark and Recapture

We return to consider Lincoln-Peterson method of mark and recapture and find its maximum likelihood estimate. Recall that

- t be the number captured and tagged,
- k be the number in the second capture,
- r be the number in the second capture that are tagged, and let
- N be the total population size.

Thus, t and k is under the control of the experimenter. The value of r is random and the populations size N is the parameter to be estimated.

Mark and Recapture

The likelihood function for N is the hypergeometric distribution

$$L(N|r) = \binom{t}{r}\binom{N-t}{k-r}\Big/\binom{N}{k}.$$

Exercise. Show that the maximum likelihood estimate

$$\hat{N} = \left[\frac{tk}{r}\right].$$

where $[\cdot]$ mean the greatest integer less than.

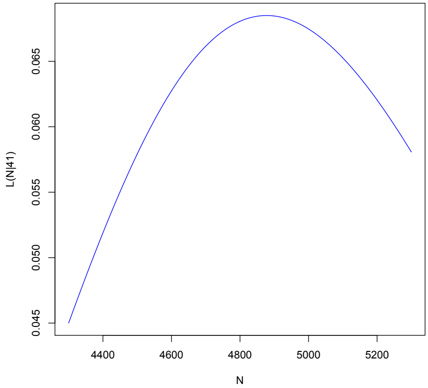
Hint: Find the values of N for which $L(N|r)/L(N-1|r) > 1$.

Thus, the maximum likelihood estimate is, in this case, obtained from the method of moments estimate by rounding down to the next integer.

Mark and Recapture

We return to the simulation of a lake having 4500 fish.

```
> N<-4500;t<-400;k<-500
> fish<-c(rep(1,t),rep(0,N-t))
> (r<-sum(sample(fish,k)))
[1] 41
> (Nhat<-floor(k*t/r))
[1] 4878
> N<-c(4300:5300)
> L<-dhyper(r,t,N-t,k)
> plot(N,L,type="l",
  ylab="L(N|41)",col="blue")
```



Plot of likelihood from the simulation with $r = 41$. The maximum $\hat{N} = 4878$.

Linear Regression

Our data are n observations. The responses y_i are linearly related to the explanatory variable x_i with an error ϵ_i,

$$y_i = \alpha + \beta x_i + \epsilon_i.$$

Here we take the ϵ_i to be independent $N(0, \sigma)$ random variables. Our model has three parameters, the intercept α, the slope β, and the variance of the error σ^2.
Thus, the joint density for the ϵ_i is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{\epsilon_1^2}{2\sigma^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{\epsilon_2^2}{2\sigma^2} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{\epsilon_n^2}{2\sigma^2} = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2$$

Since $\epsilon_i = y_i - (\alpha + \beta x_i)$, the likelihood function,

$$L(\alpha, \beta, \sigma^2 | \mathbf{y}, \mathbf{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

Linear Regression

The logarithm

$$\ln L(\alpha, \beta, \sigma^2 | \mathbf{y}, \mathbf{x}) = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

Consequently, maximizing the likelihood function for the parameters α and β is equivalent to minimizing

$$SS(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

The principle of maximum likelihood is equivalent to the least squares criterion.

Principle of Least Squares

This principle leads to a minimization problem for

$$SS(\alpha,\beta)=\sum_{i=1}^{n}\epsilon_i^2=\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2.$$

Let's the denote by $\hat{\alpha}$ and $\hat{\beta}$ the value for α and β that minimize SS.

$$\frac{\partial}{\partial\alpha}SS(\alpha,\beta)=-2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)$$

At the values $\hat{\alpha}$ and $\hat{\beta}$, this partial derivative is 0. Consequently,

$$0=\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)\qquad \sum_{i=1}^{n}y_i=\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)\qquad \bar{y}=\hat{\alpha}+\hat{\beta}\bar{x}.$$

Thus, we see that the center of mass point $(\bar{x},\bar{y})$ is on the regression line.

Principle of Least Squares

To emphasize this fact, we rewrite the line in slope-point form.

$$y_i - \bar{y} = \beta(x_i - \bar{x}) + \epsilon_i.$$

Now, the sums of squares criterion becomes a condition on β,

$$\tilde{SS}(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} ((y_i - \bar{y}) - \beta(x_i - \bar{x}))^2.$$

Now, differentiate with respect to β and set this equation to zero for the value $\hat{\beta}$.

$$\frac{d}{d\beta}\tilde{SS}(\hat{\beta}) = -2\sum_{i=1}^{n} ((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}))(x_i - \bar{x}) = 0.$$

Principle of Least Squares

$$0 = \sum_{i=1}^{n} ((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}))(x_i - \bar{x}) = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}).$$

Thus,

$$\begin{aligned}
\hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \\
\hat{\beta} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \\
\hat{\beta} \operatorname{var}(x) &= \operatorname{cov}(x, y) \\
\hat{\beta} &= \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)}
\end{aligned}$$

Linear Regression

Exercise. Show that the maximum likelihood estimator for σ^2 is

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{k=1}^{n}(y_i - \hat{y}_i)^2.$$

where $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. are the predicted values from the regression line.

Frequently, software will report the unbiased estimator. For ordinary least square procedures, this is

$$\hat{\sigma}^2_U = \frac{1}{n-2}\sum_{k=1}^{n}(y_i - \hat{y}_i)^2.$$

For the measurements on the lengths in centimeters of the femur and humerus for the five specimens of *Archeopteryx*, we have the following R output for linear regression.

```
> femur<-c(38,56,59,64,74), humerus<-c(41,63,70,72,84)
```

Linear regression

```
> summary(lm(humerus~femur))
Call:
lm(formula = humerus ~ femur)
Residuals:
      1       2       3       4       5
-0.8226 -0.3668  3.0425 -0.9420 -0.9110
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.65959    4.45896  -0.821 0.471944
femur        1.19690    0.07509  15.941 0.000537 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 1.982 on 3 degrees of freedom
Multiple R-squared: 0.9883, Adjusted R-squared: 0.9844
F-statistic: 254.1 on 1 and 3 DF,  p-value: 0.0005368
```