Topic 15

Maximum Likelihood Estimation

15.1 Introduction

The **principle of maximum likelihood** is relatively straightforward to state. As before, we begin with observations $X = (X_1, \ldots, X_n)$ of random variables chosen according to one of a family of probabilities P_θ. In addition, $\mathbf{f}(\mathbf{x}|\theta)$, $\mathbf{x} = (x_1, \ldots, x_n)$ will be used to denote the density function for the data when θ is the true state of nature.

Then, the principle of maximum likelihood yields a choice of the estimator $\hat{\theta}$ as the value for the parameter that makes the observed data most probable.

Definition 15.1. *The* **likelihood function** *is the density function regarded as a function of* θ.

$$\mathbf{L}(\theta|\mathbf{x}) = \mathbf{f}(\mathbf{x}|\theta),\ \theta \in \Theta. \tag{15.1}$$

The **maximum likelihood estimate (MLE)***,*

$$\hat{\theta}(\mathbf{x}) = \arg\max_{\theta} \mathbf{L}(\theta|\mathbf{x}). \tag{15.2}$$

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

We will learn that especially for large samples, the maximum likelihood estimators have many desirable properties. However, especially for high dimensional data, the likelihood can have many local maxima. Thus, finding the *global maximum* can be a major computational challenge.

This class of estimators has an important **invariance property**. If $\hat{\theta}(\mathbf{x})$ is a maximum likelihood estimate for θ, then $g(\hat{\theta}(\mathbf{x}))$ is a maximum likelihood estimate for $g(\theta)$. For example, if θ is a parameter for the variance and $\hat{\theta}$ is the maximum likelihood estimate for the variance, then $\sqrt{\hat{\theta}}$ is the maximum likelihood estimate for the standard deviation. This flexibility in estimation criterion seen here is not available in the case of unbiased estimators.

For independent observations, the likelihood is the product of density functions. Because the logarithm of a product is the sum of the logarithms, 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. Having the parameter values be the variable of interest is somewhat unusual, so we will next look at several examples of the likelihood function.

15.2 Examples

Example 15.2 (Bernoulli trials)**.** *If the experiment consists of* n *Bernoulli trials with success probability* p*, then*

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

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

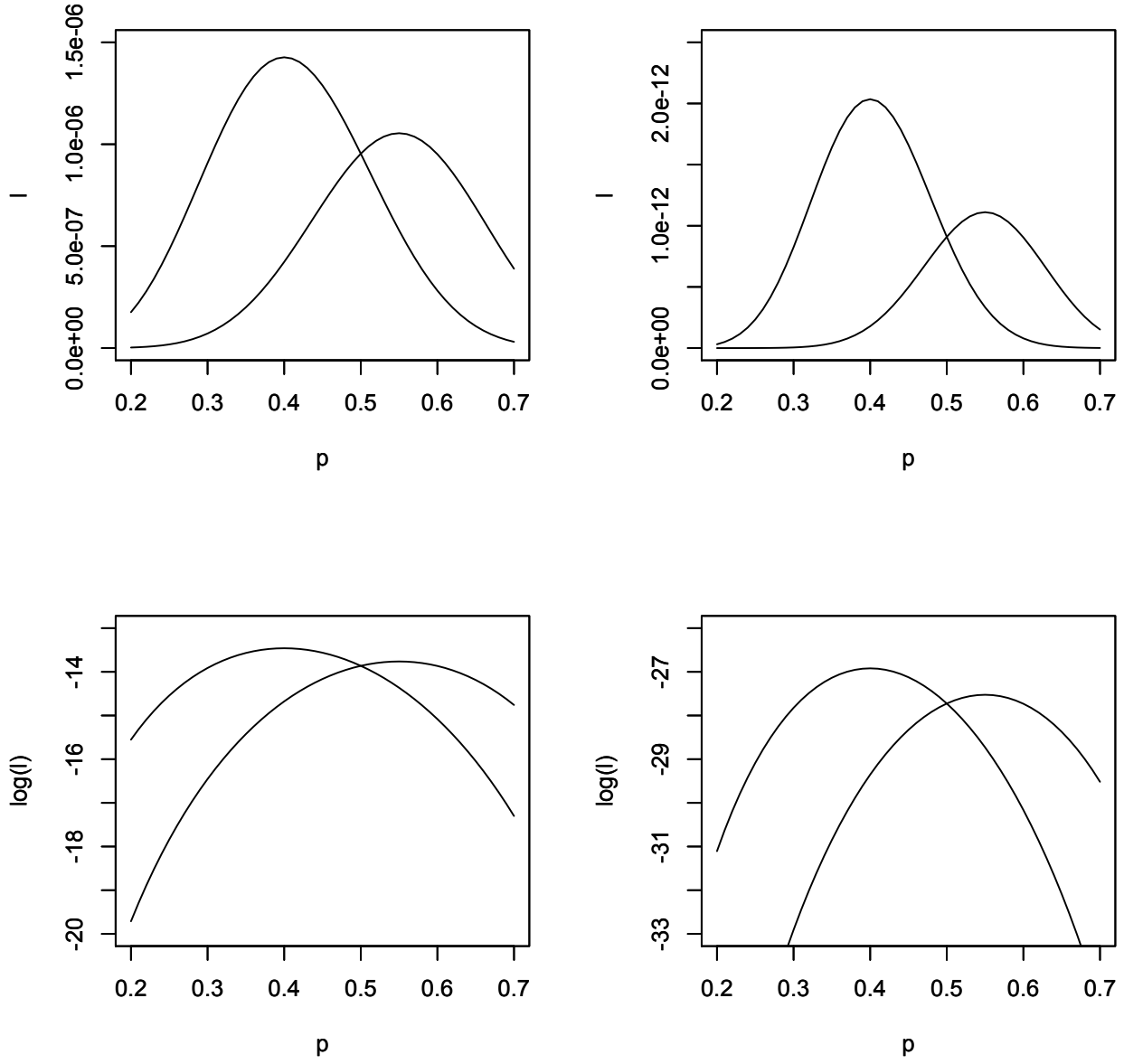

Figure 15.1: Likelihood function (top row) and its logarithm (bottom row) for Bernouli trials. The left column is based on 20 trials having 8 and 11 successes. The right column is based on 40 trials having 16 and 22 successes. Notice that the maximum likelihood is approximately 10^{-6} for 20 trials and 10^{-12} for 40. In addition, note that the peaks are more narrow for 40 trials rather than 20. We shall later be able to associate this property to the variance of the maximum likelihood estimator.

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

This equals zero when $p = \bar{x}$.

Exercise 15.3. *Check that this is a maximum.*

Thus,

$$\hat{p}(\mathbf{x}) = \bar{x}.$$

In this case the maximum likelihood estimator is also unbiased.

Example 15.4 (Normal data)**.** *Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables, we can use the properties of the exponential function to simplify the likelihood function.*

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

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 $\hat{\sigma}^2$

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

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 of the log-likelihood with respect to the parameter σ^2,

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

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

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

Note that the maximum likelihood estimator is a biased estimator.

Example 15.5 (Lincoln-Peterson method of mark and recapture)**.** *Let's recall the variables in mark and recapture:*

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

Here t and k is set by the experimental design; r is an observation that may vary. The total population N is unknown. The likelihood function for N is the hypergeometric distribution.

$$L(N|r) = \frac{\binom{t}{r}\binom{N-t}{k-r}}{\binom{N}{k}}$$

Exercise 15.6. *Show that the maximum likelihood estimator*

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

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

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

Let look at the example of mark and capture from the previous topic. There $N = 2000$, the number of fish in the population, is unknown to us. We tag $t = 200$ fish in the first capture event, and obtain $k = 400$ fish in the second capture.

```
> N<-2000
> t<-200
> fish<-c(rep(1,t),rep(0,N-t))
```

This creates a vector of length N with t ones representing tagged fish and and $N - t$ zeroes representing the untagged fish.

```
> k<-400
> r<-sum(sample(fish,k))
> r
[1] 42
```

This samples k for the recaptured and adds up the ones to obtained, in this simulation, the number $r = 42$ of recaptured fish. For the likelihood function, we look at a range of values for N that is symmetric about 2000. Here, the maximum likelihood estimate $\hat{N} = [200 \cdot 400/42] = 1904$.

```
> N<-c(1800:2200)
> L<-dhyper(r,t,N-t,k)
> plot(N,L,type="l",ylab="L(N|42)",col="green")
```

The likelihood function for this example is shown in Figure 15.2.

Example 15.7 (Linear regression)**.** *Our data are n observations with one explanatory variable and one response variable. The model is that the responses y_i are linearly related to the explanatory variable x_i with an "error" ϵ_i, i.e.,*

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

Here we take the ϵ_i to be independent mean 0 normal random variables. The (unknown) variance is σ^2. Consequently, 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

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

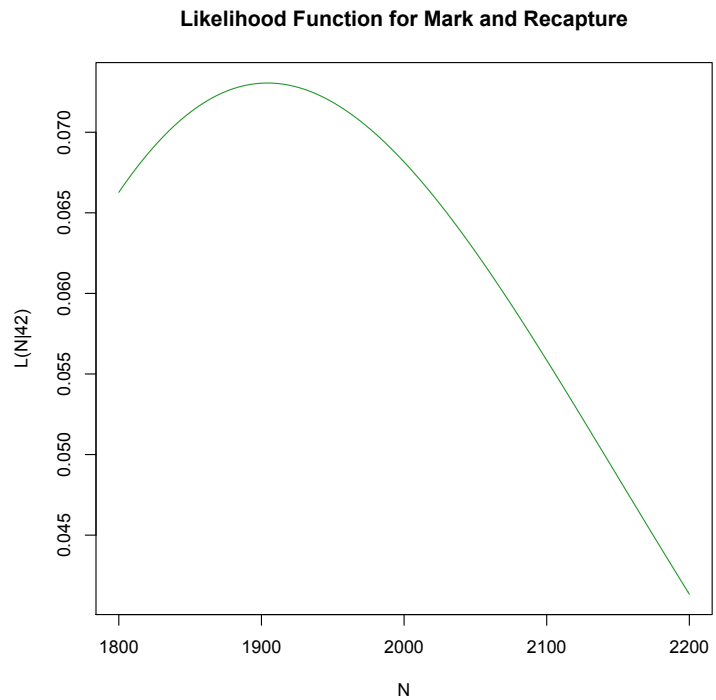

Figure 15.2: Likelihood function $L(N|42)$ for mark and recapture with $t = 200$ tagged fish, $k = 400$ in the second capture with $r = 42$ having tags and thus recapture. Note that the maximum likelihood estimator for the total fish population is $\hat{N} = 1904$.

The logarithm

$$\ln L(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{x}) = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2. \tag{15.3}$$

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

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

Thus, the principle of maximum likelihood is equivalent to the **least squares criterion** *for ordinary linear regression. The maximum likelihood estimators α and β give the regression line*

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i.$$

with

$$\hat{\beta} = \frac{\text{cov}(x, y)}{\text{var}(x)}, \quad \text{and } \hat{\alpha} \text{ determined by solving} \quad \bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}.$$

Exercise 15.8. *Show that the* **maximum likelihood estimator** *for σ^2 is*

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{k=1}^{n}(y_i - \hat{y}_i)^2. \tag{15.4}$$

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)
> 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
```

The residual standard error of 1.982 centimeters is obtained by squaring the 5 residuals, dividing by $3 = 5 - 2$ and taking a square root.

Example 15.9 (weighted least squares)**.** *If we know the relative size of the variances of the ϵ_i, then we have the model*

$$y_i = \alpha + \beta x_i + \gamma(x_i)\epsilon_i$$

where the ϵ_i are, again, independent mean 0 normal random variable with unknown variance σ^2. In this case,

$$\epsilon_i = \frac{1}{\gamma(x_i)}(y_i - \alpha + \beta x_i)$$

are independent normal random variables, mean 0 and (unknown) variance σ^2. the likelihood function

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

where $w(x) = 1/\gamma(x)^2$. In other words, the weights are inversely proportional to the variances. The log-likelihood is

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

Exercise 15.10. *Show that the maximum likelihood estimators $\hat{\alpha}_w$ and $\hat{\beta}_w$ have formulas*

$$\hat{\beta}_w = \frac{\text{cov}_w(x, y)}{\text{var}_w(x)}, \quad \bar{y}_w = \hat{\alpha}_w + \hat{\beta}_w \bar{x}_w$$

where $\bar{x}_w$ and $\bar{y}_w$ are the weighted means

$$\bar{x}_w = \frac{\sum_{i=1}^n w(x_i)x_i}{\sum_{i=1}^n w(x_i)}, \quad \bar{y}_w = \frac{\sum_{i=1}^n w(x_i)y_i}{\sum_{i=1}^n w(x_i)}.$$

The weighted covariance and variance are, respectively,

$$\text{cov}_w(x,y) = \frac{\sum_{i=1}^n w(x_i)(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^n w(x_i)}, \quad \text{var}_w(x) = \frac{\sum_{i=1}^n w(x_i)(x_i - \bar{x}_w)^2}{\sum_{i=1}^n w(x_i)},$$

The maximum likelihood estimator for σ^2 is

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

In the case of weighted least squares, the predicted value for the response variable is

$$\hat{y}_i = \hat{\alpha}_w + \hat{\beta}_w x_i.$$

Exercise 15.11. *Show that $\hat{\alpha}_w$ and $\hat{\beta}_w$ are unbiased estimators of α and β. In particular, ordinary (unweighted) least square estimators are unbiased.*

In computing the optimal values using introductory differential calculus, the maximum can occur at either critical points or at the endpoints. The next example show that the maximum value for the likelihood can occur at the end point of an interval.

Example 15.12 (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

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

Consequently, the joint density is 0 whenever any *of the $x_i > \theta$. Restating this in terms of likelihood, no value of θ is possible that is less than any of the x_i. Conseuently, any value of θ less than any of the x_i has likelihood 0. Symbolically,*

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

Recall the notation $x_{(n)}$ for the top **order statistic** *based on n observations.*

The likelihood is 0 on the interval $(0, x_{(n)})$ and is positive and decreasing on the interval $[x_{(n)}, \infty)$. Thus, to maximize $\mathbf{L}(\theta|\mathbf{x})$, we should take the minimum value of θ on this interval. In other words,

$$\hat{\theta}(\mathbf{x}) = x_{(n)}.$$

Because the estimator is always less than the parameter value it is meant to estimate, the estimator

$$\hat{\theta}(X) = X_{(n)} < \theta,$$

Thus, we suspect it is biased downwards, i. e..

$$E_\theta X_{(n)} < \theta. \tag{15.5}$$

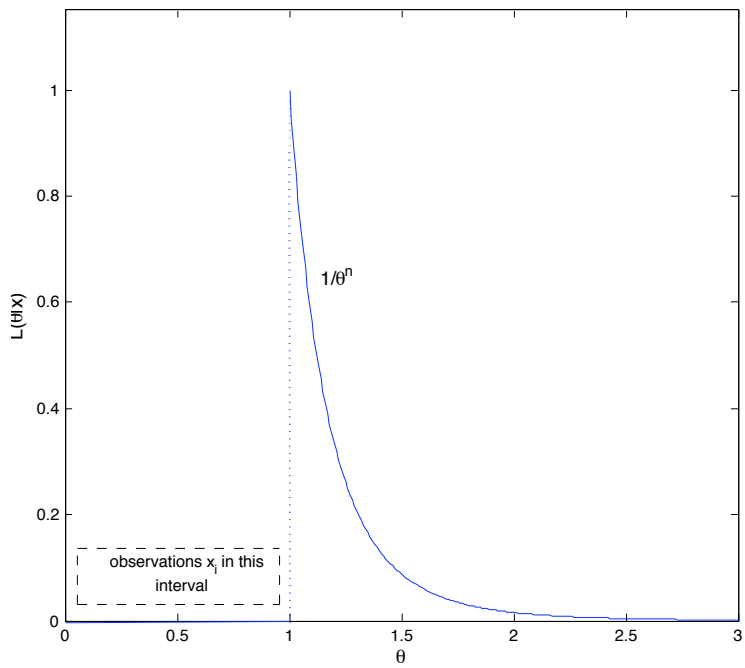

Figure 15.3: Likelihood function for uniform random variables on the interval $[0, \theta]$. The likelihood is 0 up to $\max_{1\le i\le n} x_i$ and $1/\theta^n$ afterwards.

In order to compute the expected value in (15.5), note that $X_{(n)} = \max_{1\le i\le n} X_i \le x$ *if and only if each of the* $X_i \le x$*. Thus, for* $0 \le x \le \theta$*, the distribution function for* $X_{(n)}$ *is*

$$\begin{aligned} F_{X_{(n)}}(x) &= P\{\max_{1\le i\le n} X_i \le x\} = P\{X_1 \le x, X_2 \le x, \ldots, X_n \le x\} \\ &= P\{X_1 \le x\}P\{X_2 \le x\}\cdots P\{X_n < x\} \end{aligned}$$

each of these random variables have the same distribution function

$$F_{X_i}(x) = P\{X_i \le x\} = \begin{cases} 0 & \text{for } x \le 0, \\ \frac{x}{\theta} & \text{for } 0 < x \le \theta, \\ 1 & \text{for } \theta < x. \end{cases}$$

Thus, the distribution function for $X_{(n)}$ *is the product* $F_{X_1}(x)F_{X_2}(x)\cdots F_{X_n}(x)$*, i.e.,*

$$F_{X_{(n)}}(x) = \begin{cases} 0 & \text{for } x \le 0, \\ \left(\frac{x}{\theta}\right)^n & \text{for } 0 < x \le \theta, \\ 1 & \text{for } \theta < x. \end{cases}$$

Take the derivative to find the density,

$$f_{X_{(n)}}(x) = \begin{cases} 0 & \text{for } x \le 0, \\ \frac{nx^{n-1}}{\theta^n} & \text{for } 0 < x \le \theta, \\ 0 & \text{for } \theta < x. \end{cases}$$

The mean

$$\begin{aligned} E_\theta X_{(n)} &= \int_0^\theta x f_{X_{(n)}}(x)\, dx = \int_0^\theta x \frac{n x^{n-1}}{\theta^n}\, dx \\ &= \frac{n}{\theta^n} \int_0^\theta x^n\, dx = \frac{n}{(n+1)\theta^n} x^{n+1}\Big|_0^\theta = \frac{n}{n+1}\theta. \end{aligned}$$

This confirms the bias of the estimator $X_{(n)}$ and gives us a strategy to find an unbiased estimator. Note that the choice

$$d(X) = \frac{n+1}{n} X_{(n)}$$

yields an unbiased estimator of θ.

15.3 Summary of Estimators

Look to the text above for the definition of variables.

parameter	estimate	
	Bernoulli trials	
p	$\hat{p} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$	unbiased
	mark recapture	
N	$\hat{N} = \left[\frac{kt}{r}\right]$	biased upward
	normal observations	
μ	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$	unbiased
σ^2	$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$	biased downward
	$\hat{\sigma}^2_u = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$	unbiased
σ	$\hat{\sigma}_{mle} = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$	biased downward
	linear regression	
β	$\hat{\beta} = \frac{\text{cov}(x,y)}{\text{var}(x)}$	unbiased
α	$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$	unbiased
σ^2	$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x))^2$	biased downward
	$\hat{\sigma}^2_u = \frac{1}{n-2}\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x))^2$	unbiased
σ	$\hat{\sigma}_{mle} = \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x))^2}$	biased downward
	$\hat{\sigma}_u = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x))^2}$	biased downward
	uniform $[0, \theta]$	
θ	$\hat{\theta} = \max_i x_i$	biased downward
	$\hat{\theta} = \frac{n+1}{n}\max_i x_i$	unbiased

15.4 Asymptotic Properties

Much of the attraction of maximum likelihood estimators is based on their properties for large sample sizes. We summarizes some the important properties below, saving a more technical discussion of these properties for later.

1. **Consistency**. If θ_0 is the state of nature and $\hat{\theta}_n(X)$ is the maximum likelihood estimator based on n observations from a simple random sample, then
$$\hat{\theta}_n(X) \to \theta_0 \quad \text{as } n \to \infty.$$
In words, as the number of observations increase, the distribution of the maximum likelihood estimator becomes more and more concentrated about the true state of nature.

2. **Asymptotic normality and efficiency.** Under some assumptions that allows, among several analytical properties, the use of a central limit theorem holds. Here we have
$$\sqrt{n}(\hat{\theta}_n(X) - \theta_0)$$
converges in distribution as $n \to \infty$ to a normal random variable with mean 0 and variance $1/I(\theta_0)$, the Fisher information for one observation. Thus,
$$\text{Var}_{\theta_0}(\hat{\theta}_n(X)) \approx \frac{1}{nI(\theta_0)},$$
the lowest variance possible under the Crámer-Rao lower bound. This property is called **asymptotic efficiency**. We can write this in terms of the z-score. Let
$$Z_n = \frac{\hat{\theta}(X) - \theta_0}{1/\sqrt{nI(\theta_0)}}.$$
Then, as with the central limit theorem, Z_n converges in distribution to a standard normal random variable.

3. **Properties of the log likelihood surface**. For large sample sizes, the variance of a maximum likelihood estimator of a single parameter is approximately the reciprocal of the the Fisher information
$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \ln L(\theta|X)\right].$$
The Fisher information can be approximated by the **observed information** based on the data $\mathbf{x}$,
$$J(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2} \ln L(\hat{\theta}(\mathbf{x})|\mathbf{x}),$$
giving the negative of the curvature of the log-likelihood surface at the maximum likelihood estimate $\hat{\theta}(\mathbf{x})$. If the curvature is small near the maximum likelihood estimator, then the likelihood surface is nearty flat and the variance is large. If the curvature is large, the likelihood decreases quickly at the maximum and thus the variance is small.

We now look at these properties in some detail by revisiting the example of the distribution of fitness effects. For this example, we have two parameters - α and β for the gamma distribution and so, we will want to extend the properties above to circumstances in which we are looking to estimate more than one parameter.

15.5 Multidimensional Estimation

For a multidimensional parameter space $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, the Fisher information $I(\theta)$ is now a matrix . As with one-dimensional case, the ij-th entry has two alternative expressions, namely,
$$I(\theta)_{ij} = E_\theta\left[\frac{\partial}{\partial\theta_i} \ln L(\theta|X) \frac{\partial}{\partial\theta_j} \ln L(\theta|X)\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln L(\theta|X)\right].$$

Rather than taking reciprocals to obtain an estimate of the variance, we find the matrix inverse $I(\theta)^{-1}$. This inverse will provide estimates of both variances and covariances. To be precise, for n observations, let $\hat{\theta}_{i,n}(X)$ be the maximum likelihood estimator of the i-th parameter. Then

$$\mathrm{Var}_\theta(\hat{\theta}_{i,n}(X)) \approx \frac{1}{n} I(\theta)^{-1}_{ii} \qquad \mathrm{Cov}_\theta(\hat{\theta}_{i,n}(X), \hat{\theta}_{j,n}(X)) \approx \frac{1}{n} I(\theta)^{-1}_{ij}.$$

When the i-th parameter is θ_i, the asymptotic normality and efficiency can be expressed by noting that the z-score

$$Z_{i,n} = \frac{\hat{\theta}_i(X) - \theta_i}{I(\theta)^{-1}_{ii}/\sqrt{n}}.$$

is approximately a standard normal. As we saw in one dimension, we can replace the information matrix with the observed information matrix,

$$J(\hat{\theta})_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln L(\hat{\theta}(\mathbf{x})|\mathbf{x}).$$

Example 15.13. *To obtain the maximum likelihood estimate for the gamma family of random variables, write the likelihood*

$$\mathbf{L}(\alpha, \beta|\mathbf{x}) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)} x_1^{\alpha-1} e^{-\beta x_1}\right) \cdots \left(\frac{\beta^\alpha}{\Gamma(\alpha)} x_n^{\alpha-1} e^{-\beta x_n}\right) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n (x_1 x_2 \cdots x_n)^{\alpha-1} e^{-\beta(x_1+x_2+\cdots+x_n)}.$$

and its logarithm

$$\ln \mathbf{L}(\alpha, \beta|\mathbf{x}) = n(\alpha \ln \beta - \ln \Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^n \ln x_i - \beta \sum_{i=1}^n x_i.$$

To determine the parameters that maximize the likelihood, we solve the equations

$$\frac{\partial}{\partial\alpha} \ln \mathbf{L}(\hat{\alpha}, \hat{\beta}|\mathbf{x}) = n(\ln \hat{\beta} - \frac{d}{d\alpha} \ln \Gamma(\hat{\alpha})) + \sum_{i=1}^n \ln x_i = 0$$

and

$$\frac{\partial}{\partial\beta} \ln \mathbf{L}(\hat{\alpha}, \hat{\beta}|\mathbf{x}) = n\frac{\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n x_i = 0, \quad \textit{or} \quad \bar{x} = \frac{\hat{\alpha}}{\hat{\beta}}.$$

Recall that the mean μ of a gamma distribution is α/β. Thus. by the invariance property of maximum likelihood estimators

$$\hat{\mu} = \frac{\hat{\alpha}}{\hat{\beta}} = \bar{x},$$

and the sample mean is the maximum likelihood estimate for the distributional mean.

Substituting $\hat{\beta} = \hat{\alpha}/\bar{x}$ into the first equation results the following relationship for $\hat{\alpha}$

$$n(\ln \hat{\alpha} - \ln \bar{x} - \frac{d}{d\alpha} \ln \Gamma(\hat{\alpha})) + \sum_{i=1}^n \ln x_i = 0$$

which can be solved numerically. The derivative of the logarithm of the gamma function

$$\psi(\alpha) = \frac{d}{d\alpha} \ln \Gamma(\alpha)$$

is know as the **digamma function** *and is called in* R *with* `digamma`.

For the example for the distribution of fitness effects $\alpha = 0.23$ and $\beta = 5.35$ with $n = 100$, a simulated data set yields $\hat{\alpha} = 0.2376$ and $\hat{\beta} = 5.690$ for maximum likelihood estimator. (See Figure 15.4.)

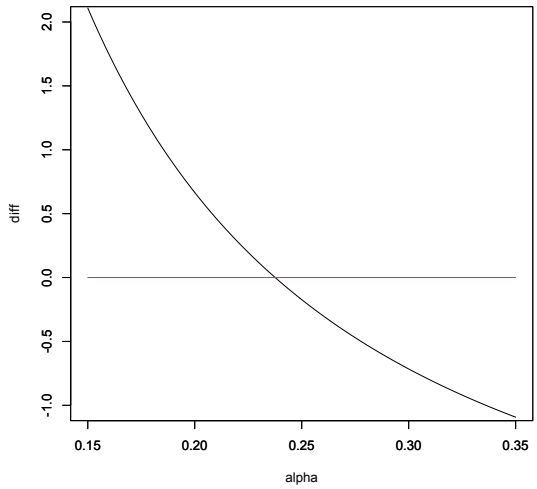

Figure 15.4: The graph of $n(\ln\hat{\alpha} - \ln\bar{x} - \frac{d}{d\alpha}\ln\Gamma(\hat{\alpha})) + \sum_{i=1}^{n}\ln x_i$ crosses the horizontal axis at $\hat{\alpha} = 0.2376$. The fact that the graph of the derivative is decreasing states that the score function moves from increasing to decreasing with α and confirming that $\hat{\alpha}$ is a maximum.

To determine the variance of these estimators, we first compute the Fisher information matrix. Taking the appropriate derivatives, we find that each of the second order derivatives are constant and thus the expected values used to determine the entries for Fisher information matrix are the negative of these constants.

$$I(\alpha,\beta)_{11} = -\frac{\partial^2}{\partial\alpha^2}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = n\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha), \quad I(\alpha,\beta)_{22} = -\frac{\partial^2}{\partial\beta^2}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = n\frac{\alpha}{\beta^2},$$

$$I(\alpha,\beta)_{12} = -\frac{\partial^2}{\partial\alpha\partial\beta}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}) = -n\frac{1}{\beta}.$$

This give a Fisher information matrix

$$I(\alpha,\beta) = n\begin{pmatrix} \frac{d^2}{d\alpha^2}\ln\Gamma(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}.$$

The second derivative of the logarithm of the gamma function

$$\psi_1(\alpha) = \frac{d^2}{d\alpha^2}\ln\Gamma(\alpha)$$

is known as the **trigamma function** *and is called in* R *with* `trigamma`.

The inverse

$$I(\alpha,\beta)^{-1} = \frac{1}{n\alpha(\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha) - 1)}\begin{pmatrix} \alpha & \beta \\ \beta & \beta^2\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha) \end{pmatrix}.$$

For the example for the distribution of fitness effects $\alpha = 0.23$ *and* $\beta = 5.35$ *and* $n = 100$*, and*

$$I(0.23, 5.35)^{-1} = \frac{1}{100(0.23)(19.12804)}\begin{pmatrix} 0.23 & 5.35 \\ 5.35 & 5.35^2(20.12804) \end{pmatrix} = \begin{pmatrix} 0.0001202 & 0.01216 \\ 0.01216 & 1.3095 \end{pmatrix}.$$

$$\text{Var}_{(0.23,5.35)}(\hat{\alpha}) \approx 0.0001202, \quad \text{Var}_{(0.23,5.35)}(\hat{\beta}) \approx 1.3095.$$

$$\sigma_{(0.23,5.35)}(\hat{\alpha}) \approx 0.0110, \quad \sigma_{(0.23,5.35)}(\hat{\beta}) \approx 1.1443.$$

Compare this to the empirical values of 0.0662 *and* 2.046 *for the method of moments. This gives the following table of standard deviations for* $n = 100$ *observation*

method	$\hat{\alpha}$	$\hat{\beta}$
maximum likelihood	0.0110	1.1443
method of moments	0.0662	2.046
ratio	0.166	0.559

Thus, the standard deviation for the maximum likelihood estimator is respectively 17% and 56% that of method of moments estimator. We will look at the impact as we move on to our next topic - interval estimation and the confidence intervals.

Exercise 15.14. *If the data are a simple random sample of 100 observations of a* $\Gamma(0.23, 5.35)$ *random variable. Use the approximate normality of maximum likelihood estimators to estimate*

$$P\{\hat{\alpha} \geq 0.2376\} \qquad P\{\hat{\beta} \geq 5.690\}.$$

15.6 Choice of Estimators

With all of the desirable properties of the maximum likelihood estimator, the question arises as to why would one choose a method of moments estimator?

One answer is that the use maximum likelihood techniques relies on knowing the density function explicitly. Moreover, the form of the density must be amenable to the analysis necessary to maximize the likelihood and find the Fisher information.

However, much less about the experiment is need in order to compute moments. Thus far, we have computed moments using the density

$$E_\theta X^m = \int_{-\infty}^{\infty} x^m f_X(x|\theta)\, dx.$$

However, consider the case of determining parameters in the distribution in the number of proteins in a tissue. If the tissue has several cell types, then we would need

- the distribution of cell types, and
- a density function for the number of proteins in each cell type.

These two pieces of information can be used to calculate the mean and variance for the number of cells with some ease. However, giving an explicit expression for the density and hence the likelihood function is more difficult to obtain. This leads to quite intricate computations to carry out the desired analysis of the likelihood function.

15.7 Technical Aspects

We can use concepts previously introduced to obtain the properties for the maximum likelihood estimator. For example, θ_0 is more likely that a another parameter value θ

$$\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta|X) \quad \text{if and only if} \quad \frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i|\theta_0)}{f(X_i|\theta)} > 0.$$

By the strong law of large numbers, this sum converges to

$$E_{\theta_0}\left[\ln \frac{f(X_1|\theta_0)}{f(X_1|\theta)}\right].$$

which is greater than 0. thus, for a large number of observations and a given value of θ, then with a probability nearly one, $\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta|X)$ and so the maximum likelihood estimator has a high probability of being very near θ_0. This is a statement of the **consistency** of the estimator.

For the asymptotic normality and efficiency, we write the linear approximation of the score function

$$\frac{d}{d\theta}\ln L(\theta|X) \approx \frac{d}{d\theta}\ln L(\theta_0|X) + (\theta-\theta_0)\frac{d^2}{d\theta^2}\ln L(\theta_0|X).$$

Now substitute $\theta = \hat{\theta}$ and note that $\frac{d}{d\theta}\ln L(\hat{\theta}|X) = 0$. Then

$$\sqrt{n}(\hat{\theta}_n(X)-\theta_0) \approx -\sqrt{n}\frac{\frac{d}{d\theta}\ln L(\theta_0|X)}{\frac{d^2}{d\theta^2}\ln L(\theta_0|X)} = \frac{\frac{1}{\sqrt{n}}\frac{d}{d\theta}\ln L(\theta_0|X)}{-\frac{1}{n}\frac{d^2}{d\theta^2}\ln L(\theta_0|X)}$$

Now assume that θ_0 is the true state of nature. Then, the random variables $d\ln f(X_i|\theta_0)/d\theta$ are independent with mean 0 and variance $I(\theta_0)$. Thus, the distribution of numerator

$$\frac{1}{\sqrt{n}}\frac{d}{d\theta}\ln L(\theta_0|X) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\theta}\ln f(X_i|\theta_0)$$

converges, by the central limit theorem, to a normal random variable with mean 0 and variance $I(\theta_0)$. For the denominator, $-d^2\ln f(X_i|\theta_0)/d\theta^2$ are independent with mean $I(\theta_0)$. Thus,

$$-\frac{1}{n}\frac{d^2}{d\theta^2}\ln L(\theta_0|X) = -\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\ln f(X_i|\theta_0)$$

converges, by the law of large numbers, to $I(\theta_0)$. Thus, the distribution of the ratio, $\sqrt{n}(\hat{\theta}_n(X)-\theta_0)$, converges to a normal random variable with variance $I(\theta_0)/I(\theta_0)^2 = 1/I(\theta_0)$.

15.8 Answers to Selected Exercises

15.3. We have found that the score function

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

Thus

$$\frac{\partial}{\partial p}\ln \mathbf{L}(p|\mathbf{x}) > 0 \quad \text{if } p < \bar{x}, \quad \text{and} \quad \frac{\partial}{\partial p}\ln \mathbf{L}(p|\mathbf{x}) < 0 \quad \text{if } p > \bar{x}$$

In words, $\ln \mathbf{L}(p|\mathbf{x})$ is increasing for $p < \bar{x}$ and decreasing for $p > \bar{x}$. Thus, $\hat{p}(\mathbf{x}) = \bar{x}$ is a maximum.

15.7. The log-likelihood function

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

leads to the ordinary least squares equations for the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$. Take the partial derivative with respect to σ^2,

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

This partial derivative is 0 at the maximum likelihood estimates $\hat{\sigma}^2$, $\hat{\alpha}$ and $\hat{\beta}$.

$$0 = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(y_i-(\hat{\alpha}+\hat{\beta}x_i))^2$$

or

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

We would like to maximize the likelihood given the number of recaptured individuals r. Because the domain for N is the nonnegative integers, we cannot use calculus. However, we can look at the ratio of the likelihood values for successive value of the total population.

$$\frac{L(N|r)}{L(N-1|r)}$$

N is more likely that $N-1$ precisely when this ratio is larger than one. The computation below will show that this ratio is greater than 1 for small values of N and less than one for large values. Thus, there is a place in the middle which has the maximum. We expand the binomial coefficients in the expression for $L(N|r)$ and simplify.

$$\begin{aligned}\frac{L(N|r)}{L(N-1|r)} &= \frac{\binom{t}{r}\binom{N-t}{k-r}/\binom{N}{k}}{\binom{t}{r}\binom{N-t-1}{k-r}/\binom{N-1}{k}} = \frac{\binom{N-t}{k-r}\binom{N-1}{k}}{\binom{N-t-1}{k-r}\binom{N}{k}} = \frac{\frac{(N-t)!}{(k-r)!(N-t-k+r)!}\frac{(N-1)!}{k!(N-k-1)!}}{\frac{(N-t-1)!}{(k-r)!(N-t-k+r-1)!}\frac{N!}{k!(N-k)!}} \\ &= \frac{(N-t)!(N-1)!(N-t-k+r-1)!(N-k)!}{(N-t-1)!N!(N-t-k+r)!(N-k-1)!} = \frac{(N-t)(N-k)}{N(N-t-k+r)}.\end{aligned}$$

Thus, the ratio

$$\frac{L(N|r)}{L(N-1|r)} = \frac{(N-t)(N-k)}{N(N-t-k+r)}$$

exceeds 1if and only if

$$\begin{aligned}(N-t)(N-k) &> N(N-t-k+r) \\ N^2 - tN - kN + tk &> N^2 - tN - kN + rN \\ tk &> rN \\ \frac{tk}{r} &> N\end{aligned}$$

Writing $[x]$ for the integer part of x, we see that $L(N|r) > L(N-1|r)$ for $N < [tk/r]$ and $L(N|r) \leq L(N-1|r)$ for $N \geq [tk/r]$. This give the maximum likelihood estimator

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

15.8 .Take the derivative with respect to σ^2 in (15.3)

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

Now set this equal to zero, substitute $\hat{\alpha}$ for α, $\hat{\beta}$ for β and solve for σ^2 to obtain (15.4).

15.9. The maximum likelihood principle leads to a minimization problem for

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

Following the steps to derive the equations for ordinary least squares, take partial derivatives to find that

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

Set these two equations equal to 0 and call the solutions $\hat{\alpha}_w$ and $\hat{\beta}_w$.

$$0 = \sum_{i=1}^{n} w(x_i)x_i(y_i - \hat{\alpha}_w - \hat{\beta}_w x_i) = \sum_{i=1}^{n} w(x_i)x_i y_i - \hat{\alpha}_w \sum_{i=1}^{n} w(x_i)x_i - \hat{\beta}_w \sum_{i=1}^{n} w(x_i)x_i^2 \tag{15.6}$$

$$0 = \sum_{i=1}^{n} w(x_i)(y_i - \hat{\alpha}_w - \hat{\beta}_w x_i) = \sum_{i=1}^{n} w(x_i)y_i - \hat{\alpha}_w \sum_{i=1}^{n} w(x_i) - \hat{\beta}_w \sum_{i=1}^{n} w(x_i)x_i \tag{15.7}$$

Multiply these equations by the appropriate factors to obtain

$$\begin{aligned} 0 = &\left(\sum_{i=1}^{n} w(x_i)\right)\left(\sum_{i=1}^{n} w(x_i)x_i y_i\right) - \hat{\alpha}_w \left(\sum_{i=1}^{n} w(x_i)\right)\left(\sum_{i=1}^{n} w(x_i)x_i\right) \\ &-\hat{\beta}_w \left(\sum_{i=1}^{n} w(x_i)\right)\left(\sum_{i=1}^{n} w(x_i)x_i^2\right) \end{aligned} \tag{15.8}$$

$$0 = \left(\sum_{i=1}^{n} w(x_i)x_i\right)\left(\sum_{i=1}^{n} w(x_i)y_i\right) - \hat{\alpha}_w \left(\sum_{i=1}^{n} w(x_i)\right)\left(\sum_{i=1}^{n} w(x_i)x_i\right) - \hat{\beta}_w \left(\sum_{i=1}^{n} w(x_i)x_i\right)^2 \tag{15.9}$$

Now subtract the equation (15.9) from equation (15.8) and solve for $\hat{\beta}$.

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

Next, divide equation (15.9) by $\sum_{i=1}^{n} w(x_i)$ to obtain

$$\bar{y}_w = \hat{\alpha}_w + \hat{\beta}_w \bar{x}_w. \tag{15.10}$$

15.10. Because the ϵ_i have mean zero,

$$E_{(\alpha,\beta)} y_i = E_{(\alpha,\beta)}[\alpha + \beta x_i + \gamma(x_i)\epsilon_i] = \alpha + \beta x_i + \gamma(x_i) E_{(\alpha,\beta)}[\epsilon_i] = \alpha + \beta x_i.$$

Next, use the linearity property of expectation to find the mean of $\bar{y}_w$.

$$E_{(\alpha,\beta)} \bar{y}_w = \frac{\sum_{i=1}^{n} w(x_i) E_{(\alpha,\beta)} y_i}{\sum_{i=1}^{n} w(x_i)} = \frac{\sum_{i=1}^{n} w(x_i)(\alpha + \beta x_i)}{\sum_{i=1}^{n} w(x_i)} = \alpha + \beta \bar{x}_w. \tag{15.11}$$

Taken together, we have that $E_{(\alpha,\beta)}[y_i - \bar{y}_w] = (\alpha + \beta x_i.) - (\alpha + \beta x_i) = \beta(x_i - \bar{x}_w)$. To show that $\hat{\beta}_w$ is an unbiased estimator, we see that

$$\begin{aligned} E_{(\alpha,\beta)} \hat{\beta}_w &= E_{(\alpha,\beta)}\left[\frac{\text{cov}_w(x,y)}{\text{var}_w(x)}\right] = \frac{E_{(\alpha,\beta)}[\text{cov}_w(x,y)]}{\text{var}_w(x)} = \frac{1}{\text{var}_w(x)} E_{(\alpha,\beta)}\left[\frac{\sum_{i=1}^{n} w(x_i)(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^{n} w(x_i)}\right] \\ &= \frac{1}{\text{var}_w(x)} \frac{\sum_{i=1}^{n} w(x_i)(x_i - \bar{x}_w) E_{(\alpha,\beta)}[y_i - \bar{y}_w]}{\sum_{i=1}^{n} w(x_i)} = \frac{\beta}{\text{var}_w(x)} \frac{\sum_{i=1}^{n} w(x_i)(x_i - \bar{x}_w)(x_i - \bar{x}_w)}{\sum_{i=1}^{n} w(x_i)} = \beta. \end{aligned}$$

To show that $\hat{\alpha}_w$ is an unbiased estimator, recall that $\bar{y}_w = \hat{\alpha}_w + \hat{\beta}_w \bar{x}_w$. Thus

$$E_{(\alpha,\beta)} \hat{\alpha}_w = E_{(\alpha,\beta)}[\bar{y}_w - \hat{\beta}_w \bar{x}_w] = E_{(\alpha,\beta)} \bar{y}_w - E_{(\alpha,\beta)}[\hat{\beta}_w] \bar{x}_w = \alpha + \beta \bar{x}_w - \beta \bar{x}_w = \alpha,$$

using (15.11) and the fact that $\hat{\beta}_w$ is an unbiased estimator of β

15.14. For $\hat{\alpha}$, we have the z-score

$$z_{\hat{\alpha}} = \frac{\hat{\alpha} - 0.23}{\sqrt{0.0001202}} \geq \frac{0.2376 - 0.23}{\sqrt{0.0001202}} = 0.6841.$$

Thus, using the normal approximation,

$$P\{\hat{\alpha} \geq 0.2367\} = P\{z_{\hat{\alpha}} \geq 0.6841\} = 0.2470.$$

For $\hat{\beta}$, we have the z-score

$$z_{\hat{\beta}} = \frac{\hat{\beta} - 5.35}{\sqrt{1.3095}} \geq \frac{5.690 - 5.35}{\sqrt{1.3095}} = 0.2971.$$

Here, the normal approximation gives

$$P\{\hat{\beta} \geq 5.690\} = P\{z_{\hat{\beta}} \geq 0.2971\} = 0.3832.$$

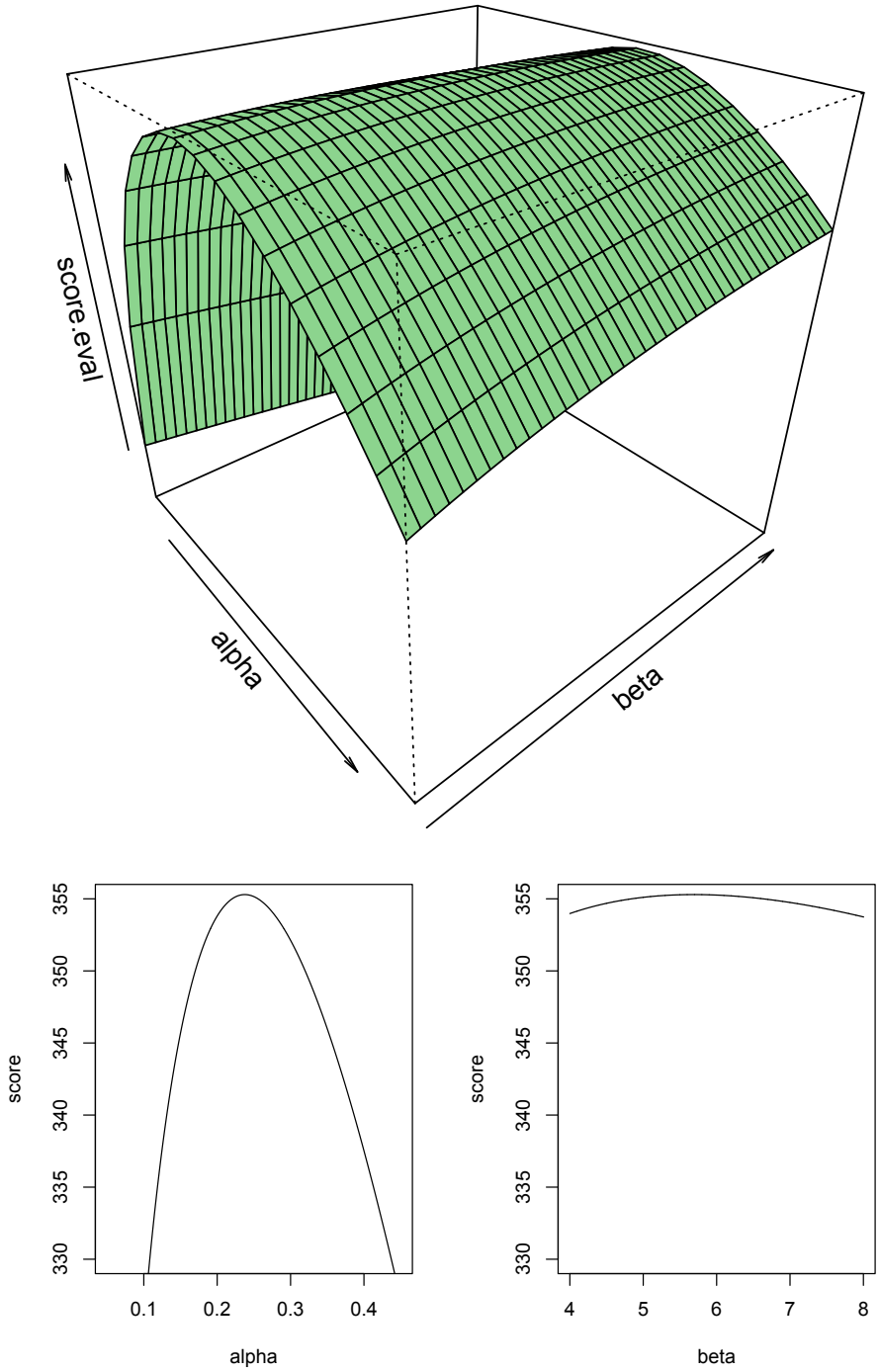

Figure 15.5: (**top**) The log-likelihood near the maximum likelihood estimators. The domain is $0.1 \leq \alpha \leq 0.4$ and $4 \leq \beta \leq 8$. (**bottom**) Graphs of vertical slices through the log-likelihood function surface. (left) $\hat{\alpha} = 0.2376$ and $4 \leq \beta \leq 8$ varies. (right) $\hat{\beta} = 5.690$ and $0.1 \leq \alpha \leq 0.4$. The variance of the estimator is approximately the negative reciprocal of the second derivative of the log-likelihood function at the maximum likelihood estimators (known as the observed information). Note that the log-likelihood function is nearly flat as β varies. This leads to the interpretation that a range of values for β are nearly equally likely and that the variance for the estimator for $\hat{\beta}$ will be high. On the other hand, the log-likelihood function has a much greater curvature for the α parameter and the estimator $\hat{\alpha}$ will have a much smaller variance than $\hat{\beta}$