Appendix A

Review of Likelihood Theory

This is a brief summary of some of the key results we need from likelihood theory.

A.1 Maximum Likelihood Estimation

Let $Y_1, \ldots, Y_n$ be n independent random variables (r.v.'s) with probability density functions (pdf) $f_i(y_i; \boldsymbol{\theta})$ depending on a vector-valued parameter $\boldsymbol{\theta}$.

A.1.1 The Log-likelihood Function

The joint density of n independent observations $\mathbf{y} = (y_1, \ldots, y_n)'$ is

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\theta}) = \mathrm{L}(\boldsymbol{\theta}; \mathbf{y}). \tag{A.1}$$

This expression, viewed as a function of the unknown parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$, is called the *likelihood* function.

Often we work with the natural logarithm of the likelihood function, the so-called *log-likelihood* function:

$$\log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} \log f_i(y_i; \boldsymbol{\theta}). \tag{A.2}$$

A sensible way to estimate the parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$ is to maximize the likelihood (or equivalently the log-likelihood) function, choosing the parameter value that makes the data actually observed as likely as possible. Formally, we define the *maximum-likelihood estimator* (mle) as the value $\hat{\boldsymbol{\theta}}$ such that

$$\log \mathrm{L}(\hat{\boldsymbol{\theta}}; \mathbf{y}) \geq \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y}) \text{ for all } \boldsymbol{\theta}. \tag{A.3}$$

Example: The Log-Likelihood for the Geometric Distribution. Consider a series of independent Bernoulli trials with common probability of success π. The distribution of the number of *failures* Y_i before the first success has pdf

$$\Pr(Y_i = y_i) = (1-\pi)^{y_i}\pi. \tag{A.4}$$

for $y_i = 0, 1, \ldots$. Direct calculation shows that $\mathrm{E}(Y_i) = (1-\pi)/\pi$.

The log-likelihood function based on n observations $\mathbf{y}$ can be written as

$$\begin{aligned} \log \mathrm{L}(\pi; \mathbf{y}) &= \sum_{i=1}^{n} \{y_i \log(1-\pi) + \log \pi\} & \text{(A.5)} \\ &= n(\bar{y} \log(1-\pi) + \log \pi), & \text{(A.6)} \end{aligned}$$

where $\bar{y} = \sum y_i/n$ is the sample mean. The fact that the log-likelihood depends on the observations only through the sample mean shows that $\bar{y}$ is a *sufficient* statistic for the unknown probability π.

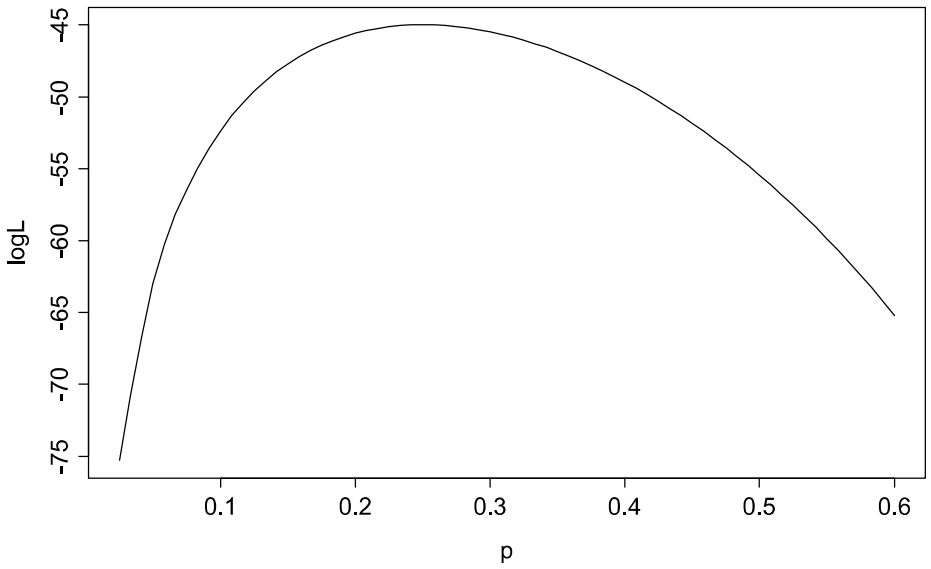

FIGURE A.1: The Geometric Log-Likelihood for $n = 20$ and $\bar{y} = 3$

Figure A.1 shows the log-likelihood function for a sample of $n = 20$ observations from a geometric distribution when the observed sample mean is $\bar{y} = 3$.□

A.1.2 The Score Vector

The first derivative of the log-likelihood function is called Fisher's *score function*, and is denoted by

$$\mathbf{u}(\boldsymbol{\theta}) = \frac{\partial \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}. \tag{A.7}$$

Note that the score is a vector of first partial derivatives, one for each element of $\boldsymbol{\theta}$.

If the log-likelihood is concave, one can find the maximum likelihood estimator by setting the score to zero, i.e. by solving the system of equations:

$$\mathbf{u}(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{A.8}$$

Example: The Score Function for the Geometric Distribution. The score function for n observations from a geometric distribution is

$$u(\pi) = \frac{\mathrm{d} \log \mathrm{L}}{\mathrm{d}\pi} = n(\frac{1}{\pi} - \frac{\bar{y}}{1-\pi}). \tag{A.9}$$

Setting this equation to zero and solving for π leads to the maximum likelihood estimator

$$\hat{\pi} = \frac{1}{1+\bar{y}}. \tag{A.10}$$

Note that the m.l.e. of the probability of success is the reciprocal of the number of trials. This result is intuitively reasonable: the longer it takes to get a success, the lower our estimate of the probability of success would be.

Suppose now that in a sample of $n = 20$ observations we have obtained a sample mean of $\bar{y} = 3$. The m.l.e. of the probability of success would be $\hat{\pi} = 1/(1+3) = 0.25$, and it should be clear from Figure A.1 that this value maximizes the log-likelihood.

Exercise: MLE Poisson

A.1.3 The Information Matrix

The score is a random vector with some interesting statistical properties. In particular, the score evaluated at the true parameter value θ has mean zero

$$\mathrm{E}[\mathbf{u}(\boldsymbol{\theta})] = \mathbf{0}$$

and variance-covariance matrix given by the *information matrix*:

$$\mathrm{var}[\mathbf{u}(\boldsymbol{\theta})] = \mathrm{E}[\mathbf{u}(\boldsymbol{\theta})\mathbf{u}'(\boldsymbol{\theta})] = \mathbf{I}(\boldsymbol{\theta}). \tag{A.11}$$

Under mild regularity conditions, the information matrix can also be obtained as minus the expected value of the second derivatives of the log-likelihood:

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}[\frac{\partial^2 \log \mathrm{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}]. \tag{A.12}$$

The matrix of negative observed second derivatives is sometimes called the *observed* information matrix.

Note that the second derivative indicates the extent to which the log-likelihood function is peaked rather than flat. This makes the interpretation in terms of information intuitively reasonable.

Example: Information for the Geometric Distribution. Differentiating the score we find the observed information to be

$$-\frac{\mathrm{d}^2 \log \mathrm{L}}{\mathrm{d}\pi^2} = -\frac{\mathrm{d}u}{\mathrm{d}\pi} = n(\frac{1}{\pi^2} + \frac{\bar{y}}{(1-\pi)^2}). \tag{A.13}$$

To find the expected information we use the fact that the expected value of the sample mean $\bar{y}$ is the population mean $(1-\pi)/\pi$, to obtain (after some simplification)

$$\mathrm{I}(\pi) = \frac{n}{\pi^2(1-\pi)}. \tag{A.14}$$

Note that the information increases with the sample size n and varies with π, increasing as π moves away from $\frac{2}{3}$ towards 0 or 1.

In a sample of size $n = 20$, if the true value of the parameter was $\pi = 0.15$ the expected information would be $I(0.15) = 1045.8$. If the sample mean turned out to be $\bar{y} = 3$, the observed information would be 971.9. Of course, we don't know the true value of π. Substituting the mle $\hat{\pi} = 0.25$, we estimate the expected and observed information as 426.7. □

A.1.4 Newton-Raphson and Fisher Scoring

Calculation of the mle often requires iterative procedures. Consider expanding the score function evaluated at the mle $\hat{\boldsymbol{\theta}}$ around a trial value $\boldsymbol{\theta}_0$ using a first order Taylor series, so that

$$\mathbf{u}(\hat{\boldsymbol{\theta}}) \approx \mathbf{u}(\boldsymbol{\theta}_0) + \frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{A.15}$$

Let $\mathbf{H}$ denote the Hessian or matrix of second derivatives of the log-likelihood function

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \log \mathrm{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{A.16}$$

Setting the left-hand-size of Equation A.15 to zero and solving for $\hat{\boldsymbol{\theta}}$ gives the first-order approximation

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \mathbf{H}^{-1}(\boldsymbol{\theta}_0)\mathbf{u}(\boldsymbol{\theta}_0). \tag{A.17}$$

This result provides the basis for an iterative approach for computing the mle known as the *Newton-Raphson* technique. Given a trial value, we use Equation A.17 to obtain an improved estimate and repeat the process until differences between successive estimates are sufficiently close to zero. (Or until the elements of the vector of first derivatives are sufficiently close to zero.) This procedure tends to converge quickly if the log-likelihood is well-behaved (close to quadratic) in a neighborhood of the maximum and if the starting value is reasonably close to the mle.

An alternative procedure first suggested by Fisher is to replace minus the Hessian by its expected value, the information matrix. The resulting procedure takes as our improved estimate

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\mathbf{u}(\boldsymbol{\theta}_0), \tag{A.18}$$

and is known as *Fisher Scoring*.

Example: Fisher Scoring in the Geometric Distribution. In this case setting the score to zero leads to an explicit solution for the mle and no iteration is needed. It is instructive, however, to try the procedure anyway. Using the results we have obtained for the score and information, the Fisher scoring procedure leads to the updating formula

$$\hat{\pi} = \pi_0 + (1 - \pi_0 - \pi_0\bar{y})\pi_0. \tag{A.19}$$

If the sample mean is $\bar{y} = 3$ and we start from $\pi_0 = 0.1$, say, the procedure converges to the mle $\hat{\pi} = 0.25$ in four iterations. □

A.2 Tests of Hypotheses

We consider three different types of tests of hypotheses.

A.2.1 Wald Tests

Under certain regularity conditions, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ has approximately in large samples a (multivariate) normal distribution with mean equal to the true parameter value and variance-covariance matrix given by the inverse of the information matrix, so that

$$\hat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})). \tag{A.20}$$

Exercise: Variance of \hat \lambda Poisson

The regularity conditions include the following: the true parameter value $\boldsymbol{\theta}$ must be interior to the parameter space, the log-likelihood function must be thrice differentiable, and the third derivatives must be bounded.

This result provides a basis for constructing tests of hypotheses and confidence regions. For example under the hypothesis

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{A.21}$$

for a fixed value $\boldsymbol{\theta}_0$, the quadratic form

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \text{var}^{-1}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{A.22}$$

has approximately in large samples a chi-squared distribution with p degrees of freedom.

This result can be extended to arbitrary linear combinations of $\boldsymbol{\theta}$, including sets of elements of $\boldsymbol{\theta}$. For example if we partition $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$, where $\boldsymbol{\theta}_2$ has p_2 elements,then we can test the hypothesis that the last p_2 parameters are zero

$$H_o : \boldsymbol{\theta}_2 = 0,$$

by treating the quadratic form

$$W = \hat{\boldsymbol{\theta}}_2{}' \, \text{var}^{-1}(\hat{\boldsymbol{\theta}}_2) \, \hat{\boldsymbol{\theta}}_2$$

as a chi-squared statistic with p_2 degrees of freedom. When the subset has only one element we usually take the square root of the Wald statistic and treat the ratio

$$z = \frac{\hat{\theta}_j}{\sqrt{\text{var}(\hat{\theta}_j)}}$$

as a z-statistic (or a t-ratio).

These results can be modified by replacing the variance-covariance matrix of the mle with any consistent estimator. In particular, we often use the inverse of the expected information matrix evaluated at the mle

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}).$$

Sometimes calculation of the expected information is difficult, and we use the observed information instead.

Example: Wald Test in the Geometric Distribution. Consider again our sample of $n = 20$ observations from a geometric distribution with sample mean $\bar{y} = 3$. The mle was $\hat{\pi} = 0.25$ and its variance, using the estimated expected information, is $1/426.67 = 0.00234$. Testing the hypothesis that the true probability is $\pi = 0.15$ gives

$$\chi^2 = (0.25 - 0.15)^2/0.00234 = 4.27$$

with one degree of freedom. The associated p-value is 0.039, so we would reject H_0 at the 5% significance level. □

A.2.2 Score Tests

Under some regularity conditions the score itself has an asymptotic normal distribution with mean 0 and variance-covariance matrix equal to the information matrix, so that

$$\mathbf{u}(\boldsymbol{\theta}) \sim N_p(0, \mathbf{I}(\boldsymbol{\theta})). \tag{A.23}$$

This result provides another basis for constructing tests of hypotheses and confidence regions. For example under

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

the quadratic form

$$Q = \mathbf{u}(\boldsymbol{\theta}_0)' \, \mathbf{I}^{-1}(\boldsymbol{\theta}_0) \, \mathbf{u}(\boldsymbol{\theta}_0)$$

has approximately in large samples a chi-squared distribution with p degrees of freedom.

The information matrix may be evaluated at the hypothesized value $\boldsymbol{\theta}_0$ or at the mle $\hat{\boldsymbol{\theta}}$. Under H_0 both versions of the test are valid; in fact, they are asymptotically equivalent. One advantage of using $\boldsymbol{\theta}_0$ is that calculation of the mle may be bypassed. In spite of their simplicity, score tests are rarely used.

Example: Score Test in the Geometric Distribution. Continuing with our example, let us calculate the score test of $H_0 : \pi = 0.15$ when $n = 20$ and $\bar{y} = 3$. The score evaluated at 0.15 is $u(0.15) = -62.7$, and the expected information evaluated at 0.15 is $\mathbf{I}(0.15) = 1045.8$, leading to

$$\chi^2 = 62.7^2/1045.8 = 3.76$$

with one degree of freedom. Since the 5% critical value is $\chi^2_{1,0.95} = 3.84$ we would accept H_0 (just). □

A.2.3 Likelihood Ratio Tests

The third type of test is based on a comparison of maximized likelihoods for nested models. Suppose we are considering two models, ω_1 and ω_2, such that $\omega_1 \subset \omega_2$. In words, ω_1 is a subset of (or can be considered a special case of) ω_2. For example, one may obtain the simpler model ω_1 by setting some of the parameters in ω_2 to zero, and we want to test the hypothesis that those elements are indeed zero.

The basic idea is to compare the maximized likelihoods of the two models. The maximized likelihood under the smaller model ω_1 is

$$\max_{\boldsymbol{\theta}\in\omega_1} \mathrm{L}(\boldsymbol{\theta}, \mathbf{y}) = \mathrm{L}(\hat{\boldsymbol{\theta}}_{\omega_1}, \mathbf{y}), \tag{A.24}$$

where $\hat{\boldsymbol{\theta}}_{\omega_1}$ denotes the mle of $\boldsymbol{\theta}$ under model ω_1.

The maximized likelihood under the larger model ω_2 has the same form

$$\max_{\boldsymbol{\theta}\in\omega_2} \mathrm{L}(\boldsymbol{\theta}, \mathbf{y}) = \mathrm{L}(\hat{\boldsymbol{\theta}}_{\omega_2}, \mathbf{y}), \tag{A.25}$$

where $\hat{\boldsymbol{\theta}}_{\omega_2}$ denotes the mle of $\boldsymbol{\theta}$ under model ω_2.

The ratio of these two quantities,

$$\lambda = \frac{\mathrm{L}(\hat{\boldsymbol{\theta}}_{\omega_1}, \mathbf{y})}{\mathrm{L}(\hat{\boldsymbol{\theta}}_{\omega_2}, \mathbf{y})}, \tag{A.26}$$

is bound to be between 0 (likelihoods are non-negative) and 1 (the likelihood of the smaller model can't exceed that of the larger model because it is *nested* on it). Values close to 0 indicate that the smaller model is not acceptable, compared to the larger model, because it would make the observed data very unlikely. Values close to 1 indicate that the smaller model is almost as good as the large model, making the data just as likely.

Under certain regularity conditions, minus twice the log of the likelihood ratio has approximately in large samples a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models. Thus,

$$-2\log\lambda = 2\log \mathrm{L}(\hat{\theta}_{\omega_2}, y) - 2\log \mathrm{L}(\hat{\theta}_{\omega_1}, y) \to \chi^2_\nu, \tag{A.27}$$

where the degrees of freedom are $\nu = \dim(\omega_2) - \dim(\omega_1)$, the number of parameters in the larger model ω_2 minus the number of parameters in the smaller model ω_1.

Note that calculation of a likelihood ratio test requires fitting two models (ω_1 and ω_2), compared to only one model for the Wald test (ω_2) and sometimes no model at all for the score test.

Example: Likelihood Ratio Test in the Geometric Distribution. Consider testing $H_0 : \pi = 0.15$ with a sample of $n = 20$ observations from a geometric distribution, and suppose the sample mean is $\bar{y} = 3$. The value of the likelihood under H_0 is $\log \mathrm{L}(0.15) = -47.69$. Its unrestricted maximum value, attained at the mle, is $\log \mathrm{L}(0.25) = -44.98$. Minus twice the difference between these values is

$$\chi^2 = 2(47.69 - 44.99) = 5.4$$

with one degree of freedom. This value is significant at the 5% level and we would reject H_0. Note that in our example the Wald, score and likelihood ratio tests give similar, but not identical, results. □

The three tests discussed in this section are asymptotically equivalent, and are therefore expected to give similar results in large samples. Their small-sample properties are not known, but some simulation studies suggest that the likelihood ratio test may be better that its competitors in small samples.

Exercise: MLE Gaussian
Observed information matrix
Expected information matrix