Applied Likelihood methods:
With examples in R, SAS and ADMB

Russell B. Millar
Department of Statistics
University of Auckland

June 18, 2010

Copyright ©2010, Russell Millar, Auckland, New Zealand

Contents

Preface

Part One: Preliminaries

Chapter 1

Likelihood

When it is not in our power to follow what is true, we ought to follow what is most probable — Rene Descartes

1.1 Introduction

The word *likelihood* has its origins in the late fourteenth century (Simpson and Weiner 1989), and examples of its usage include as an indication of probability or promise, or grounds for probable inference. In the early twentieth century, Sir Ronald Fisher (1890-1962) presented a general purpose tool for statistical inference (Fisher 1912) and some nine years later he gave this tool the name *likelihood* (Fisher 1921). Fisher's choice of terminology was ideal, because the centuries-old interpretation of the word *likelihood* is also applicable to the statistical likelihood that is used throughout this book.

Here, likelihood is used within the traditional framework of frequentist statistics. It is presented as a general-purpose tool for inference, including the evaluation of statistical significance, calculation of confidence intervals, model assessment, and prediction. The frequentist theory underlying the use of likelihoods is covered in Part Three, where it is seen that maximum likelihood estimators (MLEs) have optimal properties for sufficiently large sample sizes. It is for this reason that likelihood-based inference is the most widely used form of traditional parametric inference. The pragmatic use of likelihood-based inference is the primary focus of this book

and is contained in Part Two. The reader who is already comfortable with the concept of likelihood and its basic properties can proceed to Part Two directly.

Likelihood is a fundamental concept of other statistical paradigms, particularly Bayesian statistics. The Bayesian approach to inference is not considered here, but consideration of the philosophical distinctions between frequentist and Bayesian statistics is examined in Chapter 14.

A simple binomial example is used in Section 1.2 to motivate and demonstrate many of the essential properties of likelihood that are developed in later Chapters. In Example 1.1 the likelihood is simply the probability of observing $y = 10$ successes from 100 trials. The fundamental conceptual point is that likelihood views the probability of observing y successes not as a function of y, but as a function of the unknown success probability p. That is, the likelihood function does not consider other values of y. It takes the knowledge that $y = 10$ was the observed number of successes and it uses the binomial probability of $y = 10$, evaluated at different possible values of p, to judge the relative likelihood of different values of p.

1.2 Motivating example

Throughout this text, the true unknown value of the parameter(s) is denoted with a zero subscript. Thus, in Example 1.1, p_0 is used to denote the true unknown probability of success. Without the subscript, p, is used to denote any possible value that the parameter could take, that is, any value within the parameter space. In Example 1.1 the parameter space of the binomial probability is, of course, all values of p between 0 and 1.

Example 1.1. Binomial. A random sample of one hundred trials is performed and ten result in success. What can be inferred about the unknown probability of success, p_0?

For any value of p ($0 \leq p \leq 1$) for the unknown probability of success, the probability of 10 successes from 100 trials is given by the binomial probability formula

(Section 15.3) with $y = 10$ successes from $n = 100$ trials. This is

$$L(p) = \text{Prob}(10 \text{ successes}) = \frac{100!}{90!\ 10!}\ p^{10}(1-p)^{90} \quad , 0 \le p \le 1 \tag{1.1}$$

The above probability is the likelihood, and has been denoted $L(p)$ to make its dependence on p explicit.

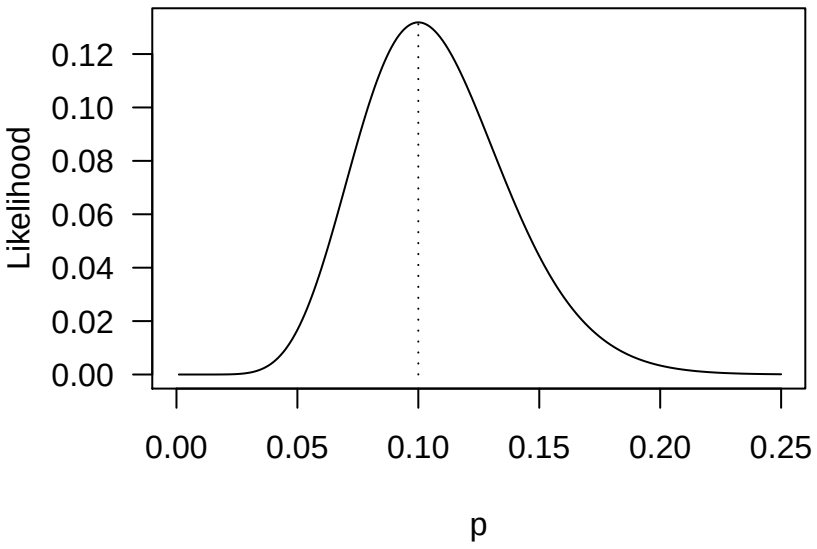


Figure 1.1: Binomial likelihood for 10 successes from 100 trials.

A plot of $L(p)$ (Fig. 1.1) shows it to be unimodal with a peak at 0.1. That is, the maximum likelihood estimate (MLE) of p_0 is simply the observed proportion of successes. This MLE will be denoted $\widehat{p}$. □

Box 1.1.

The curve in Fig. 1.1 looks somewhat like the bell-shaped curve of the normal density function. However, it is not a density (it is a likelihood function) and nor is it bell-shaped. On close inspection it can be seen that the curve is slightly right-skewed.

In the above example, the MLE $\widehat{p}$ is simply a point-estimate of p_0, and is of limited use without any sense of how reliable it is. For example, it would be more meaningful to have a range of plausible values of the unknown p_0, or to know if some pre-specified value, e.g., $p_0 = 0.5$ was reasonable. Such questions can be addressed

by examining the shape of the likelihood function, or more usually, the shape of the log-likelihood function.

The log of the likelihood function is used far more predominantly in likelihood inference than the likelihood function itself, for several good reasons:

1. The likelihood and log-likelihood are both maximized by the MLE.

2. Likelihood values are often extremely small (but can also be extremely large) depending on the model and amount of data. This can make numerical optimization of the likelihood highly problematic, compared to optimization of the log-likelihood.

3. The plausibility of parameter values is quantified by ratios of likelihood, corresponding to a constant difference on the log scale.

4. And, it is from the log-likelihood (and its derivatives) that most of the theoretical properties of MLEs are obtained, via application of the central limit theorem and related large sample results (see Part Three).

Points 3 and 4 above allude to the two most commonly used forms of likelihood inference — inference based on the likelihood ratio and inference based on asymptotic normality of the MLE, respectively. These two forms of likelihood-based inference are asymptotically equivalent (Section 12.5) in the sense that they lead to the same conclusions for sufficiently large sample sizes. However, in real situations there can be a non-negligible difference between these two approaches (Section 4.4).

Using the likelihood ratio approach in the context of Example 1.1, an interval of plausible values of the unknown parameter p_0 is obtained as all values p for which $\log(L(p))$ is above a certain threshold. In Section 3.4 it is shown that the threshold can be chosen so that the resulting interval has desirable frequentist properties. In the continuation of Example 1.1 below, the threshold is chosen so that the resulting interval is a (approximate) 95% confidence interval for the unknown value p_0.

The curvature of the log-likelihood is of fundamental importance in both the theory and practice of likelihood inference. The curvature is quantified by the second derivative. When evaluated at the MLE, the second derivative is negative and the

larger its absolute value the more sharply curved the log-likelihood at its maximum. Intuitively, a sharply curved log-likelihood is desirable because this narrows the range over which the log-likelihood is close to its maximum value, that is, it narrows the range of plausible parameter values. In Section 3.2 it is shown that the inverse of the negative of the second derivative provides an estimate of the variance of the MLE. This is particularly convenient in practice because many optimization programs calculate this second derivative as part of the optimization algorithm, in which case the second derivative is an automatic byproduct from maximizing the log-likelihood. The approximate normality of MLEs enables confidence intervals and hypothesis tests to be performed using well-established techniques.

The likelihood ratio and curvature-based methods of likelihood inference are demonstrated in the following continuation of Example 1.1.

Example 1.1 ctd. The log-likelihood function for $p, 0 < p < 1$ is

$$\begin{aligned} l(p) &= \log(L(p)) \\ &= \log\left(\frac{100!}{90!\ 10!}\right) + 10\log(p) + 90\log(1-p) \\ &= 30.48232 + 10\log(p) + 90\log(1-p) \end{aligned} \tag{1.2}$$

and the maximized value of this log-likelihood is $l(\widehat{p}) = l(0.1) \approx -2.03$.

In Section 3.4 it is seen that an approximate 95% confidence interval for p_0 is given by all values of p for which $l(p)$ is within about 1.92 of the maximized value of the log-likelihood. (The value 1.92 arises as one half of the 95% quantile of a chi-square distribution with one degree of freedom). That is, the interval is given by all values of p for which $l(p)$ is -3.95 or higher. This confidence interval can be read from Fig 1.2, or obtained numerically for greater accuracy. This interval is (0.051,0.169) to the accuracy of three decimal places. From the equivalence between confidence intervals and hypothesis tests (Section 13.1) it can be concluded that the null hypothesis $H_0 : p = p_0$ will be rejected at the 5% level for any value of p_0 outside of the interval (0.051,0.169).

To perform inference based on the curvature of the log-likelihood, the second derivative of the log-likelihood given by (1.2) is required. This second derivative is

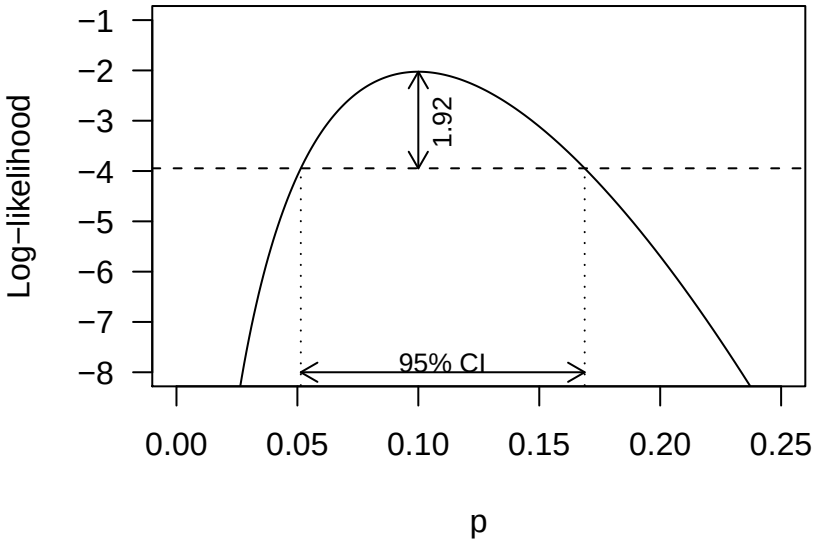


Figure 1.2: Binomial log-likelihood for 10 successes from 100 trials, and 95% likelihood ratio confidence interval.

given in equation (11.14), and for $n = 100$ trials and $y = 10$ successes it is

$$l''(p) = \frac{\partial^2 l(p)}{\partial p^2} = -\frac{10}{p^2} - \frac{90}{(1-p)^2} \; . \tag{1.3}$$

Evaluating this second derivative at the MLE $\widehat{p} = 0.1$ gives

$$l''(\widehat{p}) = -\frac{10}{0.01} - \frac{90}{0.81} = -\frac{1000}{0.9} \approx -1111.111$$

The inverse of the negative of $l''(0.1)$ is exactly 0.0009, and according to likelihood theory (Sections 3.2 and 12.2), this is the approximate variance of $\widehat{p}$. The approximate standard error is therefore $\sqrt{0.0009} = 0.03$.

Recall that for a binomial experiment, the true variance of $\widehat{p}$ is $p_0(1-p_0)/n$, which is estimated by $\widehat{p}(1-\widehat{p})/n$. For this example, this estimate of variance is also 0.0009, the same as that obtained from using $-1/l''(0.1)$. (In fact, for the binomial the two variance estimates are always the same, regardless of n and y.).

For sufficiently large n, the distribution of $\widehat{p}$ can be approximated by a normal distribution, thereby permitting approximate tests and confidence intervals for p_0 to be performed using familiar techniques. These are often called Wald tests or intervals, due to the influential work of Abraham Wald in establishing the large-sample approximate normality of MLEs (e.g., Wald 1943). The $(1-\alpha)100\%$ Wald

confidence interval for p_0 can be obtained using the familiar formula that calculates the upper (and lower) bounds as the point estimate plus (and minus) $z_{1-\alpha/2}$ times the estimated standard error, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. Thus, the approximate 95% confidence interval for p_0 is

$$\widehat{p} \pm z_{0.975}\widehat{\text{s.e}}(\widehat{p}) \tag{1.4}$$

where $z_{0.975} = 1.960$ and $\widehat{\text{s.e}}(\widehat{p}) = 0.03$. This interval is (0.041,0.159). Equivalently, this interval is the collection of all points p_0 such the null hypothesis $H_0 : p = p_0$ is not rejected at the 5% level by the test statistic, i.e.,

$$\left|\frac{\widehat{p} - p_0}{\widehat{\text{s.e}}(\widehat{p})}\right| < z_{0.975} \ . \tag{1.5}$$

□

1.2.1 Approximate normality vs likelihood ratio

The Wald form of confidence interval used in (1.4) is based on the approximate normal distribution of $\widehat{p}$. This is the most commonly used method for constructing approximate confidence intervals because of its intuitive appeal and computational ease. It was shown earlier that the likelihood ratio can be used as an alternative method of constructing confidence intervals – which should be used?

From a pragmatic point of view, there is considerable intuitive appeal in the Wald construction of a 95% (say) confidence interval, with bounds given by 1.96 standard errors each side of the point estimate. This form of CI will be the most familiar to anyone with a basic grounding in frequentist statistics. Unfortunately, when the LR and Wald intervals differ substantially, it is generally the case that the LR approach is superior, in the sense that the CIs obtained using likelihood ratio will have actual coverage probability closer to the desired coverage probability (see Section 4.4.1). In fact, the results of Brown, Cai and DasGupta (2001) question the popular usage of the Wald CI for binomial inference because of its woeful performance, even for values of n and p for which the normal approximation to the binomial distribution is generally considered reasonable (typically, $\min(n\widehat{p}, n(1-\widehat{p}) \geq 5)$. However, the LR

confidence interval is not as widely used because it requires (a little) knowledge of likelihood theory, but more importantly because it can not generally be calculated explicitly.

Application of Wald tests and CIs extends to multi-parameter inference, but becomes more cumbersome and unfamiliar when simultaneous inference about two or more parameters is required. It is then that LR-based inference tends to be used. In particular, multi-parameter inference is typical of model selection problems, and in this area LR-based inference dominates. Also, it should be noted that model selection criterion such as Akaikes Information Criterion (AIC) make direct use of the likelihood (Section 4.6).

Box 1.2.

In addition to the Wald and LR intervals, there are several other competing methods for constructing approximate confidence intervals for the unknown probability p_0 in a binomial experiment. These include the score (see Section 12.6), Agresti-Coull, and the misnamed "exact" CI's. The comparisons performed by Agresti and Coull (1998) and Brown, Cai and DasGupta (2002) suggest that the LR and score CI's are to be preferred.

Summary

To conclude, Example 1.1 is likelihood inference in a nutshell. Much of the rest of this book is devoted to providing pragmatic guidance on the use (and potential abuse) of inferential methods based on likelihood ratios and approximate normality of MLEs and their application to more complex and realistic models. These concepts extend naturally to models with two or more parameters, although the implementation can become challenging. For example, in a model with s number of parameters, the second derivative of the log-likelihood is an s-dimensional square matrix (the Hessian) and the negative of its inverse provides an approximate variance matrix for the MLEs.

1.3 Using SAS, R and ADMB

This book is not just about learning maximum likelihood inference, it is also very much about *doing* it with real data. Examples in SAS and R are provided throughout

Part Two.

Unlike the SAS and R environments, Automatic Differentiation Model Builder (ADMB) is a tool specifically designed for complex optimization problems. Its use is difficult to justify if existing functionality within SAS or R can be used instead. Other than the quick demonstration of ADMB here, it is used sparingly until Chapter 10 where it becomes the best choice for fitting latent variable models. Some of its additional abilities are noted in Sections 4.2.3 and 5.4.2.

The SAS examples presented in this text were implemented using SAS for Windows version 9.2. The SAS procedures used throughout are found in the statistics module SAS/STAT, with the exception that occasional use was made of the nonlinear optimizer PROC NLP which is in the operations research module SAS/OR. Some users of SAS/STAT will find that their licence does not extend to SAS/OR and hence will not be able to use PROC NLP. For this reason, PROC NLP is used sparingly and alternative SAS code is given where possible.

SAS procedures typically produce a lot of output. The output often includes a lot of superfluous information such as summary information about the data-set being used, computational information, and unwanted summary statistics. Throughout, the Output Delivery System (ODS) in the SAS software has been used to select only the required parts of the output produced by the SAS procedure.

For ease of readability, the SAS code presented herein follows the typographical convention used in Delwiche and Slaughter (2003). This convention is to write SAS keywords in uppercase, and to use lowercase for variable names, data-set names, comments, etc. Note that SAS code is not case sensitive.

The R examples were developed using R for Windows version 2.10.1. (Ihaka and Gentleman 1996, Anonymous 2003). R is freely available under the terms of the Free Software Foundation's GNU General Public License (see www.R-project.org). Most of the R functions used herein are incorporated in the default installation of R. Others are available within R library packages and these can be easily loaded from within the R session.

ADMB is freely available via the ADMB foundation (admb-foundation.org), where full instruction for running ADMB can also be found. In brief, ADMB is imple-

mented by programming the (negative) log-likelihood within an ADMB template file. An executable file is then created from the template file. Fortunately, much of the detail in creating the executable can now be hidden behind convenient user interfaces. Indeed, the ADMB examples in this book were run from within R using the PBSadmb package.

In many situations it will be possible to make use of existing SAS procedures and R functions that are appropriate to the type of data being modeled, notwithstanding that this convenience often comes at the loss of flexibility. Rather than using functionality specific to the binomial model, the implementations of Example 1.1 presented below demonstrate a selection of the general-purpose tools available in SAS and R. In particular, calculation of likelihood ratio confidence intervals is an application of profile likelihood, and the examples below makes use of general code for this purpose.

1.3.1 Software resources

Several small pieces of code have been written to facilitate techniques described in this text. These are listed in Section 15.4, along with a brief description of their functionality. These software resources are available for download from `www.stat.auckland.ac.nz/~millar`.

1.4 Implementation of motivating example

The code used below demonstrates how an explicit likelihood function is maximized within each of SAS, R, and ADMB, and calculation of the Wald and likelihood-ratio confidence intervals. Some efficiencies could have been gained by taking advantage of built-in functionality within the software. For example, in the SAS example, the binomial model could have been expressed using the statement `MODEL y ~ BINOMIAL(n,p)`, but the general-purpose likelihood specification has been used instead. The constant of 30.48232 is irrelevant to maximization of the log-likelihood, however, it is included in these examples for consistency with Figure 1.2.

The description of the code is relatively thorough here compared to the remainder

of this text, so as to give a clear picture of the very simple use of these tools. This level of explanation is too unwieldy to be used throughout the remainder of this text, and for more explanation the reader should refer to the abundant online resources and documentation for each of these software.

1.4.1 Binomial example in SAS

The SAS code below uses `PROC NLMIXED` to implement Example 1.1, and produces Figure 1.3. One difficulty is that this procedure does not produce likelihood ratio confidence intervals. A general purpose macro called `Plkhci` has been written for this purpose.

```
DATA binomial;
  y=10; n=100; constant=30.48232;
RUN;

*Select only parameter estimates table;
ODS SELECT ParameterEstimates;

PROC NLMIXED DF=1E6 DATA=binomial;
  PARMS p=0.5;
  BOUNDS 0<p<1;
  loglhood=constant+y*log(p)+(n-y)*log(1-p);
  MODEL y~GENERAL(loglhood);
RUN;

%INCLUDE "PlkhciMacro.sas";

%MACRO BinomialProfile(p);
  PROC NLMIXED DF=1E6 DATA=Binomial; TECH=NONE;
    loglhood=constant+y*log(&p)+(n-y)*log(1-&p);
    MODEL y~GENERAL(loglhood);
  RUN;
%MEND;
%Plkhci(BinomialProfile,0.0,0.1,-32.5082973,side="L");
%Plkhci(BinomialProfile,0.1,1.0,-32.5082973,side="R");
```

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
p	0.1000	0.03000	1E6	3.33	0.0009	0.05	0.04120	0.1588	8.566E−7

Figure 1.3: The Parameter Estimates table from `PROC NLMIXED`, including the `Lower` and `Upper` bounds of the 95% Wald confidence interval.

Some features of above code are:

- The default output includes several tables, including tables of log-likelihood values and fit statistics. The Output Delivery System statement `ODS SELECT ParameterEstimates;` is used to select only the required table.

- By default, `NLMIXED` calculates Wald intervals using a t-distribution with degrees of freedom equal to the number of observations (rows in the dataset). To get the normal-based Wald interval in (1.4) the degrees of freedom needs to be set to a large value. In this case, it was set to one million using the procedure option `DF=1E6`.

- The `PARMS` statement is an optional statement used to explicitly list the parameters and their initial values.

- The `BOUNDS` statement is an optional statement used to specify the range of the parameter values (i.e., the parameter space).

- The model is specified using the `MODEL` statement. Here, the model is given as `GENERAL(loglhood)` to specify that `PROC NLMIXED` should maximize the value of `loglhood` that is calculated by the preceding programming statement.

- The user-defined macro `BinomialProfile` contains a version of the `NLMIXED` code to be used by the profile likelihood macro `Plkhci` A brief description of this macro is found in Section 15.4.3.

- The `Plkhci` macro finds the likelihood ratio confidence bounds. It writes the lines

```
Left-sided 95% LR CI bound is 0.051413
Right-sided 95% LR CI bound is 0.168779
```

  to the log window of the SAS session.

- In the SAS output in Figure 1.3, `Gradient` gives the slope of the log-likelihood upon termination of the optimization. It should be near zero.

- The `t-Value` and `Pr>|t|` columns in Figure 1.3 should be ignored. They are the Wald test statistic and p-value for the null hypothesis $H_0 : p = 0$. This is not a relevant hypothesis here.

For SAS installations that include the operations research OR module, `PROC NLP` provides an easier option for obtaining the likelihood ratio confidence interval, via its `PROFILE statement`. Figure 1.4 shows the table that is produced from running the following code.

```
*Select only the desired table;
ODS SELECT WaldPLLimits;
PROC NLP COV=2 VARDEF=N;
  MAX loglike;
  PROFILE p / alpha=0.05;
  PARMS p=0.5;
  BOUNDS 0<p<1;
  n=100; y=10; constant=30.48232;
  loglike=constant+y*LOG(p)+(n-y)*LOG(1-p);
RUN;
```

- `PROC NLP` can produce several different estimates of variance (see Section 3.2.1) and the `COV=2` option specifies using the curvature used in the motivating example. Also, by default, `PROC NLP` makes a degrees-of-freedom adjustment to the estimate of variance. This adjustment would estimate the variance of $\widehat{p}$ as $\widehat{p}(1-\widehat{p})/(n-1)$, which is not standard practice in the context of binomial data. The procedure option `VARDEF=N` prevents this.

- The `MAX loglike` statement specifies that the value `loglike` is to be maximized.

- The `PROFILE` statement specifies a likelihood ratio confidence interval for parameter p with confidence level $(1-\alpha)100\%$.

PROC NLP: Nonlinear Maximization

Wald and PL Confidence Limits							
N	**Parameter**	**Estimate**	**Alpha**	**Profile Likelihood Confidence Limits**		**Wald Confidence Limits**	
1	**p**	0.100000	0.050000	0.051414	0.168773	0.041201	0.158799

Figure 1.4: Likelihood ratio and Wald confidence limits from `PROC NLP`

1.4.2 Binomial example in R

In the R code below, the negative of the log-likelihood is explicitly defined as function `nloglhood`, with argument p. The minimum of `nloglhood`, and its second derivative, are found using the general purpose minimizer `optim`. The likelihood ratio confidence interval is then obtained using the `plkhci` (from the `Bhat` package) function for profile likelihood confidence intervals.

```
> #Define the negative log-likelihood function
> nloglhood=function(p) return( -(30.48232+10*log(p)+90*log(1-p)) )
> #Minimize the negative log-likelihood
> binom.fit=optim(0.5,nloglhood,lower=0.0001,upper=0.9999,hessian=T)
> #MLE
> phat=binom.fit$par
> #Variance
> phat.var=1/binom.fit$hessian
> #Calculate approximate 95% Wald CI
> phat+c(-1,1)*qnorm(0.975)*sqrt(phat.var)
[1] 0.04120779 0.15879813

> #Load package Bhat
> library(Bhat)
> #Set up list for input into plkchi function
> control.list=list(label="p",est=phat,low=0,upp=1)
> #Calculate approximate 95% likelihood ratio CI
> plkhci(control.list,nloglhood,"p")
[1] 0.05141279 0.16877909
```

- In the call of `optim`, the first argument specifies that the initial parameter value to be used by the optimizer is 0.5. The `lower` and `upper` arguments specify the parameter space – in this case they were set to 0.0001 and 0.9999 because computational error occurs if bounds of 0 and 1 are used due to `nloglhood` being undefined at these values. The `hessian=T` argument requests that the value of the second derivative (calculated at the MLE) is included in `binom.fit`.

- The list object `binom.fit` has several components, including the estimated MLE `binom.fit$par` and hessian `binom.fit$hessian`.

- The first argument to the profile likelihood function `plkhci` is a list with elements giving the parameters of `nloglhood`, the MLE, and lower and upper bounds of the parameter space.

1.4.3 Binomial example in ADMB

The following ADMB template file (`BinomialMLE.tpl`) is used to find the MLE and its approximate standard error.

```
DATA_SECTION
  init_number y
  init_number n

PARAMETER_SECTION
  init_bounded_number p(0,1)
  objective_function_value negllhood

PROCEDURE_SECTION
  negllhood=-(30.48232+y*log(p)+(n-y)*log(1-p));
```

- ADMB requires a data section, and the data are contained in a file with name `BinomialMLE.dat`. This text file contains a single row, with value `10 100`.
- The parameter section specifies the parameter name and its range, and the name of the variable that will have the value of the negative log-likelihood.

An executable program is generated from the template file, and executed, using the following R code. This uses functions from the `PBSadmb` package.

```
library(PBSadmb)
readADopts()
makeAD("BinomialMLE")
runAD("BinomialMLE")
```

- `readADopts` reads a text file containing information about the ADMB installation.
- `makeAD` creates a C++ file from the template file, compiles it, and links it to produce an executable with name `BinomialMLE.exe`.
- `runAD` runs the executable. A number of files are produced, including text file `BinomialMLE.std` containing the MLE and its standard error, and text file `BinomialMLE.par` containing the value of the negative log-likelihood at the MLE.

A few additional lines of code are required to obtain the likelihood ratio confidence interval. The parameter section requires a `likeprof_number` specification to name the quantity of interest. In this case it is p, but this name is already in use, so the variable `pcopy` is used to contain a copy of p. The preliminary calculations section is used to set options for the grid of `pcopy` values over which the objective function is evaluated.

```
DATA_SECTION
  init_number y
  init_number n

PARAMETER_SECTION
  init_bounded_number p(0,1)
  objective_function_value negllhood
  likeprof_number pcopy

PRELIMINARY_CALCS_SECTION
  pcopy.set_stepnumber(500);
  pcopy.set_stepsize(0.01);

PROCEDURE_SECTION
  negllhood=-(30.48232+y*log(p)+(n-y)*log(1-p));
  pcopy=p;
```

Within R, the executable is created as before. The `runAD` function now requires the optional `lprof` argument to pass to the executable, to force calculation of the likelihood ratio confidence interval. If the template file is named `BinomialLRCI.tpl`, then the `runAD` call looks like

```
runAD("BinomialLRCI",argvec="-lprof > RunWindow.txt")
```

which also redirects a copious log file into the text file `RunWindow.txt`. This produces a text file `pcopy.plt`, in which the bounds of the 95% confidence interval, calculated to be `0.0513786` and `0.168687`, can be found.

1.5 Exercises

1.1 The Poisson distribution is the default distribution for the modeling of count data. If the value $y = 3$ is observed from a Poisson distribution with unknown parameter λ then the log-likelihood is (to within an additive constant) $l(\lambda) = -\lambda + 3\log\lambda$. This log-likelihood is maximized by $\widehat{\lambda} = 3$.

1. Plot $l(\lambda)$ for values of λ from 0.1 to 15.
2. By suitable modification to the program code in Section 1.3, use R or SAS to verify that $\widehat{\lambda} = 3$, and to calculate the 95% Wald and likelihood ratio confidence intervals for λ.

Chapter 2

A taste of likelihood

All models are wrong, but some are useful — George E. P. Box[1].

2.1 Introduction

The binomial distribution used in the motivating example (Example 1.1) is a discrete distribution and the likelihood function for the observation of ten successes from 100 trials was simply the probability of that event, regarded as a function of p. It was natural to estimate p using the value $\widehat{p}$ that maximizes this probability. However, this logic does not extend to data observed from continuous distributions, because the "probability" of continuous data is always zero[2].

For continuous data, the likelihood function is defined to be the density function evaluated at the observed data, regarded as a function of the unknown parameter(s). This has intuitive appeal, and if any justification is required, it can be argued that the measured value of an observation y is subject to rounding accuracy, and therefore it makes sense to consider the probability of observing a value "close" to y. In the scalar case, this would be the probability of an observation being in the interval $y-\epsilon, y+\epsilon$ (see Box 2.1). For small ϵ this probability is approximately proportional to the density function.

Thanks to mathematical measure theory (Billingsley 1979), it is not necessary to make any distinction between discrete and continuous data when using the ter-

[1] See Box 2.2

[2] In the continuous case there are infinitely many possible outcomes, all with probability zero!

minology "density" function. That is, this terminology can be used to refer to both the density function of a continuous random variable, or the probability function of a discrete random variable.

Box 2.1.

In practice, continuous data will be measured to a certain accuracy. For example, if a person's weight is rounded to the nearest 100 gm, then a recorded weight of $y = 74.2$ kg is actually the event that their weight is between 74.15 and 74.25 kg. To a close degree of approximation, the probability of this event is proportional to the value of the density function evaluated at the fixed value of 74.2. This is the likelihood, and will be a function of parameters associated with the experimental circumstances under which the weight was measured.

A bit of notation is required before formally presenting the definition of likelihood function.

2.1.1 Some necessary notation

Throughout, bold notation is used for vectors and matrices. If the data consist of the observation (i.e., measurement) of n numbers then these observations are denoted by the vector $\boldsymbol{y} = (y_1, y_2, ..., y_n)$, and they are considered to be the observed realization of the random vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$. In the examples of Section 2.3 the n observations are independent and identically distributed (iid), but this will not be assumed in general.

The experiment (or process) that generated the data is assumed to be described by a statistical distribution with (joint) density function $f(\boldsymbol{y}; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ denotes the true unknown value of the parameters. This density function may depend on covariates (i.e., explanatory variables) associated with the observed data $\boldsymbol{y}$. The set of all possible values of $\boldsymbol{\theta}$ is the parameter space, denoted Θ, and is assumed to be a subset of s-dimensional real space $\mathbb{R}^s$. The collection of distributions that is formed from all values of $\boldsymbol{\theta} \in \Theta$ comprises the model. Expressing this formally gives the following definition.

Definition 2.1 Parametric statistical model

A parametric statistical model is a collection of joint densities functions, $f(\boldsymbol{y}; \boldsymbol{\theta})$,

indexed by $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$.

Box 2.2.

In the vast majority of applications it is explicitly assumed that the form of the chosen model is correct, that is, the joint density of the data truly is $f(\boldsymbol{y}; \boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \Theta$. Inference is then obtained from knowledge about the behaviour of certain statistics (such as the MLE or likelihood ratio) given that the model is correctly specified. However, the quote at the start of this chapter belies this naive interpretation of the statistical model. In all but the simplest of random experiments, the statistical model can only be regarded as an approximation to truth. With judicious choice of model specification and evaluation, it should be a sufficiently good choice to be useful.
See Section 4.6 for more on this topic.

Example 2.1. The normal model. It is often assumed that "Y is normally distributed". Formally, this is specifying a parametric statistical model comprising the collection of all possible normal distributions. This is the collection of $N(\mu, \sigma^2)$ distributions over the parameter space $\boldsymbol{\theta} = (\mu, \sigma)$ with $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$, where $\mathbb{R}^+$ denotes the positive reals. The true distribution of Y is a member of this collection, denoted $Y \sim N(\mu_0, \sigma_0^2)$. □

Statistical theory (Chapter 12) requires the statistical model to satisfy certain regularity conditions. Some of these conditions have obvious interpretation. For example, condition R2 (Section 12.2) requires identifiability of the model. Identifiability simply means that the densities $f(\boldsymbol{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$ all correspond to distinct statistical distributions. That is, if $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two different parameter values then the joint density functions $f(\boldsymbol{y}; \boldsymbol{\theta}_1)$ and $f(\boldsymbol{y}; \boldsymbol{\theta}_2)$ must correspond to different statistical distributions. Clearly, if multiple values of $\boldsymbol{\theta}$ index the same statistical distribution then there is no possible way to distinguish between those parameter values on the basis of observing data from the common distribution.

The example below gives a trivial example of a non-identifiable model. Perhaps a more familiar example would be the use of aliasing when fitting ANOVA models. Aliasing is needed to avoid non-identifiability arising due to ambiguity between the interpretation of the intercept parameter and the coefficient parameters associated

with the levels of the factor(s). The binormal mixture model present in Section 2.3.5 is also non-identifiable, but this can easily be rectified by a judicious restriction of the parameter space.

Example 2.2. A silly non-identifiable model. Let $Y_1, ..., Y_n$ be iid normal with standard deviation of unity and mean equal to the product of two unknown real-valued parameters, a, and b. That is $Y_i \sim N(ab, 1)$. This model is not identifiable because all sets of parameter vectors $\boldsymbol{\theta} = (a, b)$ having the same value of their product correspond to the same statistical distribution. □

When the the data $y_1, ..., y_n$ are independent, with Y_i having distribution with density $f_i(y_i; \boldsymbol{\theta})$, then the joint density function is simply the product of the individual densities

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

In many cases, the Y_i may share a distribution of common form (e.g., normal), but differ due to explanatory variables $\boldsymbol{x}_i$, in which case it may be more convenient to replace $f_i(y_i; \boldsymbol{\theta})$ with the notation $f(y_i; \boldsymbol{x}_i, \boldsymbol{\theta})$.

Definition 2.2 Likelihood function

The likelihood function *is the joint density function evaluated at the observed data, and regarded as a function of* $\boldsymbol{\theta}$ *alone. That is,* $L(\boldsymbol{\theta}) = f(\boldsymbol{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$.

The log-likelihood function, denoted $l(\boldsymbol{\theta})$, is the (natural) log of the likelihood function.

Definition 2.3 Maximum likelihood estimate (MLE)

Given $\boldsymbol{y} = y_1, ..., y_n$, *any* $\widehat{\boldsymbol{\theta}} \in \Theta$ *that maximizes* $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta})$ *over* Θ *is called a* maximum likelihood estimate (MLE) *of the unknown true parameter* $\boldsymbol{\theta}$.

The above definition does not assume existence or uniqueness of the MLE.

In this text, the distinction between estimate and estimator will be maintained by replacing the observed $\boldsymbol{y}$ with the random variable $\boldsymbol{Y}$. So, for example, the maximum likelihood *estimator* is obtained from Definition 2.3 by replacing $\boldsymbol{y}$ by $\boldsymbol{Y}$. In the case of the binomial model $Y \sim Bin(100, p)$ used in Example 1.1, the observation was $y = 10$ and the ML *estimate* was $y/100 = 0.1$, which was the realization of the ML *estimator* Y/n. For simplicity's sake, the notation $\widehat{\boldsymbol{\theta}}$ will not make any explicit distinction between the ML estimate and the ML estimator. The distinction is implied by context, for example, when talking about the statistical properties of MLEs it is the estimator that is being considered.

Any parameter value that maximizes the likelihood also maximizes the log-likelihood and, for reasons given in the previous Chapter, it is usually the case that calculations are made using $l(\boldsymbol{\theta})$ rather than $L(\boldsymbol{\theta})$. In particular, if the log-likelihood function is differentiable then local extreme values can be found by setting its partial derivatives to zero. That is, by finding the roots of the *likelihood equation(s)*

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_i} = 0, \quad j = 1, ..., s.$$

To ensure that a root of the likelihood equations corresponds to a local maximum (rather than a local minimum or saddle point) it is necessary to check that the second derivative is negative. In the multi-parameter case ($s \geq 2$) the second derivative is a matrix and it is required to be negative-definite.

2.1.2 MLEs of functions of the parameters

It is often the case that the parameters $\boldsymbol{\theta}$ used to specify the statistical model are not of direct interest, but rather, it may be desired to estimate a function of the model parameters. This is straightforward because, in plain words, the MLE of a function of the parameters is that function of the MLE.

The above statement is intuitive, but nonetheless, it deserves a formal verification. To that end, let $g(\boldsymbol{\theta})$ (possibly vector valued) denote a function of $\boldsymbol{\theta}$. If g is a one-to-one function on Θ then the collection of densities in the statistical model can instead be indexed by $\boldsymbol{\zeta} = g(\boldsymbol{\theta}) \in g(\Theta)$. That is, the densities are indexed by $\boldsymbol{\zeta}$ where $f_{\boldsymbol{\zeta}}(\boldsymbol{y}; \boldsymbol{\zeta}) = f_{\boldsymbol{\zeta}}(\boldsymbol{y}; g(\boldsymbol{\theta})) = f(\boldsymbol{y}; \boldsymbol{\theta})$. The collection of densities comprising

the statistical model is the same in both cases, and the density that maximizes the likelihood within this collection is $f(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}) = f_{\boldsymbol{\zeta}}(\boldsymbol{y}; g(\widehat{\boldsymbol{\theta}}))$. That is, $\widehat{\boldsymbol{\zeta}} = g(\widehat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\zeta} = g(\boldsymbol{\theta})$. This argument can be extended to the situation where g is not one-to-one.

Since the statistical model is invariant to one-to-one transformations of the parameters, the modeler is free to implement the model using the parameterization of their choice. The most convenient parameterization is not always the best. Some model parameterizations will result in MLEs with better properties (e.g., MLE closer to normally distributed) than others – this is examined in Section 4.4. In conventional linear-normal models (i.e., regression, ANOVA, etc) it makes sense to parameterize a normal distribution using mean and variance because of the well established exact theory for these models. However, this is not the case more generally, especially in complex models that include multiple sources of randomness. Then it is often preferable to use the mean and log-variance as the parameters of normally distributed error terms (e.g., see Section 10.4.3).

Example 2.3. An iid sample $y_1, ..., y_n$ observed from a $N(\mu, \sigma^2)$ distribution resulted in the MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma}) = (5, 2)$. It was of interest to estimate the following two quantities.

1) The coefficient of variation (CV), $\zeta_1 = g_1(\theta) = \frac{\sigma}{\mu}$.

2) The probability that a new observation would not exceed 6,

$$\zeta_2 = g_2(\theta) = P(Y \leq 6) \ .$$

The MLE of the CV is immediate, $\widehat{\zeta_1} = g(\widehat{\theta}) = \frac{\widehat{\sigma}}{\widehat{\mu}} = 0.4$.
For part 2), $g_2(\theta)$ needs to be expressed explicitly as a function of μ and σ. This is

$$g_2(\theta) = P\left(Z \leq \frac{6-\mu}{\sigma} \ , \right)$$

where Z has a standard normal distribution. Then,

$$\widehat{\zeta_2} = g_2(\widehat{\theta}) = P(Z \leq 0.5) \approx 0.6915 \ .$$

□

In practice, it may be necessary to make inference about $\boldsymbol{\zeta}$. Section 4.2 shows how the large-sample distribution of $\widehat{\boldsymbol{\zeta}}$ can be deduced from the large-sample distribution of $\widehat{\theta}$.

2.2 Interpretation of likelihood

Chapter 14 presents some of the foundational concepts of statistical inference, including statements of the likelihood principle. The arguments in Chapter 14 are extremely controversial amongst philosopher-statisticians and they quickly become esoteric and flavoured by nuances and subtleties in meaning. This present Section steers clear of the controversies, and takes a quick look at a long-established interpretation of likelihood, namely, that it provides a measure of the relative support for different values of $\boldsymbol{\theta}$ (Edwards 1972). For simplicity, the example below uses a statistical model where θ can take only two possible values.

Example 2.4. Likelihood to the rescue.
You've just bought a fashionable Louise Vashon wristwatch for a bargain price, but now you're starting to wonder whether all is what it seems because the evening news just reported the arrest of an international gang that had been making imitation Louise Vashon watches.

Genuine Louise Vashon watches are very accurate. In fact, over a month, their deviation (seconds) from true time is well described by a $N(0, 1)$ distribution. The imitation watches are quite inaccurate and it has been found that their time deviation is well described by a $N(0, 100^2)$ distribution. A jeweler friend measured the precise accuracy of your watch, and it was found to gain 2 s per month.

Box 2.3.

The scenario posed in Example 2.4 can be considered a parametric statistical model with a parameter space containing just two values, 0 and 1, say. When $\theta = 0$ then $f(y; \theta)$ is the density of a standard normal, and when $\theta = 1$ it is the density of a normal with mean zero and standard deviation of 100.

Question: Is your watch genuine?

Argument 1: A genuine watch has time deviation that is described by a standard normal distribution, and the measured time deviation of 2 s is two standard deviations above the mean. Ouch, it looks like you were ripped off and have a fake watch!

Argument 2: Calculate the likelihood of observing $y = 2$ from a $N(0, 1)$ distribution and compare to that of a $N(0, 100^2)$ distribution? These likelihoods are 0.0540 and 0.0040 for genuine and imitation watches, respectively (Fig. 2.1). Yippee, your watch is thirteen and a half times more likely to be genuine than not!

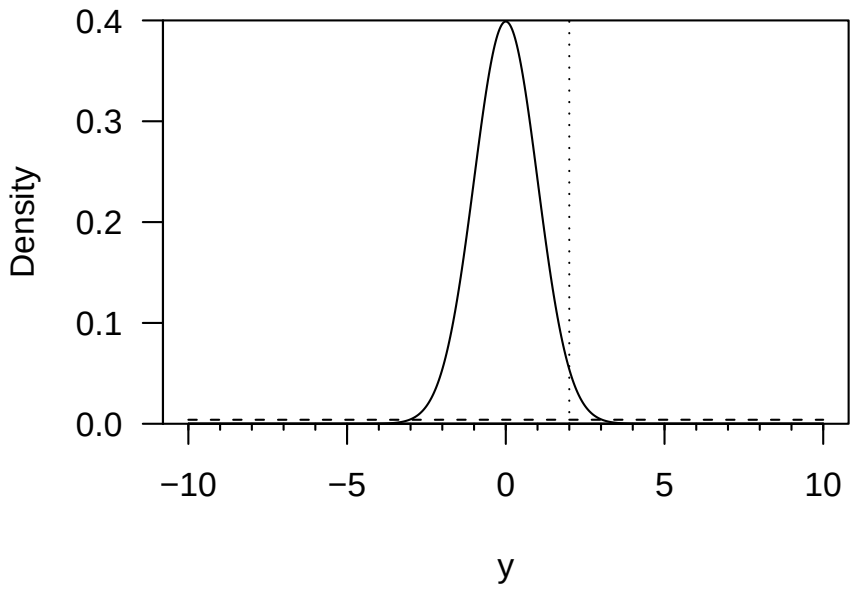


Figure 2.1: $N(0, 1)$ (solid line) and $N(0, 100^2)$ (the near-horizontal long-dashed line) densities. The likelihood principle says that it is only the values of the densities at the observed value of $y = 2$ that are relevant to the question of the authenticity of your watch.

The conflicting conclusions arise due to the fundamental difference in the framework used to address the question – *is your watch genuine?* The first argument invokes the logic of hypothesis testing. Under the null hypothesis that the watch is genuine, the observed $y = 2$ is somewhat extreme, and this null hypothesis would be formally rejected at the 5% level.

A hypothesis test performed at the 5% level is designed to falsely reject the null hypothesis five percent of the time under repetition of the experiment. In this case, although $y = 2$ has higher likelihood for a genuine watch than a fake watch, it is nonetheless a mildly extreme value to observe from a standard normal distribution and falls within the critical region of the most powerful size 0.05 test.

The second argument is the appropriate one. The falsification philosophy of hypothesis testing (e.g., Popper 1959) is irrelevant to this example. Note that the question could have been expressed equivalently as *is your watch a fake?*. Under the null hypothesis that the watch is fake, the value $y = 2$ *also* results in rejection of the null hypothesis (see the continuation of this example in Section 14.5 for full details)! □

This section closes with the confession that the above example is extremely contrived, for the purpose of demonstrating the pure interpretation of likelihood as a measure of relative support for different parameter values. The example does not fit the model framework that is typical of practice and required for likelihood-based inference to have established properties. For example, in the case of nested models, the likelihood ratio test in Section 3.4 will not reject the null hypothesis unless a sufficiently large improvement in value of the log-likelihood is given under the larger model that is specified under the alternative hypothesis. This is very much in the spirit of likelihood.

2.3 IID Examples

The examples presented here look at maximum likelihood estimation for models where the data are n iid observations, and so the joint density function for the data is simply $f(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta})$. Despite this simplicity, these examples do demonstrate many interesting phenomena. For example, small-sample bias of MLEs is noted in the Normal and Uniform examples, and it is seen that the likelihood equation for the Uniform model has no solution because the likelihood is not differentiability at the MLE. In the Cauchy example the likelihood can have multiple maxima.

In the final example, the binormal mixture model is seen to be non-identifiable, and its likelihood has multiple unbounded maxima. Nonetheless, it is possible to finding a sensible and meaningful local maximum.

2.3.1 Binomial(n, p)

Here, y is a a single binomial(n, p) observation. Nonetheless, this can be considered an iid example because, for the purpose of making inference about p, a binomial(n, p) experiment is equivalent to n iid Bernoulli(p) experiments (see Chapter 14).

Example 2.5. IID Binomial(n, p). Let Y be distributed Bin$(n, p), 0 < p < 1$. Given y, the likelihood function for p is

$$L(p) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y} \ ,$$

with log-likelihood

$$l(p) = \log \left(\begin{array}{c} n \\ y \end{array} \right) + y \log(p) + (n-y) \log(1-p) \ .$$

The likelihood equation is

$$\frac{\partial l(p)}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p} = \frac{y - np}{p(1-p)} = 0,$$

the solution of which produces the unique MLE, $\widehat{p} = y/n$.

To be thorough, we should demonstrate that $\widehat{p}$ is indeed a maximum of $l(p)$ by showing that $\frac{\partial^2 l(p)}{\partial p^2} < 0$ at $\widehat{p}$. □

2.3.2 Normal(μ, σ^2)

This example demonstrates the small-sample bias that is typical of MLEs, and moreover, that this bias is not cause for concern.

Example 2.6. IID Normal(μ, σ^2). Let Y_i be iid N$(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$. The normal density function is

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2} \right), \quad y \in \mathbb{R} \tag{2.1}$$

and so the density for the data vector $\boldsymbol{y} = (y_1, ..., y_n)$ is

$$f(\boldsymbol{y}; \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right), \quad y_i \in \mathbb{R} \ .$$

The log-likelihood is

$$l(\mu, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2.$$

and the likelihood equations are

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mu) = 0 \ , \tag{2.2}$$

and

$$\frac{\partial l(\mu, \sigma)}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n (y_i - \mu)^2 = 0 \ . \tag{2.3}$$

From (2.2) it is immediate that $\widehat{\mu} = \overline{y}$. Plugging this value of $\widehat{\mu}$ into (2.3) and simplifying gives the ML estimate $\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2}$. □

Note that the ML estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2 \ ,$$

which uses a divisor that does not adjust for the loss of one degree of freedom from estimation of μ. The usual unbiased estimator is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2 \ ,$$

and since $\widehat{\sigma}^2 = (\frac{n-1}{n})S^2$, the bias of $\widehat{\sigma}^2$ is

$$E\left[\widehat{\sigma}^2\right] - \sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = \frac{-\sigma^2}{n}.$$

The bias in $\widehat{\sigma}^2$ is typically of no great concern here. Indeed, Press, Teulolsky, Vetterling and Flannery (2007, Section 14.1 therein) have this to say on the matter,

> *We might also comment that if the difference between n and $n-1$ ever matters to you, then you are probably up to no good anyway ...*

Moreover, $\widehat{\sigma}^2$ has lower mean-squared error than S^2 (Exercise 2.4) and so arguably is a better estimator of σ^2!

The small-sample bias in the ML estimation of variance parameters can become relevant in the context of variance components models. Then, it may be justified to use a modified form of likelihood which encapsulates a degrees-of-freedom type adjustment (Chapter 9).

Box 2.4.

With minor modification, the above calculations can be extended to the linear regression model where Y_i are independently distributed $N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$ where $\boldsymbol{x}_i$ is the covariate vector associated with observation i. It can be shown that the MLE of $\boldsymbol{\beta}$ is the usual least-squares estimator (see Example 11.6), and that the MLE of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})^2 \ .$$

In contrast, the unbiased estimator of σ^2 has a $n-p$ in the denominator, where p is the dimension of $\boldsymbol{\beta}$.

2.3.3 Uniform$(0, M)$

In this example, the log-likelihood is not differentiable at the MLE, and the likelihood equation does not have a solution. Instead,the MLE is obtained from direct inspection of the likelihood function.

Example 2.7. IID Uniform$(0, M)$ Let Y_i be iid from a Uniform$(0, M)$ distribution. Note that the density function is $1/M$ only for y values in the the interval $[0, M]$, and is zero otherwise. Thus, the likelihood for the n observations is

$$L(M) = \begin{cases} \frac{1}{M^n}, & y_i \leq M \text{ for all } i \\ 0, & \text{otherwise} \end{cases}$$

For parameter M to be greater than or equal to all y_i, it is equivalent that M be greater than or equal to the maximum y value, y_{max}. That is (Figure 2.2)

$$L(M) = \begin{cases} \frac{1}{M^n}, & y_{max} \leq M \\ 0, & \text{otherwise} \end{cases}$$

It is clear that $\widehat{M} = y_{max}$. □

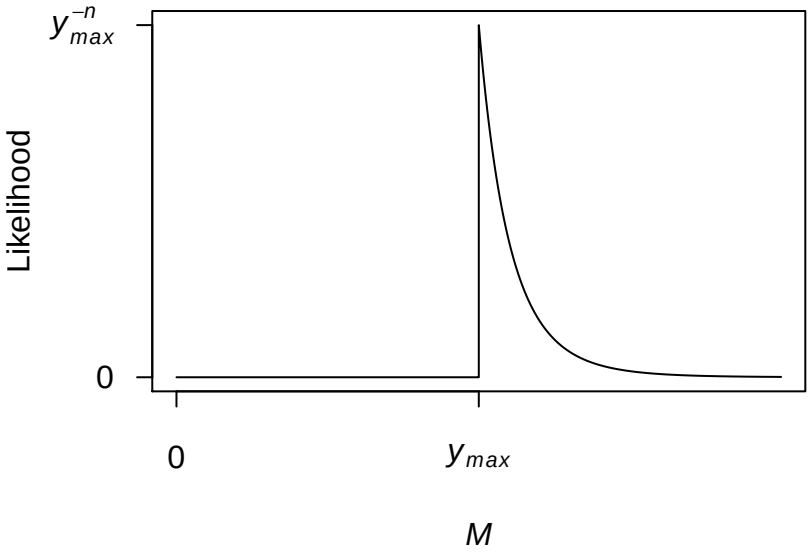


Figure 2.2: Likelihood from Uniform(0,M) data.

Since y_{max} is always less than or equal to M, the estimator Y_{max} will be a negatively biased estimator of M. It can be shown that the expected value of $E[Y_{max}] = nM/(n+1)$, i.e., that the bias of Y_{max} is $-M/(n+1)$ (Exercise 2.3).

2.3.4 Cauchy(θ)

The Cauchy log-likelihood is prone to having multiple local maxima for a small sample size, but is typically near quadratic in shape for larger n.

Example 2.8. IID Cauchy(θ). Let Y_i be iid from a one-parameter Cauchy distribution with median $\theta \in \mathbb{R}$. Each y_i has density function

$$f(y;\theta) = \frac{1}{\pi(1+(y-\theta)^2)}, \quad y \in \mathbb{R} \ ,$$

and so the log-likelihood arising from the data $\boldsymbol{y} = (y_1, ..., y_n)$ is

$$l(\theta) = -n\log(\pi) - \sum_{i=1}^{n} \log(1+(y_i-\theta)^2) \ . \tag{2.4}$$

The derivative of the log-likelihood is

$$l'(\theta) = 2\sum_{i=1}^{n} \frac{y_i - \theta}{1+(y_i-\theta)^2} \ .$$

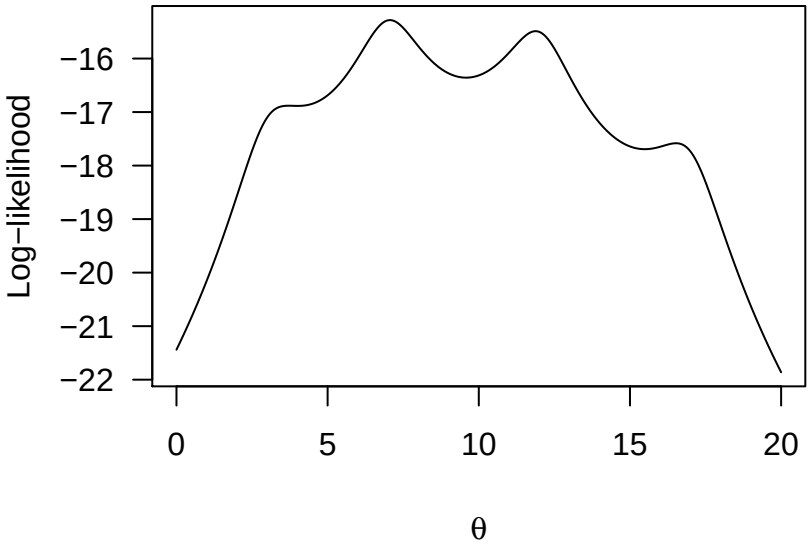


Figure 2.3: Log-likelihood of θ for the data $\boldsymbol{y} = (3, 7, 12, 17)$.

Finding roots of the above equation requires numerical solution. □

Figure 2.3 plots the Cauchy log-likelihood for the data $(y_1, y_2, y_3, y_4) = (3, 7, 12, 17)$ (these are the values used by (Edwards 1972)).

With a sample size of $n = 100$, the Cauchy likelihood function appears to become better behaved, in the sense that (under simulation of new data) it is typically unimodal and close to quadratic in shape (Fig. 2.4).

2.3.5 Binormal$(p, \mu, \sigma, \nu, \tau)$ mixture model for Old Faithful

In this example, it is shown that the likelihood can be made arbitrarily large, and hence no global MLE exists. In addition, the model is not identifiable. However, in practice these complications are generally pathological, and can be avoided by imposing sensible constraints on the parameter space.

Example 2.9. IID Binormal$(p, \mu, \sigma, \nu, \tau)$.

The binormal distribution is a mixture of two normal distributions, $N(\mu, \sigma^2)$ and $N(\nu, \tau^2)$, say. Each data point, y_i, is obtained by first choosing one of the two normal distributions at random (with probabilities p and $1 - p$, respectively), and then generating a value from that distribution. However, the identity of the chosen

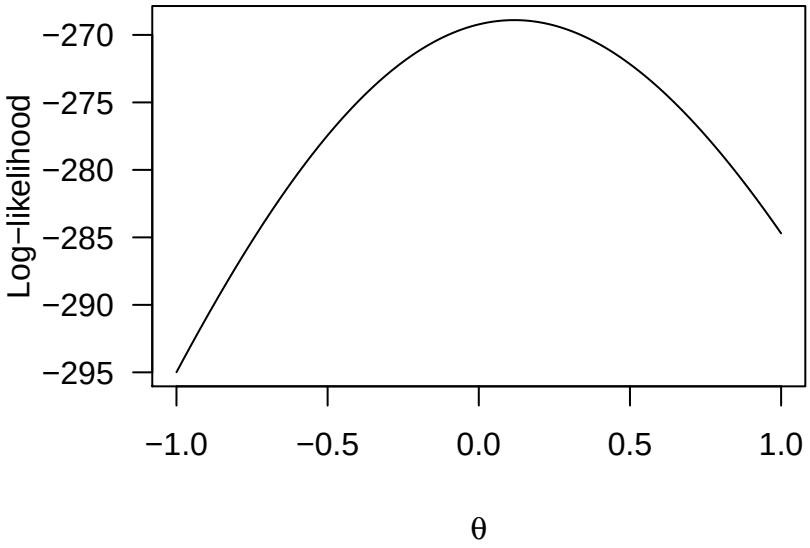


Figure 2.4: Log-likelihood of θ for a random sample of size 100 from a standard Cauchy distribution ($\theta = 0$).

distribution is never observed. The data $\boldsymbol{y}$ therefore consist of some (or possibly no) observations from the $N(\mu, \sigma^2)$ distribution, and some (or possibly no) from the $N(\nu, \tau^2)$ distribution, but it is not known which observations came from which of these two distributions. The vector of parameters is $\boldsymbol{\theta} = (p, \mu, \sigma, \nu, \tau)$ where $0 \leq p \leq 1$, $\mu, \nu \in \mathbb{R}$, and $\sigma, \tau > 0$.

To express the binormal distribution formally, let B_i be (unobserved) iid Bernoulli(p) random variables. If $B_i = 1$ then Y_i is observed from a $\mathrm{N}(\mu, \sigma^2)$ distribution, otherwise it is observed from a $\mathrm{N}(\nu, \tau^2)$ distribution. That is,

$$Y_i | B_i = b_i \sim \begin{cases} N(\mu, \sigma^2) & , b_i = 1 \\ N(\nu, \tau^2) & , b_i = 0 \end{cases}$$

The joint density of (Y_i, B_i) is therefore given by

$$\begin{aligned} f(y_i, b_i; \boldsymbol{\theta}) &= f(y_i | b_i; \mu, \sigma, \nu, \tau) P(B_i = b_i; p) \\ &= \begin{cases} \frac{p}{\sqrt{2\pi}\sigma} \exp(-(y-\mu)^2/2\sigma^2) & , b_i = 1 \\ \frac{1-p}{\sqrt{2\pi}\tau} \exp(-(y-\nu)^2/2\tau^2) & , b_i = 0 \end{cases} \end{aligned}$$

from which the marginal density of Y_i is obtained as

$$\begin{aligned} f(y_i;\boldsymbol{\theta}) &= \sum_{b_i\in\{0,1\}} f(y_i,b_i;\boldsymbol{\theta}) \qquad (2.5)\\ &= P(B_i=1;p)f(y_i|B_i=1;\mu,\sigma)+P(B_i=0;p)f(y_i|B_i=0;\nu,\tau)\\ &= \frac{p}{\sqrt{2\pi}\sigma}\exp(-(y-\mu)^2/2\sigma^2)+\frac{1-p}{\sqrt{2\pi}\tau}\exp(-(y-\nu)^2/2\tau^2)\ . \qquad (2.6)\end{aligned}$$

That is, the density function $f(y;\boldsymbol{\theta})$ of a binormal distribution is a linear combination of the density functions of the $\mathrm{N}(\mu,\sigma^2)$ and $\mathrm{N}(\nu,\tau^2)$ distributions. Given $\boldsymbol{y}$, the log-likelihood function $l(\boldsymbol{\theta})$ must be maximized numerically. □

One complication is that $l(\boldsymbol{\theta})$ is unbounded, that is, it can be arbitrarily large. To see how this can happen, fix ν,τ and $0<p<1$ at any set values you wish – this is done to ensure that the second term in (2.6) takes fixed positive values for all i, so that (2.6) will be bounded away from zero regardless of the values of μ and σ. Now set $\mu=y_i$ for any choice of i. For observation i, these parameter choices give

$$f(y_i;\boldsymbol{\theta})=\frac{p}{\sqrt{2\pi}\sigma}+\frac{1-p}{\sqrt{2\pi}\tau}\exp(-(y_i-\nu)^2/2\tau^2)\ , \qquad (2.7)$$

and note that (2.7) can be made arbitrarily large by making σ arbitrarily small.

A further wrinkle in the binormal model is that it is non-identifiable because the role of the two distributions in the mixture can be swapped. That is, the binormal distribution corresponding to parameters (p,μ,σ,ν,τ) is the same as that specified by parameters $(1-p,\nu,\tau,\mu,\sigma)$.

The unbounded likelihood and non-identifiability issues can be eliminated by suitable restriction on the parameter space. One possibility is to constrain the ratio of the two standard deviations by requiring that $0<c<\sigma/\tau<1$, where c is some suitably small constant (Hathaway 1985). In practice, despite the unbounded likelihood and non-identifiability, a sensible local maximum of the likelihood function can usually be found using an unconstrained optimizer, provided that the initial values of $\boldsymbol{\theta}$ are somewhere in the general vicinity of the local maximum. Ultimately, it is the shape of the likelihood function in the neighbourhood of this local maximum that is relevant to inference.

A well known example of a binormal mixture distribution is the waiting time between eruptions of the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The bimodal nature of these data suggest that two distinct geological processes are occurring within the geyser. The R language contains a data-frame called `faithful`, containing 272 observations.

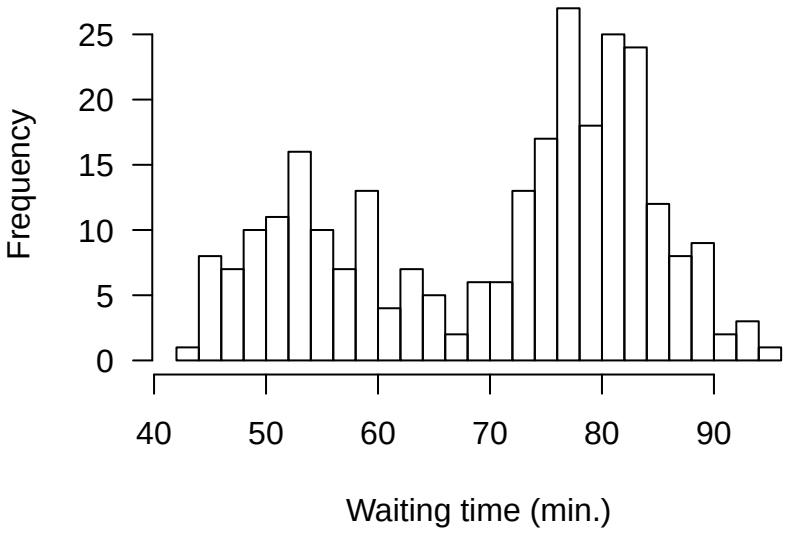


Figure 2.5: Waiting times (minutes) between eruptions of the Old Faithful geyser, from the `faithful` data-frame.

A histogram of the Old Faithful data (Fig. 2.5) shows that the waiting times look like a combination of (very roughly) 40% from a $N(52, 25)$ distribution and 60% from a $N(80, 25)$ distribution. The corresponding parameter values $\boldsymbol{\theta}^{(0)} = (p, \mu, \sigma, \nu, \tau) = (0.4, 52, 5, 80, 5)$ would make good start values for finding a local MLE using R or SAS.

2.4 Exercises

2.1 If $y_1, ..., y_n$ are iid Pois(λ), show that the MLE of λ is the sample mean, $\overline{y}$.

2.2 In example 2.3, calculate the MLE of $\zeta = g(\theta)$, defined by $P(Y \leq g(\theta)) = 0.99$.

2.3 In Example 2.7 we saw that for $Y_1, ..., Y_n$ from a Uniform$(0, M)$ distribution, the MLE of M was the maximum observed value y_{max}. The distribution function of

y_{max} is

$$\begin{aligned} F(t) &= \text{Prob}(Y_{max} \le t) \\ &= \text{Prob}(Y_i \le t, i = 1, ..., n) \\ &= \begin{cases} 0 & , t < 0 \\ \left(\frac{t}{M}\right)^n & , 0 \le t \le M \\ 1 & , M < t \end{cases} \end{aligned}$$

Differentiate $F(t)$ to obtain the density function $f(t)$ of y_{max}, and hence show that the expected value of Y_{max} is $nM/(n+1)$.

2.4 The mean-squared error of an estimator $\widehat{\theta}$ of $\theta_0 \in \mathbb{R}$ is

$$\text{mse}(\widehat{\theta}) = E[(\widehat{\theta} - \theta_0)^2] = \text{var}(\widehat{\theta}) + \text{bias}(\widehat{\theta})^2$$

and it provides an overall measure of the estimator's performance that takes into account both variability and bias.

For iid $N(\mu, \sigma^2)$ data (Example 2.6), and using the fact that $\text{var}(S^2) = 2\sigma^4/(n-1)$, calculate the mean-squared error of S^2 and $\widehat{\sigma}^2$ and hence show that $\text{mse}(\widehat{\sigma}^2) < \text{mse}(S^2)$.

2.5 Let $Y_1, ..., Y_n$ be iid from a geometric distribution with density function

$$f(y) = p^y(1-p) \quad , y = 0, 1, 2, ...$$

where $0 < p < 1$. Show that the MLE of p is $\widehat{p} = \overline{y}/(1+\overline{y})$.

2.6 Let $Y_1, ..., Y_n$ be iid Gamma(α, β) with known α, leaving only the scale parameter β to be estimated. Show that the MLE of β is $\widehat{\beta} = \overline{y}/\alpha$.

2.7 Let $Y_1, ..., Y_n$ be iid from a distribution with density function

$$f(y) = \theta y^{\theta - 1}, \quad 0 < y < 1,$$

where $\theta > 0$. Show that the MLE of θ is

$$\widehat{\theta} = \frac{-n}{\sum_{i=1}^n \log y_i} \ .$$

2.8 Let $Y_i, i = 1, ..., n$ be iid from a Pareto(α, M) distribution, for some $\alpha > 0, M > 0$. That is,

$$f(y_i; \alpha, M) = \begin{cases} \alpha \frac{M^\alpha}{y_i^{\alpha+1}} & , y_i \ge M \\ 0 & \text{otherwise} \end{cases}$$

Find the MLE of (α, M).

2.9 Iid data $Y_1, ..., Y_n$ are observed from a location-shifted exponential distribution with density function

$$f(y; \alpha, \lambda) = \frac{1}{\lambda} \exp\left(\frac{-(y-\alpha)}{\lambda}\right) \ , \qquad y \ge \alpha, \qquad \alpha \in \mathbb{R}, \ \lambda > 0 \ .$$

Determine the maximum likelihood estimator of (α, λ).

2.10 Iid data $y_1, ..., y_n$ are observed from a Laplace distribution with density function

$$f(y; \alpha) = \frac{1}{2} \exp(-|y - \alpha|), \quad y \in \mathbb{R}, \quad \alpha \in \mathbb{R} \ .$$

Assuming that n is even (i.e $n = 2m$ for some positive integer m), determine a maximum likelihood estimator of α. Is your MLE unique?

2.11 Let $\boldsymbol{Y} = Y_1, ..., Y_n$ be iid observations from a zero-inflated Poisson distribution (Section 7.5.2). This distribution arises from a mixture model where, with probability p, Y is observed from the "distribution" which has point mass at zero (i.e., $P(Y = 0) = 1$), and with probability $(1 - p)$, Y is observed from a Poisson(λ) distribution. The zero-inflated Poisson distribution has density function

$$f(y; p, \lambda) = \begin{cases} p + (1-p)e^{-\lambda} & y = 0, \\ (1-p)\frac{e^{-\lambda}\lambda^y}{y!} & y = 1, 2, 3, ... \end{cases}$$

Assuming that λ is known, show that the MLE of p is

$$\widehat{p} = \frac{n_0 - ne^{-\lambda}}{n(1 - e^{-\lambda})}$$

where n_0 is the number of y_i's that take the value zero.

Part Two: Pragmatics

Chapter 3

Tests and construction of confidence intervals and regions

...the null hypothesis is never proved or established, but is possibly disproved ... — Sir Ronald A. Fisher

3.1 Introduction

This Chapter looks at how tests and intervals/regions are constructed from the large-sample approximate normality of MLEs, and from the approximate χ^2 distribution of likelihood ratio statistics. Tests and intervals/regions are considered for a single element of the parameter vector, $\boldsymbol{\theta}$, and for a subset of (or all) elements of $\boldsymbol{\theta}$.

In Chapter 4 it is seen that inference using the likelihood ratio is generally more reliable to that based on approximate normality. Nonetheless, the primary reason for assuming approximate normality is that it allows tests and confidence intervals to be easily constructed using familiar formulae. This is the so-called Wald approach where, for example, an approximate 95% CI for θ_k is given by $\widehat{\theta}_k$ plus or minus a couple of its standard errors. Approximate normality also has the flexibility of extending to functions of the parameters via the delta method (4.2.1). However, this flexibility can be as dangerous as it is useful, because "approximate normality" can often be wishful thinking rather than a reasonable approximation to the true sampling distribution.

There is undeniable virtue in the simplicity of the Wald approach, and in many

cases it will make little difference compared to using likelihood ratio. However, one should be aware that approximate normality is based on assumptions of approximate linearity, and hence should be especially dubious when, for example, a parameter has a very non-linear influence on the model[1]. To encourage use of likelihood ratio, this Chapter includes demonstration of the R function `Plkhci` and SAS macro of the same name (both described in Section 15.4) for construction of likelihood ratio confidence intervals.

The normality-based Wald approach is somewhat less appealing for joint inference about a multi-dimensional subset of the parameter vector. The formulae then require the use of quadratic forms, and they become more sensitive to the assumption of approximate multivariate normality. In this situation the likelihood ratio is more widely used.

3.2 Approximate normality of MLEs

Let the vector $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, ..., \widehat{\theta}_s)$ denote the MLE of the unknown true parameter vector $\boldsymbol{\theta}_0 = (\theta_{01}, ..., \theta_{0s}) \in \Theta \in \mathbb{R}^s$. Assuming appropriate regularity conditions and a reasonable sample size then, under repetition of observing new data $\boldsymbol{Y}$ from the distribution with density $f(\boldsymbol{y}; \boldsymbol{\theta}_0)$, the distribution of $\widehat{\boldsymbol{\theta}}$ is approximately that of an s-dimensional normal random vector with expected value $\boldsymbol{\theta}_0$. (See Chapter 12 for formal statement of the regularity conditions, and Section 12.3 for the derivation of the approximate normality result.)

The approximate normality of the MLE can be written

$$\widehat{\boldsymbol{\theta}} \sim N_s(\boldsymbol{\theta}_0, \mathbf{V}(\boldsymbol{\theta}_0)) \tag{3.1}$$

where $\sim$ denotes "approximately distributed", and N_s denotes an s-dimensional multivariate normal. The $s \times s$ dimensional matrix $\mathbf{V}(\boldsymbol{\theta}_0)$ is the large-sample variance matrix, and is typically a function of $\boldsymbol{\theta}_0$. For element k of $\widehat{\boldsymbol{\theta}}$, this reduces to

$$\widehat{\theta}_k \sim N(\theta_{0k}, \mathbf{V}(\boldsymbol{\theta}_0)_{kk}) \tag{3.2}$$

[1] Loosely speaking, this is where a small positive change in the value of θ_k alters the model to a very different degree than a small negative change in θ_k.

where $\mathbf{V}(\boldsymbol{\theta}_0)_{kk}$ denotes the kth diagonal element of $bfV(\boldsymbol{\theta}_0)$.

3.2.1 Estimating the large-sample variance of $\widehat{\boldsymbol{\theta}}$

To use (3.1) or (3.2) to perform hypothesis tests or construct confidence intervals/regions it is standard practice to replace the large-sample variance matrix $\mathbf{V}(\boldsymbol{\theta}_0)$ by an estimate, $\widehat{\mathbf{V}}$. This leads to very convenient formulae for hypothesis tests and confidence intervals/regions (Section 3.3). In some situations there are superior alternatives to this, for example, evaluating $\mathbf{V}(\boldsymbol{\theta}_0)$ using the value of $\boldsymbol{\theta}_0$ specified under a null hypothesis. However, these alternatives are not as straightforward to apply, and are rarely seen in practice. They are not considered in this Part of the text, but are examined in Part III (e.g., see the continuation of Example 12.7 on p. 271).

There are several reasonable choices of how to estimate $\mathbf{V}(\boldsymbol{\theta}_0)$ and there will typically be little difference between them for moderate sample sizes (see Section 12.4 and Exercise 12.9). It is the case that matrix $\mathbf{V}(\boldsymbol{\theta}_0)$ can be derived from the concept of an expected (under repetition of the experiment) curvature of the log-likelihood (see Section 11.6.1), and it is most natural to use an estimate derived from the curvature that is actually realized for the observed data $\boldsymbol{y}$. This curvature is quantified by the matrix of second derivatives of the log-likelihood function, and is evaluated at $\widehat{\boldsymbol{\theta}}$. This is the so-called Hessian matrix, denoted $\mathbf{H}(\widehat{\boldsymbol{\theta}})$.

Many optimizing routines, including the popular Newton-Raphson algorithm (Section 5.2) calculate the Hessian matrix in the course of optimization. The estimate of the variance matrix is simply the negative of the inverse of $\mathbf{H}(\widehat{\boldsymbol{\theta}})$,

$$\widehat{\mathbf{V}} = -\mathbf{H}(\widehat{\boldsymbol{\theta}})^{-1} \; . \tag{3.3}$$

The negative-Hessian of the log-likelihood plays a prominent role in likelihood theory and is called the *observed* Fisher information matrix.

Using $\widehat{\mathbf{V}}$ in place of $\mathbf{V}(\boldsymbol{\theta}_0)$, the approximate normality of $\widehat{\boldsymbol{\theta}}$ is

$$\widehat{\boldsymbol{\theta}} \sim N_s(\boldsymbol{\theta}_0, \widehat{\mathbf{V}}) \; . \tag{3.4}$$

In particular, for element k of the MLE vector

$$\widehat{\theta}_k \sim N(\theta_{0k}, \widehat{v}_{kk}) \tag{3.5}$$

where $\widehat{v}_{kk}$ denotes the kth diagonal element of $\widehat{\mathbf{V}}$. These are the statements of approximate normality that are utilized in construction of Wald tests and confidence intervals/regions in Section 3.3,

Box 3.1.

The approximate normality of $\widehat{\theta}_k$ in (3.5) should be interpreted as saying that its distribution function will look very much like that of a $N(\theta_{0k}, \widehat{v}_{kk})$ random variable. This is enough to guarantee that tests and confidence intervals that are derived under the assumption of normality will have the right properties, at least approximately. However, it does not guarantee that $\widehat{\theta}_k$ has *all* of the usual properties of a normal distribution. For example, it is quite possible that the mean and variance of $\widehat{\theta}_k$ do not exist, or that $\widehat{\boldsymbol{\theta}}$ does not exist for some subset of values $\boldsymbol{y}$ in the sample space that has positive probability of occurring (see Example 12.2).

3.3 Wald tests, confidence intervals and regions

The Wald test has the same general form for tests of a single parameter value, θ_k, or of a subset of two or more parameter values. To express this general form, it is convenient to require that the null hypothesis be stated in the form $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ where $\boldsymbol{\psi} = (\theta_1, ..., \theta_r)$ is the first r elements of $\boldsymbol{\theta}_0$. That is, the value of the first $r < s$ parameters are fixed under H_0. There is no loss of generality in specifying H_0 using the first r parameters, because the parameter vector can always be re-ordered if necessary. Moreover, this form of hypothesis can be used to test functions of the parameters by re-parameterization of the model.

Under the above notation, testing the value of a single parameter value corresponds to $\psi = \theta_1$ and the null hypothesis is $H_0 : \theta_1 = \theta_{01}$ for some specified constant θ_{01}. Of course, the "1" subscript could be replaced by "k", for any $k = 1, ..., s$, by re-ordering of the parameters, and that is assumed below.

3.3.1 Test for a single parameter

Under $H_0 : \theta_k = \theta_{0k}$, the test statistic is the familiar "Z statistic" obtained from standardizing (3.2). That is

$$Z = \frac{\widehat{\theta}_k - \theta_{0k}}{\sqrt{\widehat{v}_{kk}}} \sim N(0,1) \ . \tag{3.6}$$

Equivalently, since the square of a standard normal distribution is the χ^2_1 distribution,

$$W = \frac{(\widehat{\theta}_k - \theta_{0k})^2}{\widehat{v}_{kk}} \sim \chi^2_1 \ , \tag{3.7}$$

where W is the Wald test statistic.

Equation (3.7) gives the approximate distribution of the Wald test statistic under repetition of the experiment in which the random variable $\boldsymbol{Y}$ is generated from the distribution with density $f(\boldsymbol{y}; \boldsymbol{\theta}_0)$. Values of W that are too large to have plausibly come from a χ^2_1 distribution provide evidence against H_0. Specifically, if w is the value of the Wald test statistic that is observed, the p-value is $P(\chi \geq w)$ where $\chi \sim \chi^2_1$. An approximate size α test is given by rejecting H_0 if W exceeds the $1-\alpha$ quantile $\chi^2_{1,1-\alpha}$, or equivalently, if the absolute value of Z exceeds $\sqrt{\chi^2_{1,1-\alpha}}$. Note that $\sqrt{\chi^2_{1,1-\alpha}} = z^2_{1-\alpha/2}$ because of the relationship between the $N(0,1)$ and χ^2_1 distribution.

The approximate $(1-\alpha)100\%$ confidence interval for θ_{0k} is given by all values of θ_{0k} that are not rejected by the level α Wald test. This is the familiar interval

$$(\widehat{\theta}_k - z_{1-\alpha/2}\sqrt{\widehat{v}_{kk}}, \widehat{\theta}_k + z_{1-\alpha/2}\sqrt{\widehat{v}_{kk}}) \ . \tag{3.8}$$

3.3.2 Joint test of two or more parameters

Denoting $\widehat{\boldsymbol{\psi}} = (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_r)$, it follows from (3.4) that, under H_0,

$$\widehat{\boldsymbol{\psi}} \sim N_r(\boldsymbol{\psi}_0, \widehat{\mathbf{V}}_{\boldsymbol{\psi}}) \tag{3.9}$$

where $\widehat{\mathbf{V}}_{\boldsymbol{\psi}}$ is the approximate variance matrix of $\widehat{\boldsymbol{\psi}}$ and is given by the upper-left $r \times r$ sub-matrix of $\widehat{\mathbf{V}}$. The Wald test statistic is obtained as a "standardization" of $\widehat{\boldsymbol{\psi}}$ represented in quadratic form (e.g., Seber and Lee 2003, p. 30), and is

$$W = (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T \widehat{\mathbf{V}}_{\boldsymbol{\psi}}^{-1} (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \sim \chi^2_r \ . \tag{3.10}$$

Hypothesis tests and p-values are obtained as in the single parameter case, except that now the degrees of freedom is r, corresponding to the number of parameters restricted by the null hypothesis. The approximate $(1-\alpha)100\%$ confidence region for $\boldsymbol{\psi}_0$ is an r-dimensional ellipse, determined as all value $\boldsymbol{\psi}_0$ such that $W \leq \chi^2_{r,1-\alpha}$.

3.3.3 In R and SAS: Old Faithful revisited

Here, the binormal mixture model is fitted to data on waiting times between eruptions of the Old Faithful geyser (see Fig. 2.5) using general purpose optimizers within R and SAS.

The Old Faithful data are a time series of the durations between successive eruptions of Old Faithful over a period of about two weeks, and it would be natural to suspect that the measurements may not be independent due to temporal autocorrelation. This is indeed the case – successive waiting times are strongly negatively correlated. For the sake of demonstration, this temporal autocorrelation will be overlooked, and the data will be assumed iid. It is also the case that the binormal mixture model can be implemented using functions provided in several R packages. However, the direct coding and maximization of the log-likelihood is just as easy, and moreover, it provides greater flexibility for making inference and for adjusting the model if lack of fit is indicated.

In Example 2.9 the binormal model was introduced as a mixture of $\mathrm{N}(\mu, \sigma^2)$ and $\mathrm{N}(\nu, \tau^2)$ distributions. It has five parameters $\boldsymbol{\theta} = (p, \mu, \sigma, \nu, \tau)$, where p is the probability that Y will be generated from the $\mathrm{N}(\mu, \sigma^2)$ distribution. From equation (2.6), the log-likelihood arising from observations $y_1, ..., y_n$ from the binormal is

$$\sum_{i=1}^{n} \log\left(pf(y_i;\mu,\sigma) + (1-p)f(y_i;\nu,\tau)\right) \ , \tag{3.11}$$

where $f(y;\mu,\sigma)$ and $f(y;\nu,\tau)$ are the density functions of $\mathrm{N}(\mu, \sigma^2)$ and $\mathrm{N}(\nu, \tau^2)$ random variables, respectively.

To begin, approximate standard errors and 95% confidence intervals are obtained for each parameter. Then, to demonstrate a joint hypothesis test, it will be posited that volcanologists have completed seismic tests which suggest that the component of the binormal that has smaller mean should have an expected waiting time of 55

s with a standard deviation of 5 s, and that this component should be responsible for the eruption exactly one-third of the time (at random). That is, $\boldsymbol{\psi} = (p, \mu, \sigma)$ and the null hypothesis is $H_0 : \boldsymbol{\psi} = (1/3, 55, 5)$.

Using R

In the R code below, the estimated variance matrix $\widehat{\mathbf{V}}$ is extracted from the model object returned by the `optim` optimizer, and the standard errors and confidence intervals are explicitly calculated. The `optim` function was briefly described in Section 1.4.2.

The waiting time between eruptions is variable `waiting` in data-frame `faithful`.

```
> attach(faithful)
> #Define the negative log-likelihood
> nllhood=function(theta,y) {
+   p=theta[1]; mu=theta[2]; sigma=theta[3]; nu=theta[4]; tau=theta[5]
+   lhood=p*dnorm(y,mu,sigma)+(1-p)*dnorm(y,nu,tau)
+   return(-sum(log(lhood)))
+ }

> #Use start values inferred from histogram
> Faithful.fit=optim(c(0.4,52,5,80,5),nllhood,y=waiting,hessian=T)
> MLE=Faithful.fit$par
> ObsInfo=Faithful.fit$hess
> #Calculate inverse of observed Fisher information matrix
> Vhat=solve(ObsInfo)
> Std.Errors=sqrt(diag(Vhat))
> #Output the MLEs,estimated std errors, and approx Wald 95% CIs to 4 decimal places.
> Wald.table=cbind(MLE,Std.Errors,
+   LowerBound=MLE-qnorm(0.975)*Std.Errors,UpperBound=MLE+qnorm(0.975)*Std.Errors)

> parnames=c("p","mu","sigma","nu","tau")
> rownames(Wald.table)=parnames
> round(Wald.table,4)
          MLE Std.Errors LowerBound UpperBound
p      0.3609     0.0312     0.2998     0.4220
mu    54.6145     0.6995    53.2435    55.9856
sigma  5.8698     0.5370     4.8173     6.9224
nu    80.0908     0.5046    79.1017    81.0798
tau    5.8682     0.4010     5.0822     6.6542
```

Some specific points to note include:

- The additional argument `y=waiting` in the call of the `optim` function is passed to `nllhood`.
- Since `nllhood` is the negative log-likelihood, `Faithful.fit$hess` is the negative of $\mathbf{H}(\widehat{\boldsymbol{\theta}})$, that is, it is the observed Fisher information matrix. Its inverse gives $\widehat{\mathbf{V}}$.

- The `solve` function (used with only a single matrix argument) returns the matrix inverse, and the `diag` function returns the diagonal vector.
- The call of `optim` produced some warning messages (not shown), due to being unable to evaluate `nllhood` for values of p outside of $(0, 1)$ that resulted in the likelihood being negative. This can be avoided by using `lower` and `upper` bound arguments in the `optim` call.
- It would be convenient to automate several of the steps in the above code. This convenience is provided by the `mle` function in the `stats4` package.

The MLEs obtained above certainly look very sensible with respect to the histogram of the waiting times in Figure 2.5. The waiting times were assumed to be iid, and so a quantile-quantile plot of ordered waiting times versus model quantiles will provide a strong visual assessment of model fit. However, this is not entirely straightforward because the quantiles of a binormal mixture model are a non-standard function of the model parameters. These model quantiles were found numerically (code available at www.stat.auckland.ac.nz/~millar). The quantile-quantile plot is very close to linear and indicates a good fit of the estimated model, with perhaps just some hint that there were fewer very short waiting times than might be predicted.

The additional lines of code shown below calculate the Wald test statistic and the p-value for $H_0 : (p, \mu, \sigma) = (1/3, 55, 5)$.

```
> #H0: (p,mu,sigma)=(1/3,55,5)
> psi0=c(1/3,55,5)
> psihat=MLE[1:3]
> diff=psihat-psi0
> #Wald statistic
> W=t(diff)%*%solve(Vhat[1:3,1:3])%*%diff
> W
         [,1]
[1,] 4.428055

> cat("\n p-value is",1-pchisq(W,3))

 p-value is 0.2187981
```

So, it would appear that the observed data do not provide any real evidence against the volcanologists' hypothesis.

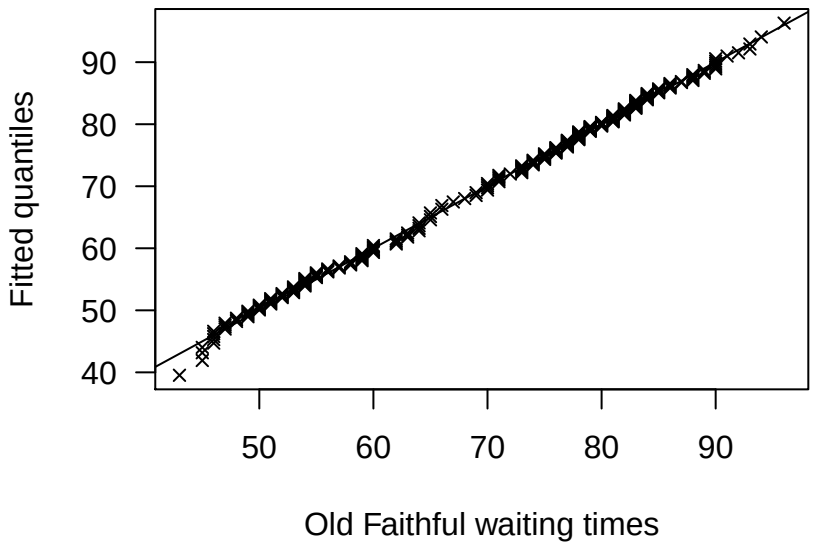


Figure 3.1: Quantile-quantile plot of Old Faithful waiting times versus model quantiles. The identity line is also shown.

Using SAS

The following SAS code assumes that dataset `OldFaithful` contains the 272 observations in the variable `waiting`. A brief description of some features of `PROC NLMIXED` was given in Section 1.4.1.

```
ODS SELECT ParameterEstimates;
PROC NLMIXED DATA=OldFaithful DF=1E6;
  PARMS p=0.5 mu=55 sigma=5 nu=80 tau=5;
  BOUNDS 0<p<1, 0<sigma, 0<tau;
  ll=log( p*PDF("NORMAL",waiting,mu,sigma)+
          (1-p)*PDF("NORMAL",waiting,nu,tau) );
  MODEL waiting ~ GENERAL(ll);
RUN;
```

The `PDF` function in the above code evaluates the form of probability density function that is specified by its first argument. Running this code produces the SAS output in Figure 3.2.

The `CONTRAST` statement of `PROC NLMIXED` uses the Wald statistic to jointly test whether specified functions of the parameters are zero. Noting that the null hypothesis $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ can be expressed

$$H_0 : \boldsymbol{\psi} - \boldsymbol{\psi}_0 = 0 \ ,$$

where $\boldsymbol{\psi} = (p, \mu, \sigma)$ and $\boldsymbol{\psi}_0 = (1/3, 55, 5)$, the previous SAS code simply requires the addition of the statement

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
p	0.3609	0.03116	1E6	11.58	<.0001	0.05	0.2998	0.4220	-4.64E-6
mu	54.6149	0.6997	1E6	78.06	<.0001	0.05	53.2435	55.9862	2.832E-7
sigma	5.8712	0.5373	1E6	10.93	<.0001	0.05	4.8181	6.9244	-4.09E-7
nu	80.0911	0.5046	1E6	158.72	<.0001	0.05	79.1021	81.0801	-7.49E-8
tau	5.8677	0.4010	1E6	14.63	<.0001	0.05	5.0819	6.6536	-3.56E-8

Figure 3.2: Parameter estimates table from using PROC NLMIXED to fit the bi-normal mixture model to the Old Faithful geyser waiting time data.

```
CONTRAST "H0" p-1/3, mu-55, sigma-5;
```

and this produces the table of output shown in Figure 3.3.

Contrasts				
Label	**Num DF**	**Den DF**	**F Value**	**Pr > F**
H0	3	1E6	1.48	0.2184

Figure 3.3: Table produced by the CONTRAST statement. The F value is W/r.

The F-value reported in Figure 3.3 is the value of W/r, and so it needs to be multiplied by r to obtain the Wald statistic W. In this example $r = 3$, and so the calculated value of W was 4.434. The p-value (shown as **Pr>F**) is obtained by comparison of the calculated F-value with an $F_{r,d}$ distribution. Here, d was explicitly set to one million using the `DF=1E6` option, and for such a large value of d $F_{r,d}$ is approximately χ^2_r/r. That is, SAS is comparing the calculated value of W/r against a χ^2_r/r distribution, or equivalently, comparing W against a χ^2_r distribution,

as desired.

Box 3.2.

PROC NLMIXED calls W/r an F-statistic because it makes an adjustment for the uncertainty in replacing $\mathbf{V}(\boldsymbol{\theta}_0)$ by $\widehat{\mathbf{V}}$ in (3.4). It is implicitly assuming that

$$\widehat{\mathbf{V}} \sim \frac{\chi^2_d}{d}\mathbf{V}(\boldsymbol{\theta}_0) \ ,$$

where d denotes the degrees of freedom (by default, the number of observations in the dataset). Substituting into (3.10) gives

$$W = (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T \widehat{\mathbf{V}}_{\boldsymbol{\psi}}^{-1} (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \dot{\sim} \frac{\chi^2_r}{\chi^2_d/d} \ .$$

Then

$$F = \frac{W}{r} \dot{\sim} \frac{\chi^2_r/r}{\chi^2_d/d} \sim F_{r,d} \dot{\sim} \frac{\chi^2_r}{r} \text{ for large } d \ .$$

3.4 Likelihood ratio tests, confidence intervals and regions

As in the previous section, it will be assumed for notational convenience that the null hypothesis has the form $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ where $\boldsymbol{\psi} = (\theta_1, ..., \theta_r)$ is the first r elements of $\boldsymbol{\theta}$. This hypothesis can alternatively be expressed as

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \ ,$$

where Θ_0 is the subset of the parameter space where the first r elements are always equal to $\boldsymbol{\psi}_0 = (\theta_{01}, ..., \theta_{0r})$.

The likelihood ratio test (LRT) statistic is simply twice the difference in the log-likelihoods between the unrestricted fit (obtained from maximization over the parameter space Θ) and the fit under H_0 (obtained from maximization over the restricted parameter space Θ_0).

It is useful to use the notation $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ where $\boldsymbol{\lambda} = (\theta_{r+1}, ..., \theta_s)$ because maximization over the restricted parameter space corresponds to maximization with respect to $\boldsymbol{\lambda}$. Then, the ML estimator over Θ_0 can be written $\widehat{\boldsymbol{\theta}}_0 = (\boldsymbol{\psi}_0, \widehat{\boldsymbol{\lambda}}_0)$ where

$\widehat{\boldsymbol{\lambda}}_0$ is obtained as

$$l(\boldsymbol{\psi}_0, \widehat{\boldsymbol{\lambda}}_0; \boldsymbol{y}) = \max_{\boldsymbol{\lambda}} \; l(\boldsymbol{\psi}_0, \boldsymbol{\lambda}; \boldsymbol{y}) \; .$$

Regarded as a function of $\boldsymbol{\psi} = \boldsymbol{\psi}_0$, the function $l^*(\boldsymbol{\psi}_0; \boldsymbol{y}) = l(\boldsymbol{\psi}_0, \widehat{\boldsymbol{\lambda}}_0; \boldsymbol{y})$ is called the profile likelihood function for $\boldsymbol{\psi}$, and is re-visited in more depth in Section 4.5.

Under appropriate regularity conditions, the LRT statistic has an approximate chi-square distribution with r degrees of freedom,

$$X = 2[l(\widehat{\boldsymbol{\theta}}; \boldsymbol{Y}) - l(\widehat{\boldsymbol{\theta}}_0; \boldsymbol{Y})] \sim \chi^2_r \; . \tag{3.12}$$

The likelihood ratio test is particular convenient under the so-called *simple* hypothesis under which $r = s$, since then $\boldsymbol{\theta}$ is fully specified under H_0. That is, $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, and Θ_0 is the zero-dimensional parameter space $\Theta_0 = \boldsymbol{\theta}_0$. Consequently, the maximized log-likelihood under H_0 is just $l(\boldsymbol{\theta}_0)$.

Large values of X constitute evidence against H_0. In particular, the $(1-\alpha)100\%$ confidence interval for a single parameter, θ_k, is the collection of values θ_{0k} such that $H_0 : \theta_k = \theta_{0k}$ is not rejected. That is, the collection of values for which

$$2[l(\widehat{\boldsymbol{\theta}}; \boldsymbol{Y}) - l(\theta_{0k}, \widehat{\boldsymbol{\lambda}}_0; \boldsymbol{Y})] < \chi^2_{1,1-\alpha} \; . \tag{3.13}$$

For 95% likelihood-ratio confidence intervals, the relevant quantile is $\chi^2_{1,0.95} \approx 3.84$. So, in plain words, the 95% confidence interval for θ_k is the collection of points θ_{0k} such that the "partially maximized" log-likelihood decreases by no more than 1.92 compared to maximization over the full parameter space. Here, "partially maximized" refers to maximization over Θ_0, that is, maximization with respect to $\theta_i, i \neq k$.

Likelihood ratio confidence regions for $\boldsymbol{\psi} \in \mathbb{R}^r$ are similarly obtained, as the collection of $\boldsymbol{\psi}_0$ values for which maximization over Θ_0 decreases the log-likelihood by no more than half of $\chi^2_{r,1-\alpha}$ compared to $l(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})$.

3.4.1 Using R and SAS: Another visit to Old Faithful

For certain classes of models, R has functions for calculating likelihood ratio confidence intervals. For example, the confidence interval function `confint` is able to calculate LR confidence intervals for some classes of models, including nonlinear

least squares and generalized linear models. However, it is not applicable to the binormal mixture model, and the general-purpose function `Plkhci` is used instead.

Calculation of likelihood ratio confidence intervals is implemented within several maximum likelihood-based SAS procedures. For example, the `MODEL` statement in `PROC GENMOD` (for generalized linear modelling) includes option `LRCI` for specifying that likelihood-ratio confidence intervals of parameter estimates are required. For general likelihoods, `PROC NLP` (within the `OR` module) contains a `PROFILE` statement for specifying the parameter of interest. For installations that do not include the `OR` module, a macro is available from www.stat.auckland.ac.nz/~millarfor use with `PROC NLMIXED`, and this is demonstrated in Section 3.4.1.

To match the examples used to demonstrate the Wald approach, here it is required to find likelihood ratio confidence intervals for each of the five parameters of the binormal mixture model of the Old Faithful waiting times, and to test the joint null hypothesis $H_0 : (p, \mu, \sigma) = (1/3, 55, 5)$.

Using R

Calculation of likelihood ratio confidence intervals can be computer intensive. For example, the 95% likelihood-ratio confidence interval for p is all values of p_0 for which the null hypothesis $H_0 : p = p_0$ is not rejected. Each such hypothesis requires a maximization over the remaining four parameters. Fortunately, the `plkhci` function (`Bhat` package) implements the efficient algorithm of (Venzon and Moolgavkar 1988). The R code below uses function `Plkhci` (with capital `P`), which is a very minor modification of `plkhci function` that permits the data $\boldsymbol{y}$ to be passed as an additional argument (see Section 15.4.1).

```
> library(Bhat)
> plist=list(label=parnames,est=MLE,
+         low=c(0,0,0,0,0),upp=c(1,100,10,100,10))

> LRCI.table=matrix(0,nrow=5,ncol=2,
+                   dimnames=list(parnames,c("LowerBound","UpperBound")))

> source("Plkhci.R") #Defines function Plkhci
> for(i in 1:5)
+   LRCI.table[i,]=Plkhci(plist,nllhood,parnames[i],prob=0.95,eps=0.000001,
+                         y=waiting)

> round(LRCI.table,4)
```

```
      LowerBound UpperBound
p         0.3013     0.4230
mu       53.2775    56.0784
sigma     4.9408     7.1116
nu       79.0478    81.0517
tau       5.1677     6.7635
```

To test $H_0 : (p, \mu, \sigma) = (1/3, 55, 5)$, function `nllhoodH0` is used to find the MLE under H_0. This function uses the fixed values of p, μ and σ, and takes $\boldsymbol{\lambda} = (\nu, \tau)$ as arguments over which `optim` performs the optimization.

```
> #H0: psi=c(1/3,55,5)
> nllhoodH0=function(lambda,y) {
+   theta=c(1/3,55,5,lambda)
+   return(nllhood(theta,y))
+ }

> Faithful.fitH0=optim(c(80,5),nllhoodH0,y=waiting)
> cat("Under H0, maximized log-likelihood is",-Faithful.fitH0$value,"\n")
Under H0, maximized log-likelihood is -1036.829

> cat("Unrestricted maximized log-likelihood is",-Faithful.fit$value,"\n")
Unrestricted maximized log-likelihood is -1034.002

> #Calculate likelihood ratio statistic
> LRStat=2*(-Faithful.fit$value+Faithful.fitH0$value)
> cat("LRT statistic is",LRStat,"and the p-value is",1-pchisq(LRStat,3),"\n")
LRT statistic is 5.655282 and the p-value is 0.1296405
```

The maximization of the log-likelihood under H_0 could have been obtained more easily using the `Profile` function, which is a utility that automatically constructs `nllhoodH0` and performs the restricted maximization (see the code for this example in Section 15.4.2).

Using SAS

The following `PROC NLP` code can be used by users with access to the SAS operations research (OR) module. A brief description of some of the features of `PROC NLP` is provided in Section 1.4.1.

```
PROC NLP DATA=OldFaithful COV=2 VARDEF=N;
MAX loglike;
PARMS p=0.5, mu=55, sigma=5, nu=80, tau=5;
BOUNDS 0<p<1, 0<sigma, 0<tau;
PROFILE p mu sigma nu tau / ALPHA=0.05;
loglike=LOG( p*PDF("NORMAL",waiting,mu,sigma)+
          (1-p)*PDF("NORMAL",waiting,nu,tau) );
RUN;
```

PROC NLP: Nonlinear Maximization

Wald and PL Confidence Limits							
N	Parameter	Estimate	Alpha	Profile Likelihood Confidence Limits		Wald Confidence Limits	
1	p	0.360886	0.050000	0.301253	0.423075	0.299804	0.421968
2	mu	54.614856	0.050000	53.277268	56.080098	53.243518	55.986194
3	sigma	5.871219	0.050000	4.941278	7.112374	4.818087	6.924352
4	nu	80.091069	0.050000	79.046875	81.051847	79.102082	81.080057
5	tau	5.867734	0.050000	5.167642	6.764319	5.081864	6.653604

Figure 3.4: Parameter estimates table from using PROC NLP to fit the binormal mixture model to the Old Faithful waiting time data. The LR confidence limits are labeled "Profile Likelihood Confidence Limits"

If `PROC NLP` is not available then `PROC NLMIXED` can be used in combination with the `Plkhci` macro. The likelihood ratio CI's are computed for each parameter separately and the following code finds the CI for parameter p only. This requires the user to specify a macro that takes p as its sole argument, and containing appropriate PROC NLMIXED code in which p is "hard-wired", with optimization over all remaining parameters. When this macro is invoked, the symbol `&p` is replaced by the value passed in argument p. The Plkhci macro is briefly described in Section 15.4.3.

```
%INCLUDE "PlkhciMacro.sas";
%MACRO OldFaithfulProfile_p(p);
PROC NLMIXED DATA=OldFaithful;
  PARMS mu=55 sigma=5 nu=80 tau=5;
  BOUNDS 0<sigma, 0<tau;
  ll=LOG( &p*PDF("NORMAL",waiting,mu,sigma)+
          (1-&p)*PDF("NORMAL",waiting,nu,tau) );
  MODEL waiting ~ GENERAL(ll);
  RUN;
%MEND;

%Plkhci(OldFaithfulProfile_p,0.0,0.3609,-1034.002,side="L");
%Plkhci(OldFaithfulProfile_p,0.3609,0.5,-1034.002,side="R");
```

The above code adds the following lines to the SAS log window.

```
Left-sided 95% LR CI bound is 0.30124929205
Right-sided 95% LR CI bound is 0.4230467286
```

For the likelihood ratio test of $H_0 : (p, \mu, \sigma) = (1/3, 55, 5)$, the maximization

under H_0 can be achieved with the following code.

```
PROC NLMIXED DATA=OldFaithful DF=1E6;
  p=1/3; mu=55; sigma=5;
  PARMS nu=80 tau=5;
  BOUNDS 0<tau;
  ll=log( p*PDF("NORMAL",waiting,mu,sigma)+
          (1-p)*PDF("NORMAL",waiting,nu,tau) );
  MODEL waiting ~ GENERAL(ll);
RUN;
```

3.5 More examples

3.5.1 Two-dimensional log-likelihood contours

This example works with the log-likelihood of the logistic regression model fitted to the binomial data in Section 7.4. There, the binomial experiment consisted of observing the number of fish (at given lengths) entering a trawl, and the proportion of those retained by it (see Table 7.1). For current purposes, it is enough to know that this is a two-parameter model, and since these parameters are regression coefficients, they will be denoted $\boldsymbol{\theta} = (\beta_0, \beta_1)$ for consistency with Section 7.4.

The MLE is $\widehat{\boldsymbol{\theta}} = (-10.632, 0.304)$ and the maximized value of the log-likelihood is $l(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})$=-37.88. A contour plot of the log-likelihood is shown in Figure 3.5. A $(1-\alpha)100\%$ LR confidence region for $\boldsymbol{\theta}$ is given by all parameter vectors (β_0^*, β_1^*) such that the null hypothesis $H_0 : (\beta_0, \beta_1) = (\beta_0^*, \beta_1^*)$ is not rejected at level α. That is, such that $l(\beta_0^*, \beta_1^*) > -37.88 - 0.5\chi^2_{2,1-\alpha}$. For a 95% confidence region, $\chi^2_{2,0.95} \approx 5.99$ and the region is given by $l(\beta_0^*, \beta_1^*) > -40.88$.

The contour plot can also be used to obtain LR confidence intervals for β_0 and β_1. For example, from (3.13), an approximate 95% CI for parameter β_0 is given by all values β_0^* such that

$$l(\beta_0^*, \widehat{\beta}_1^*) > -39.80$$

where $\widehat{\beta}_1^*$ is the value of β_1 that maximizes $l(\beta_0^*, \beta_1)$ with respect to β_1. For any β_0^* the approximate value of $l(\beta_0^*, \widehat{\beta}_1^*)$ can easily be read from Figure 3.5 by scanning along the vertical line corresponding to $\beta_0 = \beta_0^*$. If this line passes through the contour $l(\beta_0, \beta_1) = -39.80$ then β_0^* is in the 95% CI. The 95% likelihood ratio CI for parameter β_0 is seen to be approximately (-12.45,-9.05).

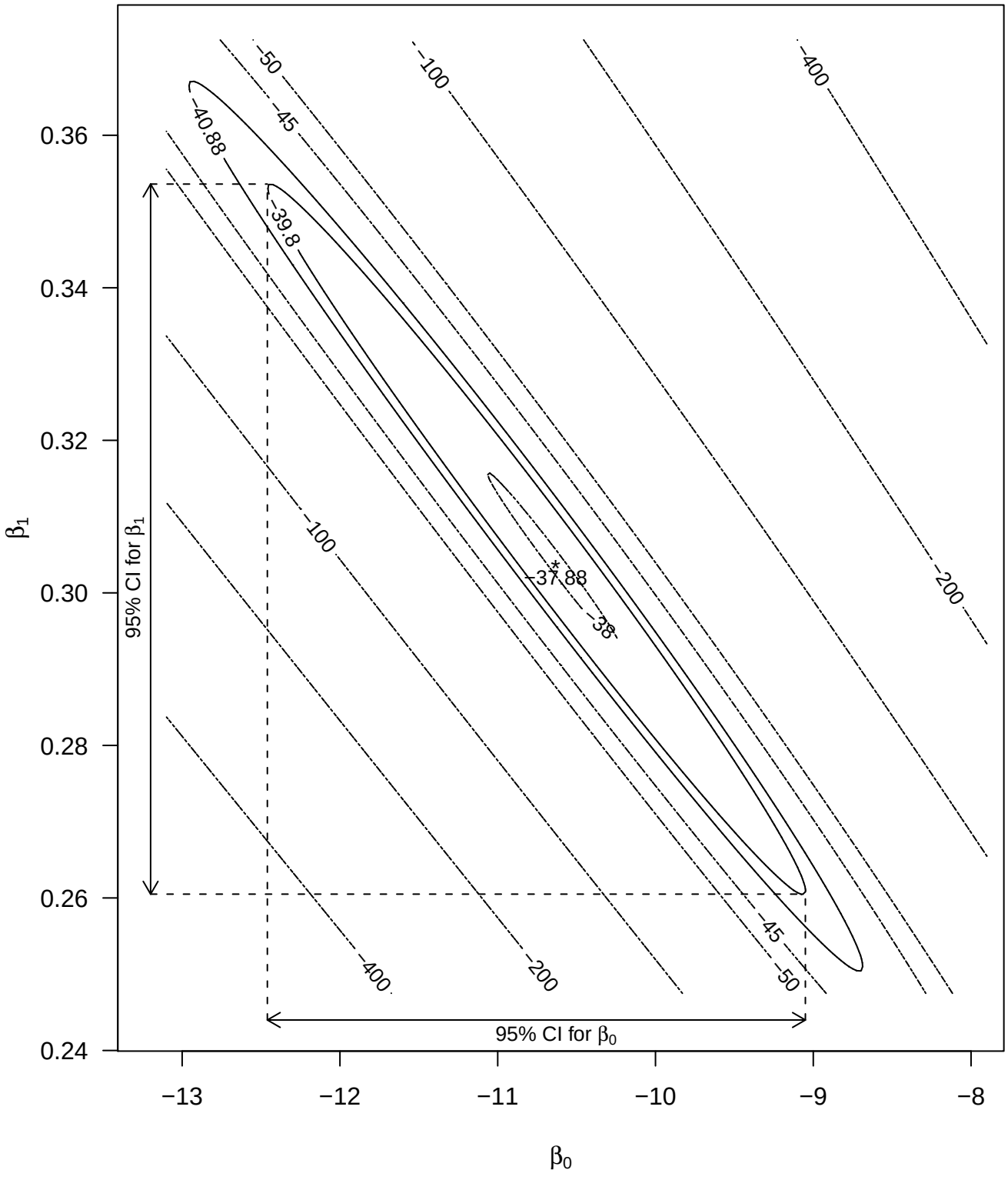


Figure 3.5: Contour plot of the two-dimensional log-likelihood $l(\beta_0, \beta_1)$. The contours corresponding to the critical values of -39.80 and -40.88 are shown using solid lines.

3.5.2 The G-test for contingency tables

The so-called G-test is the name given to the application of likelihood ratio tests to contingency tables. In general, the table could be one-way, two-way or multi-way, and the hypothesis could be a test of independence, or any other restriction (e.g., see Exercise 3.5). Here, it is not necessary to consider the layout of the table, and the observed counts will simply be denoted $n_i, i = 1, ..., m$.

The most convenient model for such data is multinomial$(n_+, p_1, ..., p_m)$ where $n_+ = \sum_i^m n_i$. The unrestricted model has parameter space that is a subset of $\mathbb{R}^{m-1}$ because of the restriction $\sum_{i=1}^m p_i = 1$. A bit of algebra confirms that the MLEs are the observed proportions $\widehat{p}_i = n_i/n_+$. The maximized log-likelihood is therefore (to within a constant)

$$l(\widehat{\boldsymbol{p}}; n_1, ..., n_m) = \sum_{i=1}^m n_i \log(\widehat{p}_i) = \sum_{i=1}^m n_i \log(n_i/n_+) \ ,$$

where $\widehat{\boldsymbol{p}} = \widehat{p}_1, ..., \widehat{p}_m$.

Let $\widehat{\boldsymbol{p}}_0 = (\widehat{p}_{01}, ..., \widehat{p}_{0m})$ denote the MLEs under H_0 and let $\widehat{n}_i = n_+\widehat{p}_{0i}$ denote the expected cell count under this fitted model. Then

$$l(\widehat{\boldsymbol{p}}_0; n_1, ..., n_m) = \sum_{i=1}^m n_i \log(\widehat{p}_{0i}) = \sum_{i=1}^m n_i \log(\widehat{n}_i/n_+).$$

The G-test statistic is the likelihood ratio statistic

$$G = 2[l(\widehat{\boldsymbol{p}}; n_1, ..., n_m) - l(\widehat{\boldsymbol{p}}_0; n_1, ..., n_m)] = 2\sum_i^m n_i \log(n_i/\widehat{n}_i)$$

with degrees of freedom given by the number of parameters restricted under H_0. □

Example 3.1. G-test of the Poisson model. In a study of micro-propagation, Marin, Jones and Hadlow (1993) measured the number of roots produced on shoots of an apple cultivar. For one particular treatment, the following data were recorded from 40 shoots.

Number	0	1	2	3	4	5	6	7	8	9
Frequency	19	2	2	4	3	1	4	3	0	2

It is desired to test whether a Poisson model is a reasonable model for these data. One approach is to treat the frequencies of counts as multinomial data[2]. To reduce sparseness (i.e., low expected cell count), the last cell of the multinomial will correspond to counts of 8 or more, giving

Number	0	1	2	3	4	5	6	7	8+
Frequency	19	2	2	4	3	1	4	3	2
Expected	3.45	8.46	10.36	8.46	5.18	2.54	1.04	0.36	0.15

Here, the expected counts are obtained from the ML fit under the null hypothesis that an iid Poisson(λ) model is correct for the number of roots. The MLE is the sample mean of the 40 observations, $\widehat{\lambda} = 2.45$.

The unrestricted multinomial model has 8 parameters, whereas the restricted multinomial (obtained from the iid Poisson model) has just one, and so the G statistic is compared against a χ^2_7 distribution. The observed $G \approx 75.2$ is enormously extreme for a χ^2_7 distribution, and the Poisson model is well and truly rejected.

Here, the first cell of this multinomial is the frequency of cultivar shoots that produced no roots, and it is clear that there are many more of these than expected under the Poisson model. It would therefore be natural to explore a zero-inflated model for these data. This is implemented in Exercise 3.7. □

3.6 Exercises

3.1 In the example of Section 3.5.1 the MLE was $(\widehat{\beta}_0, \widehat{\beta}_1) = (-10.632, 0.304)$ with estimated variance matrix

$$\widehat{\mathbf{V}} = \begin{pmatrix} 0.7477 & -0.02022 \\ -0.02022 & 0.0005584 \end{pmatrix}.$$

1. Test $H_0 : (\widehat{\beta}_0, \widehat{\beta}_1) = (-10, 0.3)$ with $\alpha = 0.05$ using the Wald test.
2. Calculate an approximate 95% Wald CI for parameter β_0.
3. Calculate an approximate 99% likelihood ratio CI for parameter β_1.

3.2 For the Old Faithful data, use R or SAS to perform Wald and LR tests of the simple hypothesis $H_0 : (p, \mu, \sigma, \nu, \tau) = (1/3, 55, 5, 80, 5)$.

3.3 Proschan (1963) fitted an exponential distribution to the times between failures of air-conditioning systems on Boeing 720 aircraft. The twelve recorded failure times for aircraft number 8044 were

[2]See the continuation of this example in Section 7.5.2 for an alternative approach.

487 18 100 7 98 5 85 91 43 230 3 130

1. Assuming these data are IID Exp(μ), show that $\widehat{\mu} = \overline{y}$.
2. Calculate the 95% Wald confidence interval for μ.
3. Draw a plot of the log-likelihood (use integer values of μ from 45 to 250, say), and add a horizontal line showing the threshold value for determining the 95% LR confidence interval for μ.
4. Determine the 95% LR confidence interval using R or SAS.

3.4 Consider a 2×2 contingency table with cell counts denoted n_{ij} and assume that the counts are multinomial with distribution Mult$(n, p_{11}, p_{12}, p_{21}, p_{22})$ where n is the total count and p_{ij} is the probability of the cell in row i and column j. Let $r_1 = p_{11} + p_{12}$ and $c_1 = p_{11} + p_{21}$ denote the row and column probabilities, respectively.

1. Under the null hypothesis that rows and columns are independent, write down the multinomial log-likelihood as a function of r_1 and c_1.
2. Show that the MLEs of r_1 and c_1 are simply the proportions of the counts in row 1 and column 1, respectively.
3. It follows that $\widehat{p}_{ij} = \frac{n_{i1}+n_{i2}}{n}\frac{n_{1j}+n_{2j}}{n}$. Hence, implement a G-test of independence for the contingency table

	Outcome	
	Good	Bad
Treatment 1	54	36
Treatment 2	46	44

3.5 Under Hardy-Weinberg equilibrium, alleles A and B occur independently and thus the probabilities of genotypes AA, AB and BB are $p^2, 2p(1-p)$ and $(1-p)^2$, respectively, where p is the probability of allele type A. McDonald, Verrelli and Geyer (1996) recorded genotype frequencies of 14 AA, 21 AB and 25 BB genotypes from 60 American oyster. Perform a G-test of Hardy-Weinberg equilibrium for these data.

3.6 The data used in the example in Section 3.5.1 are binomial data from 37 different lengthclasses, l, of fish. The data are given in Table 7.1, where for each l, n_l is the number of fish entering a trawl, and y_l is the number that were retained. The logistic regression model that is fitted to these data assumes that $Y_l \sim \text{Bin}(n_l, p_l)$ are independent. The case study in Section 7.4 parameterizes p_l using the conventional generalized linear model form

$$p_l = \frac{\exp(\beta_0 + \beta_1 l)}{1 + \exp(\beta_0 + \beta_1 l)}$$

where β_0 and β_1 are the two parameters to be estimated. However, here, p_l is to be parameterized using parameters β_1 and l_{50} in the form

$$p_l = \frac{\exp(\beta_1(l - l_{50}))}{1 + \exp(\beta_1(l - l_{50}))} .$$

Note that this is an example of an inverse prediction problem, and parameter l_{50} is so denoted because it corresponds to the value of the covariate l that predicts $p_l = 0.5$,

1. Find the MLE $(\widehat{\beta}_1, \widehat{l}_{50})$ by maximizing the likelihood using the `optim` function in R or the `NLMIXED` or `NLP` procedures in SAS.
2. If using R, add the calculations required to give the approximate standard errors of $(\widehat{\beta}_1, \widehat{l}_{50})$.
3. Use R or SAS to find approximate 95% likelihood ratio CI's for β_1 and l_{50}. (Note: If using R, this can be done using the `Plkhci` function. With SAS procedure `NLMIXED` it can be done in conjunction with the `Plkhci` macro in SAS, or with procedure `NLP` it simply requires inclusion of an appropriate `PROFILE` statement.) The CI for β_1 can be verified immediately from Figure 3.5.
4. Verify the approximate 95% likelihood ratio CI for l_{50} using only Figure 3.5.

3.7 In Example 3.1 a G-test was used to show that a Poisson model was not appropriate for the apple micro-propagation data. It was seen that the data contained many more zeros than expected under the Poisson model, and hence it would be natural to consider a zero-inflated Poisson (ZIP) model for these data. The density of the ZIP model is given in Exercise 2.11 (there it was assumed that λ was known, but here both p and λ are to be estimated).

1. Verify that $(\widehat{p}, \widehat{\lambda}) = (0.46984.6216)$ by maximizing the likelihood using the `optim` function in R or the `NLMIXED` or `NLP` procedures in SAS.
2. If using R, add the calculations required to give the approximate standard errors of $(\widehat{p}, \widehat{\lambda})$.
3. Find approximate 95% likelihood ratio CI's for p and λ. (See the note in Question 3.6, part 3.)
4. Verify that the G test statistic for the ZIP model is 4.97, and determine the p-value for the hypothesis that these data are iid ZIP.

(Note: To check your fit of the ZIP model, this model can be fitted using SAS procedure `GENMOD` or the R package `VGAM`).

3.8 Let $Y_1, ..., Y_n$ follow a first-order auto-regressive model (AR1) of the form

$$(Y_{i+1} - \mu) = \rho(Y_i - \mu) + \epsilon_{i+1} \ , \quad i = 1, ..., n-1 \ ,$$

where $\epsilon_i \sim N(0, \sigma^2)$ are iid and $-1 < \rho < 1$. Here $\boldsymbol{\theta} = (\mu, \rho, \sigma)$. It follows that the distribution of Y_{i+1} given $y_1, ...y_i$ is

$$Y_{i+1} \mid y_1, ...y_i \sim N(\mu + \rho(y_i - \mu), \sigma^2) \ , \quad i = 1, ..., n-1 \ . \tag{3.14}$$

Also, it can be shown that

$$Y_1 \sim N\left(\mu, \frac{\sigma^2}{1-\rho^2}\right) \ . \tag{3.15}$$

Since

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = f(y_1; \boldsymbol{\theta}) \prod_{i=1}^{n-1} f(y_{i+1} | y_1, ..., y_i; \boldsymbol{\theta}) \ , \tag{3.16}$$

if follows that the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$ is given by the product of the likelihoods arising from the $n-1$ terms in (3.14), and the term in (3.15).

One hundred observations from an AR1 model are contained in dataset `AR1.dat`.

1. Plot the data.
2. Maximize the log-likelihood using using the `optim` function in R or the `NLMIXED` or `NLP` procedures in SAS.
3. Verify your fitted model using the `arima0` function in R, or `PROC ARIMA` in SAS.
4. Perform a likelihood ratio test of the null hypothesis that the observations are iid. That is, test $H_0 : \rho = 0$.

Chapter 4

What you really need to know

Make everything as simple as possible, but no simpler — Albert Einstein

4.1 Introduction

This Chapter delves into some of the important issues that are encountered in practice, beginning with inference about functions of the parameters in Section 4.2 where the delta method is used for determining the approximating normal distribution of functions of the MLE. Section 4.3 presents relevant examples of application of the delta method, including the derivation of the approximate variance of a product of MLEs, and the approximate variance of the log odds-ratio in a 2 by 2 contingency table. Section 4.4 compares the Wald and likelihood ratio approaches, and comes to the conclusion that inference based on the likelihood ratio is generally more reliable. Section 4.5 develops profile likelihood as a tool for reducing the dimensionality of the likelihood function (see Section 9.5 for alternative methodology). Model selection criterion are briefly encountered in Section 4.6. Section 4.7 looks at the bootstrap as a computational-intensive method that can be useful in data-poor and non-standard situations. The bootstrap also provides one of several practical methodologies for prediction (Section 4.8). Finally, this Chapter concludes with a brief look at a variety of situations under which ML inference can go bad.

4.2 Inference about $g(\boldsymbol{\theta})$

In practice, the research question may require inference about quantities that are a function of $\boldsymbol{\theta}$. The function $g(\boldsymbol{\theta})$ could also involve covariates, for example, $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}, \boldsymbol{x}_i)$ could be the fitted value for observation i with covariate vector $\boldsymbol{x}_i$.

In some cases, inference about $\zeta = g(\boldsymbol{\theta})$ can be made by re-parameterizing the model such that ζ becomes a parameter under the re-parameterization. For example, in the logistic regression example in the case study of Section 7.4 the unknown parameter vector is $\boldsymbol{\theta} = (\beta_0, \beta_1)$. However, β_0 and β_1 are themselves of little interest to the researcher, rather it is the value of $\zeta = -\beta_0/\beta_1$ that is relevant. This model could be re-parameterized using $\boldsymbol{\theta}^* = (\zeta, \beta_1)$, say, since there is a one-to-one correspondence between values of $\boldsymbol{\theta}$ and values of $\boldsymbol{\theta}^*$. However, one major difficulty is that re-parameterizations may prevent the use of convenient software. In the logistic regression case, the model parameterized using $\boldsymbol{\theta}^*$ can no longer be expressed as a generalized linear model and can not be fitted using the software employed in Section 7.4.

The delta method provides a shortcut for inference about $g(\boldsymbol{\theta})$ that is based on the approximate normality of MLEs. The maximum likelihood estimator of $g(\boldsymbol{\theta})$ is $g(\widehat{\boldsymbol{\theta}})$, and the delta method derives the approximate distribution of $g(\widehat{\boldsymbol{\theta}})$ from the approximate distribution of $\widehat{\boldsymbol{\theta}}$. The delta method can be considered the practical application of the delta theorem seen in the Part III of this text.

4.2.1 The delta method

The delta method is not particular to MLEs, and here it is derived for an arbitrary random variable X. In Section 4.2.2, the role of X is taken by the MLE $\widehat{\theta}$.

One-dimensional case

Suppose that the scalar random variable X has mean μ and variance σ^2, and let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable function. The delta method provides approximations for the mean and variance of of the random variable $g(X)$ and is obtained from the

first-order Taylor series approximation of $g(X)$ around μ,

$$g(X) \approx g(\mu) + g^{'}(\mu)(X - \mu) \ . \tag{4.1}$$

Taking expectation and variance of the right-side of (4.1), it follows that

$$E[g(X)] \approx g(\mu)$$

and

$$\text{var}(g(X)) \approx (g^{'}(\mu))^2\sigma^2 \ .$$

These approximations will be reasonably accurate if g is close to linear at μ and X is sufficiently concentrated in the neighbourhood of μ. Example 4.4 presents a contrived example where this is not the case. In practice, the delta method should always be treated with a bit of circumspection. Note also that itcould be the case that $g(X)$ does not possess a mean and variance at all.

Example 4.1. Variance stabilizing transformation for Poisson data. Let Y be Poisson distributed with mean and variance λ, and consider the random variable $g(Y) = Y^{1/2}$. The square-root function has derivative $g^{'}(y) = y^{-1/2}/2$, and so

$$\text{var}\left(Y^{1/2}\right) \approx (g^{'}(\lambda))^2\text{var}(Y) = \frac{1}{4} \ ,$$

which does not depend on λ. □

Box 4.1.

Historically, count data have been modeled by applying linear regression models to the square-rooted counts, with the justification that the transformed data have (approximately) homogeneous variance that does not depend on the expected value. This is just an attempt to shoe-horn data into an unnatural form so as to apply an inappropriate, albeit familiar, model and it will be unclear how to interpret the regression parameters except in the simplest of analyses. With few exceptions, count data should be analyzed using the techniques in Chapter 7.

Multi-dimensional case

In the multi-dimensional case, suppose that the random vector $\boldsymbol{X} \in \mathbb{R}^s$ has mean vector $\boldsymbol{\mu}$ and $s \times s$ variance matrix Σ. The function $g : \mathbb{R}^s \to \mathbb{R}^p$ can be denoted

$$g(\boldsymbol{x}) = \begin{pmatrix} g_1(\boldsymbol{x})^T \\ . \\ . \\ . \\ g_p(\boldsymbol{x})^T \end{pmatrix}$$

where each function $g_i : \mathbb{R}^s \to \mathbb{R}$ is assumed to be differentiable with respect to each $x_i, i = 1, ..., s$. The $p \times s$ Jacobian (derivative) matrix of g is

$$G(\boldsymbol{x}) = \begin{pmatrix} g_1'(\boldsymbol{x})^T \\ . \\ . \\ . \\ g_p'(\boldsymbol{x})^T \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}, ..., \frac{\partial g_1}{\partial x_s} \\ . \\ . \\ . \\ \frac{\partial g_p}{\partial x_1}, ..., \frac{\partial g_p}{\partial x_s} \end{pmatrix}$$

and the first order Taylor series approximation of $g(\mathbf{X})$ around $\boldsymbol{\mu}$ is

$$g(\mathbf{X}) \approx g(\boldsymbol{\mu}) + G(\boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu}) \ . \tag{4.2}$$

It follows that,

$$E[g(\boldsymbol{X})] \approx g(\boldsymbol{\mu})$$

and

$$\text{var}(g(\boldsymbol{X})) \approx G(\boldsymbol{\mu})\Sigma G(\boldsymbol{\mu})^T \ .$$

Example 4.2. Let Y_1 and Y_2 be independently distributed Poisson$(\lambda_i), i = 1, 2$, respectively. Consider the random vector

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} = \begin{pmatrix} g_1(Y_1, Y_2) \\ g_2(Y_1, Y_2) \end{pmatrix} = \begin{pmatrix} \frac{Y_1}{Y_1+Y_2} \\ Y_1 + Y_2 \end{pmatrix} \ .$$

The Jacobian of this transformation is

$$G(\boldsymbol{y}) = \begin{pmatrix} \frac{y_2}{(y_1+y_2)^2} & \frac{-y_1}{(y_1+y_2)^2} \\ 1 & 1 \end{pmatrix}$$

and so, denoting $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, the approximate variance of (T_1, T_2) is

$$\begin{aligned} \text{var}\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} &\approx G(\boldsymbol{\lambda}) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} G(\boldsymbol{\lambda})^T \\ &= \begin{pmatrix} \frac{\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^3} & 0 \\ 0 & \lambda_1+\lambda_2 \end{pmatrix} = \begin{pmatrix} \frac{p(1-p)}{n} & 0 \\ 0 & n \end{pmatrix} \end{aligned}$$

where $n = \lambda_1 + \lambda_2$ and $p = \lambda_1/n$. □

4.2.2 The delta method applied to MLEs.

In Section 3.2 we used the notation $\widehat{\boldsymbol{\theta}} \sim N_s(\boldsymbol{\theta}_0, \mathbf{V}(\boldsymbol{\theta}_0))$ to make explicit that, subject to regularity conditions and sufficiently large sample size, the maximum likelihood estimator has a distribution that is well approximated by a normal distribution with mean $\boldsymbol{\theta}_0$ and variance depending on $\boldsymbol{\theta}_0$. The delta theorem can immediately be applied using this approximating mean and variance. Analogous to (4.2), the first order Taylor series approximation of $g : \mathbb{R}^s \to \mathbb{R}^p$ around $\boldsymbol{\theta}_0$ is

$$g(\widehat{\boldsymbol{\theta}}) \approx g(\boldsymbol{\theta}_0) + G(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \ . \tag{4.3}$$

Since normality is preserved by linear transformations, it follows that if $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^s$ is approximately normally distributed then so to is $g(\widehat{\theta}) \in \mathbb{R}^p$, provided that the higher order terms in (4.3) can be ignored. The approximate mean and variance matrix of $g(\widehat{\boldsymbol{\theta}})$ are given by the delta method, giving

$$g(\widehat{\boldsymbol{\theta}}) \sim N_p(g(\boldsymbol{\theta}_0), G(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)G(\boldsymbol{\theta}_0)^T) \ . \tag{4.4}$$

This can be regarded as the approximate version of the asymptotic result establish in (12.51) in Part 3 of this text. Replacing $\mathbf{V}(\boldsymbol{\theta}_0)$ by its estimate $\widehat{\mathbf{V}}$, and approximating $G(\boldsymbol{\theta}_0)$ by $G(\widehat{\boldsymbol{\theta}})$, gives a more pragmatic statement for construction of Wald tests and confidence interval/regions,

$$g(\widehat{\boldsymbol{\theta}}) \dot{\sim} N_p(g(\boldsymbol{\theta}_0), G(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{V}}G(\widehat{\boldsymbol{\theta}})^T) \ . \tag{4.5}$$

One of the caveats associated with approximate normality of MLEs (and functions of MLEs) is that the sample size must be sufficiently large. But, just how

large is *sufficiently* large? The answer to that question may differ dramatically for $\widehat{\boldsymbol{\theta}}$ and $g(\widehat{\boldsymbol{\theta}})$. For the sample size that was used, it could be that $\widehat{\boldsymbol{\theta}}$ has distribution that is very close to normal, but that $g(\widehat{\boldsymbol{\theta}})$ does not, or vice-versa. Thus, if inference based on normal approximation is to be conducted, it is of particular relevance to choose a sensible parameterization of the model. Section 4.4.1 contains some guidance regarding this issue.

Example 4.3. Let Y be distributed $\text{Bin}(n, p_0)$ and consider estimation of $\zeta = g(p) = \text{logit}(p) = \log(\frac{p}{1-p})$. For sufficiently large n, $\widehat{p} = y/n$ is approximately normal with mean p and variance $p(1-p)/n$. The derivative of the logit function is

$$\frac{\partial g(p)}{\partial p} = \frac{1}{p(1-p)}$$

and hence from (4.4), $\widehat{\zeta} = \text{logit}(\widehat{p})$ has (for sufficiently large n) a distribution close to that of a normal with mean $\text{logit}(p_0)$ and variance $1/(np_0(1-p_0))$. □

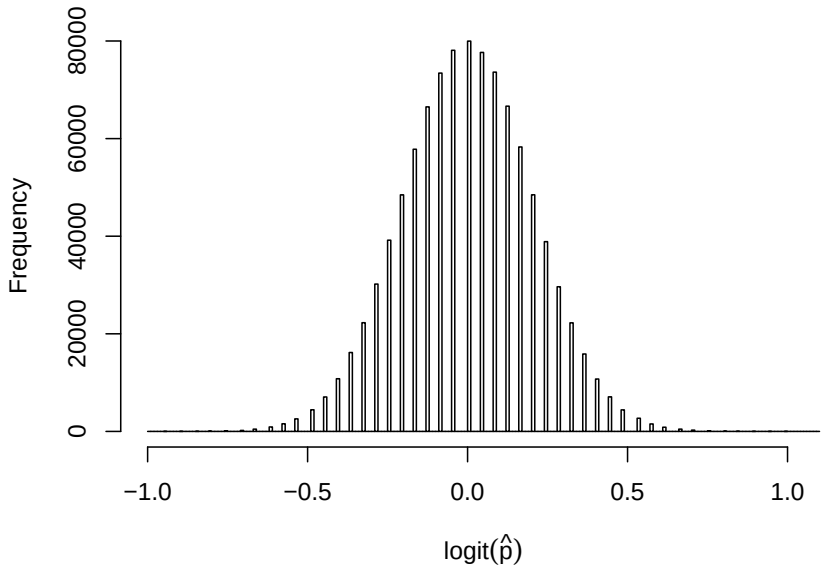


Figure 4.1: Histogram of the logit of the observed proportion from one million Binomial(100,0.5) experiment.

Figure 4.1 shows a histogram of one million values of $\text{logit}(\widehat{p})$ where Y is distributed Binomial(100,0.5). From the previous example, the approximating normal distribution has mean $\text{logit}(0.5) = 0$ and variance $1/25$, that is, standard deviation of 0.2. This approximation looks good here. However, $\text{logit}(\widehat{p})$ does not possess a mean or variance because it is undefined when $Y = 0$ or 100, which has the near-infinitesimal probability 1.6×10^{-30}. (Strictly speaking, $\text{logit}(\widehat{p})$ is not a random variable at all!)

Example 4.3 ctd. In practice, p_0 is not known and the pragmatic version of the delta method in (4.5) is used. For the motivating Example 1.1, $y = 10$ was observed from a $\text{Bin}(100, p_0)$ experiment, resulting in

$$\widehat{\zeta} = \text{logit}(\widehat{p}) \sim N(\text{logit}(p_0), 1/(n\widehat{p}(1-\widehat{p}))) = N(\text{logit}(p_0), \frac{1}{9}) \ .$$

An approximate Wald 95% confidence interval for $\text{logit}(p_0)$ is therefore

$$\text{logit}(\widehat{p}) \pm z_{0.025} \times \frac{1}{3} \approx (-2.851, -1.544) \ .$$

This confidence interval can be inverted into a confidence interval for p using the relationship $p = e^{\zeta}/(1+e^{\zeta})$, resulting in the interval $(0.055, 0.176)$. By comparison, in Chapter 1 the Wald interval calculated using the original p parameterization was (0.041,0.159) and the likelihood ratio interval was (0.051,0.169). □

4.2.3 The delta method in R, SAS and ADMB

Using R

The `msm` package includes a function `deltamethod` for general application of the delta method to simple algebraic functions of $\boldsymbol{\theta}$. The following code implements the delta method for Example 4.3 and returns the approximate standard deviation of $\text{logit}(\widehat{p})$.

```
> library(msm)
> phat=0.1 #MLE
> var.phat=0.0009 #Estimated variance of phat
> deltamethod(~log(x1/(1-x1)),mean=phat,cov=var.phat)
[1] 0.3333333
```

- The first argument to `deltamethod` is the function $g : \mathbb{R}^s \to \mathbb{R}^p$, specified as a formula (or list of formulae if $p > 1$) using parameter names $x_1, ...x_s$.
- The next two arguments are $\widehat{\boldsymbol{\theta}}$ and its estimated variance $\widehat{\mathbf{V}}$.
- `deltamethod` employs the symbolic differentiation capabilities of R (specifically, the `deriv` function) to calculate the Jacobian $G(\widehat{\boldsymbol{\theta}})$.
- `deltamethod` returns the approximate standard errors of $g(\widehat{\boldsymbol{\theta}})$ (the square-root of the diagonal of $G(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{V}}G(\widehat{\boldsymbol{\theta}})^T$) by default. The option `ses=FALSE` instead returns $G(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{V}}G(\widehat{\boldsymbol{\theta}})^T$.

Using SAS

The `NLMIXED` procedure has an `ESTIMATE` statement that is used to specify $g : \mathbb{R}^s \to \mathbb{R}^p$. The SAS code below has two changes from that seen in the binomial model in Example 4.3. For convenience, it uses the `BINOMIAL` model specification, thereby avoiding calculation of the log-likelihood as required than using the `GENERAL` form, and it uses the `ESTIMATE` statement to obtain the standard error of $\text{logit}(\widehat{p})$ from the delta method (Fig. 4.2).

```
DATA binomial;
  y=10; n=100;

PROC NLMIXED DF=1E6 DATA=binomial;
  PARMS p=0.5;
  BOUNDS 0<p<1;
  MODEL y~BINOMIAL(n,p);
  ESTIMATE "Logit(p)" log(p/(1-p));
RUN;
```

Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Logit(p)	−2.1972	0.3333	1E6	−6.59	<.0001	0.05	−2.8505	−1.5439

Figure 4.2: The Additional Estimates table from `PROC NLMIXED` showing the standard deviation of $\text{logit}(\widehat{p})$.

Alternatively, the `DeltaMethod` macro (available from www.stat.auckland.ac.nz/~millar) is provided for general implementation of the delta method, for $g : \mathbb{R}^s \to \mathbb{R}^p$ for $p \leq s \leq 2$.

```
%INCLUDE "DeltaMethodMacro.sas";
%DeltaMethod(expr1=log(x1/(1-x1)),mu1=0.1,var1=0.0009);
```

This produces a SAS table identical to Figure 4.2.

Using ADMB

The delta method is implicitly handled under automatic differentiation. In ADMB, the `sdreport` variable type is used to specify $g(\boldsymbol{\theta})$. Example 4.3 can be implemented in ADMB via addition of

```
sdreport_number logitp
```

in the `PARAMETER_SECTION`, and

```
logitp=log(p/(1-p));
```

in the `PROCEDURE_SECTION`.

4.3 Delta method examples

4.3.1 Example 1: Variance of a product.

Consider $g(\boldsymbol{\theta}) = \theta_i\theta_j, i \neq j$. This function maps from $\mathbb{R}^s$ to $\mathbb{R}$ and so its derivative, G, is a row vector of length s, with ith element equal to θ_j, jth element equal to θ_i, and all other elements equal to zero. The approximate variance of $g(\widehat{\boldsymbol{\theta}}) = \widehat{\theta}_i\widehat{\theta}_j$ is therefore

$$\begin{aligned} \widehat{\text{var}}(\widehat{\theta}_i\widehat{\theta}_j) &= G(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{V}}G(\widehat{\boldsymbol{\theta}})^T \\ &= \widehat{\theta}_j^2\widehat{\text{var}}(\widehat{\theta}_i) + 2\widehat{\theta}_i\widehat{\theta}_j\widehat{\text{cov}}(\widehat{\theta}_i, \widehat{\theta}_j) + \widehat{\theta}_i^2\widehat{\text{var}}(\widehat{\theta}_j) \end{aligned} \tag{4.6}$$

Formula (4.6) is well known and widely applied. One application is provided by Gribben, Helson and Millar (2004) in an investigation of the biomass of the geoduck *Panopea zelandica* (Fig. 4.3) at several bays in the North Island of New Zealand.

Figure 4.3: New Zealand geoduck *Panopea zelandica*

The experiment proceeded in two independent stages, the first using SCUBA to estimate the number of geoduck in the bay, and the second weighed a random sample to estimate their mean weight. In Kennedy Bay, the number of geoducks was estimated to be 22976 ($=\widehat{\theta}_1$) with approximate standard error of 4007, and the mean weight was estimated to be 242.2 ($=\widehat{\theta}_2$) g with approximate standard error of 8.6 g. The estimate of geoduck in Kennedy Bay was $22976 \times 242.2 \approx 5.6 \times 10^6 g$, that is, 5.6 tonnes. The two stages of the experiment were independent and hence $\mathrm{cov}(\widehat{\theta}_1, \widehat{\theta}_2) = 0$, and the approximate variance of the estimated biomass is therefore

$$\widehat{\theta}_j^2 \widehat{\mathrm{var}}(\widehat{\theta}_i) + \widehat{\theta}_i^2 \widehat{\mathrm{var}}(\widehat{\theta}_j) = 22976^2 \times 8.6^2 + 242.2^2 \times 4007^2 \approx 9.8 \times 10^{11}$$

and the approximate standard error is 9.9×10^5 g, that is, just under one tonne.

The above calculation is performed by the R code

```
>deltamethod(~x1*x2,mean=c(22976,242.2),cov=diag(c(4007^2,8.6^2)))
```

and the SAS code

```
%DeltaMethod(expr1=x1*x2,mu1=22976,var1=4007**2,
             mu2=242.2,var2=8.6**2,cov=0);
```

Box 4.2.

It can be shown (Goodman (1960) and Exercise 4.5) that the exact variance of the product $\widehat{\theta}_i\widehat{\theta}_j$ of two independent estimators is

$$\text{var}(\widehat{\theta}_1\widehat{\theta}_2) = E[\widehat{\theta}_1]^2\text{var}(\widehat{\theta}_2) + E[\widehat{\theta}_2]^2\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_1)\text{var}(\widehat{\theta}_2) \ . \tag{4.7}$$

Furthermore, if $\widehat{\text{var}}(\widehat{\theta}_1)$ and $\widehat{\text{var}}(\widehat{\theta}_2)$ are unbiased estimators of the true variances $\text{var}(\widehat{\theta}_1)$ and $\text{var}(\widehat{\theta}_2)$ then

$$\widehat{\text{var}}(\widehat{\theta}_1\widehat{\theta}_2) = \widehat{\theta}_1^2\widehat{\text{var}}(\widehat{\theta}_2) + \widehat{\theta}_2^2\widehat{\text{var}}(\widehat{\theta}_1) - \widehat{\text{var}}(\widehat{\theta}_1)\widehat{\text{var}}(\widehat{\theta}_2) \ . \tag{4.8}$$

is an unbiased estimator of $\text{var}(\widehat{\theta}_1\widehat{\theta}_2)$.
However, it is *not* the case that (4.8) will necessarily be a better estimator of $\text{var}(\widehat{\theta}_i\widehat{\theta}_j)$ than (4.6). The value of (4.8) can be negative and hence this estimator is not truly unbiased because negative estimators of variance are not permissable. Even ignoring this, in the present example the variance estimator of the number of geoducks was not unbiased and hence neither would be (4.8). In this example the difference between (4.6) and (4.8) is negligible and, to three significant figures, both result in an estimated standard error on the estimated biomass of $9.90 \times 10^5 g$.

4.3.2 Example 2: Vector transformation

In the context of the binormal mixture model for the Old Faithful geyser waiting time data (Section 3.3.3), suppose that it is of interest to jointly consider the ratio of the mean parameters, μ/ν, and ratio of the standard deviations, σ/τ. That is, $g(\boldsymbol{\theta}) = (\mu/\nu, \sigma/\tau)$.

The approximate variance of $g(\widehat{\boldsymbol{\theta}})$ is obtained by the R code

```
> deltamethod(list(~x2/x4,~x3/x5),MLE,Vhat,ses=F)
              [,1]          [,2]
[1,] 7.684293e-05 0.0002495297
[2,] 2.495297e-04 0.0167564703
```

That is,

$$\begin{pmatrix} \widehat{\mu}/\widehat{\nu} \\ \widehat{\sigma}/\widehat{\tau} \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_0/\nu_0 \\ \sigma_0/\tau_0 \end{pmatrix}, \begin{pmatrix} 7.684 \times 10^{-5} & 2.495 \times 10^{-4} \\ 2.495 \times 10^{-4} & 0.01676 \end{pmatrix} \right)$$

The SAS code in Section 3.3.3 requires the two additional statements

```
ESTIMATE "Ratio of means" mu/nu;
ESTIMATE "Ratio of ses" sigma/tau;
```

and it is also necessary to add the procedure option `ECOV` to the `PROC NLMIXED` statement so that SAS will produce the variance matrix of $g(\widehat{\boldsymbol{\theta}})$ rather than just the individual standard errors.

4.3.3 Example 3: Variance of log odds-ratio

Consider the two-by-two contingency table

	Col. 1	Col. 2	
Row 1	n_{11}	n_{12}	n_{1+}
Row 2	n_{21}	n_{22}	n_{2+}

Let p_i denote the probability of the column 1 event, given that the observation is in row $i, i = 1, 2$. Within row i, the odds of column 1 is defined to be $p_i/(1-p_i)$, and the odds-ratio is the ratio of the row 1 and row 2 odds. That is,

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \; .$$

The odds-ratio is commonly used to quantify the magnitude of any association between the row and column events, with an odds-ratio of unity corresponding to no association. The MLE of the odds ratio is given by replacing p_i by its MLE $\widehat{p}_i = n_{i1}/n_{i+}$, giving

$$\widehat{OR} = \frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \; .$$

Inference about the odds-ratio is usually made on the log scale, that is, by working with the log odds-ratio. The log odds ratio can be written as the difference in log odds of the two rows

$$\log(\widehat{OR}) = \log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right) - \log\left(\frac{\widehat{p}_2}{1-\widehat{p}_2}\right) \tag{4.9}$$

where the two log odds terms on the right-hand side of (4.9) are independent. Using the result from Example 4.3

$$\begin{aligned}
\widehat{\text{var}}(\log(\widehat{OR})) &= \widehat{\text{var}}\left[\log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right)\right] + \widehat{\text{var}}\left[\log\left(\frac{\widehat{p}_2}{1-\widehat{p}_2}\right)\right] \\
&= \frac{1}{n_{1+}\widehat{p}_1(1-\widehat{p}_1)} + \frac{1}{n_{2+}\widehat{p}_2(1-\widehat{p}_2)} \\
&= \frac{n_{1+}}{n_{11}n_{12}} + \frac{n_{2+}}{n_{21}n_{22}} \\
&= \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}
\end{aligned} \tag{4.10}$$

It is interesting to note that the above variance formula can be obtained more directly by using the equivalence between binomial (or multinomial) and Poisson models (see Exercise 14.6 and Section 9.1). In particular, for making inference about the log odds-ratio, one can model the cell counts, n_{ij} as independent realizations from a Poisson(λ_{ij}) distribution. Then,

$$\begin{aligned}\widehat{\text{var}}(\log(\widehat{OR})) &= \widehat{\text{var}}(\log(n_{11}) + \log(n_{12}) + \log(n_{21}) + \log(n_{22})) \\ &= \widehat{\text{var}}(\log(n_{11})) + \widehat{\text{var}}(\log(n_{12})) + \widehat{\text{var}}(\log(n_{21})) + \widehat{\text{var}}(\log(n_{22}))\ .\end{aligned}$$

where, from the delta method,

$$\widehat{\text{var}}(\log(n_{ij})) = \left(\frac{\partial \log(n_{ij})}{\partial n_{ij}}\right)^2 \widehat{\text{var}}(n_{ij}) = \frac{1}{n_{ij}^2} n_{ij} = \frac{1}{n_{ij}}$$

Confidence intervals for odds-ratios are made by exponentiating the Wald confidence interval for the log odds-ratio. In R, this approach is implemented in the `oddsratio` function (using the `method="wald"` option) within the epitools package. In SAS it is calculated by PROC FREQ when the `MEASURES` option is used in a tt TABLES statement.

4.4 Wald statistics - quick and dirty?

Are they quick? – generally yes. In the scalar parameter case the use of Wald statistics is straightforward and very familiar. For example, an approximate 95%

confidence interval is simply $\widehat{\theta}$ plus or minus a couple of standard deviations.

Box 4.3.

The Wald test statistic is not invariant to transformation. To see this, consider the scalar parameter case and suppose that $\widehat{\theta} \sim N(\theta_0, \widehat{v})$. Let $\zeta = g(\theta)$ be invertible, so that $H_0 : \theta = \theta_0$ is equivalent to $H_0 : \zeta = \zeta_0$ where $\zeta_0 = g(\theta_0)$. From (4.5) $g(\widehat{\theta}) \sim N(g(\theta_0), g'(\widehat{\theta})^2\widehat{v}) = N(\zeta_0, g'(\widehat{\theta})^2\widehat{v})$. By applying a Taylors series expansion of $g(\theta_0)$ around $\widehat{\theta}$, the Wald test statistic for $H_0 : \zeta = \zeta_0$ can be written

$$\begin{aligned} \frac{(\widehat{\zeta} - \zeta_0)^2}{g'(\widehat{\theta})^2\widehat{v}} &= \left(\frac{g'(\widehat{\theta})(\widehat{\theta} - \theta_0) + ...}{g'(\widehat{\theta})\widehat{v}^{1/2}} \right)^2 \\ &= \frac{(\widehat{\theta} - \theta_0)^2}{\widehat{v}} + ... \, , \end{aligned} \tag{4.11}$$

where the notation ... is used generically to denote remainder terms arising from the Taylor series expansion. The first term in (4.11) is the Wald statistic for $H_0 : \theta = \theta_0$.

Are they dirty? The lack of invariance to parameterization is one reason to suspect that they could be (Box 4.3). The following (contrived) example shows that the Wald statistic can be very dirty indeed.

Example 4.4. In a scalar parameter model with parameter space $\Theta = \mathbb{R}$, suppose that $\widehat{\theta} = 1$ and $\widehat{\text{var}}(\widehat{\theta}) = 0.01$. Then the Wald test statistic for $H_0 : \theta = 0$ is

$$W_\theta = \frac{(\widehat{\theta} - 0)^2}{0.01} = 100 \, .$$

Comparing $W_\theta = 100$ to a χ^2_1 distribution, the p-value for H_0 is a near-infinitesimal 8×10^{-24}.

Now, consider a re-parameterization of the model using $\zeta = g_a(\theta) = \theta^a$ for a an odd positive integer, that is, $a \in \{1, 3, 5, ...\}$. (This requirement on a ensures that $g_a : \mathbb{R} \to \mathbb{R}$ is a one-to-one mapping.) Then $\widehat{\zeta} = 1$ is the MLE and the null hypothesis can be expressed $H_0 : \zeta = 0$. The derivative is $g'(\theta) = a\theta^{a-1}$ and from the delta method,

$$\widehat{\text{var}}(\widehat{\zeta}) = g'(\widehat{\theta})^2 \widehat{\text{var}}(\widehat{\theta}) = 0.01a^2 \, .$$

The Wald test statistic for the equivalent hypothesis $H_0 : \zeta = 0$ is therefore

$$W_\zeta = \frac{(\widehat{\zeta} - 0)^2}{0.01a^2} = \frac{100}{a^2} \, .$$

Note that the statistic W_ζ can be made arbitrarily small, and hence the p-value for H_0 can be made arbitrarily close to unity, by using increasingly large values of a. For example, when $a = 25$ then $W_\zeta = 0.16$ and the p-value for H_0 is approximately 0.69. □

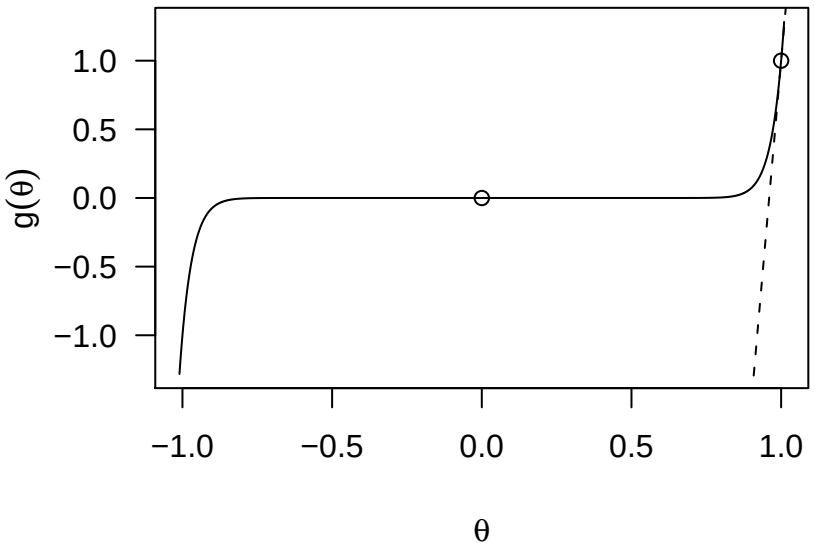


Figure 4.4: Plot of $g(\theta) = \theta^{25}$, showing the inadequacy of a linear approximation (dashed line) at $\widehat{\theta} = 1$.

The above example shows that the p-value can be made arbitrarily large using a suitably ridiculous re-parameterization of the model. In this case, the degeneration of the Wald statistic arises because the linear approximation of $g(\theta)$ around $g(\widehat{\theta})$ can be made arbitrarily bad (Figure 4.4). In contrast, inference based on the likelihood ratio is unaffected, because the statistical model is preserved under re-parameterization.

4.4.1 Wald versus likelihood ratio

Pawitan (2001, p.47) provides a heuristic argument that the likelihood ratio always performs at least as well as the Wald statistic. The argument proceeds along the following lines – when the log-likelihood is quadratic then the likelihood ratio and Wald statistics are identical (Exercise 4.8). In this case, the likelihood corresponds

to that of that of an iid normal model with mean θ and sample mean $\widehat{\theta}$ and hence it would be reasonable to suppose that

$$W = \frac{(\widehat{\theta} - \theta_0)^2}{\widehat{\text{var}}(\widehat{\theta})} \sim \chi^2_1 \ ,$$

to a high degree of approximation, and hence also for the likelihood ratio statistic.

Thus, if there exists any parameterization $\zeta = g(\theta)$ such that the log-likelihood is an approximately quadratic function of ζ then the likelihood-ratio statistic will have good performance (see Figure 4.5).

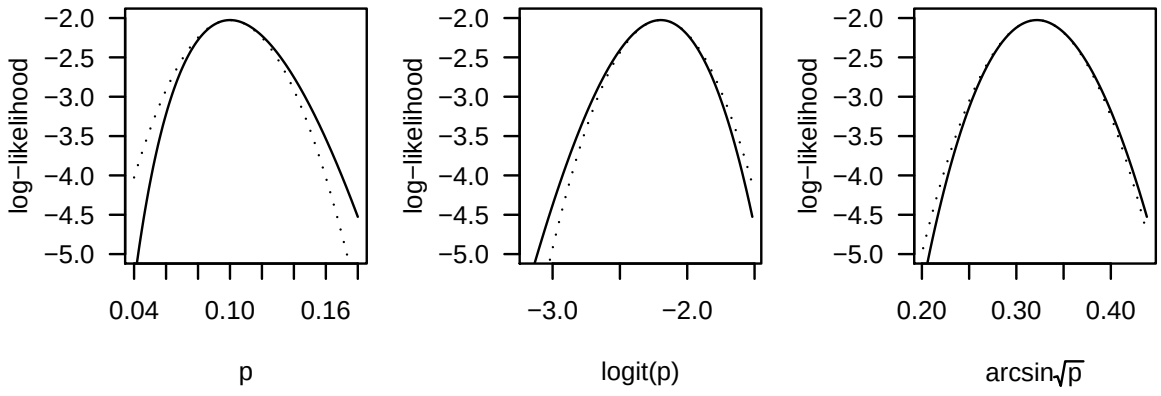


Figure 4.5: Plot of $g(p)$ (identity, logit(p), or $\sin^{-1}(\sqrt{p})$) versus $l(g(p))$ for $y = 10$ observed from a binomial$(100, p)$ experiment.

When to use Wald

Often times, it will not be possible to use likelihood ratio for pragmatic reasons. For one, standard errors of estimators are often required to be reported, regardless of any consideration to the legitimacy of these for use in hypothesis tests or in calculation of confidence intervals. Secondly, calculation of the likelihood ratio statistic may be computationally infeasible in some complex models. Thirdly, the likelihood function may not always be available – in the geoduck example of Section 4.3.1 it would have required the log-likelihood functions of the two experiments in order to obtain a likelihood ratio confidence interval for geoduck biomass.

In practice, the approximate normality of $\widehat{\theta}$ that underlies the Wald statistic will be sufficiently accurate if the sample size is sufficient to ensure that the log-

likelihood is reasonably well approximated by a quadratic. In some cases there are guidelines for this. For example, in the binomial(n, p) model it is typical to require $\min(n\widehat{p}, n(1-\widehat{p})) \geq 5$, although it should be noted that Brown et al. (2001) remain highly critical of the Wald statistic even when this criterion is satisfied.

4.5 Profile likelihood

In Section 3.4 we saw that the likelihood ratio test of the null hypothesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ required maximization over the restricted parameter space Θ_0. There, we assumed that $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ where $\boldsymbol{\psi} \in \mathbb{R}^r$, $\boldsymbol{\lambda} \in \mathbb{R}^{s-r}$, and $\boldsymbol{\theta} \in \Theta_0$ are all points in Θ for which $\boldsymbol{\psi}$ equals some specified value $\boldsymbol{\psi}_0$, and the null hypothesis can be re-expressed as $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$. The partial maximization required finding $\widehat{\boldsymbol{\theta}}_0 = (\psi_0, \widehat{\lambda}_0)$, the ML estimator in Θ_0. That is

$$l^*(\boldsymbol{\psi}_0; \boldsymbol{y}) \equiv \max_{\boldsymbol{\lambda}} \; l(\boldsymbol{\psi}_0, \boldsymbol{\lambda}; \boldsymbol{y}) \;, \tag{4.12}$$

where l^*, regarded as a function of $\boldsymbol{\psi}$, is called the profile log-likelihood (for $\boldsymbol{\psi}$).

When $\boldsymbol{\psi}$ is one or two dimensional then a plot of the profile log-likelihood is an insightful tool for assessing the support given to differing values of $\boldsymbol{\psi}$. Hypothesis tests, and hence likelihood ratio confidence intervals/regions, can be evaluated directly from the profile plot, as was done in Figures 1.2 and 3.5. This follows immediately from noting that $l(\widehat{\boldsymbol{\theta}}) = l^*(\widehat{\boldsymbol{\psi}})$ and $l(\widehat{\boldsymbol{\theta}}_0) = l^*(\boldsymbol{\psi}_0)$, and so the LRT statistic for H_0 is

$$\begin{aligned} X &= 2[l(\widehat{\boldsymbol{\theta}}) - l(\widehat{\boldsymbol{\theta}}_0)] \\ &= 2[l^*(\widehat{\boldsymbol{\psi}}) - l^*(\boldsymbol{\psi}_0)] \;. \end{aligned}$$

Box-Cox transformations are presented as a common application of profile likelihood in Section 6.2.

4.5.1 Profile likelihood for Old Faithful

Likelihood ratio confidence intervals for the five parameters of the binormal mixture model were obtained for the Old Faithful waiting time data in Section 3.4.1. In

practice, it can be informative to also examine the shape of the profile log-likelihood. Here, the profile log-likelihood for parameter p is obtained using utilities available in SAS and R. The profile can also be obtained in ADMB, using the `likeprof_number` declaration, as demonstrated in Section 1.4.3.

Using R

The following code uses the `Profile` function to calculate $l^*(p)$ over the sequence of values of p from 0.27 to 0.46 in steps of 0.005. A brief description of the `Profile` function is found in Section 15.4.2.

```
> #nllhood and parnames have been previously defined
> source("Profile.R")
> p.seq=seq(0.27,0.46,0.005)
> Profile.p=NULL
> for(i in 1:length(p.seq))
+   Profile.p[i]=Profile(parnames,nllhood,label="p",psi=p.seq[i],
+         lambda=c(50,5,80,5),y=waiting)$value

> #Display profiled log-likelihood for first 3 values in sequence
> cbind(p.seq,Profile.p)[1:3,]
     p.seq Profile.p
[1,] 0.270 -1038.584
[2,] 0.275 -1038.074
[3,] 0.280 -1037.597
```

The profile is shown in Figure 4.6, along with the 95% LR confidence interval of (0.3013,0.4230) that was obtained in Section 3.4.1.

Using SAS

The following SAS code creates a dataset `Profile_p` containing $l^*(p)$ for 39 evenly-space values of p from 0.27 to 0.46. That is 0.27 to 0.46 in steps of 0.005. This code requires the user-defined macro `OldFaithfulProfile_p` defined in Section 3.4.1.

```
%INCLUDE "ProfileMacro.sas";
%Profile(OldFaithfulProfile_p,lower=0.27,upper=0.46,n=39,dsname=Profile_p);
```

4.6 Model selection

So far, the statistical models used in examples and exercises have been pre-specified. This specification has been based on consideration of the type of data being measured

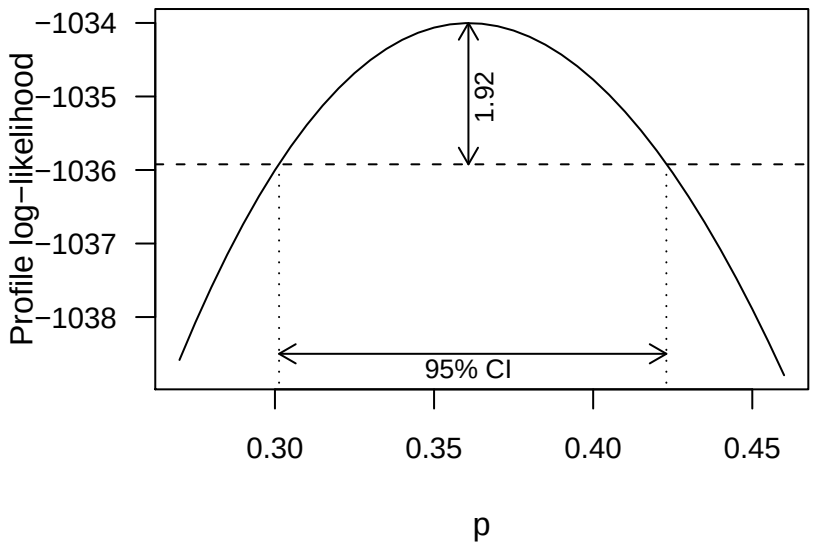


Figure 4.6: Profile log-likelihood for parameter p for the binormal mixture model of the Old Faithful waiting time data.

(e.g., continuous data, counts, proportions) and perhaps a preliminary assessment of the observed $\boldsymbol{y}$. However, selection of an appropriate statistical model is a very important consideration because inference is conditional on the model being correctly specified, although the practical reality is that the statistical model is assumed to be good enough to be useful. An excellent overview of the relevance of model choice is provided by Harrell (2001, Section 1.4).

Model selection commonly arises in the context of regression modeling where it is necessary to choosing an appropriate subset from the set of all possible explanatory covariates and their interactions. In this case the models are nested, in the sense that all possible models can be obtained by sequentially removing terms from the full model in which all possible terms are used. Model selection could then proceed via backward elimination or forward selection, using hypothesis testing to decide whether a term should be removed or added. While this may seem a reasonable approach, it has already been seen that hypothesis testing is not a tool that is suited to the task of choosing the preferred model (Example 2.4). This form of model selection is heuristic, but is nonetheless still widely used in practice.

4.6.1 AIC

Akaike (1974) provided a holistic approach to model selection by using the concept of information loss. Specifically, he considered the loss of statistical information that is incurred by using the fitted model to approximate the true unknown model. This loss is quantified by the Kullback-Leibler divergence between the true model and fitted model, and can be regarded as the loss of predictive ability from using the fitted model rather than the actual true model. Under this framework, the preferred model is the one that minimizes this loss amongst the collection of all candidate models.

In the context of maximum likelihood estimation, Akaike (1974) obtained a simple formula for estimation of the predictive loss from using $f(\boldsymbol{y}; \widehat{\boldsymbol{\theta}})$ as an estimate of the true density of $\boldsymbol{Y}$. This lead to what is now known as Akaike's information criterion (AIC)

$$\text{AIC} = -2l(\widehat{\boldsymbol{\theta}}) + 2s \ , \tag{4.13}$$

where s is the number of model parameters. The preferred statistical model is the one resulting in smallest AIC. Note that AIC can be interpreted as achieving a parsimonious balance between model fit, as quantified by $l(\widehat{\boldsymbol{\theta}})$, and model complexity, as quantified by the number of parameters.

Box 4.4.

As a general rule-of-thumb, if the difference in AIC between two models is 2 or more then it can be said that the model with smaller AIC is strongly preferred. However, if the difference is less than 2 then it could be argued that both models are worthy of consideration. See Burnham and Anderson (2002) for discussion of the notion of model averaging, whereby inference explicitly incorporates the uncertainty in picking the preferred model.

Model selection using AIC does not require the competing models to be nested. For example, the competing models could specify different error structure on the data (e.g., normal versus lognormal) or could use different functional forms to describe the effect of explanatory variables (e.g., GLMs with different link function).

Example 4.5. A zero-inflated Poisson (ZIP) was fitted to the micro-propogation

count data in Exercise 3.7. However, the negative binomial (NB) distribution can also be very effective at modeling count data that contain more zeroes than expected under a Poisson model (e.g., Warton 2005). The ZIP and NB models can both be fitted using `PROC GENMOD` in SAS 9.2 (see Section 7.6.3), or using a choice of functions provided by several R packages (e.g., the `vglm` function in the `VGAM` package). The fitted ZIP and NB models had log-likelihood of -74.88 and -81.00, respectively. These two models both have two parameters, and the respective AICs are therefore 153.76 and 166.00. The difference in AICs is 12.24, and the ZIP model is very strongly preferred. □

There are numerous variants to the AIC, including a version that uses a correction for small sample size (AICC), and quasi-AIC (QAIC) for use with over-dispersed models (Section ss:QuasiIntro). Also, Bayesian application of the minimum information loss argument (Schwarz 1978) has lead to the Bayesian information criterion (BIC). The BIC is analogous to AIC, except that the $2s$ term is replaced by $s\log(n)$ where n is the sample size. Many studies of the relative performance of AIC have appeared in the published literature, and it has been established that AIC tends to over-fit when sample size is large (e.g., Zheng and Loh 1995). Consequently, some modelers prefer BIC to AIC, because (for $n \geq 8$), BIC imposes a stronger penalty on model complexity than AIC. See Burnham and Anderson (2002) for a comprehensive presentation of the AIC and several of its variants.

The number of candidate models can be large, especially in the situation of determining the best set of terms to use in a regression model. For such purposes, R and SAS provide some stepwise functionality for moving through the collection of all possible models. For example, see R function `step`, and the `SELECTION` option in the `MODEL` statements of the regression procedure `PROC REG` and logistic regression procedure `PROC LOGISTIC`.

4.7 Bootstrapping

Frequentist inferential procedures are based on the notion of repeat sampling and so, in the likelihood context, it is necessary to determine the properties and behaviour of ML based inference under repetition of the experiment. The general tools and techniques that we have been using up to this point have been obtained from a well-established body of theory that required large doses of calculus, probability theory and mathematical statistics (see chapters 12 and 13). The bootstrap effectively replaces this calculus and theory with computational effort.

The essential concept of bootstrapping is to emulate repetition of the experiment by simulating new data on the computer, followed by recalculation of the MLE using the simulated data. The appellation "bootstrap" was chosen by Efron (1979) in the first comprehensive account of this computer intensive methodology. The terminology was inspired by the adage "pulling oneself up by one's bootstrap", because the bootstrap achieves success from its own computational efforts.

The bootstrap has many potential advantages over the large-sample tools constructed in Chapter 12 and used up to this point. First and foremost, simulating the experiment, and subsequent model fitting process, is an intuitive thing to do. This simulation-based approach allows inference to be extended to situations where the theory has difficulties, such as including parameter uncertainty in prediction (Section 4.8.1). Moreover, bootstrapping can be valid for situations where the set of regularity conditions specified in Chapter 12 do not hold. For example, in Section 4.9.3 it is shown how the bootstrap can be used when the parameter lies on the boundary of the parameter space under the null hypothesis.

Bootstrapping is also applicable beyond the arena of likelihood-based inference and can be used to investigate the properties of a wide spectrum of estimators. However, to be valid, the bootstrap does requires that the estimator be *consistent*. This property is formally defined in Section 12.2, but an adequate working definition of consistency is that the estimator will get closer and closer to the true parameter value as sample size is increased to infinity.

Application of bootstrap methodology has evolved into many variants (e.g., Efron

1987). Herein, attention will be restricted to the simplest form, because of the virtue of simplicity and because this form has the most general application. Extensive coverage of the bootstrap can be found in Efron and Tibshirani (1993) and Davison and Hinkley (1997), and for a more applied focus see Manly (1997) and Chernick (2008).

4.7.1 Bootstrap simulation

The bootstrap emulates the sampling distribution of $\widehat{\boldsymbol{\theta}}$ by emulating the data generation and model fitting processes. It does this by generating artificial data $\boldsymbol{y}^* = (y_1^*, ..., y_n^*)$ from a distribution that approximates the true unknown sampling distribution of the actual data, followed by recalculating the MLE using these artificial data. This is done a large number of times, B say, resulting in a large collection of bootstrap MLEs, denoted $\widehat{\boldsymbol{\theta}}^*_{(j)}, j = 1, ..., B$. The distribution of these artificially generated bootstrap MLEs can then be used to infer the sampling distribution of $\widehat{\boldsymbol{\theta}}$ (Section 4.7.2).

One obvious choice for a distribution to approximate the true unknown sampling distribution of the data is to use the fitted model. That is, to generate $\boldsymbol{y}^*$ from the model $f(\cdot; \widehat{\boldsymbol{\theta}})$. This is parametric bootstrapping.

In experiments where the data-generating process can be emulated directly, non-parametric bootstrapping is an option. For example, if the the data $y_i, i = 1, ..., n$ are iid then the *empirical distribution function*, EDF, can be used as a discrete approximation to the true unknown cumulative distribution function of $\boldsymbol{Y}$. That is,

$$\begin{aligned} \widehat{F}(y) &= \frac{\text{Number of } y_i \leq y}{n} \\ &\approx P(Y \leq y) \ , \end{aligned}$$

where $\widehat{F}(y)$ denotes the EDF. The non-parametric bootstrap generates new data $\boldsymbol{y}^*$ by random sampling from the EDF. Note that the EDF assigns probability mass $\frac{1}{n}$ to the value of every observed data point (and if two or more data points take the same value then the probability mass for that value is summed over those data points). Consequently, random sampling from the EDF can be accomplished by sampling with replacement from the observed data $(y_1, ..., y_n)$.

More generally, it is sometimes the case that the data can be partitioned into distinct subsets such that the data are iid within each subset, in which case the non-parametric bootstrap is applied separately within each subset. It would also be appropriate to re-sample the collection of subsets (with replacement) if this emulated the experimental design used to obtain the data (e.g., see Section 3.8 of Davison and Hinkley 1997).

4.7.2 Bootstrap confidence intervals

The most commonly used form of bootstrap confidence interval is the percentile method CI. This method is the most simple, and it performs well under most situations.

Suppose that a $100(1-\alpha)\%$ CI is required for θ_i (component i of $\boldsymbol{\theta}$). Let $\widehat{\theta}^*_{i,\alpha/2}$ and $\widehat{\theta}^*_{i,1-\alpha/2}$ denote the $\alpha/2$ and $1-\alpha/2$ empirical quantiles of the collection of values $\widehat{\theta}^*_{i,(j)}, j = 1, ..., B$, where $\widehat{\theta}^*_{i,(j)}$ denotes component i of the bootstrap MLE from bootstrap simulation j. The $100(1-\alpha)\%$ percentile method CI is simply

$$(\widehat{\theta}^*_{i,\alpha/2}, \widehat{\theta}^*_{i,1-\alpha/2}) \ . \tag{4.14}$$

Justification

The percentile method interval in (4.14) is justified with the following argument based on the parametric bootstrap. The estimator, $\widehat{\boldsymbol{\theta}}$, is the MLE obtained from observing data $\boldsymbol{y}$ distributed according to the density $f(\boldsymbol{y}, \boldsymbol{\theta})$. Analogously, for each bootstrap simulation $\widehat{\boldsymbol{\theta}}^*$ is the MLE obtained from observing simulated data $\boldsymbol{y}^*$ distributed according to the density $f(\boldsymbol{y}, \widehat{\boldsymbol{\theta}})$. So, subject to appropriate regularity conditions, it will be reasonable to assume that the behaviour of $\widehat{\boldsymbol{\theta}}$ as an estimator of $\boldsymbol{\theta}$ should be well approximated by the behaviour of $\widehat{\boldsymbol{\theta}}^*$ as an estimator of $\widehat{\boldsymbol{\theta}}$. That is, for component i of the parameter vector,

$$\widehat{\theta}^*_i - \widehat{\theta}_i \approx_D \widehat{\theta}_i - \theta_i \ , \tag{4.15}$$

where $\approx_D$ denotes approximately equal in distribution.

Suppose, for now (see Box 4.5), that the distributions in (4.15) are approximately

symmetric around zero. Then

$$\begin{aligned} \widehat{\theta}_i^* - \widehat{\theta}_i &\approx_D -(\widehat{\theta}_i - \theta_i) \\ &= \theta_i - \widehat{\theta}_i \ . \end{aligned} \tag{4.16}$$

Thus, for any interval $(a, b), a < b$

$$P_*(a < \widehat{\theta}_i^* - \widehat{\theta}_i < b) \approx P_{\boldsymbol{\theta}}(a < \theta_i - \widehat{\theta}_i < b) \ , \tag{4.17}$$

where P_* denotes probability with respect to the bootstrap distribution of $\widehat{\boldsymbol{\theta}}^*$ and $P_{\boldsymbol{\theta}}$ denotes that the sampling distribution of $\widehat{\boldsymbol{\theta}}$ is with respect to repeat observation of data from the true unknown model $P_{\boldsymbol{\theta}}$.

Equation (4.17) can be rewritten,

$$P_*(a + \widehat{\theta}_i < \widehat{\theta}_i^* < b + \widehat{\theta}_i) \approx P_{\boldsymbol{\theta}}(a + \widehat{\theta}_i < \theta_i < b + \widehat{\theta}_i) \ .$$

and so, if a and b are such that $P_*(a+\widehat{\theta}_i < \widehat{\theta}_i^* < b+\widehat{\theta}_i) = 1-\alpha$ then it is also the case that $P_{\boldsymbol{\theta}}(a+\widehat{\theta}_i < \theta_i < b+\widehat{\theta}_i) = 1-\alpha$. Since a and b are arbitrary, this establishes that an interval containing $\widehat{\theta}_i^*$ with probability $(1-\alpha)$ is also a $(1-\alpha)100\%$ confidence interval for θ_i.

Box 4.5.

The assumption of symmetry used to obtain (4.16) can be weakened by virtue of the invariance of confidence intervals and quantiles to monotone transformations (e.g., if $\zeta = g(\theta_i)$ is monotone increasing then the percentile method confidence interval $(\widehat{\theta}^*_{i,\alpha/2}, \widehat{\theta}^*_{i,1-\alpha/2})$ for θ_i is equivalent to the confidence interval $(g(\widehat{\theta}^*_{i,\alpha/2}), g(\widehat{\theta}^*_{i,1-\alpha/2}))$ for ζ.) It is enough that approximate symmetry holds for any monotone transformation, $\zeta = g(\theta_i)$. Note that it is not necessary to know g.

4.7.3 Bootstrap estimate of variance

The approximate equivalence in (4.15) suggests using the variance of $\widehat{\theta}_i^* - \widehat{\theta}_i$ as an estimator of the variance of $\widehat{\theta}_i - \theta_i$. The variance of $\widehat{\theta}_i^* - \widehat{\theta}_i$ is over the bootstrap sampling and will be denoted $\text{var}^*(\widehat{\theta}_i^* - \widehat{\theta}_i)$. Now, $\widehat{\theta}_i$ is a constant with respect to the bootstrap sampling and so it follows that $\text{var}^*(\widehat{\theta}_i^* - \widehat{\theta}_i) = \text{var}^*(\widehat{\theta}_i^*)$. Similarly, the variance of $\widehat{\theta}_i - \theta_i$ is over repeat experimentation under the fixed value of the

unknown paramater θ, and so $\text{var}(\widehat{\theta}_i - \theta_i) = \text{var}(\widehat{\theta}_i)$. That is, the bootstrap estimate of variance is simply the variance of the bootstrap.

The bootstrap values $\widehat{\theta}_i^*$ are iid, and hence the bootstrap variance is given by

$$\widehat{\text{var}}(\widehat{\theta}_n)^* = \sum_{j=1}^{B} \frac{(\widehat{\theta}_{i,(j)}^* - \overline{\theta}_i^*)^2}{B-1} \tag{4.18}$$

where $\overline{\theta}_i^*$ is the average of the B bootstrap values $\widehat{\theta}_{i,(j)}^*, j = 1, ..., B$.

4.7.4 Bootstrap pragmatics

In practice, it happens more often then you would like that the bootstrap simulation will not finish due to a computational error. This typically occurs when one of the $\boldsymbol{y}^*$ is sufficiently "unusual" that the optimizer struggles to find the MLE. Such errors can usually be fixed by better specification of start values and explicit specification of bounds on the parameter space, to prevent negative variance parameters or probabilities outside of the unit interval, say. It may also be necessary to impose bounds to prevent the optimizer from venturing into regions of the parameter space where calculation of the log-likelihood would result in numerical underflow of overflow. These bounds must not be allowed to affect the ability of the optimizer to find $\widehat{\boldsymbol{\theta}}^*$.

Box 4.6.

Experience has taught that a SAS macro will usually continue to run a bootstrap simulation even if an error message is produced from a failed attempt to fit the model to one or more $\boldsymbol{y}^*$. These errant fits can later be investigated on a case-by-case basis. In R, the bootstrap simulation will often come to a premature end unless the error is trapped. This can be done using the `try` function.

4.7.5 Bootstrapping Old Faithful

The Old Faithful geyser data (Figure 2.5) are modeled as iid from a binormal mixture distribution, and both the parametric and non-parametric bootstraps are demonstrated below.

Using R

The convergence code from the `optim` function was checked at each iteration to ensure that the algorithm had successfully converged. This code is integer-valued and takes the value 0 if successful convergence occurred, the value 1 if the iteration limit is reached, and other positive values corresponding to potential problems with the optimization. Several percent of the optimizations returned a convergence code of 1 using the default iteration limit of 500. With the iteration limit increased to 5000, all ten thousand bootstrap optimizations returned a convergence code of 0. The ten thousand bootstraps took a few minutes on a 2GHz computer.

```
> ###Parametric bootstrap of binormal model for Old Faithful waiting times
> nboots=100
> PBootstrapMLEs=matrix(NA,nrow=nboots,ncol=length(MLE))
> colnames(PBootstrapMLEs)=parnames
> ConvergenceCode=NULL
> n=length(waiting)
> for(i in 1:nboots) {
+   b=rbinom(n,1,MLE[1])
+   ystar=b*rnorm(n,MLE[2],MLE[3])+(1-b)*rnorm(n,MLE[4],MLE[5])
+   fitstar=optim(c(0.5,55,5,80,5),nllhood,y=ystar,hessian=T,
+                 control=list(maxit=5000))
+   PBootstrapMLEs[i,]=fitstar$par
+   ConvergenceCode[i]=fitstar$conv
+ }

> #Check convergence codes
> table(ConvergenceCode)
ConvergenceCode
  0
100

> std.err=sd(PBootstrapMLEs)
> BStrap.CI=apply(PBootstrapMLEs,2,quantile,prob=c(0.025,0.975))
> round(cbind(MLE,std.err,t(BStrap.CI)),3)
         MLE std.err   2.5%  97.5%
p      0.361   0.031  0.305  0.414
mu    54.615   0.678 53.478 56.076
sigma  5.870   0.427  5.111  6.630
nu    80.092   0.486 79.222 81.030
tau    5.869   0.354  5.215  6.526
```

The nonparametric bootstrap is implemented using the following code. Sampling with replacement was implemented using the `sample` function with the `replace=T` option.

```
> round(cbind(MLE,std.err,t(BStrap.CI)),3)
         MLE std.err   2.5%  97.5%
p      0.361   0.030  0.301  0.416
mu    54.615   0.733 53.202 56.067
sigma  5.870   0.478  4.938  6.668
```

```
nu    80.092   0.544 79.165 81.189
tau    5.869   0.393  5.144  6.612
```

It is more informative to look at the histogram of bootstrapped values because this shows the shape of the confidence intervals corresponding to the quantiles of the bootstrap MLEs.

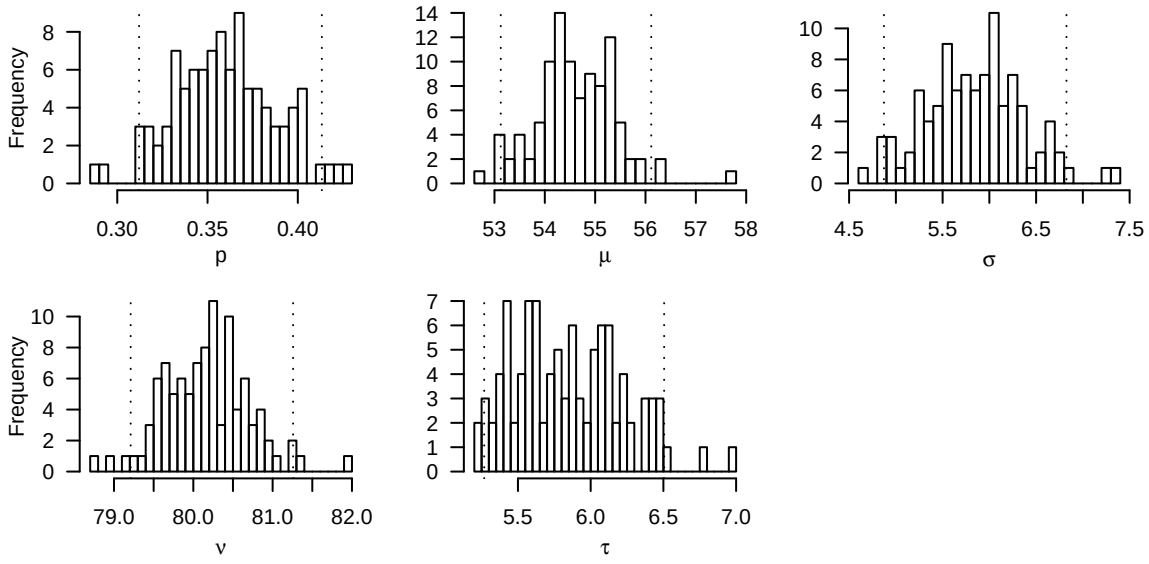


Figure 4.7: Bootstrap MLEs from ten thousand parametric bootstraps of the Old Faithful waiting time data. Vertical dashed lines show the 2.5% and 97.5% quantiles and hence are the percentile method 95% confidence intervals.

Using SAS

Bootstrapping in SAS can be implemented using the SAS macro language (and macros are available for download from SAS support at `http://support.sas.com`), however the code will be simpler and quicker to execute if macros can be avoided. In the SAS implementation below, ten thousand bootstrapped samples, $\boldsymbol{y}^*_{(j)}, j = 1, ..., 10000$, are written to a 10 000 by 272 SAS dataset. `PROC NLMIXED` then processes this dataset, row by row, using a `BY` statement.

The parametric bootstrap code below assumes that the SAS dataset `OldFaithfulMLE` contains the parameter estimates table generated from running the code on page 45.

SAS code for parametric bootstrap of Old Faithful

```
PROC TRANSPOSE DATA=saslib.OldFaithfulPars Out=Pars;
  VAR Estimate;
  ID Parameter;
RUN;

DATA P_bootstraps(KEEP=iter ystar1-ystar272);
  ARRAY ystar {272};
  SET Pars;
  DO iter=1 TO 10000;
    DO i=1 TO 272;
      b=RANBIN(0,1,p);
      ystar[i]=b*(mu+sigma*RANNOR(0))+(1-b)*(nu+tau*RANNOR(0));
    END;
  OUTPUT;
  END;
RUN;

PROC TRANSPOSE DATA=P_bootstraps
          OUT=saslib.PBootData (RENAME=(COl1=ystar) );
  BY iter;
RUN;

ODS SELECT ParameterEstimates;
ODS OUTPUT ParameterEstimates=saslib.PBootPars;
*ODS SELECT ALL;

*Using FDIGITS=10 to get similar std errors as R;
PROC NLMIXED DATA=saslib.PBootData DF=1E6 FDIGITS=10;
  BY iter;
  PARMS p=0.5 mu=55 sigma=5 nu=80 tau=5;
  BOUNDS 0<p<1, 0<sigma, 0<tau;
  ll=log( p*PDF("NORMAL",ystar,mu,sigma)+
          (1-p)*PDF("NORMAL",ystar,nu,tau) );
  MODEL ystar ~ GENERAL(ll);
RUN;
*********************************************************;

PROC UNIVARIATE DATA=saslib.PBootPars NOPRINT;
  CLASS parameter (ORDER=DATA);
  VAR estimate;
  OUTPUT OUT=BootStrapCIs PCTLPTS=2.5 97.5 PCTLPRE=Q;
RUN;

ODS PS FILE="&almhome\Outputs\OldFaithfulBootstrapCIs.ps";
PROC PRINT NOOBS;
RUN;
ODS PS CLOSE;
```

Parameter	Q2_5	Q97_5
p	0.3023	0.4221
mu	53.2726	56.0688
sigma	4.8239	6.9774
nu	79.1221	81.0789
tau	5.1190	6.5974

Figure 4.8: Parametric bootstrap 95% confidence intervals

The non-parametric bootstrap is accomplished with the change

SAS code for generating non-parametric bootstrap data

```
******Generate non-parametric bootstrap data************;
PROC TRANSPOSE DATA=saslib.geyser OUT=faithful PREFIX=y;
RUN;

DATA NP_bootstraps(KEEP=iter ystar1-ystar272);
  ARRAY y {272} y1-y272;
  ARRAY ystar {272};
  SET faithful;
  DO iter=1 TO 10000;
    DO i=1 TO 272;
      ystar[i]=y[ ROUND(0.5+272*RANUNI(0)) ];
    END;
  OUTPUT;
  END;
RUN;

PROC TRANSPOSE DATA=NP_bootstraps
               OUT=saslib.NPBootData (RENAME=(COl1=ystar) );
  BY iter;
RUN;
```

4.7.6 How many bootstrap simulations is enough?

The number of bootstrap simulations, B, must be sufficiently large that the observed quantiles will be close to the bootstrap distribution quantiles. This will ensure that the bootstrap confidence intervals are not materially affected by the randomness inherent in the bootstrap.

Quantiles in the extreme tails of a distribution require a large sample size to be estimated well. For example, for a sample size of 100, the 0.025 quantile can be calculated as the average of the 2nd and 3rd smallest observations and hence will be very sensitive to the random occurrence of one or two extremely small values. Indeed, Davison and Hinkley (1997) recommend performing at least 1000 bootstrap simulations when calculating 95% bootstrap confidence intervals.

In practice it is a good idea to perform several tentative bootstrap runs with a modest value of B and to examine the variability in the observed quantiles. The standard error of the quantiles will decrease at a $\sqrt{B}$ rate and so it will be possible to determine a suitable value of B for the full bootstrap simulation. This approach can be applied post-bootstrap. For example, the full run of B simulations can be partitioned into K subsets of size B/K, and the quantiles calculated for each sub-

set. The standard deviations of the quantiles over the K subsets gives an unbiased estimate of the standard deviation of bootstrap quantiles from a bootstrap run of length B/K. Dividing these standard deviations by $\sqrt{K}$ then gives an estimate of the standard error of the quantiles from the full bootstrap of length B. The SAS code below applies this approach by splitting the 10 000 bootstraps into ten subsets of 1 000.

SAS code for estimating quantile standard errors

```
PROC SORT DATA=saslib.PBootPars OUT=work;
  BY parameter iter;

DATA work;
  SET work;
  group=CEIL(iter/1000);

PROC UNIVARIATE DATA=work NOPRINT;
  BY parameter group;
  VAR estimate;
  OUTPUT OUT=GroupMeans PCTLPTS=2.5 97.5 PCTLPRE=Q;

PROC MEANS DATA=GroupMeans NOPRINT;
  BY parameter;
  VAR Q2_5 Q97_5;
  OUTPUT OUT=ses STD=seQ2_5 seQ97_5;

DATA ses;
  SET ses;
  seQ2_5=seQ2_5/SQRT(10);
  seQ97_5=seQ97_5/SQRT(10);

PROC PRINT DATA=ses;
  VAR parameter seQ2_5 seQ97_5;
RUN;
```

Estimated std errors of bootstrap quantiles

Obs	Parameter	seQ2_5	seQ97_5
1	p	0.000799	0.000482
2	mu	0.017348	0.016317
3	sigma	0.012104	0.014819
4	nu	0.011343	0.013023
5	tau	0.010755	0.011784

Figure 4.9: Estimated standard errors of the bootstrap quantiles obtained from partitioning the 10 000 bootstrap simulations

4.8 Prediction

So far we have been concerned with estimation and inference about parameters $\boldsymbol{\theta}$ and functions of those parameters $g(\boldsymbol{\theta})$. In the statistical model, these are regarded as fixed but unknown quantities. However, there may be occasions when the question of interest requires inference about a random quantity. It is usual to refer to this as a *prediction* problem, to distinguish it from estimation of a fixed unknown quantity.

In many cases, the random quantity to be predicted will be a future observation of the response variable. For example, it could be tomorrow's price of a company share, or the number of animals in a population next year under a particular management scheme, or the number of customers that will arrive to be served. Prediction also arises naturally in the context of models having unobserved random effects when it is desired to predict the value of these effects (see Chapter 10). Unlike a future observation, these so-called latent variables are never directly observed.

Let Z denote the random quantity (possibly vector valued) to be predicted. Our goal is to make statements about Z that have sound frequentist properties. In particular, if $Z \in \mathbb{R}$ then it is natural to calculate a $100(1-\alpha)\%$ *prediction interval* for Z. This is defined analogously to a confidence interval, that is, under hypothetical repetition of the experiment that generates both the observed data $\boldsymbol{y}$ and a realization z of the random variable Z, in the long run approximately $100(1-\alpha)\%$ of the prediction intervals will contain the value of z that was realized.

In special cases, such as prediction of a new observation from a linear regression model, it is possible to obtain exact prediction intervals. This can be done in R using the `predict.lm` function, and in SAS via the `PREDICT` option in the `OUTPUT` statement of `PROC GLM` or `PROC REG`. Example 4.6 shows an example of an exact prediction interval for the iid normal model.

Example 4.6. Prediction in the IID normal model. Let $y_i, i = 1, ..., n$ be observed iid $\mathrm{N}(\mu, \sigma^2)$ data, and suppose that a $100(1-\alpha)\%$ prediction interval is required for an additional iid observation Y_{n+1}. Now,

$$\overline{Y} - Y_{n+1} \sim N(0, (1 + \tfrac{1}{n})\sigma^2) \ ,$$

and so

$$\frac{\overline{Y} - Y_{n+1}}{S\sqrt{1+\frac{1}{n}}} \sim t_{n-1} \ , \tag{4.19}$$

where $S^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)$ denotes the sample variance. Then

$$P\left(-t_{n-1,1-\alpha/2} \leq \frac{\overline{Y} - Y_{n+1}}{S\sqrt{1+\frac{1}{n}}} \leq t_{n-1,1-\alpha/2}\right) = 1-\alpha \ ,$$

and hence $\overline{Y} \pm t_{n-1,1-\alpha/2} S\sqrt{1+\frac{1}{n}}$ gives a $100(1-\alpha)\%$ prediction interval for Y_{n+1}.

□

In Example 4.6 the calculation of an exact predictive interval is possible because the expression on the left side of (4.19) is a pivotal statistic, that is, its distribution does not depend on any unknown parameters. However, in the general case such constructions are not possible and it becomes necessary to explicitly consider the distribution of Z. Now, Z may not be independent of $\boldsymbol{Y}$, and since $\boldsymbol{Y} = \boldsymbol{y}$ has been observed, the density of interest is the conditional density $f(z|\boldsymbol{y};\boldsymbol{\theta})$. In many situations, such as prediction of a future independent observation (e.g., Example 4.6), Z will be independent of $\boldsymbol{Y}$ and then of course $f(z|\boldsymbol{y};\boldsymbol{\theta}) = f(z;\boldsymbol{\theta})$.

The challenge for prediction is that $f(z|\boldsymbol{y};\boldsymbol{\theta})$ depends on the unknown $\boldsymbol{\theta}$. It is natural to use the plug-in approach, and work with the density $f(z|\boldsymbol{y};\widehat{\boldsymbol{\theta}})$, but this fails to take into account the sampling variability. This approach, and three others, are presented below.

4.8.1 Prediction in Practice

The four approaches listed below could be used in situations where exact predictive inference is not possible. However, it is only pseudo-Bayes or bootstrap prediction that are recommended in practice.

The plug-in approach

The *plug-in* approach simply replaces θ with the MLE, so that predictive inference about Z is obtained using the density $f(z|\boldsymbol{y};\widehat{\boldsymbol{\theta}})$. This approach is also often called

the *estimative* or the *naive* approach. The latter appellation is due to the fact that it takes no account of the sampling variability in $\widehat{\boldsymbol{\theta}}$. Consequently, the predictions will tend to have too little variability and so prediction intervals will tend to have coverage that is too small.

One dramatic example of the danger of the plug-in approach is provided by Ludwig (1996) in a comparison of plug-in and fully Bayesian methodology for predicting the population of bird species into the future. Very few of the plug-in predictions corresponded to a population collapse, but collapse was much more frequent from the Bayesian forecasts. In the case of Palila (*Loxioides balleui*), an endangered finch-billed honeycreeper, less than 0.3% of the plug-in forecasts showed population collapse, compared to more than 17% from the Bayesian forecasts. However, Ludwig (1996) did not examine sensitivity of the Bayesian forecasts to prior specification, and did not include pseudo-Bayes or bootstrap prediction in his comparison.

Predictive likelihood

The general idea behind predictive likelihood is to consider the unobserved value of z as an additional parameter via an extended definition of likelihood (Bjørnstad 1996). However, this approach is computationally challenging and not widely used, and is not without controversy. Indeed, Hall, Peng and Tajvidi (1999) argued that predictive likelihood did not improve on the naive approach, and instead recommended using a bootstrap approach to prediction.

Pseudo-Bayesian prediction

Under the Bayesian paradigm, both Z and $\boldsymbol{\theta}$ are random unobserved quantities, with joint posterior density denoted $f(z, \boldsymbol{\theta}|\boldsymbol{y})$. Realizations of Z can be generated in two steps.

1. Generate $\boldsymbol{\theta}$ from its posterior density $f(\boldsymbol{\theta}|\boldsymbol{y})$.

2. Generate z from $f(z|\boldsymbol{\theta}, \boldsymbol{y})$, where $\boldsymbol{\theta}$ is from step 1. In many applications, this step will be straightforward, for example, when Z is a future independent observation then z is just a realization from the sampling model $f(\cdot; \boldsymbol{\theta})$.

Pseudo-Bayesian prediction uses the same steps, but in step 1 an approximation to the posterior density $f(\boldsymbol{\theta}|\boldsymbol{y})$ is used. Subject to appropriate regularity conditions and sample size the posterior can be approximated by a (multivariate) normal distribution centered at the MLE and with variance equal to the approximate variance of the MLE Berger (1985, p. 224). That is,

$$\boldsymbol{\theta}|\boldsymbol{y} \sim N_s(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}) \tag{4.20}$$

where $\widehat{\mathbf{V}}$ denotes the inverse of the observed or expected Fisher information. In practice, for the approximation in (4.20) to be as accurate as possible, the model should be parameterized so that likelihood profiles are close to quadratic (see Example 4.7). A formal evaluation of the accuracy of (4.20) in the context of approximating Bayesian credible intervals can be found in Severini (1994).

In the case of linear regression and generalized linear models, step 1 above can be performed by the `sim` function in the `arm` package. Gelman and Hill (2007) give examples of its use. More generally, simulated values from a multivariate normal can be obtained using the `mvrnorm` function in the `MASS` package of R, or using the `MVN` SAS macro available from `http://support.sas.com`.

Bootstrap prediction

Bootstrap prediction is analogous to pseudo-Bayesian prediction, except that the bootstrap distribution of $\widehat{\theta}^*$ is used in place of the pseudo-posterior $f(\boldsymbol{\theta}|\boldsymbol{y})$. Harris (1989) showed that this approach performs well and is superior to the naive plug-in approach when making prediction for the class of exponential family models (Section 7.1.1), and the large-sample equivalence of bootstrap and Bayesian prediction is examined in Fushiki, Komaki and Aihara (2004).

Example 4.7. Binomial prediction In Example 1.1, $y = 10$ was observed from a binomial$(100, p)$ experiment. Suppose that it is desired to predict a future observation from this same distribution.

Figure 4.10 shows a relative frequency histogram of ten million simulated predicted values $Y \sim \text{Bin}(100, p)$, for five different predictors. The first three (plug-in,

pseudo-Bayes, and bootstrap) are described above, and the last two are Bayesian predictors. The first of these uses the Jeffreys reference prior, $f(p) \propto p^{-1/2}(1-p)^{-1/2}$, and the second uses the uniform prior $p \sim \mathrm{U}(0,1)$. The Jeffreys prior is the most commonly used for p, but there are plausible arguments in support of the uniform prior. Indeed, Bayes himself preferred the uniform prior for reasons derived from considering the predictive distribution of Y, and this choice has gained further support from the work of Tuyl, Gerlach and Mengersen (2009).

The naive approach simply uses the plug-in predictive distribution binomial$(100, \widehat{p})$ with $\widehat{p} = 0.1$. It has much less variability compared to the other predictors, due to the failure to incorporate uncertainty in $\widehat{p}$. The pseudo-Bayes approach simulated $\zeta = \mathrm{logit}(p) \sim N(\mathrm{logit}(\widehat{p}), 1/9)$ to achieve a better normal approximation in (4.20) than would be obtained by simulating $p \sim N(\widehat{p}, 0.0009)$. In this example, the parametric and non-parametric bootstrap are identical, since both are generating $Y^* \sim \mathrm{Bin}(100, 0.1)$. These last four predictors give similar results, with the greatest difference being between that the bootstrap predicts relatively more small values of Y and relative less large values, in comparison to the uniform-prior Bayesian predictor. □

4.9 Things that can mess you up

4.9.1 Multiple maxima of the likelihood

For some familiar families of models (including linear regression) a unique MLE can be obtained as the unique explicit solution to the likelihood equations. In the case of generalized linear models, the MLE can not in general be obtained explicitly, however it has been proved that the MLE is unique under very general conditions (Wedderburn 1976).

In general, when implementing a novel model there may be no guarantees of the existence or uniqueness of an MLE. In this case, it may be prudent to explore the shape of the likelihood function using profile and contour plots. When using numeric

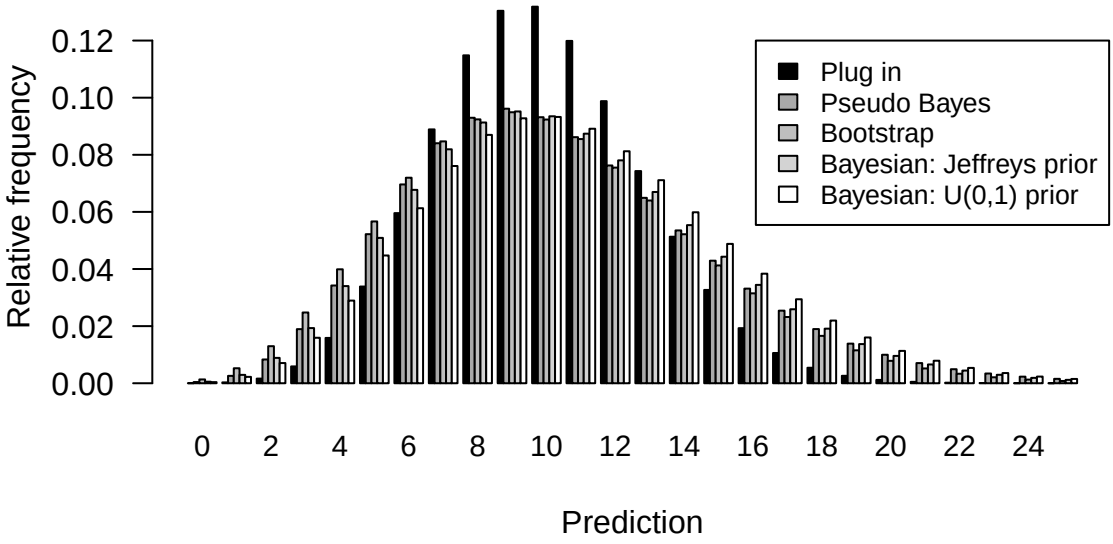


Figure 4.10: Side-by-side relative frequency histograms of ten million simulated predicted value of $Y \sim \text{Bin}(100, p)$ for five choices of prediction method.

optimizers to find an MLE, it is always good practice to repeat the optimization using several different starting values to see whether the optimizer will converge to different local maxima.

A lack of uniqueness is not necessarily problematic. For example, it was seen in Section 2.3.5 that the binormal mixture model possesses unbounded likelihood. This did not prevent a sensible local maxima being obtained when this model was fitted to the Old Faithful data in Section 3.3.3.

4.9.2 Lack of convergence

If the chosen optimizing algorithm does not work then of course another could be used. Another possible cure is to use better initial values $\boldsymbol{\theta}^{(0)}$ for the optimizer, perhaps by inspecting the data usual suitable exploratory plots. Optimizing software typically allow specification of many options, including specification of the convergence criterion to decide when iteration should cease.

4.9.3 Parameters on the boundary of the parameter space

In Chapter 12, several regularity conditions were required in order to proceed with the proofs of the asymptotic distribution of MLEs and the likelihood ratio test-statistic. One of these conditions, R4, prevents θ_0 from being on the boundary of the parameter space.

To see the problem that arises with parameters on the boundary of Θ, consider the case where the parameter space is the non-negative reals, $\Theta = [0, \infty)$, and suppose that $\theta_0 = 0$. Since $\widehat{\theta} \in \Theta$, then $\widehat{\theta}$ can never be less than θ_0, and hence its distribution can not be approximated by a normal centered at $\theta_0 = 0$. It is also the case that the LRT-statistic for $H_0 : \theta = \theta_0$ can no longer be assumed to have an asymptotic χ^2_1 distribution.

The boundary phenomenon is something to watch out for when working with mixture models and variance components models (e.g., Section 10.4). In the latter, the boundary issue arises because variances must, of course, be non-negative.

Example 4.8. The binormal mixture model used to describe waiting times of the Old Faithful geyser (Example 2.9) has a five-dimensional parameter vector $\boldsymbol{\theta} = (p, \mu, \sigma, \nu, \tau)$.

Consider the null hypothesis $H_0 : p = 0$. At first glance this appears to be a one-dimensional restriction on the parameter space. However, under H_0 the waiting times are from a single normal distribution, and so it could also be argued that the restriction is of three dimensions. In fact, due to the unusual geometry of the parameter space under H_0, the distribution of the LRT-statistic is complex, and has been shown to be close to that of a χ^2_6 random variable (McLachlan 1987, Lo 2005). Also, see Exercise 4.9. □

The theoretical behaviour of the LRT-statistic when $\boldsymbol{\theta}_0$ is on the boundary of Θ has been established for some classes of model (e.g., Self and Liang 1987, Lo 2005). However, except in special models, these theoretical results can be difficult to apply, and the asymptotic distributions can be poor approximations to the true sampling

distribution for moderate sample sizes (Pinheiro and Bates 2000, Crainiceanu and Ruppert 2004). For linear mixed models with one variance component, Crainiceanu and Ruppert (2004) obtained an exact representation of the finite-sample LRT statistic. This has been implemented in R, including an extension to more general linear mixed models, by Scheipl, Greven and Küchenhof (2008). Application of this test is demonstrated in Section 10.4.2.

Bootstrap simulation provides a general strategy for assessing the sampling distribution of $\widehat{\theta}$, and other statistics such as the likelihood ratio test statistic.

To be completed

In many situations, a formal hypothesis test may not be relevant. In the linear mixed model example of Section 10.4, the goal is to seek a parsimonious model, and AIC preferable...

4.9.4 Non-arrival at Asymptopia

Perhaps the most well-known case of this arises when the number of parameters in the model increases as n increases. Some methodology for coping with this situation is the focus of Chapter 9.

4.10 Exercises

4.1 Show that $X = \sin^{-1}\sqrt{\widehat{p}} = \sin^{-1}\sqrt{Y/n}$ is the variance stabilizing transformation for the binomial $Y \sim \text{binomial}(n, p)$, i.e., that the approximate variance of X does not depend on p.

4.2 In part b) of Example 2.3, the MLE of $P(Y \leq 6)$ was seen to be $g(\widehat{\theta}) = \Phi((6-\widehat{\mu})/\widehat{\sigma})$ where Φ is the distribution function of the standard normal and $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma}) = (5, 2)$. Suppose now that the approximate variance matrix of $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma})$ is

$$\widehat{\mathbf{V}} = \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix}$$

Use the delta method to obtain the approximate standard deviation of $g(\widehat{\theta})$. (Hint: Note that Φ is implemented as the `pnorm` function in R and `PROBNORM` function in SAS.)

4.3 For the two-parameter model used in Section 3.5.1 the MLE is $(\widehat{\beta}_0, \widehat{\beta}_1) = (-10.632, 0.304)$ and its approximate variance matrix is

$$\widehat{\mathbf{V}} = \begin{pmatrix} 0.7477 & -0.02022 \\ -0.02022 & 0.0005584 \end{pmatrix}.$$

In the fisheries trawling context of this example, the parameters of interest are $l_{50} = -\widehat{\beta}_0/\widehat{\beta}_1$ and $SR = 2\log(3)/\widehat{\beta}_1$. These correspond to the length of 50% retention probability, and the difference between the lengths of 75% and 25% retention (the so-called "selection range"), respectively. Use the delta method to determine the approximate distribution of the MLE of $g(\widehat{\beta}_0, \widehat{\beta}_1) = (-\widehat{\beta}_0/\widehat{\beta}_1, 2\log(3)/\widehat{\beta}_1)$, and hence calculate approximate 95% Wald CIs for l_{50} and SR.

4.4 Use the delta theorem to determine the approximate variance of $\widehat{\theta}_1/\widehat{\theta}_2$ (as a function of $\widehat{\theta}_1$, $\widehat{\theta}_2$, $\text{var}(\widehat{\theta}_1)$, $\text{var}(\widehat{\theta}_2)$ and $\text{cov}(\widehat{\theta}_1, \widehat{\theta}_2)$).

4.5 a. Verify (4.7) using one (or both) of the following approaches:

i. Using (13.7), or

by an exact Taylor series expansion (it includes one second order term) of $g(\widehat{\theta}_1, \widehat{\theta}_2)$ about $g(E[\widehat{\theta}_1], E[\widehat{\theta}_2])$.

ii. Verify (4.8).

4.6 In the Old Faithful example (Section 3.3.3), calculate the 95% Wald CI for $\zeta = \mu - \nu$.

4.7 *Batch sampling.* Randomly selected soil samples are combined into batches and then the batches are tested for the presence or absence of a toxin. The observable is whether or not the batch contained toxin, i.e., soil samples are not tested individually. From 100 batches of 10 samples each, 12 tested positive.

a. Calculate an MLE of the probability p of the toxin in an individual sample of soil.

b. Determine the approximating normal distribution of $\widehat{p}$ and calculate an approximate 95% Wald confidence interval for p.

c. Determine the approximate 95% likelihood ratio CI for p.

Hint: You may find it easiest to parameterize the likelihood using the parameter $\theta = (1-p)^{10}$.

4.8 Assuming $\theta \in \mathbb{R}$, show that the Wald and LR test statistics for $H_0 : \theta = \theta_0$ are identical when the log-likelihood is quadratic.

4.9 For the binormal mixture model fitted to the Old Faithful waiting time data used in Section 4.7.5, implement a parametric bootstrap simulation to assess the behaviour of the likelihood ratio test statistic under the null hypothesis $H_0 : p = 0$. That is,

- Simulate data $\boldsymbol{y}^*_{(j)}, j = 1, ..., 1000$ from the fitted model under H_0, and calculate the LRT statistic of H_0 for each.
- Compare the actual value of the LRT to the simulated values.
- Plot the empirical distribution function of your simulated LRT statistics, and overlay the distribution function of a χ^2_6 random variable (by way of example, see Fig. 2 of McLachlan (1987)).

Note: for some simulated data, $\boldsymbol{y}^*_{(j)}$, it may be that the optimizer will not find a sensible local MLE. Your simulation must include the flexibility to manage this possibility should it arise.

Chapter 5

Maximizing the likelihood

5.1 Introduction

In the general case, finding an MLE will require numerical maximization of the log-likelihood $l(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$. One of the key steps in practical maximum likelihood inference is to determine the easiest way to implement this optimization. In the best case, with classes of model that are well established, there may be a well documented piece of software that will accomplish this task with minimal fuss. For example, fitting of generalized linear models (Chapter 7) is achieved using R function glm or SAS procedure GENMOD. For less standard models, it may require user specific implementation, perhaps via SAS procedures `NLMIXED` or `NLP`, or R function `optim`. In the toughest cases, it may even be that the likelihood can not be expressed explicitly (e.g., Chapter 10) and maximization could be a more challenging task.[1]

For some classes of models there are specialized optimization approaches that are known to work well. For example, generalized linear models are often fitted using iteratively re-weighted least squares, as is the case with the glm function in R. The `VGAM` package in R has extended the application of iteratively re-weighted least squares to a much wider range of models, including an assortment of zero-inflated models for count data. It has been seen that simple mixture models, such as the binormal mixture first encountered in Example 2.9, can quickly be fitted using `NLMIXED` or `optim` (Section 3.3.3). However, these general-purpose optimizers are

[1] This is the primary reason for the popularity of MCMC techniques.

unlikely to be successful when applied to mixture models with many components, especially if the component distributions are not well distinguished. For this purpose the EM (expectation maximization) algorithm (Section 5.3) is better suited.

In SAS 9.2, the `NLMIXED` procedure uses a quasi-Newton optimization algorithm (a variant of the Newton-Raphson algorithm, Section 5.2) by default. When the log-likelihood is sufficiently smooth then these algorithms work well and find the MLE quickly. They achieve this performance by utilizing the first and second derivatives of the log-likelihood. `NLMIXED` also provides several alternatives including conjugate gradient, double dogleg, Nelder-Mead, standard Newton-Raphson, and ridge-stabilized Newton-Raphson, and the choice can be specified using the `TECHNIQUE=` procedure option. The `NLP` procedure provides an even wider choice of algorithms.

In R, function `optim` uses the Nelder-Mead method by default. The Nelder-Mead algorithm is derivative free and uses an algorithm based on local extrapolation based on the log-likelihood evaluated at a set of neighbouring points. Other possibilities include conjugate gradient, quasi-Newton and simulated annealing. These are chosen using the `method=` argument. See Nocedal and Wright (2006) for more detailed description of optimization algorithms.

When using Newton-Raphson type algorithms, most optimizers can approximate the required first and second derivatives numerically. Some, including `optim` and `NLP` provide the facility for the user to provide one or both of these derivatives as analytical formulae. In `optim` this is via the `gr=` option, and in `NLP` through use of the `GRADIENT` and `HESSIAN` statements. ADMB uses a quasi-Newton optimizer, and exact (to computer precision) derivatives are algorithmically generated using automatic differentiation.

Section 5.2 gives a brief introduction to the Newton-Raphson algorithm. Section 5.3 demonstrates a version of the EM algorithm that is widely applicable in practice. The EM algorithm was brought to the widespread attention of the statistical community by Dempster, Laird and Rubin (1977), and Wu (1983) subsequently corrected some errors in that work regarding convergence of the EM algorithm. An extensive treatment of the EM algorithm and its subsequent extensions is provided by McLachlan and Krishnan (2008), while Navidi (1997) presents a gentler presen-

tation using a graphical representation of the algorithm for the case of maximizing the likelihood of a one-parameter model.

Section 5.4 looks at maximizing the likelihood in stages. It is seen that the use of profile likelihood can sometimes be an extremely efficient method for reducing the dimensionality of a numerical optimization (Section 5.4.1). In high dimensional problems it can be a challenge to obtain good starting values for an optimizer. One way to overcome this is to build up to maximization of the log-likelihood in steps (Section 5.4.2). Each step, other than the last, maximizes a reduced form of the log-likelihood for the purpose of proving good starting parameter values for the next step. The final step maximizes the full log-likelihood using the starting parameter values obtained from the penultimate step. This stepwise approach is a feature of ADMB, and a simple example is provided.

5.2 The Newton-Raphson algorithm

The Newton-Raphson algorithm is based on quadratic approximation of the objective function (the log-likelihood), via linear approximation of the score function $s(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}$. Given the current estimate $\boldsymbol{\theta}^{(k)} \in \mathbb{R}^s$ of $\widehat{\boldsymbol{\theta}}_n$ at iteration k of the algorithm, the Newton-Raphson algorithm approximates $s(\boldsymbol{\theta})$ using a Taylor's series expansion about $\boldsymbol{\theta}^{(k)}$,

$$s(\boldsymbol{\theta}) \approx s(\boldsymbol{\theta}^{(k)}) + \mathbf{H}(\boldsymbol{\theta}^{(k)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)})$$

where $\mathbf{H}(\boldsymbol{\theta}^{(k)})$ is the s by s Hessian matrix of second derivatives of $l(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^{(k)}$.

An MLE is obtained as a solution of the likelihood equation, $s(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. That is,

$$\mathbf{0} \approx s(\boldsymbol{\theta}^{(k)}) + \mathbf{H}(\boldsymbol{\theta}^{(k)})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(k)}) \ . \tag{5.1}$$

Solving for $\widehat{\boldsymbol{\theta}}$ in (5.1) provides the updating algorithm

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{H}(\boldsymbol{\theta}^{(k)})^{-1} s(\boldsymbol{\theta}^{(k)}) \ . \tag{5.2}$$

The Newton-Raphson algorithm (5.2) provides no guarantees that $l(\boldsymbol{\theta}^{(k+1)})$ will be greater than $l(\boldsymbol{\theta}^{(k)})$. Moreover, it can be the case that $\boldsymbol{\theta}^{(k+1)}$ from (5.2) is not

in the parameter space Θ. In practice, the algorithm is used in the modified form (Nocedal and Wright 2006, see Chap 3)

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \lambda_k \mathbf{H}(\boldsymbol{\theta}^{(k)})^{-1} s(\boldsymbol{\theta}^{(k)})$$

where $\lambda_k \in \mathbb{R}$ is the value that maximizes $l(\boldsymbol{\theta}^{(k+1)})$ over all possible values of λ. This is a one-dimensional optimization and is computationally fast.

The Newton-Raphson algorithm requires the computation and inversion of the $s \times s$ hessian matrix $\mathbf{H}(\boldsymbol{\theta}^{(k)})$ at each iteration and this can be computationally demanding in high dimensional models. For the most widely used matrix inversion algorithms, the computational demand of inverting an s by s matrix increases with the cube of s (Press et al. 2007). The quasi-Newton algorithm reduces the computational demand by using first derivative updates to the Hessian and its inverse at each iteration (Nocedal and Wright 2006).

Another variation of the N-R algorithm is given by replacing $\mathbf{H}(\boldsymbol{\theta})$ with its expected value for classes of models in which the formulae for the expected value of this matrix are known. This variation is called Fisher's method of scoring, which takes its name from the fact that the expected value of $\mathbf{H}(\boldsymbol{\theta})$ is the negative of the expected Fisher information matrix (see formula 11.20 in Chapter 11). Finally, note that the usual approximate variance matrix of $\widehat{\boldsymbol{\theta}}$, $\widehat{\mathbf{V}}$, is readily available from application of the N-R algorithm, because it is just the negative of the inverse of $\mathbf{H}(\widehat{\boldsymbol{\theta}})$ (Section 3.2.1).

5.3 The EM (Expectation - Maximization) algorithm

The EM algorithm can be used in situations where the statistical model can be posed as one where the observed data $\boldsymbol{y}$ are, in some sense, "incomplete". That is, where it can be conceived that there are underlying unobserved random quantities, $\boldsymbol{b}$, say. A surprisingly diverse variety of models can be expressed in this way (see Section 5.3.2).

In the form in which it is presented in Section 5.3.1, the EM algorithm is suitable when the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$ is unwieldy to work with (perhaps not even being

expressible in closed form), but the log-likelihood of the completed data $L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{b}^*)$ is relatively easy to maximize. Here $\boldsymbol{b}^*$ denotes an estimate of $\boldsymbol{b}$. Mixture models provide a good example. Their likelihood is moderately complex because it is obtained as a weighted average of the densities of each component distribution. For each observation, there is conceptually an unobserved random variable corresponding to the identity of the component distribution that generated that observation. If these random variables were to be observed then the model would cease to be a mixture model, but rather would just be a set of independent models for the samples from each of the component distributions. This example is developed more fully in Example 5.1.

Only brief account of the EM algorithm and its properties is given here. The text by McLachlan and Krishnan (2008) provides exhaustive coverage of the EM algorithm and its many variants.

5.3.1 The simple EM algorithm

The first step of the EM algorithm requires calculating the expected value of the completed-data log-likelihood, but it takes a much more intuitive form in many applications and it is this form that is presented here. This simplified form is appropriate when the log-likelihood $l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{b})$ of the completed data is linear in $\boldsymbol{b}$. With minor modification this can be weakened to the requirement that linearity is with respect to appropriate functions of the completed data (see Box 5.1).

The EM algorithm when the completed-data log-likelihood is linear in $\boldsymbol{b}$. Let $\boldsymbol{\theta}^{(k)}$ denote the value of $\boldsymbol{\theta}$ at iteration k.

Step 1 (Expectation). Under the complete-data model with parameter value $\boldsymbol{\theta}^{(k)}$, calculate $\boldsymbol{b}^*$ as the expected value of $\boldsymbol{b}$ given $\boldsymbol{y}$. That is,

$$\boldsymbol{b}^* = E_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{b}|\boldsymbol{y}) \ .$$

Step 2 (Maximization). Using the values of $\boldsymbol{b}^*$ calculated in Step 1, maximize the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{b}^*)$ to obtain a new parameter estimate $\boldsymbol{\theta}^{(k+1)}$. Return to Step 1.

Box 5.1.

If the completed data $\boldsymbol{y}^{(c)} = (\boldsymbol{y}, \boldsymbol{b})$ belong to the class of exponential family distributions of the form given by (14.4) then the log-likelihood is linear with respect to sufficient statistics $T_j(\boldsymbol{y}^{(c)})$ (where T_j is defined in Exercise 14.3 for the iid case). Consequently, the above EM algorithm is modified such that it is the expected values of $T_j(\boldsymbol{y}^{(c)})$ that are calculated in the expectation step.
The general version of the EM algorithm obtains $\boldsymbol{\theta}^{(k+1)}$ by maximizing $E_{\boldsymbol{\theta}^{(k)}}[\log L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{b})|\boldsymbol{y}]$. See Dempster et al. (1977) or Pawitan (2001) for more details.

Example 5.1. Mixture of two known distributions. Recall that the model for the Old Faithful geyser waiting times (Example 2.9) was a mixture whereby Y_i was observed from either a $N(\mu, \sigma^2)$ or $N(\nu, \tau^2)$ distribution according to an unobserved Bernoulli random variable, B_i. Since B_i is unobserved we can view the data as incomplete, and the complete data for each observation is the bivariate pair (Y_i, B_i). Here, the two components densities are assumed known (this assumption is dropped in Example 5.2) and will be denoted $f_0(y)$ and $f_1(y)$. This leaves only p to be estimated.

The complete-data pair (y, b) has density

$$f(y, b; p) = \begin{cases} pf_1(y), & b = 1 \\ (1-p)f_0(y), & b = 0 \end{cases} \tag{5.3}$$

and hence the marginal density for Y is

$$\begin{aligned} f(y; p) &= f(y, 1; p) + f(y, 0; p) \\ &= pf_1(y) + (1-p)f_0(y) \ . \end{aligned}$$

In Step 1 of the EM algorithm, the unobserved $b_i, i = 1, ..., n$ are replaced by their expected values. These expected values are real-valued quantities in the interval [0,1], and so it is necessary to extend the complete-data likelihood in (5.3) so that it is also defined for any value b between 0 and 1. A natural way to do this is to write the complete-data likelihood in the form

$$L(p; y, b) = f(y, b; p) = (pf_1(y))^b((1-p)f_0(y))^{(1-b)}. \tag{5.4}$$

Note that the log of this complete-data likelihood is linear in b, and hence the simple EM algorithm is applicable here.

EM step 1. Given $p^{(k)}$, we want to calculate $b_i^* = E(B_i|y_i), i = 1, ..., n$. Since B_i is Bernoulli,

$$\begin{aligned} b_i^* = E(B_i|y_i) &= P(B_i = 1|y_i) \\ &= \frac{f(y_i, 1)}{f(y_i; p^{(k)})} \\ &= \frac{p^{(k)} f_1(y_i)}{f(y_i; p^{(k)})} \ . \end{aligned}$$

EM step 2. If we denote $b^* = \sum b_i^*$ then, from equation (5.4), the likelihood from the completed data is

$$\begin{aligned} L(p; \boldsymbol{y}, \boldsymbol{b}^*) &= \prod_{i=1}^{n} f(y_i, b_i^*; p) \\ &= \prod_{i=1}^{n} \left((p f_1(y_i))^{b_i^*} ((1-p) f_0(y_i))^{(1-b_i^*)} \right) \\ &= p^{b^*} (1-p)^{(n-b^*)} \prod_{i=1}^{n} f_1(y_i)^{b_i^*} f_0(y_i)^{(1-b_i^*)} \ . \end{aligned} \tag{5.5}$$

In (5.5), parameter p occurs only in the term $p^{b^*}(1-p)^{(n-b^*)}$. To within a constant, this term is equivalent to the likelihood arising from a binomial(n, p) experiment in which $y = b^*$ is "observed" Here, b^* is not necessarily integer, but nonetheless, by analogy with the binomial likelihood it is immediate that (5.5) is maximized by $p^{(k+1)} = b^*/n$ That is, the EM algorithm is simply

$$p^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{p^{(k)} f_1(y_i)}{f(y_i; p^{(k)})}. \tag{5.6}$$

□

In the above example the Newton-Raphson algorithm would also be quite easy to implement. However, if the number of known components in the mixture distribution were increased to $k > 2$, then the EM algorithm would simply consist of $k - 1$ simple update formulae for the probabilities $p_1, p_2, ..., p_{k-1}$, all of similar

form to (5.6). In comparison, the computational demand of a conventional implementation of the Newton-Raphson algorithm would increase roughly at a rate proportional to the cube of k. Also, the Newton-Raphson algorithm may give infeasible solutions since it does not take into account that the p_i are bounded by 0 and 1. This could be circumvented by a suitable re-parameterization, such as using $\eta_i = \log(p_i/(1-p_k)), i = 1, ..., k-1$, However, this parameterization can be prone to numerical instability due to probabilities close to zero, and this would be particularly problematic for moderately high k. In contrast, note that equation (5.6) cannot give infeasible values of $p^{(k+1)}$.

Example 5.2. Mixture of two unknown normals.

In the binormal mixture model where μ, σ^2, ν and τ^2 are also unknown, the maximization step of the EM algorithm must be extended to also include maximization over these four parameters (Exercise 5.1). This results in the additional updating equations

$$\mu^{(k+1)} = \frac{\sum_{i=1}^n b_i^* y_i}{np^{(k+1)}}, \qquad \nu^{(k+1)} = \frac{\sum_{i=1}^n (1-b_i^*) y_i}{n(1-p^{(k+1)})}$$

and

$$\sigma^{2(k+1)} = \frac{\sum_{i=1}^n b_i^* (y_i - \mu^{(k+1)})^2}{np^{(k+1)}}, \qquad \tau^{2(k+1)} = \frac{\sum_{i=1}^n (1-b_i^*)(y_i - \nu^{(k+1)})^2}{n(1-p^{(k+1)})}.$$

□

5.3.2 Properties of the EM algorithm

The main advantage of the EM algorithm is its relative ease of implementation, especially in high-dimensional models. Under weak regularity conditions it can be shown that $L(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{y}) \geq L(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ with strict inequality unless $\boldsymbol{\theta}^{(k)}$ is a local maxima of the likelihood (Dempster et al. 1977). Under additional weak conditions, including the requirement that the likelihood is bounded on Θ, it can be shown that $\boldsymbol{\theta}^{(k)}$ converges to a local maximum of $L(\boldsymbol{\theta}; \boldsymbol{y})$ (Wu 1983)[2].

There are a diverse range of applications of the EM algorithm, including

[2]Wu (1983) corrects some invalid convergence statements made in Dempster et al. (1977)

- Situations where the data are grouped, truncated, or censored (e.g., see Exercise 5.6).

- In estimation of animal abundance (Section 6.4), where the number of unobserved animals is the missing observation (e.g., Van Deusen 2002).

- In unbalanced experimental designs, completing the design by filling in the missing cells (the E-step) results in a simpler completed-data likelihood.

- Random effects and state-space models. In general, the likelihood can not be expressed in closed form, but the completed-data likelihood, using the estimated random effects from the E-step, is usually straightforward.

- When multivariate data contain missing values, these can be imputed (the E-step), so that the completed data contain the full multivariate vector of observations on every individual. For example, see (Schafer 1997), R package `mirf`, and SAS procedure `MI`.

- In maximum likelihood modeling of evolutionary trees, the likelihood of observed DNA sequences is routine to compute if the evolutionary tree has been completed by specification of ancestry. For example, see Felsenstein (1981).

The chief drawback of the EM algorithm is that convergence is typically linear and can be very slow (see Box 5.2), especially once $\boldsymbol{\theta}^{(k)}$ gets close to $\widehat{\boldsymbol{\theta}}$. A further drawback is that the EM provides no direct way to obtain approximate standard

errors of the MLE (but see Section 5.3.4).

Box 5.2. Convergence rates of algorithms.

Suppose that the algorithmic sequence $\{\boldsymbol{\theta}^{(k)}\}$ converges to $\widehat{\boldsymbol{\theta}}$. The algorithm has convergence rate $q \geq 1$ if, for all k sufficiently large,

$$||\boldsymbol{\theta}^{(k+1)} - \widehat{\boldsymbol{\theta}}|| \leq C||\boldsymbol{\theta}^{(k)} - \widehat{\boldsymbol{\theta}}||^q \tag{5.7}$$

where $0 < C$ and $||\cdot||$ denotes Euclidean distance. If the value of $\widehat{\boldsymbol{\theta}}$ is to be determined to high accuracy then an algorithm with larger value of q will require fewer iterations. For sufficiently smooth log-likelihoods the Newton-Raphson algorithm converges at a quadratic rate, $q = 2$. The EM algorithm typically converges at a linear rate, $q = 1$, and the smallest value of C for which the above inequality holds may be only slightly less than unity (see Example 5.3), making it extremely slow to converge. In some situations the EM algorithm converges at a sublinear rate, which corresponds to $q = 1$, but with $C = 1$ being the smallest value of C for which (5.7) holds.

For the case $\theta \in \mathbb{R}$, and for sufficiently large k, the linear convergence of the EM algorithm is typically such that

$$\theta^{(k+1)} - \widehat{\theta} \approx a(\theta^{(k)} - \widehat{\theta}) \ . \tag{5.8}$$

It then follows that the increments, $\theta^{(k+1)} - \theta^{(k)}$, decrease in magnitude at the same rate. That is,

$$\theta^{(k+1)} - \theta^{(k)} \approx a(\theta^{(k)} - \theta^{(k-1)}) \ , \tag{5.9}$$

and hence a can be estimated from the ratio of increments. The linear rate of decrease in the increments is exploited in Section 5.3.3 for the purpose of accelerating the EM algorithm.

In the example below, the completed-data model is multinomial and hence has log-likelihood that satisfies the linearity requirement of the simple form of the EM algorithm.

Example 5.3. EM algorithm applied to a contingency table with grouped data.

The objective is to estimate the MLE's of row and column probabilities in a 2 by 3 contingency table in which rows and columns are independent. The unknown parameters are the probability of row 1, column 1 and column 2, denoted $\boldsymbol{\theta} =$

(r_1, c_1, c_2), respectively. For the sake of this example, 100 subjects are put into a 2 by 3 table, but the experimenter confused the counts in the [1,3] and [2,2] cells. Sixty five subjects were put into the other cells and so 35 subjects are known to have been assigned to the [1,3] and [2,2] cells.

The observed table was

10	5	
20		30

Here, the observed data are the observed cells counts $\boldsymbol{n} = (n_{11}, n_{12}, n_{21}, n_{22})$. The value in the unobserved cells will be denoted N_{13} and N_{22}, where the capitalization is used as a reminder that these remain random quantities by virtue of not being observed.

The expectation step of the EM algorithm requires determination of the expected values in the unobserved cells, with the expectation calculated using the current value of $\boldsymbol{\theta}$. Since the total count is 100, and the four observed counts sum to 65, it must be the case that $N_{13} + N_{22} = 35$. It can be shown (directly from the definition of conditional probability) that

$$N_{13}|\boldsymbol{n} \sim \text{Bin}(35, \rho) \ ,$$

where

$$\rho = p_{13}/(p_{13} + p_{22}) \ ,$$

and the cell probabilities are $p_{13} = r_1 c_3$, $p_{22} = r_2 c_2$ due to the assumption of row and column independence.

So, at iteration k with numerical estimates of the parameters $\boldsymbol{\theta}^{(k)} = (r_1^{(k)}, c_1^{(k)}, c_2^{(k)})$, the expectation step of the EM algorithm estimates the missing values as

$$\begin{aligned} n_{13}^* = E[N_{13}|\boldsymbol{n}] &= 35\rho^{(k)} \\ &= 35\frac{p_{13}^{(k)}}{p_{13}^{(k)} + p_{22}^{(k)}} \\ &= 35\frac{r_1^{(k)}(1 - c_1^{(k)} - c_2^{(k)})}{r_1^{(k)}(1 - c_1^{(k)} - c_2^{(k)}) + r_2^{(k)} c_2^{(k)}} \ , \end{aligned}$$

and of course $n_{22}^* = 35 - n_{13}^*$.

The maximization step simply calculates new values $r_1^{(k+1)}, c_1^{(k+1)}$, and $c_2^{(k+1)}$ as the row and column proportions from the completed contingency table. That is,

$$r_1^{(k+1)} = \frac{15 + n_{13}^*}{100} , \quad c_1^{(k+1)} = 0.3 , \quad c_2^{(k+1)} = \frac{n_{22}^* + 5}{100} . \tag{5.10}$$

The following R code implements 200 iterations of this simple algorithm, and prints out $\boldsymbol{\theta}^{(k)}$ for the first and last five of these iterations.

```
> #Observed data
> n11=10; n12=5; n21=20; n23=30
> #Initial parameter guesses
> r1=1/3; r2=1-r1; c1=1/3; c2=1/6; c3=1-c1-c2
> for(k in 1:200) {
+ # E step (apportions the 35 missing obs to the [1,3] and [2,2] cells)
+   p13=r1*c3;    p22=r2*c2
+   n13=35*p13/(p13+p22);   n22=35-n13
+ #M step
+   r1=(n11+n12+n13)/100; r2=1-r1
+   c1=(n11+n21)/100; c2=(n12+n22)/100; c3=1-c1-c2
+   if(k<6|k>195)
+     cat("\n","Iter",k,": P(Row 1)=",r1,", P(Col 1)=",c1,", P(Col 2)=",c2) }

 Iter 1 : P(Row 1)= 0.36 , P(Col 1)= 0.3 , P(Col 2)= 0.19
 Iter 2 : P(Row 1)= 0.3605505 , P(Col 1)= 0.3 , P(Col 2)= 0.1894495
 Iter 3 : P(Row 1)= 0.3610844 , P(Col 1)= 0.3 , P(Col 2)= 0.1889156
 Iter 4 : P(Row 1)= 0.361602 , P(Col 1)= 0.3 , P(Col 2)= 0.1883980
 Iter 5 : P(Row 1)= 0.3621036 , P(Col 1)= 0.3 , P(Col 2)= 0.1878964
 Iter 196 : P(Row 1)= 0.3749976 , P(Col 1)= 0.3 , P(Col 2)= 0.1750024
 Iter 197 : P(Row 1)= 0.3749977 , P(Col 1)= 0.3 , P(Col 2)= 0.1750023
 Iter 198 : P(Row 1)= 0.3749978 , P(Col 1)= 0.3 , P(Col 2)= 0.1750022
 Iter 199 : P(Row 1)= 0.3749979 , P(Col 1)= 0.3 , P(Col 2)= 0.1750021
 Iter 200 : P(Row 1)= 0.374998 , P(Col 1)= 0.3 , P(Col 2)= 0.1750020
```

The sequence $c_1^{(k)}$ converged on the value 0.3 at the first iteration, and $r_1^{(k)}$ and $c_2^{(k)}$ appear to be converging towards 0.375 and 0.175, respectively. It is left as an exercise (Exercise 5.4) to verify that the MLE is indeed $\boldsymbol{\theta} = (\widehat{r}_1, \widehat{c}_1, \widehat{c}_2) = (0.375, 0.3, 0.175)$. □

5.3.3 Accelerating the EM algorithm

The linear rate of convergence of the EM algorithm can be used to accelerate the algorithm by extrapolation. To see how this works, note that the convergence of $\boldsymbol{\theta}^{(k)}$ to $\widehat{\boldsymbol{\theta}}$ as $k \to \infty$ allows $\boldsymbol{\theta}^{(k)}$ to be expressed as

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(k)} + \sum_{j=k}^{\infty} (\boldsymbol{\theta}^{(j+1)} - \boldsymbol{\theta}^{(j)}) .$$

Restricting attention to $\theta \in \mathbb{R}$, it follows from the linear convergence in (5.9) that

$$\begin{aligned} \widehat{\theta} &\approx \theta^{(k)} + (\theta^{(k+1)} - \theta^{(k)}) \sum_{j=0}^{\infty} a^j \\ &= \theta^{(k)} + \frac{(\theta^{(k+1)} - \theta^{(k)})}{1-a} \ . \end{aligned} \tag{5.11}$$

Formula (5.11) is a form of Aitken acceleration and extension to the multi-parameter case is described in McLachlan and Krishnan (2008, Chap. 4).

Example 5.3 ctd.

Restricting attention to just the single parameter r_1, a plot of k versus $r_1^{(k)}$ shows the slow convergence to $\widehat{r}_1 = 0.375$ (Fig. 5.1). Furthermore, for large k it appears that

$$r_1^{(k+1)} - r_1^{(k)} \approx 0.955(r_1^{(k)} - r_1^{(k-1)}) \ . \tag{5.12}$$

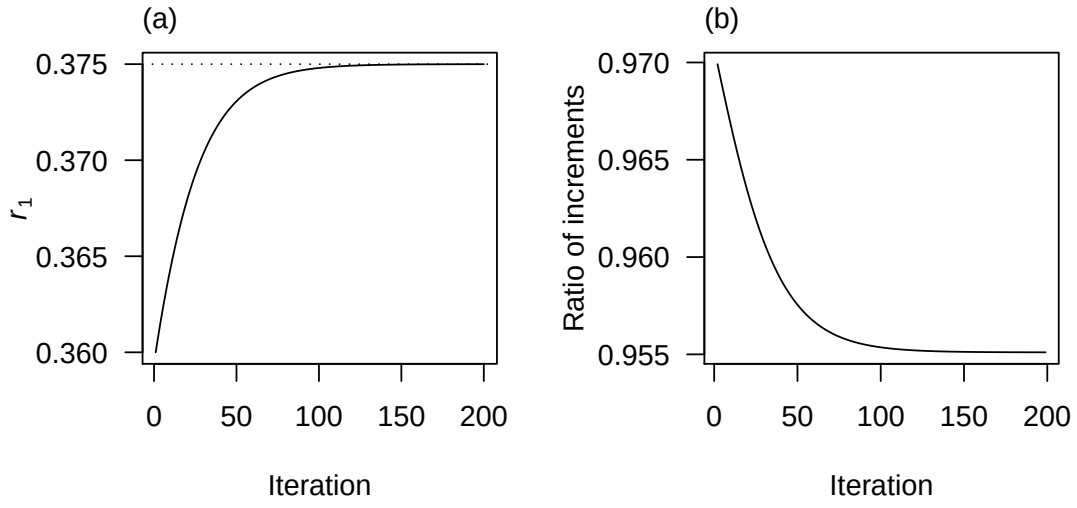


Figure 5.1: (a) The sequence $r_i^{(k)}, k = 1, ..., 200$. (b) Ratio of increments, $(r_1^{(k+1)} - r_1^{(k)})/(r_1^{(k)} - r_1^{(k-1)})$, $k = 1, ..., 199$.

Suppose that it is decided to apply acceleration after 50 iterations of the EM algorithm. The acceleration calculation requires use of the values $(r_1^{(48)}, r_1^{(49)}, r_1^{(50)}) = (0.372890, 0.372982, 0.373070)$. An estimate of the constant a in (5.9) is required, and this can be obtained as

$$a = \frac{r_1^{(50)} - r_1^{(49)}}{r_1^{(49)} - r_1^{(48)}} \approx 0.957 \ .$$

Using $k = 49$ in (5.11), the accelerated estimate is

$$r_1^{(49)} + \frac{r_1^{(50)} - r_1^{(49)}}{a} = 0.372982 + \frac{0.000088}{0.043} = 0.3750285 \ ,$$

which differs from $\widehat{r}_1$ by just 0.0000285. In comparison, the EM algorithm takes until iteration 144 to get this close to $\widehat{r}_1$. □

5.3.4 Inference from the EM algorithm

In the above examples, the MLE was obtained from application of simple updating formulae derived via the EM algorithm. However, in practice obtaining the MLE may not be enough to answer the questions of interest and it may be required to provide approximate standard errors and/or p-values and confidence intervals. There are a number of pragmatic approaches that can be taken.

If the likelihood for the observed (i.e., incomplete) data can be explicitly calculated then it may be possible to calculate the observed information matrix, thereby obtaining approximate standard errors. This calculation could be facilitated by use of optimizers such as `optim` and `NLMIXED`. In situations where these optimizers are very slow or unstable, it may nonetheless be the case that they will find the MLE if given a start value (obtained from the EM algorithm) that is sufficiently close to the MLE. When the observed-data likelihood is intractable then it may be possible to calculate the observed information matrix from the simpler complete-data likelihood (e.g., Louis 1982, Jamshidian and Jennrich 2000).

Alternatives include making inference using the bootstrap or likelihood-ratio methods (e.g., Exercise 5.5). These are likely to be the better approaches in practice, particularly because the EM algorithm is often used in complex models that may be highly parameterized, in which case approximate normality of the maximum likelihood estimator may be questionable.

5.4 Multi-stage maximization

The general idea of multi-stage maximization is to break the maximization into two (or more) maximizations of lower dimension. That is, if the s-dimensional vector

of parameters is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})^T$, then separate maximizations are done with respect to $\psi \in \mathbb{R}^r$ and $\boldsymbol{\lambda} \in \mathbb{R}^{s-r}$. The reader may recognize profile likelihood as an instance of this. Indeed, profile likelihood is revisited in Section 5.4.1 but from the viewpoint of the approach being a potential device for effective optimization rather than a convenient tool for inference on parameters of interest (as presented in Section 4.5).

Example 5.5 presents a variant of profile likelihood that maximizes an approximation to the likelihood, and hence the maximizing value of this approximate likelihood, $\widetilde{\boldsymbol{\theta}}$, will not in general equal $\widehat{\boldsymbol{\theta}}$. This approach is used in practice when it can be argued that the approximate likelihood effectively contains the relevant information contained in the true likelihood, so that it can be argued that $\widetilde{\boldsymbol{\theta}}$ will be negligibly different from $\widehat{\boldsymbol{\theta}}$. This approach wears a lot of different names within the published literature because it can be justified using arguments based on marginal or conditional likelihood as used in the presence of nuisance parameters (Section Z.ZZZ). Here, it will be called profile approximate-likelihood to reinforce the fact that this it is used as an optimization technique rather than a technique for mitigating the effects of nuisance parameters.

Section 5.4.2 introduces the idea of performing the optimization in phases. A simplified version of the desired model is initially fitted, whereby a subset of the parameters are fixed as constants. The subset of estimated parameters from the simplified model then provides good start values for the next phase of modeling, where the previously fixed parameters are now included as parameters to be estimated.

5.4.1 Efficient maximization via profile likelihood

With the parameter vector partitioned as $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, recall that the profile likelihood for $\boldsymbol{\psi}$ is defined as

$$l^*(\boldsymbol{\psi}; \boldsymbol{y}) \equiv \max_{\boldsymbol{\lambda}} \; l(\boldsymbol{\psi}, \boldsymbol{\lambda}; \boldsymbol{y}) \; .$$

This two-stage maximization provides a particularly effective means of obtaining $\widehat{\boldsymbol{\theta}}$ if $\boldsymbol{\psi}$ is of relatively low dimension, and the maximization with respect to $\boldsymbol{\lambda}$ is fast.

Profile likelihood allows for the speedy fitting of so-called transformation-regression models in which both the observations $\boldsymbol{y}$ and covariates may be subject to nonlinear transformation. A example of this is the linear regression model applied to a Box-Cox transformation of $\boldsymbol{y}$, and is presented in Section 6.2. Profile likelihood effectively reduces this to a one-dimensional numerical optimization. Similarly, fitting of negative binomial models within the GLM framework is accomplished by the fact that, for a fixed value of the shape parameter m, the negative binomial distribution is of exponential dispersion family form (see Exercise 7.3).

Example 5.4. Nonlinear (least-squares) regression model. Let each $Y_i, i = 1, ..., n$ be independently distributed as $\mathrm{N}(\mu_i(\boldsymbol{\beta}), \sigma^2)$, where each μ_i is differentiable with respect to $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$. The MLE of $\boldsymbol{\beta}$ is found using nonlinear least squares, which is implemented in `PROC NLIN` or the R function `nls`. In particular, `nls` has the facility to take advantage of any partial linearity (see Golub and Pereyra 1973, for details) in the specification of $\mu_i(\boldsymbol{\beta})$.

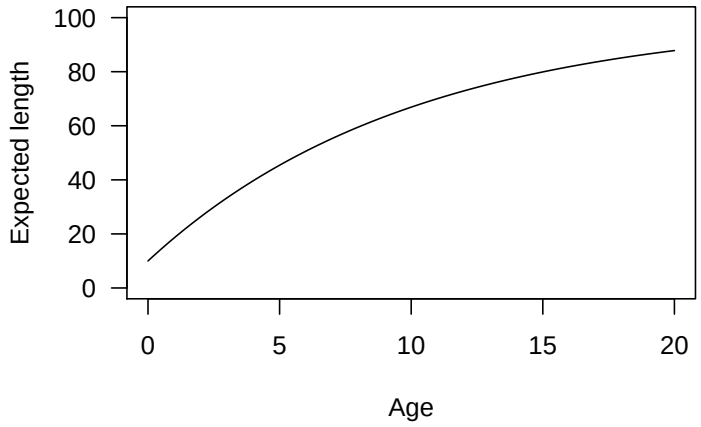


Figure 5.2: Von-Bertalanffy curve for parameter values $\boldsymbol{\beta} = (100, 90, 0.1)$.

For example, the 3-parameter von-Bertalanffy growth curve[3] for the expected length of an animal of age a_i can be written

$$\mu_i(\boldsymbol{\beta}) = \beta_1 - \beta_2 e^{-\beta_3 a_i} \ ,$$

[3]The von-Bertalanffy curve is commonly used to model the growth of fish.

where $\beta_i > 0,\ i = 1, 2, 3$. This is an increasing curve with upper asymptote of β_1 (Fig. 5.2).

If the value of β_3 is specified then the von-Bertalanffy curve has the form of a simple linear regression with respect to β_1 and β_2. Specifically, the expected length of animal i is

$$E[Y_i] = a + bx_i\ , \tag{5.13}$$

where $a = \beta_1,\ b = -\beta_2$, and $x_i = \exp(-\beta_3 a_i)$ is the explanatory variable. Thus, the maximizing values of β_1 and β_2 can be obtained explicitly for any given value of β_3.

The first call of `nls` in the R code below demonstrates its standard use. The second call takes advantage of the partial linearity of $\mu_i(\boldsymbol{\beta})$ and performs numerical optimization with respect to only β_3. The code segment `y~cbind(1,exp(-beta3*age))` is used to specify (5.13). That is, given β_3, the linear model to be fitted includes an intercept term and the explanatory variable $\exp(-\beta_3 a_i)$.

```
#Data frame vonB contains variables age and y (animal length)
#Standard use of nls
nls(y~beta1-beta2*exp(-beta3*age),data=vonB,
       start=list(beta1=90,beta2=80,beta3=0.2))

#Taking advantage of partial linearity
nls(y~cbind(1,exp(-beta3*age)),data=vonB,
       start=list(beta3=0.2),algorithm="plinear")
```

□

The next example demonstrates the profiling of an approximation to the likelihood.

Example 5.5. Copulas. Copula models are a popular class of models for specifying dependency structure in multivariate data, $\boldsymbol{y}_i = (y_{i1}, ..., y_{im})^T \in \mathbb{R}^m, i = 1, ..., n$. They are especially relevant to non-normal data where traditional multivariate-normal theory does not apply. See Trivedi and Zimmer (2007) for a comprehensive introduction to copulas.

The versatility of the copula model is achieved from the property that the distribution of $\boldsymbol{Y}_i$ can be derived from specification of a joint distribution on the m-dimensional unit cube, $[0, 1]^m$, and specification of the m marginal distributions

$F_{i1}(y_{i1}), ..., F_{im}(y_{im})$. The joint distribution specified on $[0,1]^m$ is required to have $U(0,1)$ marginal distributions. This joint distribution is the copula, and will be denoted $C(u_1, ..., u_m)$, with corresponding density denoted $c(u_1, ..., u_m)$. Let $\boldsymbol{\lambda}$ denote the parameters of the marginal distributions $F_{im}, i = 1, ..., m$, and $\boldsymbol{\psi}$ denote the parameters of the copula. Assuming independence of $\boldsymbol{Y}_i$, the log-likelihood has the form

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \log f_{ij}(y_{ij}; \boldsymbol{\lambda}) \tag{5.14}$$

$$+ \sum_{i=1}^{n} \log c(F_{ij}(y_{ij}; \boldsymbol{\lambda}), ..., F_{im}(y_{im}; \boldsymbol{\lambda}); \boldsymbol{\psi}) \ , \tag{5.15}$$

where (5.14) arises from the marginal distributions of the m elements of $\boldsymbol{y}_i$, and hence depends only on parameters $\boldsymbol{\lambda}$. The term in (5.15) is due to the copula distribution specified on $[0,1]^m$, and contains the information about the copula parameters $\boldsymbol{\psi}$.

Direct maximization of the log-likelihood generally requires an s-dimensional numerical optimization to determine the MLE $\widehat{\boldsymbol{\theta}}$. This can be challenging when s is large, however, an approximate MLE, $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\psi}}, \widetilde{\boldsymbol{\lambda}})$ can more readily be obtained via an approximation of the profile likelihood. This approximation obtains $\widetilde{\boldsymbol{\lambda}}$ by maximization over the (5.14) term only. This is typically relatively straightforward because this term is equivalent to the log-likelihood from assuming all Y_{ij} are independent. The estimate $\widetilde{\boldsymbol{\psi}}$ is then obtained from maximization of $\log c(F_{ij}(y_{ij}; \widetilde{\boldsymbol{\lambda}}), ..., F_{im}(y_{im}; \widetilde{\boldsymbol{\lambda}}); \boldsymbol{\psi})$. This is commonly known as the inference function for margins (IFM) approach and it can be shown that the estimator $\widetilde{\boldsymbol{\theta}}$ is a consistent estimator of the unknown $\boldsymbol{\theta}_0$, but is less efficient than the MLE (Joe and Xu 1996). See Yan (2007) for demonstration of fitting copulas using the `copula` package. □

5.4.2 Multi-stage optimization

The successful global maximization of a likelihood often requires specification of a good starting value $\boldsymbol{\theta}^{(0)}$. In simple models this can often be deduced from graphical inspection of the data. However, in high-dimensional cases this may not be possible,

and this section presents a simple form of multi-stage optimization that obtains a value of $\boldsymbol{\theta}^{(0)}$ from a sequence of partial optimizations. This section concludes with a brief description of the automatic implementation of multi-stage optimization within the ADMB software.

For convenience, two-stage optimization is presented, since this permits use of the existing notation whereby the parameter vector is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$. At the first stage, a restricted form of the log-likelihood is optimized with respect to $\boldsymbol{\lambda}$ only. Let $\widehat{\boldsymbol{\lambda}}^*$ denote the resulting maximizing value of $\boldsymbol{\lambda}$. The second stage performs a full maximization of the log-likelihood, using the start value $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\psi}^{(0)}, \widehat{\boldsymbol{\lambda}}^*)$, where $\boldsymbol{\psi}^{(0)}$ denotes suitable start values for this subset of parameter values.

It is often convenient to implement the first-stage maximization by simply maximizing the log-likelihood $l(\boldsymbol{\theta})$ with $\boldsymbol{\psi}$ fixed at values $\boldsymbol{\psi}^{(0)}$. Then the first stage optimization has the form profile likelihood form

$$\max_{\boldsymbol{\lambda}} l(\boldsymbol{\psi}^{(0)}, \boldsymbol{\lambda}) \ ,$$

from which $\widehat{\boldsymbol{\lambda}}^*$ is obtained. However, any sensible modification of $l(\boldsymbol{\theta})$ can be used. For example, it may be the case that some component of the log-likelihood contains most of the relevant information about $\boldsymbol{\lambda}$, and this would be a suitable candidate if it were convenient to maximize with respect to $\boldsymbol{\lambda}$ (with other parameters held fixed if present).

Example 5.6. Two-stage optimization of the copula log-likelihood. The copula log-likelihood in Example 5.5 is the sum of terms (5.14) and (5.15). The first term involves only $\boldsymbol{\lambda}$, and maximization of this term gives the so-called IFM estimate of $\boldsymbol{\lambda}$, which would be denoted $\widehat{\boldsymbol{\lambda}}^*$ using the above notation. In a two-stage optimization of the full log-likelihood, $\widehat{\boldsymbol{\lambda}}^*$ is a natural start value for $\boldsymbol{\lambda}$. □

The extension to three or more stages is immediate. Specifically, the parameter vector $\boldsymbol{\theta}$ is partitioned into a collection of parameter subsets and these are sequentially added in a series of restricted optimizations. At each stage, the maximizing values of those parameters from the previous stage are used as initial values.

Multi-stage optimization in ADMB

Multi-stage optimization is incorporated within ADMB, where the terminology "multi-phase" is used instead of "multi-stage". In the `PARAMETER_SECTION`, an optional integer value can be associated with the declaration of a parameter, to indicate the stage at which the parameter enters the optimization. Within the `PROCEDURE_SECTION`, program code is provided to return the value of the objective function to be minimized at each stage. This is enabled by use of the `current_phase()` function, which returns the integer value of the stage currently being processed. See ADMB-project (2008a) for full details.

The toy example below demonstrates multi-stage maximization of the log-likelihood from iid $Y_i \sim N(\mu, \sigma^2)$ data. The declaration `init_number mu(1)` specifies that parameter μ is to be included at the first stage, and the objective function at this stage is just the sum of squared residuals. The declaration of the initial bounded number `sigma(0,100,2)` specifies that σ is bounded between 0 and some large upper bound (here taken to be 100), and is to be made active at the second stage. The second-stage objective function is the full negative log-likelihood (notwithstanding that the constant terms have been omitted).

```
DATA_SECTION
  init_int n
  init_vector y(1,n)

PARAMETER_SECTION
  init_number mu(1)                    //Phase 1 parameter
  init_bounded_number sigma(0,100,2) //Phase 2 parameter
  objective_function_value f           //Negative log-likelihood

PROCEDURE_SECTION
  int i;
  f=0.0;
  f=f+norm2(y-mu);  //norm2() returns the sum-of-squares
  if(current_phase()==2) f=n*log(sigma)+f/(2*sigma*sigma);
```

5.5 Exercises

5.1 Obtain the updating formulae in Example 5.2 by maximizing the complete data likelihood $\prod_{i=1}^{n} f(y_i, b_i^*; p, \mu, \sigma, \nu, \tau)$. (This likelihood has the form given by equation (5.5), but is now a function of all five parameters.)

5.2 Using R or SAS, apply the EM algorithm of Example 5.2 to the Old-Faithful waiting time data (Figure 2.5), and determine (approximately) the value of a in (5.9).

5.3 The following question is based on the lead example used in Dempster et al. (1977), and is derived from a model for recombination in gene mapping studies (e.g., see Lange 2002).

a. Let $(X_1, X_2, X_3, X_4, X_5)$ be a multinomial random vector with cell probabilities $(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4})$. Given $(x_1, x_2, x_3, x_4, x_5)$ show that the maximum likelihood estimate of θ is

$$\widehat{\theta} = \frac{x_2 + x_5}{x_2 + x_3 + x_4 + x_5}.$$

b. Suppose now that the observations in cells 1 and 2 are pooled, so that we observe only the incomplete data $(y_1, y_2, y_3, y_4) = (x_1 + x_2, x_3, x_4, x_5)$. Given the values $(y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$, use the result of part a. to apply the the EM algorithm for calculation of $\widehat{\theta}$. Perform three iterations of the EM algorithm by hand, starting with the initial value $\theta^{(0)} = 0.5$.

5.4 For Example 5.3, verify that $\widehat{\boldsymbol{\theta}} = (\widehat{r}_1, \widehat{c}_1, \widehat{c}_2) = (0.375, 0.3, 0.175)$ is the MLE using one of the following methods.

a. Show that $\widehat{\boldsymbol{\theta}}$ is the fixed point solution of the EM algorithm. That is, show that if $\boldsymbol{\theta}^{(k)} = (0.375, 0.3, 0.175)$ then $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$. (From the convergence results in Wu (1983), it follows that this is the unique MLE.)

b. Determine the likelihood for the the observed (incomplete) data and maximize using R function `optim` or SAS procedure `NLMIXED`.

5.5 In Example 5.3 the observed (incomplete) data $y = c(10, 5, 20, 30, 35)$ are multinomial, denoted Mult$(5, p_1, p_2, ..., p_5)$) where $p_5 = 1 - \sum_{i=1}^{4} p_i$, and hence the incomplete likelihood can be explicitly calculated.

a. Under the null hypothesis of row and column independence, the MLE's of $(p_1, p_2, ..., p_5)$ are obtained from $(\widehat{r}_1, \widehat{c}_1, \widehat{c}_2) = (0.375, 0.3, 0.175)$. Calculate the maximized log-likelihood under this hypothesis.

b. Calculate the maximized log-likelihood without restriction on the parameter space. That is, assuming y is an observation from a Mult$(5, p_1, p_2, ..., p_5)$ distribution.

c. Use the G-test to obtain a p-value for the null hypothesis of row and column independence.

5.6 Suppose that $Y_i, i = 1, 2, 3, 4$ are iid from an exponential distribution with mean μ. However, only $y_1 = 5, y_3 = 12$ and $y_4 = 17$ are observed. The value of y_2 is not observed, but it is known that $y_2 > 10$.

The data are incomplete because y_2 is not observed. The data can be completed by estimating y_2 in the E-step, in which case the completed data model is simply that $y_i, i = 1, ..., 4$ are iid observations from the $Exp(\mu)$ distribution. Note that the completed data likelihood is linear in y_2, and hence the simple version of the EM algorithm is applicable.

a. Use the lack-of-memory property of the exponential distribution to construct the E-step. This property states that, if Y has an exponential distribution then, for any $s \geq 0$,

$$E[Y \mid Y > s] = s + E[Y].$$

(See also Exercise 6.1).

b. Choose a start value $\mu^{(0)}$ and apply two iterations of the EM algorithm.

c. Implement Aitken acceleration using formula (5.11) for $k = 1$.

d. Regardless of your choice of $\mu^{(0)}$, the valued of μ obtained after Aitken acceleration should be $\mu = 14\frac{2}{3}$. Verify that this is the MLE by showing that it is the fixed point solution to the EM-algorithm. That is, if $\mu^{(k)} = 14\frac{2}{3}$, then $\mu^{(k+1)} = 14\frac{2}{3}$ also.

5.7 Let $\boldsymbol{Y} = Y_1, ..., Y_n$ be iid observations from a zero-inflated Poisson distribution (see Exercise 2.11). That is, with probability p, Y necessarily takes the value 0, and with probability $(1 - p)$, Y is observed from a Poisson(λ) distribution.

a. Determine the updating formulae of the EM algorithm for calculation of $(\widehat{p}, \widehat{\lambda})$.

b. Use R or SAS to apply the updating formulae to the apple micro-propagation data in Exercise 3.7.

Chapter 6

Some widely used applications of ML

6.1 Introduction

This Chapter presents three examples chosen from the diverse myriad of applications of maximum likelihood inference. The aim is to give a flavour for the nature of the likelihoods used in each of these applications, and some very simple examples of their use.

Section 6.2 presents the Box-Cox transformation as a very elegant application of profile likelihood. The total number of model parameters may be large, but numerical optimization is required over only the sole parameter that determines the optimal power transformation. Section 6.3 looks very briefly at survival analysis. A notable feature of likelihoods constructed from survival data is that they contain terms arising from both continuous and discrete forms of data. Section 6.4 presents some simple forms of mark-recapture models. These models have the unusual feature that the parameters of primary interest correspond to the unknown number of animals in a population, and therefore are integer valued.

6.2 Box-Cox transformations

The method of Box and Cox (1964) can be useful for the modeling of non-negative continuous data that violate the normality and homogeneous variance assumptions required of the normal linear regression model (Example 11.6). Rather, it is assumed

that a suitable transformation of the data can be found such that the transformed data will satisfy the assumptions of a normal linear model.

Box 6.1.

As a general rule, one should always attempt to use a model that is appropriate for the measured data without any need for their manipulation or transformation. Indeed, the poison data that are used in Section 6.2.1 are from an example in Box and Cox (1964), but these data (Table 6.1) can now be analyzed directly (Exercise 6.3) using the techniques in Section 6.3. Nonetheless, there are many situations in which continuous data do not conform to any standard model and for which the Box-Cox transformation may be necessary.

The standard Box-Cox transformation is of the form

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ \log(y_i) & , \lambda = 0 \end{cases}$$

for $\lambda \in \mathbb{R}$. For the intended purpose of linear modeling of the transformed data, it is equivalent to use the simpler power transformation

$$\zeta_i^{(\lambda)} = \begin{cases} y_i^\lambda & , \lambda \neq 0 \\ \log(y_i) & , \lambda = 0 \end{cases} .$$

However, the Box-Cox form has the advantage that it can be considered a smooth family of transformations with respect to λ, since the log function is the limit of the transformation $\frac{y_i^\lambda - 1}{\lambda}$ for λ sufficiently close to zero[1].

Under the Box-Cox model, the transformed data $\boldsymbol{y}^{(\lambda)}$ are assumed to follow a normal linear model. That is, the transformed observation $y_i^{(\lambda)}$ is assumed to be observed from a $\mathrm{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$ distribution where $\boldsymbol{x}_i$ is a vector of known covariates and $\boldsymbol{\beta} \in \mathbb{R}^p$ and σ^2 are to be estimated. The parameters to be estimated are therefore $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \lambda)$.

In Box 13.1 it was noted that maximum likelihood estimation is invariant to transformation of the data, but this assumes that the transformation is fixed and not a function of unknown parameters. This is not the case here because the parameter λ determines which transformation from the Box-Cox family is to be used. It is therefore necessary to work with the likelihood of the raw data, and this is obtained

[1] This limit result can be established by application of L'Hôpital's rule. See, for example, Stewart (1999)

from the assumed normal linear regression model for $y_i^{(\lambda)}$. Using the transformation of variables formulae (Section 13.2), the density function for y_i is

$$f(y_i) = f(y_i^{(\lambda)}) \left| \frac{dy_i^{(\lambda)}}{dy_i} \right| = f(y_i^{(\lambda)}) y_i^{(\lambda-1)} \ .$$

and the log-likelihood for the observed data $\boldsymbol{y}$ is therefore

$$l(\boldsymbol{\beta}, \sigma^2, \lambda) = l^{(\lambda)}(\boldsymbol{\beta}, \sigma^2) + (\lambda - 1) \sum_i^n \log(y_i)$$

where

$$l^{(\lambda)}(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\boldsymbol{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})^T (\boldsymbol{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})$$

Note that for any fixed value λ^*, $l(\boldsymbol{\beta}, \sigma^2, \lambda^*)$ is maximized with respect to $\boldsymbol{\beta}$ and σ^2 by maximizing $l^{(\lambda^*)}(\boldsymbol{\beta}, \sigma^2)$. Maximization of $l^{(\lambda^*)}(\boldsymbol{\beta}, \sigma^2)$ is easily obtained from the least-squares fit of $\mathbf{X}$ on $\boldsymbol{y}^{(\lambda^*)}$. This makes it computational easy to produce a likelihood profile plot for λ. In practice, one uses this plot to suggest a "convenient" value of λ. For example, if $\widehat{\lambda}$ is close to zero (i.e., the 95% CI for λ includes zero) then a log transformation might be chosen.

6.2.1 Example: The Box and Cox poison data

Example 1 of Box and Cox (1964) uses a two-way ANOVA to model the survival time of animals[2] after being administered with a poison. Each of 48 animals receives one of three poisons and one of four treatments according to a replicated two-way design. The data are the hours until death of the animals (Table 6.1).

These data were entered into a tab-delimited text file, `SurvTimes`, in standard rectangular form where each row corresponds to one observation. The first five lines of this file are:

```
poison  trmt    time
1        A       3.1
1        A       4.5
1        A       4.6
1        A       4.3
1        B       8.2
```

[2]Box and Cox do not disclose the source of the data, or the species of animal involved.

Poison 1				Poison 2				Poison 3			
Treatment				Treatment				Treatment			
A	B	C	D	A	B	C	D	A	B	C	D
3.1	8.2	4.3	4.5	3.6	9.2	4.4	5.6	2.2	3.0	2.3	3.0
4.5	11.0	4.5	7.1	2.9	6.1	3.5	10.2	2.1	3.7	2.5	3.6
4.6	8.8	6.3	6.6	4.0	4.9	3.1	7.1	1.8	3.8	2.4	3.1
4.3	7.2	7.6	6.2	2.3	12.4	4.0	3.8	2.3	2.9	2.2	3.3

Table 6.1: Survival times (hr) after poisoning of an unidentified species of animal. Data are from Box and Cox (1964).

The R and SAS programs below both specify the profile log-likelihood to be calculated for λ from -2 to 2, in increments of 0.01.

Using R

The `boxcox` function belongs to the `MASS` library.

```
> library(MASS)
> bc.fit=boxcox(time~as.factor(poison)*trmt,lambda=seq(-2,2,0.01),data=SurvTimes)
> lambdahat=bc.fit$x[bc.fit$y==max(bc.fit$y)]
> cat("\n MLE of lambda is",lambdahat)

 MLE of lambda is -0.82
```

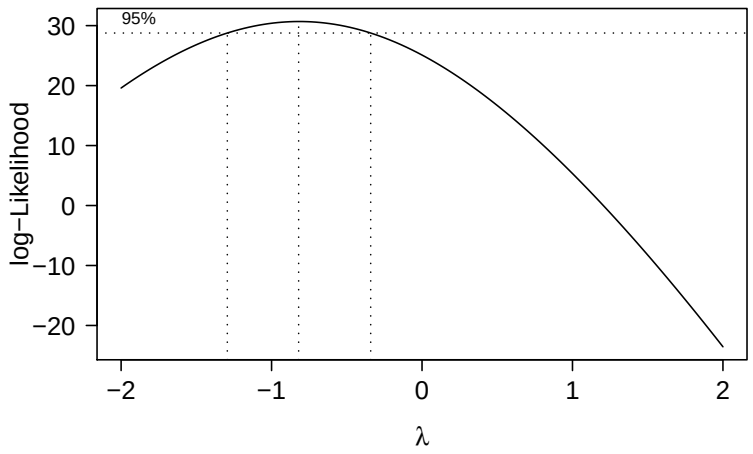


Figure 6.1: The profile log-likelihood plot for λ produced by the `boxcox` function.

Using SAS

`PROC TRANSREG` (transformation regression) is capable of many types of transformations, including splines and alternating least squares In the `MODEL` statement

used below, the `CLASS` keyword is used to expand the variables to form the familiar linear model design matrix.

```
PROC TRANSREG DATA=SurvTimes;
  MODEL BOXCOX(time / LAMBDAS=-2 to 2 by 0.01)=CLASS(poison|trmt);
```

The likelihood profile confidence interval suggests that a reciprocal transformation would be appropriate. Box and Cox (1964) note that the reciprocal transformation "has a natural appeal for the analysis of survival times since it is open to the simple interpretation that it is the rate of dying which is to be considered".....which leads us to the next application of ML inference.

6.3 Models for survival data

The general objective is to model the time until "failure" of a component or individual. This could be the lifetime of a piece of equipment (e.g., lightbulb, bearing, microchip etc), or duration until death, or duration for which a disease remains in remission after treatment. Typically, we wish to make inference about the effect of covariates $\boldsymbol{x}_i$ associated with individual i.

The method of analysis must be able to accommodate censored data. Such data commonly arise due to an individual being removed from the study prior to failure. In a clinical trial this could happen if a patient moves address and is no longer available for observation, or is "removed" for some other reason (such as being hit by a bus!). Also, the study will typically be of fixed duration and there may be individuals who have not failed by the time the study is concluded. In such cases, the actual failure time of a censored individual is not observed, but it is known that the individual had not failed at the time of censoring. This is known as right censoring. Other forms are possible, including left censoring where it is known only that an individual failed prior to a certain time, and interval censoring where the failure is known to have occurred within some interval of time. For simplicity, only right censoring is considered here. Furthermore, it is assumed that individuals are independent, and that the censoring process is independent of survival times.

Let t_i be the time at which individual i was either observed to fail, or right

censored, and let w_i be the indicator variable for the type of observation,

$$w_i = \begin{cases} 1, & \text{failure at time } t_i \\ 0, & \text{right censoring at time } t_i \end{cases} . \tag{6.1}$$

The data observed on individual i is the pair (t_i, w_i). The corresponding contribution to the likelihood function is therefore the joint density function of this pair, regarded as a function of the model parameters. However, under the assumption that censoring times are independent of survival times, it follows[3] that, for the purpose of modeling survival times, no relevant information is lost by treating the w_i as fixed. That is, the fact that the w_i may actually be realizations of a Bernoulli random variable can be ignored, and they can be treated as fixed covariates. The likelihood function is therefore obtained solely from the observed times, $\boldsymbol{t} = (t_1, ..., t_n)$, at which failure or censoring occurred.

If the event recorded on individual i is a failure ($w_i = 1$), then t_i is the failure time and this observation contributes $f_i(t_i; \boldsymbol{\theta})$ to the likelihood, where $f_i(t; \boldsymbol{\theta})$ denotes the density function for the failure time of individual i. However, if the event recorded at time t_i is a censoring event ($w_i = 0$), then it is the case that individual i had not yet failed at time t_i, but was removed from the experiment at that time. Then, all that is known is that the realized value of the survival time is in excess of time t_i at which the censoring occurred. In effect, the censoring at time t_i corresponds to observing the event that $T_i > t_i$, and the contribution to the likelihood is the probability of this event. Letting $F_i(t; \boldsymbol{\theta})$ denote the distribution function for the survival time of individual i, this probability is $1 - F_i(t_i; \boldsymbol{\theta})$.

The function $S_i(t; \boldsymbol{\theta}) = 1 - F_i(t; \boldsymbol{\theta})$ is called the survivor function because it gives the probability of surviving beyond time t. From above, observed times of failure contribute $f_i(t_i; \boldsymbol{\theta})$ to the likelihood, and right-censored times contribute $S_i(t_i; \boldsymbol{\theta})$ Thus, the log-likelihood function for right-censored data can be written

$$l(\boldsymbol{\theta}; \boldsymbol{t}) = \sum_{i=1}^{n} \{w_i \log f_i(t_i; \boldsymbol{\theta}) + (1 - w_i) \log S_i(t_i; \boldsymbol{\theta})\} . \tag{6.2}$$

This likelihood is used by the parametric survival time models developed in Sections 6.3.1 and 6.3.2. In contrast Section 6.3.3 employs a semi-parametric likelihood that

[3] From application of the conditionality principle in Section 14.3.

is obtained using arguments that are presented in Section 9.ZZZ.

The coverage of survival models in the next three sections is necessarily brief, and does not cover model assessment or comparison. Also, it is assumed that the covariates $\boldsymbol{x}_i$ associated with individual i are constant over time. This would not be a valid assumption in a long-term health study in which, say, body-mass index (BMI) was used as a covariate. For greater coverage of the modeling of survival data, the reader is referred to Kleinbaum and Klein (2005). Also, survival analysis using R is a topic covered in Everitt and Hothorn (2006), and using SAS by Der and Everitt (2009).

6.3.1 Accelerated failure time model

The accelerated failure time model (AFT) can intuitively be described using the popular convention that one dog year is equivalent to seven human years. That is, if $F_h(t)$ and $F_d(t)$ are the distribution functions for the lifespan of humans and dogs, respectively, this convention assumes

$$F_d(t) = F_h(7t) \ .$$

For example, the probability that a dog dies by the age of 10, $F_d(10)$, is equal to the probability that a human dies by the age of 70, $F_h(70)$.

The AFT model has the form

$$F_i(t) = F(\gamma_i t; \boldsymbol{\psi}) \ ,$$

where γ_i depends on covariates $\boldsymbol{x}_i$ associated with individual i, and F (with no subscript) denotes the "baseline" distribution function determined by parameters $\boldsymbol{\psi}$. The multiplier γ_i is positive, and a natural choice is to model $\log(\gamma_i)$ as a linear function of $\boldsymbol{x}_i$. This can be written

$$F_i(t; \boldsymbol{\theta}, \boldsymbol{x}_i) = F(\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})t; \boldsymbol{\psi}) \ ,$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi})$. Using this notation, the AFT model to compare dog longevity to human longevity could be expressed

$$F_d(t) = F_h(\exp(\beta)t) \ ,$$

where F_h is taken to be the baseline distribution. The popular convention corresponds to $\beta = \log(7) \approx 1.95$.

Box 6.2.

The AFT model that arises from the popular convention that a dog year is equivalent to seven human years is not a good model. With respect to physiology, the first year of a dog's life is equivalent to about 15 human years, and later years are equivalent to about 4 or 5 human years.

The Weibull distribution (see Exercise 6.4) is a popular choice for the form of the baseline failure time distribution, F. Other available choices include exponential, lognormal, log-logistic, and generalized gamma.

6.3.2 Parametric proportional hazards model

The hazard function at time t is $h(t) = f(t)/S(t)$, and is the instantaneous rate of failure at that time (Box 6.3). It many ways it is much more intuitive to think about the hazard function than the failure time distribution. By way of example, the hazard function of humans is reasonably well modeled by the bathtub shaped hazard function (Fig. 6.2). This is of course a gross simplification, and one of its omissions is that it does not show the temporary increase in hazard that occurs when the legal age of driving is first reached. In medical studies it may be natural to assume that a successful treatment results in an immediate reduction in hazard. There is intuitive appeal in modeling survival data via specification of a hazard function, and this section shows how the log-likelihood for survival data can be constructed from a specified hazard function.

Box 6.3.

To see that the hazard function, $h(t) = f(t)/S(t)$, is the instantaneous risk at time t, let Δt denote some small interval of time, and note that

$$h(t)\Delta t = \frac{f(t)\Delta t}{1 - F(t)} = \frac{F'(t)\Delta t}{1 - F(t)} \approx \frac{F(t + \Delta t) - F(t)}{1 - F(t)} \; .$$

That is $h(t)\Delta t$ is the probability of failing in the the interval $(t, t + \Delta t]$, given survival to time t.

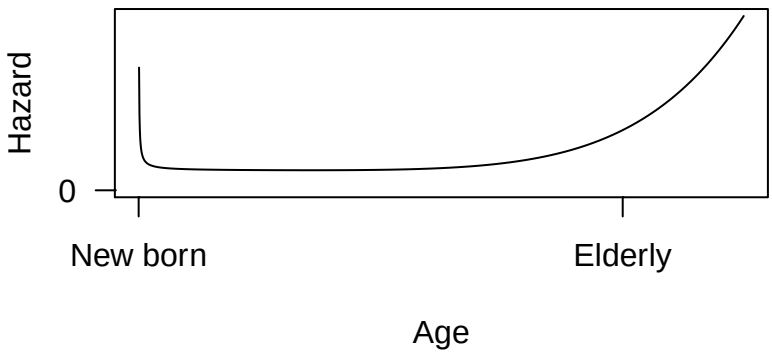


Figure 6.2: A typical bathtub hazard function for humans.

In the proportional hazards model the hazard functions of any two individuals are assumed to be proportional, with the constant of proportionality being constant over time and depending only on the (fixed) covariates. The usual formulation for the hazard function of individual i is

$$h_i(t;\boldsymbol{\lambda},\boldsymbol{\beta}) = h_0(t;\boldsymbol{\lambda})\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \ , \tag{6.3}$$

where $h_0(t;\boldsymbol{\lambda})$ is the baseline hazard function and depends on parameters $\boldsymbol{\lambda}$. Note that the hazard depends on time only through $h_0(t;\boldsymbol{\lambda})$, and hence the ratio of hazards for any two individuals is a constant.

To compute the log-likelihood function in (6.2) it is necessary to determine both the survival time density function, $f_i(t)$, and the survivor function, $S_i(t) = 1 - F_i(t)$, from specification of the hazard function. The survivor function can be obtained from the hazard function by noting that

$$\begin{aligned} h_i(t) = \frac{f_i(t)}{S_i(t)} &= \frac{\frac{\partial F_i(t)}{\partial t}}{1 - F_i(t)} \\ &= \frac{-\partial \log(S_i(t))}{\partial t} \ . \end{aligned}$$

After re-arrangement and integration this gives

$$\begin{aligned} S_i(t) &= \exp\left(-\int_0^t h_i(u)du\right) \\ &= \exp\left(-\int_0^t h_0(u)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})du\right) \\ &= \exp\left(-H_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\right) \ , \end{aligned}$$

where $H_0(t) = \int_0^t h_0(u)du$ is the cumulative baseline hazard function. With $S_i(t)$ obtained as above, and $f_i(t)$ given by $f_i(t) = h_i(t)S_i(t)$, the log-likelihood function in expression (6.2) can be computed and maximized with respect to model parameters $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\beta})$.

Example 6.1. Exponential survival times. Suppose that an individual has a hazard function that is constant over time, $h(t) = \lambda$. The survivor function is then

$$S(t) = \exp\left(-\int_0^t \lambda du\right) = \exp(-\lambda t) \ ,$$

and the density function of survival times is

$$f(t) = h(t)S(t) = \lambda \exp(-\lambda t) \ .$$

That is, a constant hazard rate of λ corresponds to exponentially distributed survival times with mean $\mu = \lambda^{-1}$. Note that the constant hazard corresponds to the lack-of-memory property of the exponential distribution (see Exercise 6.1). □

The above example shows that a constant baseline hazard corresponds to exponentially distributed failure times. Other parametric forms of the baseline hazard are possible. However, the parametric proportional hazards model is not widely used in practice because the AFT is generally more robust to model mis-specification (Hutton and Monaghan 2002). This can be mitigated to some extent by employing flexible classes of models, such as that of Royston and Parmar (2002) who used cubic splines to fit a smooth cumulative baseline hazard function. In practice, proportional hazard models are most commonly fitted using the semi-parametric approach developed by Cox (1972) and presented in Section 6.3.3. This was the original application of the proportional hazards model, and the reader should be aware that the term "proportional hazards model" is often used to mean Cox's implementation of this model.

6.3.3 Cox's proportional hazards model

Cox's implementation (Cox 1972) of the proportional hazards model has the advantage that it requires no specification of the baseline hazard function $h_0(t;\boldsymbol{\lambda})$, and hence can be considered a semi-parametric model.

The argument proceeds as follows: in the absence of knowledge about the hazard function, there is no relevant information in the actual survival times because the hazard could be zero over intervals that are free of failures, say. Rather, the relevant information is provided solely by consideration of which individual failed at each failure time. The likelihood so obtained is called the partial likelihood – see Example 9.4 and Cox (1975) for more details.

For simplicity, it is assumed that there are no tied failures times, and that the m individuals that failed have been re-ordered such that $t_1 < t_2 < ... < t_m$ are the distinct failure times. Let $R(t_j)$ denote the risk set at failure time t_j, that is, the set of individuals still in the experiment immediately prior to t_j.

Given that one failure is to occur at time t_j, the relative probabilities of failure for the individuals in $R(t_j)$ are proportional to the values of their hazard functions at t_j. To see this, note that the probability that individual j fails in time interval $[t_j, t_j + \Delta t)$, given that a member of $R(t_j)$ fails in this interval is

$$\begin{aligned} P(\text{indiv } j \text{ fails in } [t_j, t_j+\Delta t) \quad &| \quad \text{one failure in } [t_j, t_j+\Delta t)) \\ &\approx \quad \frac{h_j(t_j;\boldsymbol{\lambda})\Delta t}{\sum_{i\in R(t_j)} h_i(t_j;\boldsymbol{\lambda})\Delta t} \\ &= \quad \frac{h_j(t_j;\boldsymbol{\lambda})}{\sum_{i\in R(t_j)} h_i(t_j;\boldsymbol{\lambda})} \ . \end{aligned}$$

Under the proportional hazards formulation in (6.3), with $\boldsymbol{x}_j$ denoting the covariate vector of the individual failing at time t_j, the above probability is

$$P_j(\boldsymbol{\beta}) = \frac{h_0(t_j;\boldsymbol{\lambda})\exp(\boldsymbol{x}_j^T\boldsymbol{\beta})}{\sum_{i\in R(t_j)} h_0(t_j;\boldsymbol{\lambda})\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} = \frac{\exp(\boldsymbol{x}_j^T\boldsymbol{\beta})}{\sum_{i\in R(t_j)}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} \ , \tag{6.4}$$

which depends only on $\boldsymbol{\beta}$ and not on $h_0(t;\boldsymbol{\lambda})$.

The MLE of $\boldsymbol{\beta}$ is given by maximizing the "partial likelihood", given by the product over the m failure times of the probabilities given in (6.4).

Example 6.2. Two groups of four patients each were used in a study of the effectiveness of a drug. The first group received the drug and the second group received a placebo. The times until relapse were

Placebo group: 3 7 8 (15)

Drug group: 5 (10) 12 27

where values in parentheses denote censored observations.

A natural parameterization would be to set the hazard function of the placebo group equal to the baseline hazard function, $h_0(t)$. The hazard of the drug group is then $h_0(t)e^{\beta}$, where β is the sole parameter to be estimated.

Letting $t_1 < t_2 < .... < t_6$ denote the six relapse times, the partial likelihood is

$$L^p(\beta) = \prod_{j=1}^{6} P_j(\beta) \ .$$

At the first relapse time the risk set includes all eight individuals, and it is an individual from the placebo group who relapses. At the second relapse time the risk set includes the four drug group individuals and the remaining three from the placebo group, and an individual from the drug group relapses. Applying this logic to all six failure times results in the partial likelihood

$$L^p(\beta) = \frac{1}{4+4e^{\beta}} \times \frac{e^{\beta}}{3+4e^{\beta}} \times \frac{1}{3+3e^{\beta}} \times \frac{1}{2+3e^{\beta}} \times \frac{e^{\beta}}{1+2e^{\beta}} \times \frac{e^{\beta}}{e^{\beta}} \ . \tag{6.5}$$

The partial log-likelihood function $\log L^p(\beta)$ can be treated as a conventional log-likelihood. That is, it can be maximized to obtain $\widehat{\beta}$, and used to perform subsequent inference. Here $\widehat{\beta} = -0.689$, which has the interpretation that the hazard of patients receiving the drug is estimated to be $\exp(-0.689) = 0.502$ that of untreated patients. However, the 95% likelihood ratio confidence interval (Fig. 6.3) for β is $(-2.73, 1.12)$, which includes zero. It is perhaps not surprising that there is little statistical evidence of a difference between placebo and drug – this is a consequence of the very small sample sizes used in this example. □

6.3.4 Example in R and SAS: Leukemia data

The data in Table 6.2 are times to remission of 21 patients in each of two groups. One group received an experimental drug treatment, and the control group was given

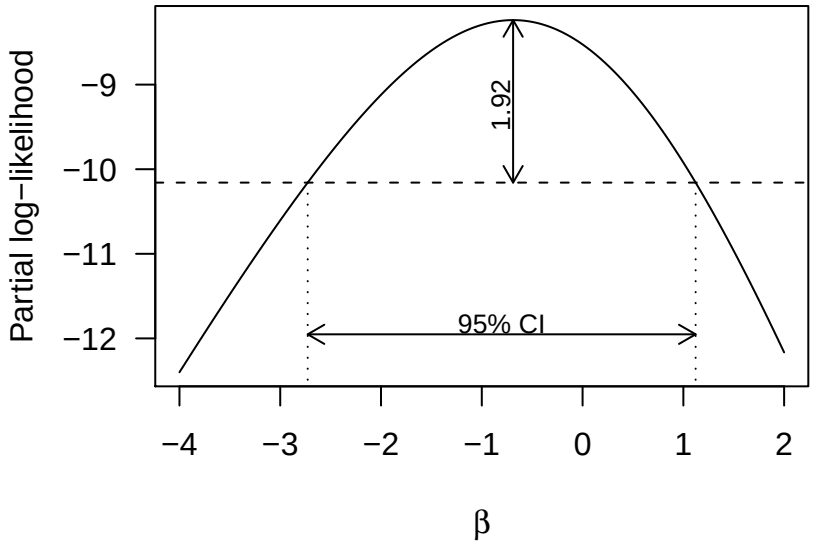


Figure 6.3: Partial log-likelihood from Cox's proportional hazards model.

a placebo treatment. It will be assumed that these 42 observations are in a dataset containing the three variables `trmt`, `time`, and the censoring indicator variable `w`.

AFT and Cox proportional hazards models can be fitted using the `survreg` and `coxph` functions in the `survival` package, or using SAS procedures `LIFEREG` and `PHREG`, respectively. Program code for fitting the Weibull AFT model using `LIFEREG` is shown below, and produces the parameter estimates shown in Figure 6.4. Note that the code uses the model statement `MODEL time*w(0)=trmt;` to specify that right censoring of `time` corresponds to `w` = `0`.

```
*Accelerated failure time model with Weibull failure time distribution;
PROC LIFEREG DATA=Leukemia;
  CLASS trmt;
  MODEL time*w(0)=trmt;
RUN;
```

The AFT model (Fig. 6.4) fits a baseline Weibull distribution $F(t;\widehat{\phi},\widehat{k})$ to the placebo group with MLEs $\widehat{\phi} = \exp(2.2484) = 9.4726$ and $\widehat{k} = 1.3658$ (See Exercise 6.4 for definition of $F(t;\phi,k)$.) The drug group has fitted Weibull distribution $F(t;\exp(1.2673)\widehat{\phi},\widehat{k}) = F(t;3.5514\widehat{\phi},\widehat{k})$. This is equal to the baseline distribution evaluated at time $3.5514t$, $F(3.5514t;\widehat{\phi},\widehat{k})$. That is, one week of remission in the placebo group equates to about 3.55 weeks of remission in the group that received the drug.

Drug						Placebo					
t	w	t	w	t	w	t	w	t	w	t	w
6	0	10	1	22	1	1	1	5	1	11	1
6	1	11	0	23	1	1	1	5	1	12	1
6	1	13	1	25	0	2	1	8	1	12	1
6	1	16	1	32	0	2	1	8	1	15	1
7	1	17	0	32	0	3	1	8	1	17	1
9	0	19	0	34	0	4	1	8	1	22	1
10	0	20	0	35	0	4	1	11	1	23	1

Table 6.2: Duration of remission, t (weeks), for two groups of patients with acute leukemia, from Table 10 of Freireich et al. (1963). Variable w is the censoring indicator variable that takes the value 1 if relapse was observed to occur at time t, and otherwise is zero if censoring occurred at time t.

Analysis of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept		1	2.2484	0.1660	1.9231	2.5737	183.51	<.0001
trmt	Drug	1	1.2673	0.3106	0.6585	1.8762	16.64	<.0001
trmt	Plac	0	0.0000	.	.	.	.	.
Scale		1	0.7322	0.1078	0.5486	0.9772		
Weibull Shape		1	1.3658	0.2012	1.0233	1.8228		

Figure 6.4: Parameter estimates table from using PROC LIFEREG to fit a Weibull accelerated failure time model to the leukemia data.

The R code below shows the Cox proportional hazards fit using the `coxph` function. The Wald test statistic indicates strong evidence of a treatment effect. The placebo treatment is the baseline, and the hazard function of the drug treatment is estimated to be about 21% that of the placebo.

```
> #Cox's proportional hazards model fitted to leukemia data
> library(survival)
> Leukemia=Surv(times,w) #Create a survival object
> PHfit=coxph(Leukemia~trmt)
> coef(summary(PHfit))
              coef exp(coef)  se(coef)         z     Pr(>|z|)
trmtDrug -1.572125 0.2076035 0.4123967 -3.812167 0.0001377538
```

For a more in-depth analysis of these data, including model assessment and consideration of other covariates, see Kleinbaum and Klein (2005).

6.4 Mark-recapture models

Mark-recapture models are one of a multitude of methods (Seber 1982, Borchers, Buckland and Zucchini 2002) that can be used to estimate the number of animals, N, in a population. It will be assumed here that the population is *closed*, that is, the number of animals is not changed by births, deaths, immigration or emigration over the duration of the study, and so N can be regarded as constant throughout the duration of the study.

In a simple mark-recapture experiment, a sample of animals of size n_1 is captured, marked, and returned to the population. Some time later (having allowed enough time for the marked animals to randomly disperse within the population) a second sample of size n_2 is taken. The number of marked animals in the second sample is recorded, and will be denoted m_2. Intuitively, one could argue that the proportion of marked animals in the second sample, $\tilde{p} = \frac{m_2}{n_2}$, is an obvious estimator of the proportion of marked animals in the population $p = \frac{n_1}{N}$. Therefore, since $n_1 = Np$, an intuitive estimator of N is

$$\tilde{N} = \frac{n_1}{\tilde{p}} = \frac{n_1 n_2}{m_2}.$$

This is commonly known as the Petersen estimate (Petersen 1896) or Lincoln index.

The Petersen estimator is positively biased and the relative bias can be large (or infinite!) for small sample sizes. In practice it is common to use Chapman's bias adjusted variant of the Petersen estimate,

$$N^* = \frac{(n_1 + 1)(n_2 + 1)}{m_2 + 1} - 1 \ .$$

Approximate standard errors of $\tilde{N}$ and N^* can be obtained from application of the delta method (see Exercise 6.5) and related methods (Seber 1982).

The above mark-recapture experiment presents some interesting challenges for application of likelihood-based inference. First, the unknown parameter, N is integer valued and the log-likelihood is undefined for non-integer values of N (but see Section 6.4.2). Thus, it is not possible to differentiate the log-likelihood (i.e., regularity conditions R4–R7 in Chapter 12 are not applicable). Moreover, the range of possible

values that n_1, n_2 and m_2 can take varies with N (i.e., condition R3 is also violated). Nonetheless, subject to due caution, likelihood-based inference works well here, and also as a general purpose tool for estimation of animal abundance in more complex situations. In particular, in Section 6.4.2 the definition of the likelihood is extended to non-integer values of N, thereby enabling standard likelihood-based estimation and inference to be utilized.

Different likelihood formulations for the above mark-recapture experiment can be postulated. Section 6.4.1 uses a likelihood formulated from the hypergeometric distribution. This likelihood is arguably a truer description of the statistical model than the multinomial likelihood demonstrated in Section 6.4.3. However, the latter is more readily extendable to other variations of the mark-recapture experiment, and it forms the basis of the general application of ML for estimation of animal abundance from these types of experiments.

6.4.1 Hypergeometric likelihood for integer valued N.

In this model the values of n_1 and n_2 are regarded as fixed. This is certainly appropriate if n_1 and n_2 were indeed specified in advance of the experiment.[4] With n_1 and n_2 fixed, only m_2 is random and the likelihood is determined by the probability of the observed value of m_2 as a function of N. Now, m_2 is the number of marked animals recorded when a sample of size n_2 is taken from a population with n_1 marked animals and $N-n_1$ unmarked animals. Assuming all possible samples of n_2 animals are equally probable, m_2 is distributed according to the hypergeometric distribution (Seber 1982, p. 59),

$$L(N; m_2) = \frac{\binom{n_1}{m_2}\binom{N-n_1}{n_2-m_2}}{\binom{N}{n_2}}, \quad N \geq \max(n_1, n_2)\ . \tag{6.6}$$

Note that the terms in (6.6) that do not involve N are constant terms since N is the sole parameter to be estimated. Thus, to within an additive constant, the

[4] More generally, this conditioning is appropriate if it can be argued that n_1 and n_2 contain no useful information about N. This would be reasonable in the absence of knowledge about the amount of sampling effort or effective proportion of habitat surveyed that was required to obtain the two samples.

log-likelihood function for N is

$$\begin{aligned} l(N; m_2) = \log(N - n_1)! &+ \log(N - n_2)! \\ &- \log(N - n_1 - (n_2 - m_2))! - \log N! \ . \end{aligned} \tag{6.7}$$

The argument used below finds an explicit formula for the MLE of N. It proceeds by determining the incremental change in log-likelihood resulting from an incremental change in N.

Note that

$$\log n! - \log(n-1)! = \log n \ . \tag{6.8}$$

It follows that for any integer $N > \max(n_1, n_2)$,

$$\begin{aligned} \Delta l(N; m_2) &\equiv l(N; m_2) - l(N-1; m_2) \\ &= \log\left(\frac{(N - n_1)(N - n_2)}{N(N - n_1 - (n_2 - m_2)}\right) \\ &= \log\left(\frac{N^2 - Nn_1 - Nn_2 + n_1 n_2}{N^2 - Nn_1 - Nn_2 + Nm_2}\right) \ . \end{aligned} \tag{6.9}$$

The numerator and denominator in 6.9 differ only in the last terms, $n_1 n_2$ and Nm_2 respectively. It follows that

$$l(N-1; m_2) > l(N; m_2) \quad \text{if and only if} \quad Nm_2 > n_1 n_2 \ . \tag{6.10}$$

So, if one starts sequentially with an initial choice of N that is arbitrarily large, it is the case that the log-likelihood is higher at $N - 1$, providing that N is greater than the Petersen estimate $\tilde{N} = \frac{n_1 n_2}{m_2}$. This can be denoted

$$\widehat{N} = [\tilde{N}] = \left[\frac{n_1 n_2}{m_2}\right] \ ,$$

where $[x]$ denotes the integer part of x.

Note that if the Petersen estimate is integer valued then equality holds in (6.10) with $N = \tilde{N}$. It is then that case that the MLE is not unique and $\tilde{N}$ and $\tilde{N} - 1$ are both MLEs.

6.4.2 Hypergeometric likelihood for $N \in \mathbb{R}^+$.

The unknown number of animals is necessarily a non-negative integer value. However, if one is prepared to think laterally, it can be very convenient to extend the

log-likelihood to positive real values. This can be achieved using the gamma function $\Gamma(N)$ for $N > 0$. The gamma function is a smooth function that has the property that $\Gamma(N+1) = N!$ when N is integer valued. In that sense, it provides an extension of the factorial function to the positive real numbers (6.5).

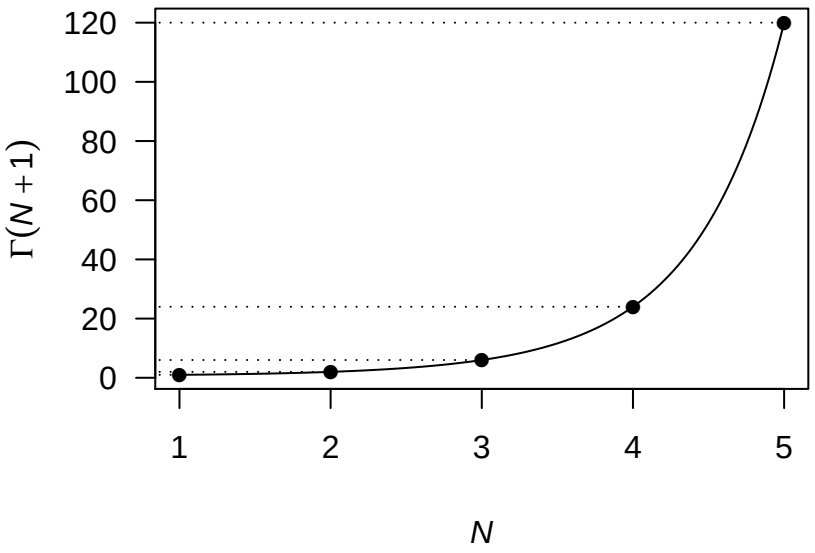


Figure 6.5: Plot of $\Gamma(N+1)$ for real valued N between 1 and 5. Note that $\Gamma(N+1) = N!$ for integer valued N.

The hypergeometric log-likelihood in (6.7) can be defined on $\mathbb{R}^+$ by replacing the factorial terms by their gamma equivalents. That is,

$$\begin{aligned} l(N; m_2) = \log\Gamma(N - n_1 + 1) \quad &+ \quad \log\Gamma(N - n_2 + 1) \\ &- \quad \log\Gamma(N - n_1 - (n_2 - m_2) + 1) \\ &- \quad \log\Gamma(N + 1)\ . \end{aligned} \tag{6.11}$$

This log-likelihood can easily be programmed in both R and SAS because they both have a log-gamma function `lgamma` for evaluation of $\log\Gamma()$.

Example 6.3. Hypergeometric likelihood for mark-recapture experiment. Borchers et al. (2002, p. 107) use the example of a mark-recapture experiment where $n_1 = 64, n_2 = 67$ and $m_2 = 34$. The log-likelihood from (6.11) is unimodal with real-valued MLE $\widehat{N} = 125.615$ and integer-valued MLE of 126. The approximate standard error (obtained from the second derivative of the log-likelihood at

$\widehat{N}$) is 10.4. This gives a 95% Wald CI of (105.2,146.0). The 95% likelihood ratio CI is (109.7,152.3) (Fig. 6.6). □

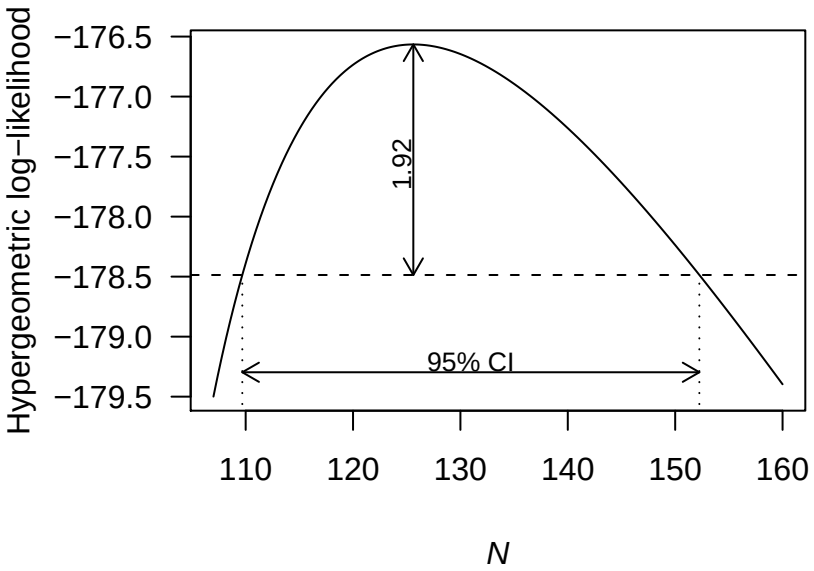


Figure 6.6: Hypergeometric log-likelihood from a mark-recapture experiment with $n_1 = 64, n_2 = 67$ and $m_2 = 34$.

6.4.3 Multinomial likelihood

The hypergeometric likelihood in (6.6) is exact when the sample sizes n_1 and n_2 are fixed and the assumptions of the hypergeometric distribution are satisfied, that is, when all animals (marked and unmarked) are equally likely to be caught on the second sampling occasion, independently of all other animals. However, it is cumbersome to extend the hypergeometric model to more complex mark recapture experiments, and a more flexible alternative is to employ a likelihood derived from a multinomial model for the counts.

The multinomial model for the counts is approximate because it assumes that the outcomes in the second sample (i.e., whether the animal is marked or unmarked) are all statistically independent. It can be argued that this is not the case in most situations, because the proportion of marked animals in the population will change throughout the implementation of the second sample, assuming that this sample is conducted without replacement. However, this approximation works well in practice

(e.g., compare Examples 6.3 and 6.4).

The multinomial likelihood necessarily contains additional parameters corresponding to capture probabilities. These additional parameters can often be explicitly calculated for any given value of N, and hence profile likelihood can be employed (see Section 4.5). For example, for any given value of N, the data from the simple mark-recapture experiment can be arranged in the following two-way table

	Recaptured	Not recaptured	
Marked	m_2	$n_1 - m_2$	n_1
Unmarked	$n_2 - m_2$	u_2	$N - n_1$
	n_2	$N - n_2$	N

where $u_2 = N - n_1 - (n_2 - m_2)$ is the (unobserved) number of animals that are not caught on either capture occasion. The likelihood to be maximized is a function of N and the cell probabilities in this table.

Example 6.4. Multinomial likelihood for mark-recapture experiment. For the values $n_1 = 64, n_2 = 67$ and $m_2 = 34$ used in Example 6.3, the above table is

	Recaptured	Not recaptured	
Marked	34	30	64
Unmarked	33	u_2	$N - 64$
	67	$N - 67$	N

Assuming that the capture and recapture events are independent, this is a two-way contingency table with independent rows and columns. The parameters to be estimated are therefore (r_1, c_1, N) where r_1 and c_1 are the probabilities of row 1 and column 1, respectively. For any N, the MLEs of r_1 and c_1 are their respective row and column proportions $\widehat{r}_1 = \frac{64}{N}$ and $\widehat{c}_1 = \frac{67}{N}$ (see Exercise 4). This enables quick calculation of the profile likelihood for N (Fig. 6.7).

Using an extended multinomial likelihood for $N \in \mathbb{R}^+$, the real-valued MLE is $\widehat{N} = 124.692$ and the integer-valued MLE is 125. The approximate standard error is 10.1, and the 95% likelihood ratio CI is (109.2,150.7). □

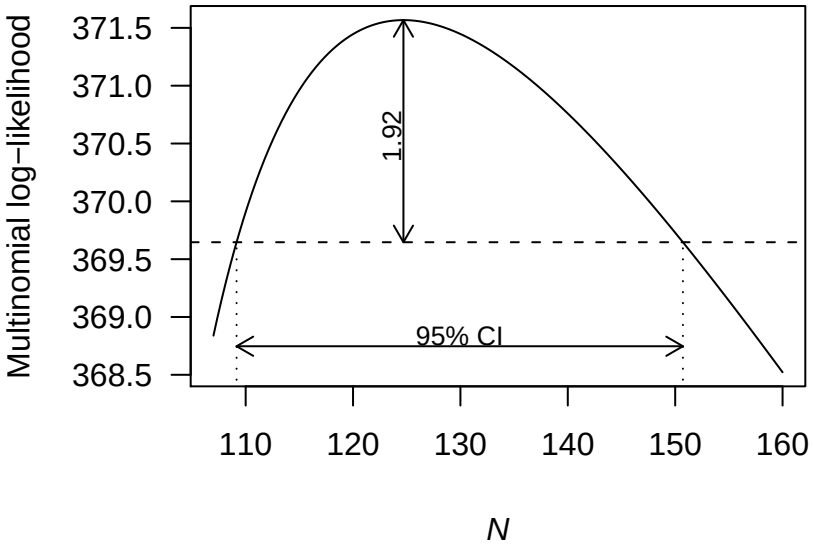


Figure 6.7: Multinomial log-likelihood from a mark-recapture experiment with $n_1 = 64, n_2 = 67$ and $m_2 = 34$.

The multinomial likelihood has the advantage that it can be readily generalized to a multitude of variations on the mark-recapture theme, including modeling experiments with multiple recapture phases and experiments where distinct tags are used on all individuals. This permits the capture history of individual animals to be observed and permits the construction of models that can describe the entire catch history of each animal observed. Such models can allow for the possibility of trap avoidance by animals that have previously been caught (e.g. Pollock 1975, Pledger 2000). See also Borchers et al. (2002, p. 108) and references therein.

6.4.4 Closing remarks

The sampling distribution of MLEs of abundance can be severely right skewed for small sample sizes, and likelihood ratio or bootstrap techniques are generally to be preferred (Cormack 1992, Evans, Kim and O'Brien 1996) over methods that assume approximate normality.

R package `Rcapture` provides a general suite of functions for the estimation of abundance from mark-recapture experiments based on the multinomial model. The implementation in `Rcapture` is based on the equivalence between multinomial and Poisson models, and hence is able to express the multinomial as a log-linear model.

The R package wisp is a comprehensive wildlife simulation package and contains functions for estimation of abundance from a wide variety of experimental designs, including mark-recapture, removal and distance-based methods.

6.5 Exercises

6.1 Lack-of-memory property of the exponential distribution: Let T be distributed exponentially with hazard rate λ, that is, $T \sim \text{Exp}(\lambda^{-1})$ with density function $f(\lambda) = \lambda \exp -\lambda t$. For any $s \geq 0$, let f_s denote the density of the remaining time to failure given survival to time s. That is, f_s is the density of $T - s$ given $T > s$. Show that

$$f_s(t) = f(t) \ .$$

That is, the remaining time to failure is also distributed $\text{Exp}(\lambda^{-1})$.

6.2 Suppose that observations $(t_i, w_i), i = 1, ..., n$ are generated from a survival experiment with exponentially distributed survival times with hazard rate rate λ. The indicator variable w_i takes the value 1 if t_i is an observed failure time, or the value 0 if observation i was censored at time t_i. Maximize the log-likelihood for λ to show that

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} t_i} \ .$$

6.3 Fit a Cox's proportional hazards model to the poison data in Table 6.1 assuming additive effects of poison and treatment. Repeat using the full model with interaction, and compare these two models using AIC.

6.4 Weibull(ϕ, k) distributed survival times have density that can be written in the form

$$f(t; \phi, k) = k\phi^{-k} t^{k-1} \exp(-(t/\phi)^k) \ , \quad y > 0, \ \phi > 0, \ k > 0 \ .$$

Show that the distribution function is

$$F(t; \phi, k) = 1 - \exp(-(t/\phi)^k) \ ,$$

and hence that the hazard function is

$$h(t; \phi, k) = k\phi^{-k} t^{k-1} \ .$$

6.5 In a mark-recapture experiment, if N is large compared to n_2 then it will be reasonable to assume $m_2 \sim \text{Bin}(n_2, p)$, where p is the proportion of animals that have been marked. Under this model for m_2, and assuming that n_1 and n_2 are fixed, apply the delta method to obtain the following approximate variance for the Petersen estimator,

$$\widehat{\text{var}}(\tilde{N}) = \frac{n_1^2 n_2 (n_2 - m_2)}{m_2^3} \ .$$

6.6 The unknown number, N, of animals in a closed population can be estimated by the method of removal. In this type of experiment, n_1 animals are removed on the first removal occasion, leaving $N - n_1$ animals. A further n_2 are removed on

a second removal occasion. It will be assumed that n_1 is $\text{Bin}(N, p)$ and that n_2 is $\text{Bin}(N - n_1, p)$, where p is common to both removal occasions.

Assume that p is known, leaving only N to be estimated. The likelihood for N is given by the product of the binomial likelihoods from the two removal occasions.

a. Show that an integer-valued MLE of N is

$$\widehat{N} = \left[\frac{n_1 + n_2}{2p - p^2}\right] ,$$

where $[x]$ denotes the integer part of x.

b. Is the above MLE the unique integer-value MLE?

6.7 Implement and maximize the hypergeometic log-likelihood for the mark-recapture experiment in Example 6.3.

6.8 Implement and maximize the multinomial log-likelihood for the mark-recapture experiment in Example 6.4.

Chapter 7

Generalized linear models and extensions

From the likelihood point of view, generalized linear models (GLMs) can be regarded as a flexible class of models that all share a convenient form of likelihood. GLMs are in wide use because this class includes models that are natural for the modeling of count and binomial data. The likelihood of a GLM, and its derivatives, have simple explicit form and optimization of the likelihood is fast. The existence and uniqueness of the MLE are guaranteed under very weak general conditions (e.g., Wedderburn 1976), Furthermore, by virtue of the linear model structure, the regression parameters of a typical[1] GLM are not constrained, in the sense that they can take any value on the real line. This avoids the "parameters on the boundary of the parameter space" problem that can lead to the sampling distribution being poorly approximated by a normal distribution (Section 4.9.3).

Software for fitting generalized linear models is well developed and widely available. GLMs can be fitted using the `glm` function in R, or `GENMOD` procedure in SAS. `PROC GENMOD` includes numerous options for model evaluation and inference, including automatic calculation of likelihood ratio confidence intervals. Similar functionality is provided in R, using functions that can apply appropriate methods to the fitted `glm` object.

A generalized linear model has two key features. The first is that the data have a distribution belonging to the exponential family (Section 7.1.1). This includes the

[1] But, see Box 7.2

normal, Poisson, binomial, gamma, inverse-Gaussian, and fixed-dispersion negative binomial. The second feature is that covariates $\boldsymbol{x}_i$ (associated with observation i) are included using a linear model, but this is a linear model for some appropriate transformation of $E[Y_i]$ (Section 7.1.2). For example, in the case of count data the standard model assumes that the data are Poisson distributed and that the log of $E[Y_i]$ follows a linear model. This is commonly known as log-linear regression.

Section 7.2 briefly presents the general form of the likelihood equations that underly the maximum likelihood modeling of exponential family data. Section 7.3 presents methods for model evaluation and comparison. The model used in the case study in Section 7.4 is a standard logistic regression, and no obvious problems with the model fit are found. However, one interesting facet of the case study is that the regression coefficients of the logistic regression are not of direct interest. Rather, it is a form of an inverse prediction (or calibration) problem whereby it is desired to estimate the covariate value that corresponds to a particular expected value of the response variable.

7.1 Specification of a GLM

Readers interested only in application of GLMs can skip Section 7.1.1.

7.1.1 Exponential family distribution†

In a generalized linear model, it is required that the distribution of $Y \in \mathbb{R}$ is such that the log of the density function can be written in the form

$$\log f(y; \psi, \phi, w) = \frac{y\psi - b(\psi)}{a(\phi, w)} + c(y, \phi, w) \tag{7.1}$$

where $\psi, \phi \in \mathbb{R}$ are parameters, and w is a known "weight". Parameter ϕ is the dispersion parameter, and function $a(\phi, w)$ is required to be of the form ϕ/w. Formula (7.1) is a special case of the general class of exponential family distributions, and some texts refer to it as the exponential dispersion model. This distinction will not be made here, and in what follows it is implicitly understood that the terminology "exponential family distribution" corresponds to the form in (7.1).

To maintain consistency with previous notation, it is convenient to drop w from the above notation because it is not a parameter to be estimated. That is, the log-density will be written

$$\log f(y;\psi,\phi) = \frac{y\psi - b(\psi)}{a(\phi)} + c(y,\phi) \tag{7.2}$$

where dependence on the known weights is implicit.

The data $Y_i \in \mathbb{R}$, $i = 1,...,n$ are assumed to be independent where Y_i has exponential family distribution with density $f(y_i,\psi_i,\phi)$. Any covariates associated with Y_i are modeled through ψ_i (Section 7.1.2). The dispersion parameter ϕ is common to all observations. However, the known weights, w_i, may vary between observations.

The following examples demonstrate that normal data and binomial proportions have distributions that are of exponential family type. In the binomial case it is seen that parameter ϕ is redundant, and it can be fixed at unity by setting the weight equal to the number of trials.

Example 7.1. If $Y \sim N(\mu,\sigma^2)$ then the log of the density function can be written as

$$\log f(y;\mu,\sigma^2) = \frac{y\mu - \mu^2/2}{\sigma^2} - 0.5\left(\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right) .$$

This establishes that the $N(\mu,\sigma^2)$ distribution is of exponential family form under the parameterization $\psi = \mu$, $\phi = \sigma^2$, and where $a(\phi) = \phi, b(\psi) = \psi^2/2$. □

Box 7.1.

The linear-regression model for normal data is included amongst the class of GLMs. However, there is no need to use maximum likelihood theory for these models because the sampling properties of the least-squares estimators of the regression co-efficients (which are also the MLEs – see Example 11.6) and associated test statistics are known exactly.

Example 7.2. For binomial data it is the proportion of "successes", rather than the number of "successes", that can be expressed in exponential family form. So, let $Y = T/n$ where $T \sim \text{Bin}(n,p)$. Then, for $y \in \{0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1\}$, the log of the

density function (in this case, probability mass function) is

$$\begin{aligned}\log f(y;p) &= \log\left(\begin{array}{c} n \\ ny \end{array}\right) + ny\log p + n(1-y)\log(1-p) \\ &= ny\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\left(\begin{array}{c} n \\ ny \end{array}\right).\end{aligned}$$

Setting $\psi = \log(\frac{p}{1-p}) = \text{logit}(p)$, and noting that $\log(1-p) = -\log(1+e^{\psi})$, the log-density function can be written in the one-parameter exponential family form by writing

$$\log f(y;\psi) = \frac{y\psi - \log(1+e^{\psi})}{1/n} + \log\left(\begin{array}{c} n \\ ny \end{array}\right).$$

Here, $b(\psi) = \log(1+e^{\psi})$ and $a(\phi) = 1/n$. The ϕ parameter is absent, but it is convenient to set $\phi = 1$ so that $a(\phi)$ has the required form of $a(\phi) = \phi/w$, with $w = n$. □

7.1.2 GLM formulation

In Section 7.1.1 it was necessary to express the exponential family density as a function of parameters ψ and ϕ. However, this parameterization is not a practical one for the purpose of specifying models for exponential family data. It is much more natural to specify the model using the mean $\mu = E[Y]$ as a parameter, and this is the approach taken by a GLM[2].

The GLM generalizes the linear regression model by fitting a linear model to a transformation of the mean. That is, if $\mu_i = E[Y_i]$ and $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ are the covariates associated with Y_i then the GLM specifies

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j \tag{7.3}$$

where $\beta_j, j = 1, ..., p$ are parameters to be estimated. The linear model term on the right-hand side of (7.3) is called the linear predictor, and will be denoted by

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \ .$$

[2]See Exercise 7.4 for the relationship between μ and ψ, and Example 11.5 for more detail.

The transformation g is called the link function because it links the mean to the linear predictor, as $g(\mu_i) = \eta_i$.

The link function is required to be invertible. Denoting $h = g^{-1}$, this gives

$$h(\eta_i) = \mu_i \ .$$

Parameters $\beta_j, j = 1, ..., p$ are regression coefficients and it is natural (although not absolutely necessary – see Box 7.2) that they be permitted to take any values in $\mathbb{R}$, and consequently the linear predictor also ranges over $\mathbb{R}$. However, it may be that μ_i is not defined on all of $\mathbb{R}$. For example, it Y_i are Poisson distributed then it is necessarily the case that $\mu_i > 0$. It is therefore standard practice to choose h to be a function that maps from $\mathbb{R}$ into the domain of feasible values of μ_i. In the Poisson case, a natural choice for h is the exponential transformation, $h(\eta_i) = e^{\eta_i}$, corresponding to the log link function $g(\mu_i) = \log \mu_i$.

Box 7.2.

Most GLM software will permit the use of non-standard forms of the link function. For example, an identity link can be used to model Poisson data if it is felt that it is appropriate to model the mean directly using a linear model (see the case study in Section 7.6). However, this model must be treated with care, because it induces constraints on $\boldsymbol{\beta}$ by virtue of the fact that the likelihood will be undefined for values of the regression coefficients that result in a negative value of η_i for any i.

Example 7.2 ctd. Here Y_i are binomial proportions, $Y_i \sim \text{Bin}(n_i, p_i)/n_i$ and so $E[Y_i] = p_i$. If we model $p_i = h(\eta_i)$ then it is natural to require that h be an invertible function and $h : \mathbb{R} \to [0, 1]$. Note that these requirements are satisfied by the distribution function of any continuous random variable with support on $\mathbb{R}$ (that is, having density function that is positive on the entire real line).

Logistic regression makes the choice of taking h to be the distribution function of a logistic random variable. With this choice,

$$p_i = h(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \ , \tag{7.4}$$

and its inverse is the link function

$$\begin{aligned} \eta_i = g(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \text{logit}(p_i)\ . \end{aligned} \tag{7.5}$$

Probit regression uses the distribution function of a normal random variable as the choice of h. The inverse of the normal distribution function is the probit function, from which this model inherits its name. Another choice that is commonly provided in many software packages is the complementary log-log link $g(p_i) = \log(-\log(1 - p_i))$, which is obtained as the inverse of the extreme value distribution function.

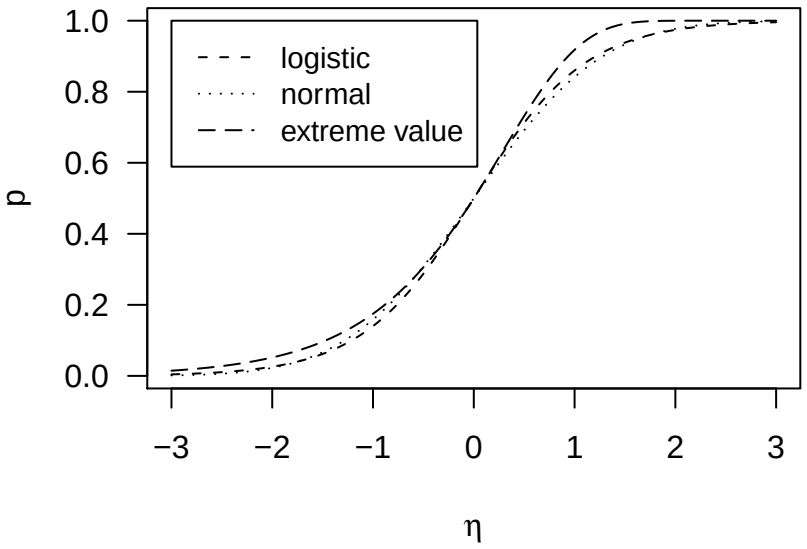


Figure 7.1: Comparison of the distribution functions of the logistic, normal, and extreme value distributions. Their inverses are the logit, probit, and complementary log-log link functions, respectively. The distribution functions have been standardized to be those of random variables with median equal to zero, and variance equal to unity.

The distribution functions of the logistic and normal are both symmetric and are very similar in shape (Fig 7.1). In practice, logistic and probit regression will give similar fits (Chambers and Cox 1967). The logistic is generally to be preferred due to ready interpretability of regression coefficients in terms of log odds (see Example 7.4.1), and because the logit is the natural link to use for binomial data (see Box 7.3). □

7.2 Likelihood calculations

Here is is assumed that the data $Y_i, i = 1, ..., n$ are independently distributed from an exponential family distribution where $\mu_i = E[Y_i]$ depends on parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$. In the previous section it was assumed that the relationship between μ_i and $\boldsymbol{\beta}$ was specified using the generalized linear model of equation (7.3). However, the likelihood calculations hold more generally and it is enough to require only that μ_i is a smooth (i.e., differentiable) function of $\boldsymbol{\beta}$. For example, the likelihood calculations are also applicable to nonlinear regression models for exponential family data.

The parameters to be estimated are $\boldsymbol{\beta}$, and the dispersion parameter ϕ if it is present. The dispersion parameter is not usually of direct interest and here focus is on estimation of the MLE $\widehat{\boldsymbol{\beta}}$. It can be shown (see Example 11.5) that the likelihood equations to be solved for finding $\widehat{\boldsymbol{\beta}}$ are

$$\frac{\partial l(\boldsymbol{\beta}, \phi; \mathbf{y})}{\partial \beta_j} = \sum_{i=i}^{n} \frac{\frac{\partial \mu_i}{\partial \beta_j}(y_i - \mu_i)}{\text{var}(Y_i)} = 0 \ , \quad j = 1, ..., p \ . \tag{7.6}$$

Note that this equation depends only on specification of the mean and variance of each Y_i. This is also true of the second derivative calculations that are required to obtain the approximate variance matrix of $\widehat{\boldsymbol{\beta}}$.

Box 7.3.

For GLMs, the second derivatives of the log-likelihood have an especially convenient form when the canonical form of the link function is used (Example 11.8). In particular, these second derivatives do not depend on the data. The canonical link functions for binomial proportion and Poisson data are the logit and log links, respectively.

7.3 Model evaluation

Many of the model evaluation tools of linear regression have analogous variants for evaluation of models fitted to exponential family data. Deviance (Section 7.3.1) plays much the same role as residual sums-of-squares, and analysis of deviance (Section 7.3.2) is the analogue of analysis of variance. Pearson and deviance residuals are introduced in Section 7.3.3.

Section 7.3.4 uses the Pearson chi-square statistic and/or the deviance statistic to provide omnibus tests of goodness of fit for models fitted to binomial and Poisson data. These tests do not have a linear regression counterpart, because they assume a known value of the dispersion parameter ϕ.

7.3.1 Deviance

To within an additive constant, the deviance of an exponential family model is just the negative of twice the log-likelihood of that model. To mimic the behaviour of residual sums-of-squares, the additive constant is chosen so that that the deviance of a model with perfect fit is zero. For this to make sense (so that deviance can be used for testing within nested sets of models, say) it is necessary to require that the value of the dispersion parameter ϕ is constant across the collection of models. Recall that $\phi = 1$ for both Poisson and binomial distributions.

Let $l(\boldsymbol{\mu}, \phi; \boldsymbol{y})$ denote the log-likelihood of an exponential family model expressed as a function of the vector of means $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ and ϕ. A model with perfect fit is one that sets the fitted mean of each observation equal to its observed value, and has log-likelihood $l(\boldsymbol{y}, \phi; \boldsymbol{y})$. Here this will be called the saturated model, though other texts may refer to it as the full model (but more commonly, *full* model is used to describe a model that includes all possible explanatory terms).

Note that the deviance of a model with fitted means $\widehat{\boldsymbol{\mu}}$ is

$$D = 2[l(\boldsymbol{y}, \phi; \boldsymbol{y}) - l(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})] \ . \tag{7.7}$$

The model deviance is sometimes also called residual deviance. In the case of a GLM, the fitted values $\widehat{\boldsymbol{\mu}}$ are obtained from the fitted value of the linear predictor, that is, $\widehat{\mu}_i = h(\widehat{\eta}_i) = h(\sum_j x_{ij}\widehat{\beta}_j)$.

The deviance of a model is simply twice the difference between the log-likelihood of the saturated model and the log-likelihood of that model. Since the saturated model maximizes the likelihood over all possible values of $\widehat{\boldsymbol{\mu}}$ (Box 7.4), it follows that model deviance is always non-negative. The deviance of the model that fits a common mean to all observations is commonly called the null deviance. The null

deviance is the GLM analogue to the total sums-of-squares in a linear regression.

Box 7.4.

The saturated model is obtained from fitting n parameters $\mu_i, i = 1, ..., n$, and can be represented using the formulation in Section 7.2 by setting $\mu_i = \beta_i, i = 1, ..., n$. The likelihood equation in (7.6) then takes the form

$$\frac{\partial l}{\partial \mu_i} = \frac{(y_i - \mu_i)}{\text{var}(Y_i)} = 0 \ , \quad i = 1, ..., n \ ,$$

and hence it follows that $\widehat{\boldsymbol{\mu}} = \boldsymbol{y}$ is a solution to these equations.

Example 7.3. Poisson deviance. If y_i are the observed values of independently distributed Poisson(λ_i) random variables then the deviance of the model with fitted values $\widehat{\lambda}_i$ is

$$D = 2 \sum_{i=1}^{n} \left[(\widehat{\lambda}_i - y_i) + y_i \log \left(\frac{y_i}{\widehat{\lambda}_i} \right) \right] \ . \tag{7.8}$$

□

Example 7.4. Negative binomial deviance. For a given value of parameter m, the negative binomial is an exponential family distribution with $\phi = 1$ (Exercise 7.3). Thus, using the parameterization of the negative binomial in (15.1), the negative binomial deviance for a model with fitted values $\widehat{\mu}_i$ is

$$D = 2 \sum_{i=1}^{n} \left[y_i \log \left(\frac{y_i}{\widehat{\mu}_i} \right) - (m + y_i) \log \left(\frac{y_i + m}{\widehat{\mu}_i + m} \right) \right] \ . \tag{7.9}$$

Note that this deviance is obtained by assuming that the fitted negative binomial model and saturated negative binomial model share the same common value of m.

□

7.3.2 Model selection

The general tools of model selection were covered in Section 4.6, and are equally applicable to GLMs and other exponential family models. In particular, likelihood ratio tests can be used for selecting between nested models, and AIC or BIC can

be used if choosing within a collection of competing models that is not necessarily nested.

In the case of binomial or Poisson data, using the likelihood ratio to perform hypothesis tests on nested models can conveniently be implemented by comparison of model deviances. If model A with deviance D_A is a subset of model B with deviance D_B, then the difference in deviances is

$$\begin{aligned} D_A - D_B &= 2[l(\boldsymbol{y};\boldsymbol{y}) - l(\widehat{\boldsymbol{\mu}}_A;\boldsymbol{y})] - 2[l(\boldsymbol{y};\boldsymbol{y}) - l(\widehat{\boldsymbol{\mu}}_B;\boldsymbol{y})] \\ &= 2[l(\widehat{\boldsymbol{\mu}}_B;\boldsymbol{y}) - l(\widehat{\boldsymbol{\mu}}_A;\boldsymbol{y})] \ , \end{aligned}$$

which is the likelihood ratio test statistic of the hypothesis that model A is true. Here, $\widehat{\boldsymbol{\mu}}_A$ and $\widehat{\boldsymbol{\mu}}_B$ denote the fitted means under models A and B, respectively. This test statistic has an approximate chi-square distribution with degrees of freedom equal to the difference in number of parameters between the two models.

The deviances of a collection of nested binomial or Poisson models can be partitioned in a similar way to partitioning of sums-of squares in an ANOVA table. Rather than using the decrease in sums-of-squares, it is the reduction in deviance that is used to measure the improvement in fit from adding a term to the model. This procedure is commonly known as analysis of deviance.

7.3.3 Residuals

Pearson and deviance residuals are commonly used in the evaluation of GLMs and other exponential family models. The Pearson residual for observation i is the raw residual divided by the estimated standard deviation of Y_i. That is,

$$r_i^P = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{Var}(Y_i)}} \ .$$

The deviance residual for observation i is the signed square-root of that observation's contribution to the deviance. That is

$$r_i^D = \text{sign}(y_i - \widehat{\mu}_i)\ \sqrt{2}\,[l(y_i,\phi;y_i) - l(\widehat{\mu}_i,\phi;y_i)]^{\frac{1}{2}} \ ,$$

where sign() is the function that takes the value 1 if its argument is positive, and the value -1 otherwise. Note that the sum of the squared deviance residuals is equal to the deviance.

Example 7.5. Poisson residuals If y_i is the observed value of a Poisson random variable and $\widehat{\lambda}_i$ is the fitted mean, then the Pearson residual is

$$r_i^P = \frac{y_i - \widehat{\lambda}_i}{\sqrt{\widehat{\lambda}_i}} \ ,$$

and the deviance residual is

$$r_i^D = \text{sign}(y_i - \widehat{\lambda}_i) \ \sqrt{2} \left[(\widehat{\lambda}_i - y_i) + y_i \log \left(\frac{y_i}{\widehat{\lambda}_i} \right) \right]^{\frac{1}{2}} \ .$$

□

7.3.4 Goodness of fit

When the dispersion parameter ϕ is known, then the log-likelihood $l(\boldsymbol{\mu}, \phi; \boldsymbol{y})$ is well defined for the saturated model having $\boldsymbol{\mu} = \boldsymbol{y}$. (This is in contrast to the case of normally distributed data with unknown variance, because $l(\boldsymbol{y}, \sigma^2; \boldsymbol{y})$ can be made arbitrarily large by making σ^2 arbitrarily small.) An omnibus goodness-of-fit test statistic for a given model can therefore be obtained as the LRT statistic for the test of that model nested within the saturated model. This test statistic is none other than the model deviance defined in (7.7). The saturated model has one parameter for each of the n observations, and so the degrees of freedom is $n - p$ where p is the number of parameters in the given model. However, caution is advised because the approximate distribution of the deviance (assuming the current model is true) may not be well approximated by a χ^2_{n-p} distribution (see Example 7.6). This is because the large-sample theory of likelihood ratio statistics does not hold since the number of parameters in the saturated model is not fixed, but instead is equal to the number of observations.

An alternative omnibus goodness-of-fit test statistic is provided by the Pearson chi-square. This is the sum of the squared Pearson residuals,

$$P_{\chi^2} = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\text{var}(Y_i)} \ , \tag{7.10}$$

and the degrees of freedom is again $n - p$.

The deviance and Pearson chi-square statistics will have sampling distributions that are approximately χ^2_{n-p} provided that the data are not overly sparse. There are

various rules of thumb, for example, in an analysis of contingency tables (a Poisson GLM in which all explanatory variables are factors) the `FREQ` procedure in SAS can be used. This gives a warning message if less than 80% of the fitted cell means exceed 5. For binomial data, $Y_i \sim \text{Bin}(n_i, p_i)$, an analogous rule of thumb would be to to require $\min(n_i\widehat{p}_i, n_i(1-\widehat{p}_i)) \geq 5$ for most of the observations.

In practice, the Pearson chi-square is more robust to sparseness than deviance (McCullagh and Nelder 1989), and is generally preferred. However, unlike deviance, it can not be used for model selection, because it does not partition additively.

Example 7.6. Deviance and P_{χ^2} for sparse binomial data. As an extreme example, suppose that $Y_i, i = 1, ..., n$ are iid $\text{Bin}(1, p)$, that is, iid Bernoulli with "success" probability p. Assuming that $0 < \widehat{p} < 1$, the Pearson chi-square statistic is

$$\begin{aligned}\sum_{i=1}^{n} \frac{(y_i - \widehat{p})^2}{\widehat{p}(1-\widehat{p})} &= \sum_{\{i:y_i=0\}} \frac{\widehat{p}}{1-\widehat{p}} + \sum_{\{i:y_i=1\}} \frac{1-\widehat{p}}{\widehat{p}} \\ &= n(1-\widehat{p})\frac{\widehat{p}}{1-\widehat{p}} + n\widehat{p}\frac{1-\widehat{p}}{\widehat{p}} \\ &= n\ .\end{aligned}$$

The saturated model has likelihood of unity, and so the deviance statistic is

$$D = -2l(\widehat{p}; \boldsymbol{y}) = -2n\left(\widehat{p}\log\widehat{p} + (1-\widehat{p})\log(1-\widehat{p})\right)\ .$$

The Pearson statistic is the constant n, which is close to its nominal $n-1$ degrees of freedom and, if compared to the upper quantiles of a χ^2_{n-1} distribution, provides no evidence against the true model. The sampling distribution of the deviance depends on p. In particular, if $p = 0.5$ then $D/(n-1)$ converges (in probability) to $2\log(2) = 1.386$ as n increases. It follows that, for large n and $p = 0.5$, the deviance statistic will invariably lead to rejection of the true model if it is assumed that the deviance possesses an approximate χ^2_{n-1} sampling distribution. □

Example 7.7. Fish counts The following fifteen counts of snapper (*Pagrus*

auratus) are a subset of the data used in the case study in Section 7.6, and were obtained from independent replicate deployments of an underwater video camera.

2 9 6 2 17 2 9 8 9 11 3 7 3 2 1

It was desired to test the null hypothesis $H_0 : Y_i \sim \text{Pois}(\lambda)$, that is, whether the observations are iid Poisson with common mean λ. Under H_0, $\widehat{\lambda} = \overline{y} \approx 6.0667$. From (7.8), the deviance statistic for the goodness of fit under H_0 is 45.910.

With count data, the Pearson chi-square test statistic is frequently written using an alternative notation where $O_i = y_i$ and $E_i = \widehat{\lambda}_i$. That is, O_i and E_i denote observed and expected value, respectively. This results in the commonly used notation

$$P_{\chi^2} = \sum_i \frac{(O_i - E_i)^2}{E_i} \ ,$$

which evaluates to 46.967 for the above counts. Both test statistics are very extreme in comparison to the χ^2_{14} distribution, and it is concluded that the iid Poisson model is not appropriate for these data. □

7.4 Case study: Logistic regression and inverse prediction

To avoid redundancy, R is used here and SAS is used in the next case study in Section 7.6. The R code in Section 7.4.1 fits a logistic regression model to binomial data. This model has been used previously in this text, and in particular, it is the log-likelihood of this model that is displayed in the contour plot of Figure 3.5.

In this example it is not the parameters β_0 and β_1 that are of direct interest. Rather, inference is required about the negative of their ratio $\zeta = -\beta_0/\beta_1$. The quantity ζ corresponds to the value of the covariate that gives rise to a probability equal to 0.5. This is an example of an "inverse prediction" problem, because rather than being used to predict values of the response variable for a given covariate value, the model is being used to estimate the covariate value for a given expected value of the response. In this example the covariate is fish length and ζ is given the more descriptive name l_{50}, to denote that it is the length at which the binomial probability

is 0.5, which in this particular application is the length at which 50% of fish will be retained.

More generally, inverse prediction arises in bioassay studies where it is desired to quantify the effectiveness of a substance when used at different dosages (the covariate). In these experiments, the quantity ζ is often given a more descriptive name such as ED50 or LD50, the former denoting the effective dose corresponding to a 50% response rate, and the latter denoting the lethal dose corresponding to a 50% mortality rate in the context of studies where the dichotomous response variable is survival or death.

7.4.1 Size selectivity modeling in R

An experiment was conducted to investigate the ability of a trawl net to release under-sized haddock. The aft part of a trawl is the codend, and this was constructed from 113 mm diamond mesh. A fine mesh cover was place around the codend, and held clear of the codend by the use of plastic hoops (Fig. 7.2). The data (Table 7.1) are from Clark (1957), and are the length frequencies of fork lengths (i.e., length of fish measured from nose to fork of the tail) in both the codend and cover, measured to the nearest cm. Small fish can escape the codend by swimming through the 113 mm mesh and will be caught in the cover, but large fish will be unable to squeeze through the codend mesh. One responsibility of fisheries managers is to legislate a mesh size for the fishing fleet that is commensurate with the minimum legal size (MLS) at which fish can be captured, because discarded small fish can incur substantial mortality. For example, it may be desired that the mesh size in the codend is such that the length at which the retention probability is 0.5, l_{50}, is equal to the MLS[3].

The R code below reads the tab-delimited data file `haddock.dat` which contains the data in Table 7.1 arranged in three columns. The column names are `forklen`, `codend` and `cover`.

[3]This policy varies with fishery agency. For example, the International Council for the Exploration of the Sea recommends (ICES 1979) that MLS correspond to the length of 25% retention, l_{25}. This length is derived from β_0 and β_1 as $l_{25} = -\frac{\beta_0 - \log(3)}{\beta_1}$.

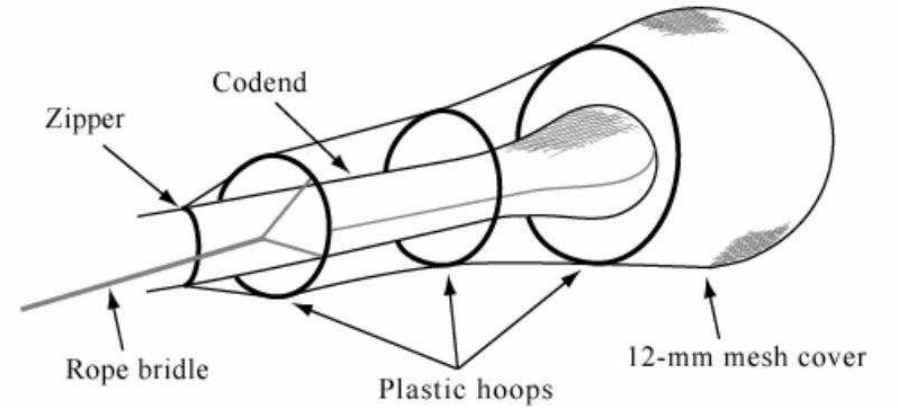


Figure 7.2: Trawl codend enclosed by a fine mesh cover. Small fish are able to pass through the meshes of the codend, but will be retained in the cover.

length	codend	cover	length	codend	cover	length	codend	cover
19.5	0	2	32.5	3	2	45.5	12	0
20.5	0	5	33.5	4	4	46.5	9	0
21.5	0	11	34.5	5	12	47.5	3	0
22.5	0	28	35.5	8	9	48.5	5	1
23.5	1	53	36.5	13	14	49.5	3	0
24.5	5	46	37.5	29	15	50.5	5	0
25.5	1	35	38.5	29	8	51.5	2	0
26.5	3	27	39.5	34	9	52.5	2	0
27.5	1	5	40.5	30	3	53.5	1	0
28.5	0	3	41.5	29	3	54.5	1	0
29.5	1	2	42.5	18	2	55.5	4	0
30.5	0	0	43.5	16	1			
31.5	0	2	44.5	11	0			

Table 7.1: Length frequencies of haddock caught in the codend and its cover.

The raw data are counts and could be modeled using a Poisson log-linear regression (Exercise 7.6). However, it is equivalent, and much more intuitive, to regard the data as binomial where, for each length, `codend` is the number of "successes" (from the fisherman's point of view) out of a total of `codend` + `cover` trials. Note that although `haddock.dat` has 37 lines of data, the codend and cover counts for fork length of 30.5 cm are both zero, and so there is no binomial observation for this fork length. The number of binomial observations is therefore 36.

A plot of `forklen` versus the proportion of haddock in the codend shows that a logistic curve may be a reasonable choice to model these data (Fig. 7.3). The

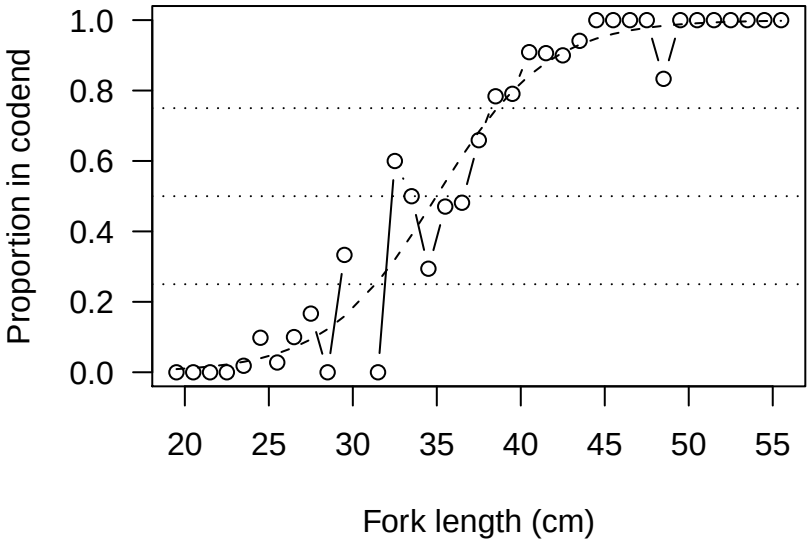


Figure 7.3: Proportion of total haddock catch (by length) that was retained in the codend.

equation for the logistic curve is given by equation (7.4), which here is

$$p_l = \frac{\exp(\beta_0 + \beta_1 l)}{1 + \exp(\beta_0 + \beta_1 l)} \ .$$

That is,

$$\text{logit}(p_l) = \beta_0 + \beta_1 l \ . \tag{7.11}$$

The ML fit of this model is achieved using the following code.

```
> Haddock.df=read.table("Haddock.dat",head=T)
> #Rename codend as y, and calculate total catch at length
> Haddock.df=transform(Haddock.df,y=codend,n=codend+cover)
> attach(Haddock.df)
> #Take a quick look at the data
> plot(forklen,y/n,las=1,type="b",
+      xlab="Fork length (cm)",ylab="Proportion in codend")

> #Logistic model: logit link is default when family is binomial
> Haddock.glm=glm(y/n~forklen,family=binomial,weight=n)
> #Overlay fitted retention probabilities
> lines(forklen[n>0],fitted(Haddock.glm),type="l",lty=2)
> abline(h=c(0.25,0.5,0.75),lty=3)
> summary(Haddock.glm)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.63219    0.86468  -12.30   <2e-16 ***
forklen       0.30396    0.02363   12.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 432.464  on 35  degrees of freedom
```

```
Residual deviance:  23.436  on 34  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 79.75
```

The estimated value of β_1 is 0.304, and from (7.11) it follows that each 1-cm increase in fork length results in an estimated increase of 0.304 in the log-odds of a haddock being retained in the codend.

The next code segment evaluates the Pearson chi-square for goodness of fit. It also evaluates the log-likelihoods of the null, logistic, and saturated models.

```
> #Calculate the Pearson chi-square statistic
> cat("\n Pearson chi-square=",sum(resid(Haddock.glm,type="pearson")^2))

 Pearson chi-square= 27.09488
> #Get log-likelihoods of null, logistic, and saturated models
> Null.lhood=logLik(glm(y/n~1,family=binomial,weight=n))
> Haddock.lhood=logLik(Haddock.glm)
> Satd.lhood=logLik(glm(y/n~as.factor(forklen),family=binomial,weight=n))
> cat("\n Null model log-likelihood=",Null.lhood,
+     "\n Logistic model log-likelihood=",Haddock.lhood,
+     "\n Saturated model log-likelihood=",Satd.lhood,"\n")

 Null model log-likelihood= -242.3891
 Logistic model log-likelihood= -37.87508
 Saturated model log-likelihood= -26.15727
```

The data here are sparse for some fork lengths, especially for larger fish where p_l is close to unity, and so the 34 degrees of freedom for the Pearson chi-square test is likely to be inappropriately high. It is reassuring to see that the observed chi-square of 27.1 is well below 34, and so a somewhat inflated d.o.f will be of little consequence. Combining the counts into wider length classes where the data are sparse would be one way of more formally mitigating this issue.

The null deviance and residual model deviance were obtained earlier and are seen to equal twice the difference between the log-likelihood of the saturated model, and the log-likelihood of the null or logistic model, respectively. The difference between the deviances of the null and logistic models is approximately 409.0. and this is the LRT test statistic of $H_0 : \beta_1 = 0$. With an approximate χ^2_1 distribution under H_0, this hypothesis is soundly rejected.

The `plot` function draws several diagnostic plots for the fitted model. In particular, a plot of the linear predictor ($\eta_l = \beta_0 + \beta_1 l$) versus deviance residual is obtained

from the following code.

```
> #Use which=1 to select only the plot of linear predictors vs resids.
> plot(Haddock.glm,which=1)
> abline(h = 0)
```

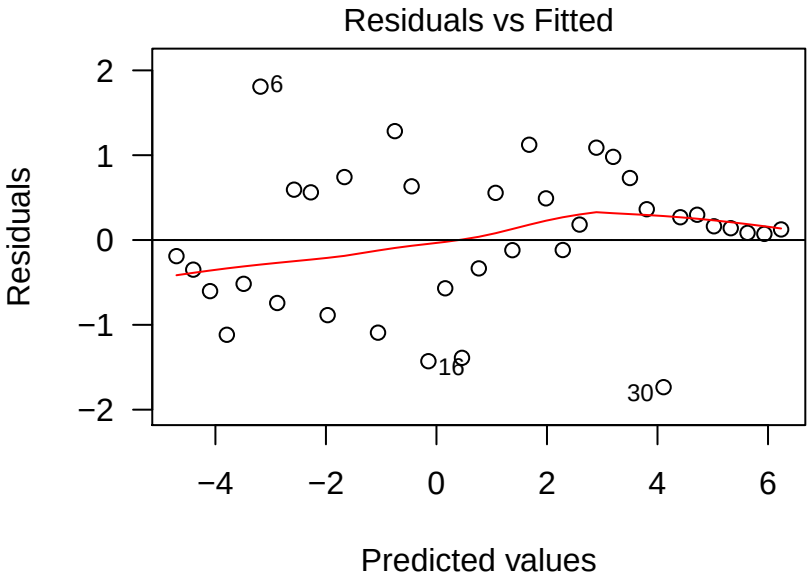


Figure 7.4: Plot of linear predictor values versus deviance residuals.

The residual plot (Fig. 7.4) could be criticized for the over-representation of negative residuals for small predicted values, and positive residuals for large predicted values.

However, it should be borne in mind that, unlike residuals from linear regression, the residuals from this model do not generally have median of zero. For example, for sufficiently large lengths the probability of retention is so close to unity that the most likely outcome is that all fish of that length will be retained, resulting in a small positive residual.

The fitted logistic regression model appears to be adequate, and so it is reasonable to proceed with further inference. Likelihood ratio confidence intervals are preferable to the Wald-based intervals (Section 4.4) given by the `summary` function. The likelihood ratio intervals for parameters β_0 and β_1 can be obtained from the `confint` function. and 95% intervals are calculated below.

It remains to make inference about the quantity of primary interest, the length of 50% retention, $l_{50} = -\beta_0/\beta_1$. The MLE is $\widehat{l}_{50} = -\widehat{\beta}_0/\widehat{\beta}_1 = 34.98$. The approximate

standard error of $\widehat{l}$ can be obtained via the delta method (Section 4.2.3), as shown below.

```
> #Likelihood ratio 95% confidence interval for regression coefs
> confint(Haddock.glm,level=0.95)
                 2.5 %     97.5 %
(Intercept) -12.4563256 -9.0501022
forklen       0.2604867  0.3535532

> #Wald 95% confidence interval for length of 50% retention
> MLE=coef(Haddock.glm)
> Vhat=vcov(Haddock.glm)
> L50hat=-MLE[1]/MLE[2]
> library(msm)
> L50.se=deltamethod(~-x1/x2,MLE,Vhat)
> L50.se
[1] 0.4246743

> cat("\n L50 is estimated to be",L50hat,"with 95% Wald CI (",
+     round(L50hat+c(-1,1)*qnorm(0.975)*L50.se,2),") \n")

 L50 is estimated to be 34.97927 with 95% Wald CI ( 34.15 35.81 )
```

The likelihood ratio confidence interval for l_{50} is preferred to the Wald, however, it is not as straightforward to obtain and requires use of a nonlinear optimizer (see Exercise 3.6). The 95% likelihood ratio confidence interval is $(34.12, 35.79)$.

Box 7.5.

A confidence interval for $\zeta = -\beta_0/\beta_1$ can be obtained using Fieller's method (Fieller 1954, Finney 1971) of calculating confidence intervals for ratios of normally distributed random variables. Calculation of the Fieller confidence interval is relatively easy, and is obtained explicitly as the solution of a quadratic equation. It is implemented in both R (`mratios` package) and SAS procedure `PROBIT`.

A number of authors have compared the relative performance of the confidence intervals constructed from use of the delta method, Fieller's method, and likelihood-ratio. In general, while Fieller's method tended to be preferable to the delta method, it was found to be conservative (i.e., have greater than the required coverage probability) and produce wider confidence intervals. Overall, likelihood-ratio confidence intervals are preferred (Williams 1986, Faraggi, Izikson and Reiser 2003).

7.5 Beyond binomial and Poisson models

The binomial and Poisson distributions provide the starting point for modeling of proportion and count data, respectively. However, it is frequently the case that a goodness-of-fit test will provide strong evidence against a model, yet there will be little indication (from diagnostic plots, say) that the specification of $E[Y_i]$ is at fault.

Indeed, in the next case study, the full model includes all three factor variables and their interactions, but nonetheless it is seen to exhibit gross "lack of fit" when the counts are modeled as Poisson distributed. The assumption that the counts are Poisson distributed is at fault here.

The inadequacy of the binomial or Poisson distributions is often due to lack of independence of the underlying Bernoulli trials in the case of proportion data, or of the occurrence of the events being enumerated in the case of count data. For example, if $B_j, j = 1, ..., m$ are iid Bernoulli(p) and Y is the proportion of "successes", $Y = \frac{1}{m}\sum_{j=1}^{m} B_j$, then Y is distributed as a binomial proportion, $Y \sim \frac{1}{m}\text{Bin}(m, p)$. However, if the B_j are all positively correlated then Y will still have mean p, but variance in excess of $p(1-p)/m$. A real-world example would be that of an insurance company whereby financial risk would be grossly underestimated if it assumed that the proportion of house insurance policies making claims was distributed according to the binomial distribution. A claim from one policy could be due to an extreme weather event that was also experienced by many other policy holders.

Similarly, count data on animal abundance is notoriously non-Poisson, because individuals of many species interact and congregate. That is, the presence of an animal tends to be positively correlated with the presence of others of that species, and hence the count of animals has variance in excess of that assumed by the Poisson distribution. Hospitals can not assumed that emergency patient arrivals will be Poisson distributed because, in the event of an epidemic (say), the arrival of one infected patient is a portend that large numbers of future arrivals may be imminent.

Note that here, as throughout this entire chapter, the data $Y_i, i = 1, ..., n$ are assumed to be independent (methods for dependent data are given in Chapters 8 and 10). The above arguments describe violation in the independence assumption of the underlying processes that gives rise to binomial or Poisson random variables. This violation typically results in the variance of the counts (or proportions) being greater than that expected under the Poisson (or binomial) model. The data are then said to be over-dispersed, or variance-inflated, or to have extra-Poisson variability in the case of count data.

Section 7.5.1 presents the convenient quasi-likelihood approach for fitting to over (or under) dispersed count and proportion data. The dispersion parameter ϕ of the Poisson and binomial distributions is unity, and quasi-likelihood instead utilizes an estimated value of ϕ. However, the resulting likelihood may not correspond to a formal statistical model (since it may not integrate to unity when integrated over the sample space), and hence the terminology *quasi*-likelihood. One limitation of this is that it can not be used when it is required to make prediction, because there is no underlying statistical model from which predictions can be postulated.

Box 7.6.

With proportion data, it can sometimes be the case that there is no clear notion of the number of trials, n. This can happen when the proportion is obtained from the percentage of a surface or volume that is composed of some constituent. For example, it could be the percentage of a bodily organ that is tumorous, or the percentage cover of a field by a particular grass, or the percentage of a leaf's surface that is covered in blotch. This is a situation to which quasi-likelihood (Sections 7.5.1 and 8) is well suited. Indeed, analysis of an historical set of barley-blotch data is given in Section 8.1.1.

Section 7.5.2 returns to the familiar framework of the parametric statistical model, and introduces the zero-inflated Poisson (ZIP) and negative binomial models as candidates for the modeling of zero-inflated and/or over-dispersed count data using R or SAS. Zero-inflation is another frequent cause of departure from the Poisson or binomial distributions. The ZIP model can be represented as a mixture of a Poisson distribution and a degenerate zero distribution, the latter taking the value zero with probability of unity. Bøhning, Dietz and Schlattmann (1999) used the zero-inflated Poisson to model the number of damaged teeth in seven year old school children. Conceptually, one might infer that some children took care of their teeth and incurred no damage to them, but for children who neglected their teeth the number that were damaged could be modeled by a Poisson distribution. The negative binomial distribution can also be very effective at modeling data that have excess zeros (compared to the Poisson), but its primary strength lies in modeling extra-Poisson variability in the non-zero counts.

7.5.1 Quasi-likelihood and quasi-AIC

Here, $\widehat{v}_i$ is used to denote the estimated value of the variance of Y_i under the fitted Poisson or binomial model. These assumed variances are, of course, $v_i = \mu_i$ for counts modeled as Poisson, and $v_i = p_i(1-p_i)/n_i = \mu_i(1-\mu_i)/n_i$ for proportion data modeled as binomial proportions, $Y_i \sim \frac{1}{n_i}\text{Bin}(n_i, p_i)$. Then, the Pearson chi-square statistic can be written

$$P_{\chi^2} = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{v}_i} \ .$$

If the data $y_i, i = 1, .., n$ are not overly sparse, and the model is correct, then the Pearson chi-square statistic will be approximately distributed χ^2_{n-p}, where p is the number of estimated parameters in the fitted model.

If the Pearson chi-square provides evidence of lack of fit, but specification of μ_i is believed to be satisfactory, then this suggests that v_i is not correctly specifying the variance of Y_i. A convenient adjustment is obtained by assuming that the true variance of Y_i, $\text{var}(Y_i)$, is a common multiplicative constant of the assumed variance v_i, for all i. That is,

$$\text{var}(Y_i) = kv_i \ . \tag{7.12}$$

With this adjustment the Pearson chi-square becomes

$$\begin{aligned} P^*_{\chi^2} &= \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{k\widehat{v}_i} \\ &= \frac{P_{\chi^2}}{k}. \end{aligned}$$

Since the χ^2_{n-p} distribution has expected value of $n-p$, it is natural to choose k so that $P^*_{\chi^2} = n - p$. That is, $\widehat{k} = P_{\chi^2}/(n-p)$.

Exponential family models have variance of the form in (7.12), where v_i is a function of μ_i and k is the dispersion parameter ϕ (this follows from equation (11.25) in Example 11.5). Whereas the dispersion parameter is unity for Poisson and binomial models, the quasi-likelihood version of these models sets the dispersion parameter equal to $\widehat{k}$. That is, the quasi-likelihood model estimates the dispersion parameter as

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{v}_i} \ . \tag{7.13}$$

The square-root of $\widehat{\phi}$ is called the estimated scale parameter. The dispersion parameter can also be estimated using the model deviance in place of the Pearson chi-square statistic. However, the Pearson is more robust to sparsity in the data and is generally preferred.

The above variance adjustment can be interpreted as extending the Poisson or binomial model by using $\phi = \widehat{\phi}$ (rather than $\phi = 1$) in their exponential family formulation $f(y; \psi, \phi)$ in equation 7.2. However, care must be taken with this interpretation because it is no longer the case that $f(y; \psi, \widehat{\phi})$ corresponds to a density function, since it does not integrate to unity. Hence, this approach is called quasi-likelihood. More general application of the notion of quasi-likelihood is covered in Section 8.

It is a trivial matter for GLM software to include the variance adjustment. For GLMs the derivatives of the log-likelihood are (from equation 7.6)

$$\frac{\partial l(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \beta_j} = \sum_{i=i}^{n} \frac{\frac{\partial \mu_i}{\partial \beta_j}(y_i - \mu_i)}{v_i} \ , \quad j = 1, ..., p \ .$$

The adjustment in (7.12), with k estimated b $\widehat{\phi}$, is simply scaling the derivative by a multiplicative constant,

$$\frac{\partial l(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \beta_j} = \frac{1}{\widehat{\phi}} \sum_{i}^{n} \frac{\frac{\partial \mu_i}{\partial \beta_j}(y_i - \mu_i)}{v_i} \ , \quad j = 1, ..., p \ . \tag{7.14}$$

Consequently, this does not alter the MLE $\widehat{\boldsymbol{\beta}}$.

In effect, the multiplicative correction to the derivatives in (7.14) is saying that inference should be based on using the log-likelihood divided by $\widehat{\phi}$ (albeit there may not be any underlying statistical model). Therefore, under this adjustment, likelihood ratio test statistics (and differences in deviance) are divided by $\widehat{\phi}$ before being evaluated against the appropriate χ^2 distribution. Since the estimated standard errors of the MLEs are inversely proportional to the magnitude of the second derivatives of the log-likelihood, it follows that the standard errors of the MLEs are multiplied by $\widehat{\phi}^{\frac{1}{2}}$ under this form of quasi-likelihood.

The quasi-likelihood correction to the AIC model selection criterion gives the

quasi-AIC

$$\text{QAIC} = -2\frac{l(\widehat{\boldsymbol{\theta}})}{\widehat{\phi}} + 2p \ . \tag{7.15}$$

Use of QAIC for model comparison makes sense only if a common value to $\widehat{\phi}$ is used (Burnham and Anderson 2002). This will typically be obtained by fitting the full model, provided this retains sufficient degrees of freedom. QAIC can not be used for comparing across different classes of model, but rather, can be used for selecting the appropriate explanatory terms to be used in a quasi-likelihood GLM for over-dispersed count or proportion data. In Section 7.6 it is also used to select between two choices of the link function.

7.5.2 Zero inflation and the negative binomial

If count or proportion data are over-dispersed, it is sometimes the case that they contain more zero observations than is plausible under the Poisson or binomial model. Then it may be sensible to modify the model to explicitly include the excess of zeros. For example, a zero-inflated Poisson (ZIP) is obtained as a probabilistic mixture of the zero distribution with a Poisson distribution. Specifically, a random variable $Y \sim \text{ZIP}(p, \lambda)$ is generated as

$$\begin{aligned} Y &= 0, && \text{with probability } p \\ Y &\sim \text{Pois}(\lambda), && \text{with probability } 1-p \ . \end{aligned}$$

Letting $1_{[y=0]}$ denote the indicator function that takes the value of unity if $y = 0$, and value of zero otherwise, the ZIP density function is (from equation 2.5, with $B_i = 1$ corresponding to observing Y from the zero distribution),

$$\begin{aligned} f(y; p, \lambda) &= p1_{[y=0]} + (1-p)\frac{e^{-\lambda}\lambda^y}{y!} \\ &= \begin{cases} p + (1-p)e^{-\lambda} & y = 0, \\ (1-p)\frac{e^{-\lambda}\lambda^y}{y!} & y = 1, 2, 3, ... \end{cases} \end{aligned}$$

The ZIP model can be fitted using `PROC GENMOD`, and in R, the ZIP and zero-inflated binomial can be fitted using the vglm function within the `VGAM` package. These software are able to model independent data $Y_i \sim \text{ZIP}(p_i, \lambda_i)$ using separate generalized linear models specifications for both the probabilities p_i, and Poisson means $\lambda_i, i = 1, ..., n$.

The negative binomial (NB) distribution is another alternative for modeling of over-dispersed count data. There are several different parameterizations of this distribution, but for current purposes it is convenient to use the NB(m, μ_i) parameterization (equation 15.1) with m fixed across all observations. Then var(Y_i) is a quadratic function of $\mu_i = E[Y_i]$,

$$\text{var}(Y_i) = \mu_i \left(1 + \frac{\mu_i}{m}\right) \ . \tag{7.16}$$

For a known value of m, the negative binomial is exponential family (Exercise 7.3), and hence it can be fitted efficiently using a one-dimensional numerical optimization with respect to m (Section 5.4.1). Negative binomial models, using the (m, μ_i) parameterization, can be fitted using `PROC GENMOD`, and using the `glm.nb` function in the `MASS` package within R or the `vglm` function in the `VGAM` package. Function `vglm` includes additional flexibility such as allowing non-constant m.

Example 7.7 ctd. If the fish counts are modelled as iid an NB(m, μ) then the resulting deviance is 15.29. The degrees of freedom remains 14, since the NB$(\widehat{m}, \widehat{\mu})$ and saturated models share the same value of $\widehat{m}$. The Pearson chi-square is 14.71. Neither of these statistics suggests lack of fit. □

Box 7.7.

When fitting negative binomial models, caution should be used when using the reported deviances of the fitted models. The likelihood of the saturated model is calculated using the estimated $\widehat{m}$ from the current model (see Example 7.4). Thus, the difference in deviance between two models no longer corresponds to twice the difference in log-likelihoods of those models.

A lesser used parameterization of the negative binomial uses the parameters (μ_i, p) (Section 15.3.1) where p is held constant. Then,

$$\text{var}(Y_i) = \frac{\mu_i}{p} \ ,$$

and so this parameterization is useful if it is desired to model var(Y_i) as proportional to μ_i. However, this parameterization corresponds to allowing the m parameter to

vary across observations, and can no longer be fitted as a conventional GLM[4], and hence may require the use of a general purpose optimizer.

The negative binomial is effective at inflating the number of zeros compared to the Poisson (Warton 2005). Nonetheless, this may not be always be enough. `PROC GENMOD` does not currently (in SAS 9.2) implement the zero-inflated negative binomial, but it can be fitted using the `vglm` function.

Example 3.1 ctd. In Example 3.1 a multinomial likelihood ratio test (the G-test) was applied to a frequency table of 40 counts of roots on apple cultivars. It was found that the Poisson model was definitely not appropriate for these counts (the G-test statistic was 75.2 on 7 degrees of freedom), but there was no evidence against the ZIP model (4.97 on 6 d.o.f.), which had MLE $(\widehat{p}, \widehat{\lambda}) = (0.4698, 4.6216)$.

Alternatively, models for these data can be assessed directly from the fit of the Poisson model, taking into account their complexity. One complication is that deviance is not strictly defined for the ZIP model (because it is not clear how the saturated model is specified), but Pearson chi-square and AIC can be used to aid in model assessment and comparison. The mean and variance of $Y \sim \text{ZIP}(p, \lambda)$ are $\mu = (1-p)\lambda$ and $\text{var}(Y) = (1+p\lambda)\mu$ (see Exercise 13.10), giving $P_{\chi^2} = 42.7$ with 38 d.o.f., and hence no evidence against lack of fit of the ZIP model.

Further validation of the ZIP model is obtained by comparison against the negative binomial and zero-inflated negative binomial (ZINB). The AICs for the fitted ZIP, NB and ZINB models are 153.76, 166.00 and 155.20, respectively (Exercise 7.8). □

7.6 Case study 2: Multiplicative vs additive model of over-dispersed counts

Like many (if not most) analyses of count data, extra-Poisson variation was detected in the data used here, and this case study presents an application of quasi-likelihood

[4]This model can be fitted using the `vglm` function in the `VGAM` package, by specifying appropriate parameter constraints.

and negative binomial modeling. ZIP and ZINB modeling could also have been investigated, and this is left as an exercise.

This analysis includes comparison of two link functions, the log and identity, because it was a primary research question to determine whether the effects of the covariates were best modeled as multiplicative or additive. That is, whether $E[Y_i] = \exp(\eta_i)$ or $E[Y_i] = \eta_i$, where η_i is the linear predictor.

7.6.1 Background

Willis and Millar (2005) counted snapper (*Pagrus auratus*) using a baited underwater video camera. The experiment was conducted in three marine reserves located on the NE coast of the North Island of New Zealand at four sampling times (Nov 1997, May 1998, Nov 1998 and May 1999). Between 12 and 24 replicate deployments of the camera were used at each combination of location and time. The total number of deployments was 204.

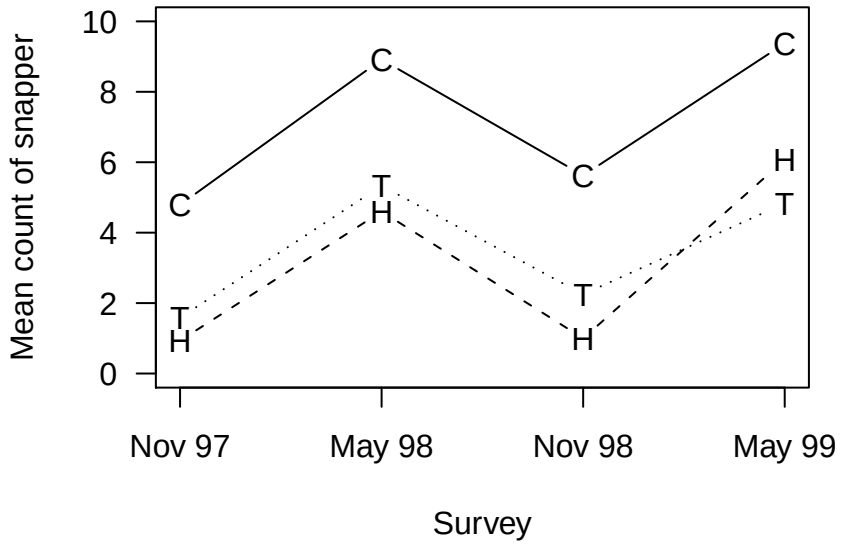


Figure 7.5: Mean video counts of snapper in three marine reserves at four sampling times.

The strong seasonal pattern in the counts (Fig. 7.5) is due to the seasonal migration of snapper moving from deep offshore waters into the coastal reserves. This movement is initiated by the increase in coastal water temperatures that begins in

the austral spring, typically in November or December. To gain greater understanding of the influence of marine reserves on the population dynamics of snapper, it was of interest to determine whether the seasonal migration had a multiplicative or additive effect on snapper density.

The data set `SnapperCounts.dat` contains the four variables, `reserve` (taking values "CROP", "Hahei" or "TAWH"), `season` ("Autumn" or "Spring", austral seasons corresponding to the May and November measurements, respectively). `year` ("97-98" and "98-99"), and the snapper count, `y`. The analysis begins by fitting a Poisson GLM with the full three-way interaction model[5]. The number of terms in the model was then reduced according to the sequence below, where R, S, and Y denote reserve, season and year effects, respectively, and $*$ denotes interaction.

1. $R * S * Y$: Full model.
2. $R * Y + Y * S$: Season (i.e., migration) effect that is not dependent on reserve.
3. $R * Y + S$: Season effect that is not dependent on reserve or year.
4. $R + Y + S$: All main effects only.
5. $R + S$: Reserve and season main effects only.
6. S: Season effect only.
7. 1: Null model (intercept only).

Each of the above seven specifications of effects was fitted using both a log link (i.e., multiplicative model) and identity link (i.e., additive model). The null and full models are identical under both links.

7.6.2 Poisson and quasi-Poisson fits

`PROC GENMOD` was used to fit the Poisson GLM under the full model.

```
TITLE "Poisson fit of full model";
PROC GENMOD;
  CLASS Reserve Season Year;
  MODEL y=Reserve*Season*Year / DIST=POISSON;
```

[5] At the time of the experiment, the three reserves were the totality of marine reserves of this kind on the NE coast of New Zealand, and it is therefore appropriate to treat them as fixed effects. Year is undoubtedly a random effect (Section 9.6), but with only two replicates, will be treated here as fixed.

Poisson fit of full model

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	192	702.3815	3.6582
Scaled Deviance	192	702.3815	3.6582
Pearson Chi-Square	192	614.0835	3.1984
Scaled Pearson X2	192	614.0835	3.1984
Log Likelihood		836.6711	
Full Log Likelihood		-627.9418	
AIC (smaller is better)		1279.8836	
AICC (smaller is better)		1281.5171	
BIC (smaller is better)		1319.7010	

The Pearson chi-square of 614.1 on 192 degrees of freedom clearly indicates lack of fit. The full model corresponds to modeling the replicate count data within each of the twelve reserve-year-season combinations as iid Poisson, and hence the lack of fit must be due to the inadequacy of the Poisson to describe the sampling variability in these replicate counts. Using the Pearson chi-square, dispersion is estimated to be $\widehat{\phi} = 614.0835/192 \approx 3.1984$ from equation (7.13).

It is desired to compare the seven models using quasi-AIC, and for this it is necessary to use the common value of $\widehat{\phi} = 3.1984$ from the fit of the full model. The SAS code below implements the additive quasi-Poisson fit of model 5. It uses the option `SCALE=1.7884` to specify $\widehat{\phi}^{\frac{1}{2}} = 1.7884$. The `IDENTITY` option specifies that the link function is the identity. That is, the Poisson means are linear (additive) functions of the regression coefficients.

```
TITLE "Quasi-Poisson additive fit of R+S model with fixed scale";
PROC GENMOD;
  CLASS Reserve Season Year;
  MODEL y=Reserve Season /  DIST=POISSON LINK=IDENTITY SCALE=1.7884;
```

The AIC value reported in the resulting goodness-of-fit table (not shown) is the desired QAIC. The QAIC values for the seven models and two link functions are shown in Table 7.2, and it is seen that the additive fit of model 5 is preferred.

		Quasi-likelihood		Negative binomial	
		Multiplicative	Additive	Multiplicative	Additive
Model	p	QAIC	QAIC	AIC	AIC
1: Full	12	416.66	416.66	1057.31	1057.31
2: $R*Y+Y*S$	8	420.00	410.58	1061.76	1050.90
3: $R*Y+S$	7	418.19	408.60	1059.91	1048.97
4: $R+Y+S$	5	414.57	404.71	1056.06	1045.25
5: $R+S$	4	413.51	403.64	1055.07	1044.01
6: S	2	456.70	456.70	1091.39	1091.39
7: Null	1	502.76	502.76	1116.06	1116.06

Table 7.2: Model selection results for the marine reserve data. The quasi-AIC values are computed using $\widehat{\phi} = 3.198$, and are not comparable to the AIC values from the negative binomial fit. p denotes the number of regression parameters.

The quasi-Poisson fit of model 5 was then repeated, but with the `SCALE=1.7884` option replaced by `PSCALE`. This instructs `GENMOD` to use the $\widehat{\phi}$ value estimated from the actual model, which is $\widehat{\phi} = 3.1187$ corresponding to a scale of $\widehat{\phi}^{\frac{1}{2}} = 1.7660$.

Quasi-Poisson additive fit of R+S model

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.7422	0.4186	0.9218	2.5626	17.32	<.0001
Reserve	CROP	1	3.4773	0.6224	2.2575	4.6972	31.22	<.0001
Reserve	Hahei	1	-0.6953	0.5088	-1.6926	0.3020	1.87	0.1718
Reserve	TAWH	0	0.0000	0.0000	0.0000	0.0000	.	.
Season	Autumn	1	3.9067	0.5178	2.8918	4.9215	56.92	<.0001
Season	Spring	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.7660	0.0000	1.7660	1.7660		

Note: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

The quasi-likelihood fit has identical parameter estimates to the Poisson fit, but the approximate standard errors have been inflated by the factor of 1.7660. The model is additive, so for example, the fitted expected count of snapper at the TAWH reserve in spring is approximately 1.74. This increases to approximately 1.74+3.91=5.65 in autumn.

7.6.3 Negative binomial fits

The above quasi-Poisson model inflated the Poisson variance by the multiplicative factor $\widehat{\phi}$, and hence is implicitly assuming $\text{var}(Y) \propto E[Y]$. However, a plot of variance versus mean within each of the 12 reserve-year-season combinations reveals that this is questionable. The plot suggests a quadratic relationship between variance and mean, and hence it would be natural to explore use of a negative binomial model.

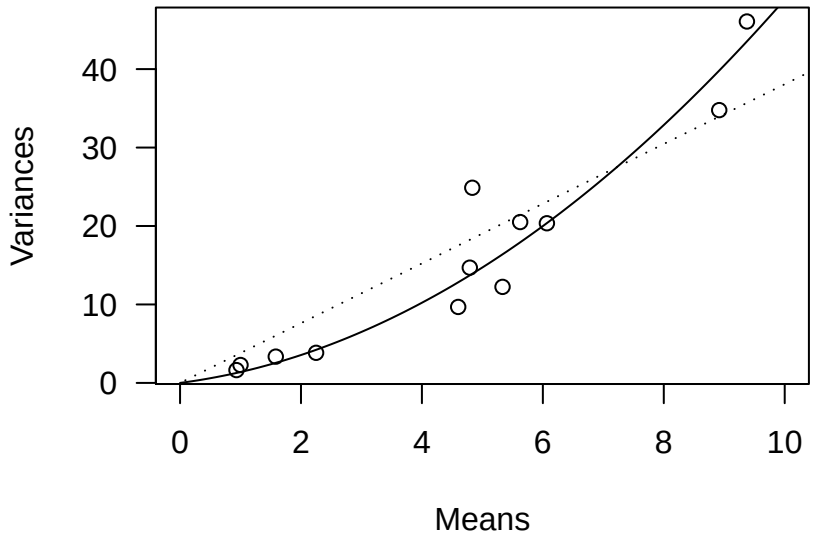


Figure 7.6: Plot of sample mean versus sample variance of the snapper count data within each of the 12 combinations of reserve, year and season. The dashed line shows the least-squares fit of $\text{var}(Y) = b\mu$, and the solid line shows the least squares fit of $\text{var}(Y) = \mu + b\mu^2$.

The negative binomial model is easily fitted with `PROC GENMOD`, simply by specifying the option `DIST=NEGBIN`.

```
TITLE "Negative binomial additive fit of R+S model";
PROC GENMOD;
  CLASS Reserve Season Year;
  MODEL y=Reserve Season /  DIST=NEGBIN LINK=IDENTITY;
```

Negative binomial additive fit of R+S model

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	200	241.6623	1.2083
Scaled Deviance	200	241.6623	1.2083
Pearson Chi-Square	200	165.1128	0.8256
Scaled Pearson X2	200	165.1128	0.8256
Log Likelihood		947.6084	
Full Log Likelihood		-517.0045	
AIC (smaller is better)		1044.0089	
AICC (smaller is better)		1044.3120	
BIC (smaller is better)		1060.5995	

The total number of parameters in the fitted model has increased from four to five, due to the addition of the negative binomial shape parameter, m. However, `PROC GENMOD` regards m as fixed in its calculation of the degrees of freedom in the above table. This is because the likelihood for the current model and saturated model are both calculated using the same value of m (see Box 7.7).

The Pearson chi-square does not show evidence against the fit, but the deviance does (p-value$\approx$ 0.23). It would be good practice to explore this further, and one possibility would be to determine if the zero-inflated negative binomial model provided a sufficient improvement in fit to warrant the extra complexity. At present (SAS version 9.2), the ZINB model is not implemented in `PROC GENMOD`, and the most expeditious way to fit this model might be to use the `vglm` function in the `VGAM` package of R. (The additive ZINB fit of model 5 has AIC of 1028.4, Exercise 7.9).

Finally, the estimates from the negative binomial additive fit of model 5 are shown below. Note that there is relatively good agreement of the MLEs and their standard errors between the quasi-Poisson and negative binomial fits.

Negative binomial additive fit of R+S model

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.8177	0.3672	1.0980	2.5374	24.50	<.0001
Reserve	CROP	1	3.3925	0.6845	2.0508	4.7342	24.56	<.0001
Reserve	Hahei	1	-0.8180	0.4269	-1.6547	0.0188	3.67	0.0554
Reserve	TAWH	0	0.0000	0.0000	0.0000	0.0000	.	.
Season	Autumn	1	3.9304	0.5894	2.7753	5.0855	44.47	<.0001
Season	Spring	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.5967	0.0942	0.4121	0.7813		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

7.7 Exercises

7.1 Show that the Poisson(λ) distribution is an exponential family distribution.

7.2 Show that the Gamma(α, β) distribution is an exponential family distribution.

7.3 Show that the negative binomial distribution NB(m, p) with *known* m,

$$f(y;p) = \frac{\Gamma(y+m)}{\Gamma(m)y!}p^m(1-p)^y, \quad y = 0, 1, ..., \quad m > 0, \quad 0 \leq p \leq 1$$

is an exponential dispersion family distribution with $\psi = \log(1-p)$ and $\phi = 1$.

7.4 In Exercise 11.8 it is seen that for exponential family distributions, $\mu = b'(\psi)$ and $\text{var}(Y) = b''(\psi)a(\phi)$. Using these equivalences, and the results of Exercises 7.2 and 7.3, calculate the mean and variance of the following random variables

1. $Y \sim \text{Gamma}(\alpha, \beta)$
2. $Y \sim \text{NB}(m, p)$

7.5 Determine the form of the deviance when y is observed from a Bin(n, p) distribution.

7.6 Repeat the analysis of the haddock length frequency data in Table 7.1 by modeling these data as counts using a Poisson log-linear GLM, and confirm that $\widehat{\beta}_0$, $\widehat{\beta}_1$, and their estimated standard errors are unchanged. Note that under the Poisson model, the data are 74 counts and `forklen` and `gear` (codend or cover) are fitted as factors (see Examples 9.1 and 9.2). The logistic selection curve corresponds to the interaction between `gear` and `forklen`.

7.7 Fit an iid negative binomial model to the fish count data in Example 7.7, and verify the deviance and Pearson chi-square statistics given in the continuation of that example in Section 7.5.2.

7.8 Use the `vglm` function in the `VGAM` package to verify the AIC values of the ZIP, NB and ZINB models stated in the continuation of Example 3.1 in Section 7.5.2.

7.9 Extend Model 5 in Section 7.6 by using the `vglm` function in the R package `VGAM` to fit a zero-inflated negative binomial model to the snapper count data in `SnapperCounts.dat`. Fit the model with constant zero-inflation probability p, and compare both log and identity links.

Chapter 8

Quasi-likelihood and Estimating functions

Section 7.5.1 introduced a special case of quasi-likelihood for the situation where the variance of the observations clearly differed from that specified under the assumed generalized linear model. There, it was assumed that the true variance of each observation, $\text{var}(Y_i)$, was proportional to the variance specified under the GLM. That is, letting v_i denote the variance specified by the GLM (e.g., $v_i = \mu_i$ if Y_i is modeled as Poisson(μ_i)), it was assumed that $\text{var}(Y_i) = \phi v_i$. More generally, quasi-likelihood is a flexible methodology for parameter estimation that requires only specification of $\mu_i = E[Y_i]$ and $\text{var}(Y_i)$, yet it enjoys many of the features of maximum likelihood estimation. It achieves this through the use of estimating functions in place of the score function (the derivative of the log-likelihood). An estimating function possesses the salient properties of the score function, thereby ensuring that the resulting quasi-likelihood estimator (QLE) will be consistent and have an approximate normal distribution under weak conditions. Here the focus is on application of quasi-likelihood and estimating equations, with the formal details and properties provided in Sections 12.2.4 and 12.3.2.

Section 8.1 looks at the implementation of quasi-likelihood that is due to Wedderburn (1974). This encapsulates the application of quasi-likelihood in Section 7.5.1, but removes the requirement that $E[Y_i]$ and $\text{var}(Y_i)$ be obtained from a dispersion-adjusted generalized linear model. Instead, it is enough that $E[Y_i]$ and $\text{var}(Y_i)$ are smooth functions of parameters $\boldsymbol{\theta}$. When $E[Y_i]$ does follow the structure of a gener-

alized linear model, but $\text{var}(Y_i)$ is otherwise arbitrary, then the QLE can be obtained using the `glm` function in R or the SAS procedure `GENMOD`

Liang and Zeger (1986) showed that the quasi-likelihood approach could be applied to the analysis of grouped data, as might be obtained from repeated measurements on randomly chosen experimental subjects. This methodology is commonly known as generalized estimating equations (GEEs), and is demonstrated in an analysis of drug efficacy data that is grouped within clinics (Section 8.2). These data are subsequently re-analyzed using a mixed-effects model in Section 10.6, where comparison is made with the GEE approach.

In the form in which it is applied in this chapter, quasi-likelihood enables the modeler to make inference about $\boldsymbol{\theta}$ from specification of only $E[Y_i]$ and $\text{var}(Y_i)$. This is advantageous in situations there may be no convenient statistical model from which to construct a likelihood function from the data, or when there is uncertainty over the choice of model. The analysis of leaf blotch data (Section 8.1.1) demonstrates a situation where it is unclear how to specify a likelihood for proportion data. A binomial model does not suit these data, because the proportion is judged by visual inspection and there is no concept of the "number of trials", n. Quasi-likelihood provides a robust alternative to maximum likelihood in situations where there is uncertainty over the form of the likelihood. Indeed, for the QLE to be a consistent estimator it is enough that $E[Y_i]$ be correctly specified (see condition (12.35) in Section 12.2.4).

There is a price to be paid for the flexibility and robustness offered by quasi-likelihood. For example, without the notion of an underlying statistical model it is not possible to predict future observations. There is limited ability to compare models using quasi-AIC. Moreover, robustness comes at the price of efficiency, and the QLE has higher large-sample variance than the MLE. Of course, this is a rather moot point if the true form of the likelihood is not well specified.

8.1 Wedderburn's quasi-likelihood

Let $Y_i, i = 1, ..., n$ be independent with distribution depending on $\boldsymbol{\theta} \in \mathbb{R}^p$. It is assumed that $\mu_i = E[Y_i]$ and $\text{var}(Y_i)$ are smooth (i.e., differentiable) functions of $\boldsymbol{\theta}$.

The quasi-likelihood estimate $\widetilde{\boldsymbol{\theta}}$ of the true unknown $\boldsymbol{\theta}_0$ is obtained as the solution of the p-dimensional set of estimating equations

$$\sum_i^n \frac{\frac{\partial \mu_i}{\partial \theta_j}(y_i - \mu_i)}{v_i} = 0 \ , \qquad j = 1, ..., p. \tag{8.1}$$

Wedderburn (1974) assumed $v_i \propto \text{var}(Y_i)$ and showed that, for sufficiently large n, $\widetilde{\boldsymbol{\theta}}$ is approximately normal distributed with mean $\boldsymbol{\theta}_0$. More generally, this large sample approximation holds even when v_i is mis-specified (Section 12.3.2), but at the cost of an increase in the variance of $\widetilde{\boldsymbol{\theta}}$ (Exercise 8.1). Note that v_i need only be specified up to a multiplicative constant because the variance-scaling technique of Section 7.5.1 can be employed.

In the following example, μ_i has generalized linear model form and the quasi-likelihood estimate was obtained using the `glm` function in R, but with a user-defined form of $\text{var}(Y_i)$. Implementation in `PROC GENMOD` is analogous, with the specification of $\text{var}(Y_i)$ being implemented via the `VARIANCE` statement.

8.1.1 Barley blotch data

The data used herein are from the original quasi-likelihood example of Wedderburn (1974), which was subsequently revisited by McCullagh and Nelder (1989). The observations are the proportion of leaf afflicted by blotch, measured on ten varieties of barley at nine sites. These data are available as the dataset `barley` in R package `gnm`, where the 90 observations are given in three columns labeled, `y`, `site`, and `variety`, respectively.

```
> data(barley,package="gnm")
> #Put data in table format for presentation
> BarleyTable=matrix(barley$y,ncol=10,
+    dimnames=list(paste("site",1:9,sep=""),paste("vrty",1:10,sep="")))

> BarleyTable
       vrty1  vrty2  vrty3  vrty4 vrty5 vrty6 vrty7 vrty8 vrty9 vrty10
site1 0.0005 0.0150 0.0100 0.2000 0.250 0.080 0.050  0.05 0.050  0.250
site2 0.0000 0.0000 0.1270 0.3750 0.550 0.165 0.050  0.50 0.050  0.425
```

```
site3 0.0000 0.0005 0.0125 0.2625 0.050 0.295 0.100  0.10 0.250  0.500
site4 0.0010 0.0005 0.0125 0.0250 0.400 0.200 0.050  0.50 0.750  0.375
site5 0.0025 0.0030 0.0250 0.0050 0.055 0.435 0.500  0.25 0.500  0.950
site6 0.0005 0.0075 0.1660 0.0001 0.010 0.010 0.750  0.50 0.750  0.625
site7 0.0050 0.0030 0.0250 0.0300 0.060 0.050 0.050  0.75 0.750  0.950
site8 0.0130 0.0300 0.0250 0.0250 0.011 0.050 0.001  0.05 0.750  0.950
site9 0.0150 0.0750 0.0000 0.0001 0.025 0.050 0.050  0.10 0.175  0.950
```

It was desired to fit a logistic model of the form

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})} \ ,$$

where $p_{ij} \equiv \mu_{ij}$ denotes the expected proportion of blotch at site i on variety j, and η_{ij} is the linear predictor containing the main effects of site and variety. The logistic model is fitted using the `glm` function in the R code below. The absence of a `weight` variable in the call of `glm` reflects the fact that there is no notion of the number of trials, n_{ij}, associated with each measurement of the proportion of leaf blotch. The quasi-binomial fit implicitly assumes all n_{ij} are equal, that is, $\text{var}(Y_{ij}) \propto p_{ij}(1-p_{ij})$.

```
> quasibinom.fit=glm(y~variety+site,family=quasibinomial,data=barley)
> #Use which=1 to select the plot of deviance resids vs linear predictors
> plot(quasibinom.fit,which=1)
> summary(quasibinom.fit)

Call:
glm(formula = y ~ variety + site, family = quasibinomial, data = barley)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.64435  -0.13537  -0.02028   0.09500   0.81003

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.0546     1.4219  -5.665 2.84e-07 ***
variety2      0.1501     0.7237   0.207 0.836289
variety3      0.6895     0.6724   1.025 0.308587
variety4      1.0482     0.6494   1.614 0.110910
variety5      1.6147     0.6257   2.581 0.011895 *
variety6      2.3712     0.6090   3.893 0.000219 ***
variety7      2.5705     0.6065   4.238 6.58e-05 ***
variety8      3.3420     0.6015   5.556 4.39e-07 ***
variety9      3.5000     0.6013   5.820 1.51e-07 ***
varietyX      4.2530     0.6042   7.039 9.38e-10 ***
siteB         1.6391     1.4433   1.136 0.259870
siteC         3.3265     1.3492   2.466 0.016066 *
siteD         3.5822     1.3444   2.664 0.009510 **
siteE         3.5831     1.3444   2.665 0.009493 **
siteF         3.8933     1.3402   2.905 0.004875 **
siteG         4.7300     1.3348   3.544 0.000697 ***
siteH         5.5227     1.3346   4.138 9.38e-05 ***
siteI         6.7946     1.3407   5.068 3.00e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasibinomial family taken to be 0.08877777)
```

```
    Null deviance: 40.803  on 89  degrees of freedom
Residual deviance:  6.126  on 72  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 8
```

The coefficients from the above fit are identical to those from a standard logistic regression, however the dispersion parameter has been estimated from the Pearson chi-square statistic, $P_{\chi^2} = 6.392$. This gives $\widehat{\phi} = P_{\chi^2}/720.0888$. That is, it has been assumed that $\text{var}(Y_{ij}) = 0.0888 p_{ij}(1 - p_{ij})$.

The residual plot from the quasi-binomial logistic regression (Fig. 8.1) shows a clear increase in the magnitude of the residual with increasing $\widehat{\eta}_{ij}$, followed by a possible decrease for values of $\widehat{\eta}_{ij}$ in excess of zero. This suggests that, compared to the quasi-binomial, there may be relatively more variability in the data for moderate values of η (corresponding to p around 0.5) than for extreme values of η (corresponding to p close to 0 or 1). Consequently, Wedderburn (1974) suggested using the variance function $\text{var}(Y_{ij}) \propto p_{ij}^2(1 - p_{ij})^2$.

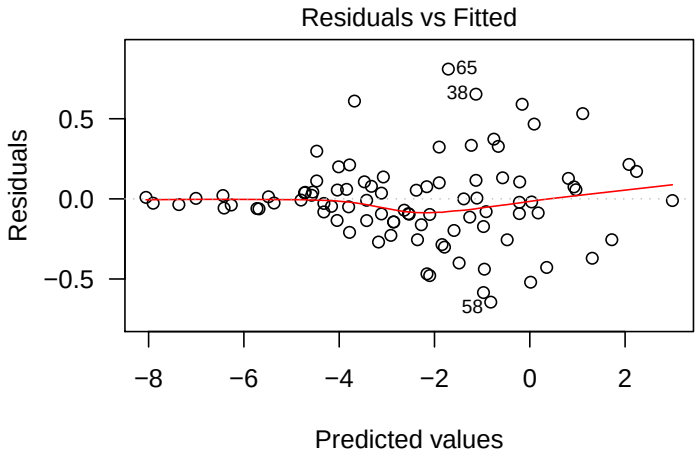


Figure 8.1: Plot of deviance residuals (uncorrected for dispersion) from a logistic regression fit to the barley blotch data.

Wedderburn's variance function is the first component of the list object created in the R code below. The list includes several other components, including a function definition for a formula `dev.resids` for calculation of the squared deviance residual, that is, the contribution to the model deviance. Here, since no explicit notion of a likelihood is utilized, it is more convenient to substitute this with the binomial form

of the Pearson chi-square statistic.

```
> BarleyVar=list(
+   varfun=function(mu) (mu*(1-mu))^2,
+   validmu=function(mu) all(mu > 0) && all(mu < 1),
+   dev.resids=function(y, mu, wt) wt * ((y - mu)^2)/(mu*(1-mu))^2,
+   initialize=expression({
+             n <- rep.int(1, nobs)
+             mustart <- pmax(0.001, pmin(0.999, y)) }),
+   name="(mu(1-mu))^2" )
```

The quasi-likelihood fit with Wedderburn form of variance is obtained by using the `quasi` function in the `family` option of `glm`. The arguments of the `quasi` function include specification of the link and variance functions. The residual plot (Fig. 8.2) from the Wedderburn quasi-likelihood fit shows a big improvement.

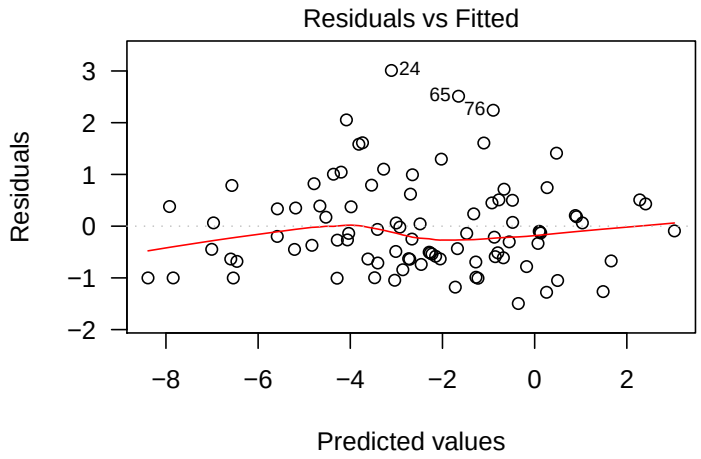


Figure 8.2: Plot of Pearson residuals from quasi-likelihood fit of logistic model to the barley blotch data, using Wedderburn form of variance.

```
> Wedderburn.fit=glm(y~site+variety,
+               family=quasi(link="logit",variance=BarleyVar),data=barley)

> plot(Wedderburn.fit,which=1) #Deviance resids vs linear predictors
> summary(Wedderburn.fit)

Call:
glm(formula = y ~ site + variety, family = quasi(link = "logit",
    variance = BarleyVar), data = barley)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4952  -0.6342  -0.1397   0.4450   3.0101

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.92238    0.44465 -17.817  < 2e-16 ***
siteB        1.38312    0.44465   3.111  0.00268 **
siteC        3.86006    0.44465   8.681 8.20e-13 ***
```

```
siteD        3.55700    0.44465   8.000 1.54e-11 ***
siteE        4.10786    0.44465   9.239 7.53e-14 ***
siteF        4.30536    0.44465   9.683 1.13e-14 ***
siteG        4.91810    0.44465  11.061  < 2e-16 ***
siteH        5.69489    0.44465  12.808  < 2e-16 ***
siteI        7.06763    0.44465  15.895  < 2e-16 ***
variety2    -0.46735    0.46870  -0.997  0.32204
variety3     0.07881    0.46870   0.168  0.86695
variety4     0.95408    0.46870   2.036  0.04547 *
variety5     1.35263    0.46870   2.886  0.00515 **
variety6     1.32854    0.46870   2.835  0.00595 **
variety7     2.34007    0.46870   4.993 4.01e-06 ***
variety8     3.26258    0.46870   6.961 1.30e-09 ***
variety9     3.13549    0.46870   6.690 4.10e-09 ***
varietyX     3.88727    0.46870   8.294 4.34e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasi family taken to be 0.9885464)

    Null deviance: 252.155  on 89  degrees of freedom
Residual deviance:  71.175  on 72  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 20
```

The reported residual deviance of 71.175 is the Pearson residual as defined in the `BarleyVar` function, which implicitly assumed $n_{ij} = 1$. It is pure coincidence that this provided a value very close to the 72 degrees of freedom. The standard errors in the above output assumed a dispersion parameter of $71.175/72 \approx 0.989$, that is, $\text{var}(Y_{ij}) = 0.989 p_{ij}^2 (1 - p_{ij})^2$.

8.2 Generalized estimating equations

Suppose that m independent subjects are randomly chosen and that n_i measurements $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})$ are recorded from subject i. Here the terminology "subject" is used generically to represent the independent experimental units from which the measurements are taken. The subject could be an individual, but in the example below, the "subjects" are medical clinics from which observations on the success of a drug treatment are recorded.

Typically, there will be random variability between the subjects. Any random variability in subject i will affect all measurements made on that subject, and so it is to be anticipated that $y_{ij}, j = 1, ..., n_i$ will be correlated, and so it may not be appropriate to model all observations $y_{ij}, i = 1, ..., m, j = 1, ..., n_i$ as independent. The generalized estimating equations approach assumes that $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})$ is

a multivariate random vector with mean $(\mu_{i1}, ..., \mu_{in_i})$ depending on parameters $\boldsymbol{\theta}$, and variance matrix var$(\boldsymbol{Y}_i)$ that may depend on both $\boldsymbol{\theta}$ and additional *association* parameters $\boldsymbol{\alpha}$ that specify the correlation structure.

Box 8.1.

GEEs model $\boldsymbol{Y}_i$ at the marginal (or population) level through specification of the mean and variance structure. In contrast, hierarchical models specify the distribution of $\boldsymbol{Y}_i$ conditional on the random variability in subject i, that is, at the subject level. It may be that $Y_{ij}, j = 1, ..., n_i$ are independent conditional on subject i variability, yet unconditionally (i.e., marginally) they are correlated due to that shared variability.

The correlation structure that is imposed on the observations within each subject should be chosen to match the manner in which the observations are made. For example, if observations $y_{ij}, j = 1, ..., n_i$ on subject i are collected sequentially (with increasing j) then it might be appropriate to assume an auto-regressive correlation structure of the form $\text{cor}(y_{ij}, y_{ik}) = \alpha^{|j-k|}$ where $|\alpha| < 1$. In the absence of an ordering then it may be reasonable to assume exchangeability, corresponding to $\text{cor}(y_{ij}, y_{ik}) = \alpha$ for all $j \neq k$. An unstructured correlation matrix, $\text{cor}(y_{ij}, y_{ik}) = \alpha_{jk}$ is another possibility, but one that should be used cautiously due to the large number of large number of correlation parameters required. It can also be useful to use a GEE model with as assumed independence structure, $\text{cor}(y_{ij}, y_{ik}) = 0, j \neq k$, to take advantage of the robust variance estimator.

Given the assumed specification of the within-subject correlation structure, denoted cor$(\boldsymbol{Y}_i) = \mathbf{R}_i(\boldsymbol{\alpha})$ the assumed variance matrix for $\boldsymbol{Y}_i$ is

$$\mathbf{V}_i = \mathbf{D}(\boldsymbol{\theta})\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{D}(\boldsymbol{\theta}) \ ,$$

where $\mathbf{D}(\boldsymbol{\theta})$ is a $n_i \times n_i$ diagonal matrix having the standard deviation of Y_{ij} in the jth diagonal position, for $j = 1, ..., n_i$. These standard deviations are determined from a specified exponential family distribution. For example, if the Poisson is specified then $\text{sd}(Y_{ij}) = \mu_{ij}^{\frac{1}{2}}$.

In the special case where the independence structure is assumed, then $\mathbf{V}_i$ is precisely the variance under the specified exponential family distribution. It follows that the estimating equation in (12.39) is equivalent to **??**eq:ExpoScore), and hence

that $\widetilde{\boldsymbol{\theta}}$ is identical to the MLE $\widehat{\boldsymbol{\theta}}$ under that exponential family model. However, the GEE approach permits use of the robust variance estimator $\widehat{\text{var}}(\widetilde{\boldsymbol{\theta}})$ given by **??**eq:MEstimatorVar).

For a given value of the association parameters $\boldsymbol{\alpha}$, an estimator $\widetilde{\boldsymbol{\theta}}$ can be obtained from a multivariate version of the estimating equations in Section 8.1 (see Example 12.6) . The residuals from this fit can then be used to produce a new value of $\boldsymbol{\alpha}$. Hardin and Hilbe (2003) give examples of explicit formulae for calculation of $\boldsymbol{\alpha}$ under several choices of the assumed correlation matrix. The iterative process of estimating $\boldsymbol{\alpha}$ and $\widetilde{\boldsymbol{\theta}}$ is continued until convergence.

The variance of $\widetilde{\boldsymbol{\theta}}$ is typically estimated using the sandwich estimator in (12.43), which is also called the robust or empirical estimator. This estimator is valid under mis-specification of the variance structure. In the case of an assumed independence structure, the GEE estimator $\widetilde{\boldsymbol{\theta}}$ is necessarily identical to the GLM (i.e., maximum likelihood) estimator $\widehat{\boldsymbol{\theta}}$. However, the GEE method provides an estimate of variance that is robust to failure of the independence assumption.

GEEs are implemented in SAS using the `REPEATED` statement in the `GENMOD` procedure. Within R, the `geepack` package provides function `geeglm`. Hardin and Hilbe (2003) note that competing software packages may give differing fits due to nuances in the denominator degrees-of-freedom term used in calculation of the association parameters. This could be problematic when the number of subjects m is small, especially if a large number of association parameters are used to model the correlation structure. For the correlated binomial data investigated in Section 8.2.1, `GENMOD` would not converge using default options. However, it gave $\widetilde{\boldsymbol{\theta}}$ very close to that obtained by R function `geeglm` when the `V6CORR` option was used in the `REPEATED` statement to force different calculation of the correlation parameter. In any case, the example below uses another variant (Carey, Zeger and Diggle 1993) which is more appropriate for binomial data.

8.2.1 Multi-center trial

The data (Beitler and Landis 1985) are from testing the effectiveness of a cream for treating infection. Trials using treatment and control creams were performed at

eight clinics, and the binomial outcome is the number of favourable outcomes (Fig. 8.3). It is seen that the treatment resulted in a higher proportion of favourable outcomes, with the sole exception of the clinic having the smallest total sample size (Clinic 8).

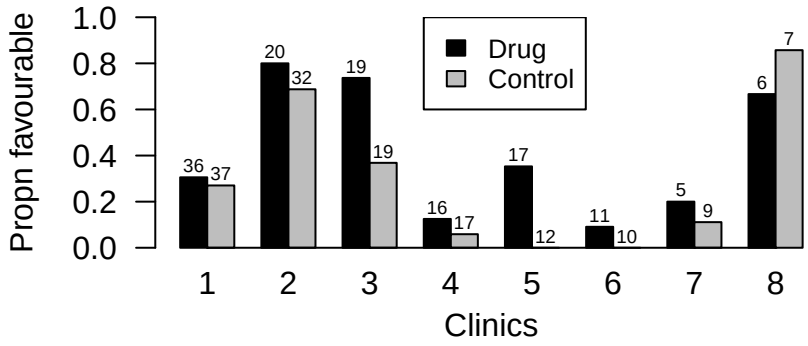


Figure 8.3: Proportion of favourable results from a multi-center trial investigating the effectiveness of two topical cream treatments for curing infection. Numbers above the bars give the sample size.

Between-clinic variability in the proportions is also visually evident. The research objective (Beitler and Landis 1985, p. 992) required "extended inferences from these data", meaning that it is necessary to model randomness in the clinics so as to make wider inference in the efficacy of the treatment.

Under the GEE approach, the clinics are considered to be randomly chosen experimental "subjects". Within each clinic, the GEE allows the two binomial responses (treatment and control) to be correlated. In this particular case, positive correlation is suggested (Fig. 8.3). That is, the within-clinic measurements are modeled as a bivariate binomial observation. Being bivariate, there is only one choice of a non-independence correlation structure within each clinic,

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} . \tag{8.2}$$

A quasi-binomial logistic model is fitted by the code shown below. In the `DATA` step, this code uses the `@@` option on the `INPUT` statement so that multiple observations can be read from a single input line. Also, the `REF=FIRST` option in `GENMOD`'s `CLASS` statement is used so that lowest level of `trmt`, zero, will correspond to the

intercept parameter, and the treatment parameter will correspond to the treatment effect.

```
DATA infection;
INPUT clinic trmt y n @@;
LINES;
1 1 11 36   1 0 10 37   2 1 16 20   2 0 22 32
3 1 14 19   3 0 7 19    4 1 2 16    4 0 1 17
5 1 6 17    5 0 0 12    6 1 1 11    6 0 0 10
7 1 1 5     7 0 1 9     8 1 4 6     8 0 6 7
RUN;

TITLE2 "Quasi-Poisson fit to Multi-center data";
*Use REF so that intercept coefficient is the control term;
PROC GENMOD DATA=infection;
      CLASS clinic trmt / PARAM=REF REF=FIRST;
      MODEL y/n = trmt / DIST=BINOMIAL PSCALE;
RUN;
```

The logistic model estimates the probability of a favourable outcome to be $e^{-0.7142}/(1+e^{-0.7142}) = 0.329$ for the control treatment, and $e^{-0.3102}/(1+e^{-0.3102}) = 0.423$ for the drug treatment. These are simply the observed proportions of favourable outcomes by the two treatments when the data are combined over clinics. The Pearson χ^2 from the logistic fit is 81.51, and dividing by the 14 degrees of freedom gives an estimated over-dispersion of $\widehat{\phi} = 5.822$. The quasi-binomial fit therefore inflates the estimated standard errors of $\widehat{\boldsymbol{\theta}}$ from the logistic fit by a factor of $\widehat{\phi}^{\frac{1}{2}} = 2.413$.

Quasi-Poisson fit to Multi-center data

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.7142	0.4296	-1.5561	0.1277	2.76	0.0964
trmt	1	1	0.4040	0.6066	-0.7850	1.5931	0.44	0.5054
Scale		0	2.4130	0.0000	2.4130	2.4130		

Note: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

A better analysis would use the grouping structure in the data to explain the over-dispersion. The following code fits a binomial GEE using the independence working correlation structure.

```
TITLE2 "Independence GEE fit to Multi-center data";
PROC GENMOD DATA=infection;
      CLASS clinic trmt / PARAM=REF REF=FIRST;
      MODEL y/n = trmt / DIST=BINOMIAL PSCALE;
  REPEATED SUBJECT=clinic / CORR=INDEP;
RUN;
```

The estimated $\widetilde{\boldsymbol{\theta}}$ is necessarily identical to the MLE $\widehat{\boldsymbol{\theta}}$ from the quasi-binomial fit. However, standard errors now differ due to use of the robust variance estimator.

Independence GEE fit to Multi-center data

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-0.7142	0.4505	-1.5971	0.1687	-1.59	0.1129
trmt	1	0.4040	0.2462	-0.0785	0.8866	1.64	0.1008

Hardin and Hilbe (2003, p. 89) note limitations in the ability to specify correlation structure on binomial proportions, and recommend instead using the approach of Carey et al. (1993) whereby the correlation structure is applied to log-odds ratios. To use this approach, `PROC GENMOD` requires that the binomial data be in Bernoulli form with one row per outcome. The `LOGOR=EXCH` option can then be used to specify the correlation structure in (8.2).

```
DATA binary;
 SET infection;
 KEEP clinic trmt outcome;
 DO i=1 TO y;
   outcome=1;
   OUTPUT;
 END;
 DO i=1 TO n-y;
   outcome=0;
   OUTPUT;
 END;
RUN;


TITLE2 "Exhangeable correlation GEE fit to Multi-center data";
*Use DESCENDING so that trmt effect is for outcome=1;
PROC GENMOD DATA=binary DESCENDING;
      CLASS clinic trmt / PARAM=REF REF=FIRST;
      MODEL outcome = trmt / DIST=BINOMIAL PSCALE;
  REPEATED SUBJECT=clinic / LOGOR=EXCH;
RUN;
```

The correlation is estimated to be $\widehat{\alpha} = 0.8965$ and the GEE estimate $\widetilde{\boldsymbol{\theta}}$ is no longer equal to $\widehat{\boldsymbol{\theta}}$. However, the most important change is that, by appropriate modeling of the grouping structure, a more efficient estimator is obtained. This is seen here by the p-value for `trmt`, which now shows strong evidence for a treatment effect.

Exhangeable correlation GEE fit to Multi-center data

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-0.8740	0.4694	-1.7940	0.0460	-1.86	0.0626
trmt	1	0.5532	0.2318	0.0989	1.0074	2.39	0.0170
Alpha1		0.8965	0.4198	0.0738	1.7192	2.14	0.0327

In particular, note that compared to the quasi-binomial fit, the GEE fits see an increase in standard error of the intercept parameter, but a decrease in that of the treatment effect. This can be reasoned as follows. The high positive correlation within clinics essentially reduces the amount of information that each clinic provides about the expected proportion of favourable outcomes, hence the relatively higher standard error on the intercept. However, the within-clinic correlation does not obfuscate the treatment differences.

8.3 Exercises

8.1 For independent Y_i and $\theta \in \mathbb{R}$, let $\widetilde{\theta}$ be the QLE obtained from Wedderburn's quasi-likelihood equation. For sufficiently large n, it can be shown (see Section **??**) that $\text{var}(\widetilde{\theta}) \approx A/B^2$ where

$$A = \sum \frac{\dot{\mu}_i^2 \text{var}(Y_i)}{v_i^2} \text{ , and } \qquad B = \sum \frac{\dot{\mu}_i^2}{v_i} \text{ ,}$$

where $\dot{\mu}_i = \partial \mu_i / \partial \theta$.

Show that this approximate variance is minimized when $v_i = \text{var}(Y_i)$.

Hint: Denote $x_i = \dot{\mu}_i / \sqrt{\text{var}(Y_i)}$ and $y_i = \dot{\mu}_i \sqrt{\text{var}(Y_i)}/v_i$, and apply the Cauchy-Schwarz inequality.

8.2 The method of moments uses an estimating equation that equates (non-central) sample moments to the population moments. For example, for Y iid from the zero-inflated Poisson distribution (see Exercise 2.11), it follows from Exercise 13.10 that $E[Y] = (1-p)\lambda$ and $E[Y^2] = E[Y](1+\lambda)$. An estimating equation for parameters λ and p is obtained as

$$\begin{pmatrix} \overline{y} - E[Y] \\ \overline{y^2} - E[Y^2] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} .$$

a) Solve this estimating equation to obtain the method-of-moments estimates, $\widetilde{\lambda}$ and $\widetilde{p}$, and apply to the ZIP data in Exercise 3.7.

b) Apply the bootstrap to the method-of-moments estimator, to obtain approximate 95% confidence intervals for λ and p.

Chapter 9

ML inference in the presence of incidental parameters

The models encountered so far have possessed likelihoods that were relatively easy to maximize. These models either contained no more than a handful or parameters (and so were easily maximized using numerical optimizers) or were conventional models that could be fitted using readily available software. Moreover, it has generally been safe to assume that the large-sample properties of ML inference would apply for sufficiently large sample sizes[1].

Cross ref.

This chapter has two purposes. The first is to present a few techniques for maximizing the likelihood when the dimensionality of $\boldsymbol{\theta}$ is large. The second is a more diverse purpose, but can loosely be stated as a presentation of methodology that may protect the estimators of selected parameters from undesirable consequences arising from the estimation of other parameters. This is the genesis of the title for this Chapter.

In many applications, it will be the case that only a small number of parameters are directly relevant to the underlying research question. These will be called the parameters of interest. However, a complex model with many additional parameters may be required to appropriately model the sampling distribution of the observed data. This situation frequently arises when the observations are not independent and the model is therefore required to capture correlation structure. This includes

[1]One exception was Example 4.8 where the hypothesized parameter value, $\boldsymbol{\theta}_0$, was on the boundary of the parameter space, and hence the theory underlying the large-sample behaviour of the MLE did not apply

repeated measures, longitudinal, temporal and spatial models. This correlation structure may or may not be of interest, depending on the research question. In the worst circumstances, naive use of maximum likelihood inference can lead to inferential disaster.

The remaining parameters are commonly called incidental or nuisance parameters. Let $\boldsymbol{\theta} = (\psi, \lambda)$ where λ contains the incidental parameters (ψ and λ are both possibly vectors).

In addition to the added numerical complexity of maximizing a likelihood with incidental parameters, if the number of incidental parameters increases with sample size then the consistency and asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$ need not hold.

Quasi-likelihood and estimating functions(Chapter **??**) are approaches that can be used in some situations, particularly analysis of longitudinal data where it would be very difficult to model the complex covariance structure of the repeat observations on the experimental subjects.

In this Chapter we seek a modified form of likelihood inference that does not depend on the incidental parameters, and uses as much information in the data as possible. We look at methods that involve working with a modified likelihood function and look at the theoretical arguments justifying their use.

9.1 Conditional likelihood

This method can be used if a "suitable" sufficient statistic $S_\lambda(\boldsymbol{Y})$ for λ exists.

Suppose that for any fixed value of ψ, S_λ is sufficient for λ, i.e., the conditional distribution of $\boldsymbol{Y}$ (given S_λ) does not depend on λ. That is

$$f_{Y|s_\lambda}(\boldsymbol{y}|s_\lambda;\psi,\lambda) \equiv f_c(\boldsymbol{y}|s_\lambda;\psi) = \frac{f_Y(\boldsymbol{y};\psi,\lambda)}{f_{s_\lambda}(s_\lambda;\psi,\lambda)}$$

and so, given the data, the conditional (on $S_\lambda(\boldsymbol{y}) \equiv s_\lambda$) log-likelihood for ψ

$$l_c(\psi) = \log f_c(\boldsymbol{y}|s_\lambda;\psi) = l(\psi,\lambda) - \log f_{s_\lambda}(s_\lambda;\psi,\lambda)\ ,$$

where $l(\psi,\lambda) = \log f_Y(\boldsymbol{y};\psi,\lambda)$ is the full-data likelihood. The conditional likelihood, $l_c(\psi)$ does not depend on λ, and the conditional MLE, $\widehat{\psi}^{(c)}$ is obtained by

maximizing $l_c(\psi)$. In general $\widehat{\psi}^{(c)}$ may be different from the MLE that would be obtained from the full likelihood.

The conditional log-likelihood l_c differs from the full-data likelihood by the term $\log f_{s_\lambda}(s_\lambda; \psi, \lambda)$. Conditional ML estimation works best when "f_{s_λ} contains little or no information concerning ψ in the absence of knowledge about λ." The precise meaning of this statement is difficult to express (e.g., see Sprott 1975, Jørgensen 1993, Pace and Salvan 1997)

As a special case, if f_{s_λ} is such that for any parameter vector (ψ_1, λ_1) and any ψ_2 there exists a corresponding a λ_2 such that

$$f_{s_\lambda}(s_\lambda; \psi_1, \lambda_1) = f_{s_\lambda}(s_\lambda; \psi_2, \lambda_2) \tag{9.1}$$

then $\widehat{\psi} = \widehat{\psi}^{(c)}$. However, this is not enough to guarantee that the information content (about ψ) of l_c is the same as that of l.

Example 9.1. Equivalence of log-linear and multinomial response models.

Multinomial data are often fitted using a generalized linear model in which the data are modelled as Poisson with fixed marginals (see STATS 330). Here we see that a multinomial model is obtained by conditioning on Poisson marginals.

Let Y_{ij} be independent Poisson's with mean μ_{ij} given by the log-linear model

$$\log \mu_{ij} = \phi_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}, \qquad i = 1, ..., n, \quad j = 1, ..., b$$

where the $\boldsymbol{x}_{ij}$ are p-dimensional vectors of explanatory variables, $\boldsymbol{\beta}$ is the p-dimensional parameter of interest, and $\phi_i,\ i = 1, ..., n$ are scalar incidental parameters.

If the block size, b, is fixed then increasing the sample size is achieved by increasing n. The parameter space has dimension $n + p$, increasing with n.

The log-likelihood is

$$l(\boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_{ij} (y_{ij} \log \mu_{ij} - \mu_{ij})$$

where $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)^T$. Since each $y_{i+} = \sum_j y_{ij}$ is Poisson with mean $\mu_{i+} = \sum_j \mu_{ij} = \sum_j \exp(\phi_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta})$ the log-likelihood for $\mathbf{y}_+ = (y_{1+}, ..., y_{n+})^T$ is

$$l_{\mathbf{y}_+}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_i (y_{i+} \log \mu_{i+} - \mu_{i+}) \ .$$

Conditioning on $\mathbf{y}_+$ gives

$$l_c(\boldsymbol{\beta}) = l(\boldsymbol{\beta}, \boldsymbol{\phi}) - l_{\mathbf{y}_+}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_i \left(\sum_j y_{ij} \boldsymbol{x}_{ij}^T \boldsymbol{\beta} - y_{i+} \log(\sum_j \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta})) \right)$$

which doesn't involve $\boldsymbol{\phi}$. Writing

$$l_c(\boldsymbol{\beta}) = \sum_i \left(\sum_j y_{ij} \log \left(\frac{\exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta})}{\sum_k \exp(\boldsymbol{x}_{ik}^T \boldsymbol{\beta})} \right) \right) \tag{9.2}$$

we see that the conditional likelihood is the same as that of a multinomial response model with, for each i, y_{i+} trials and cell probabilities

$$p_{ij} = \frac{\exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta})}{\sum_k \exp(\boldsymbol{x}_{ik}^T \boldsymbol{\beta})} \qquad j = 1, ..., b.$$

Note that the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ appear in $l_{\mathbf{y}_+}(\boldsymbol{\beta}, \boldsymbol{\phi})$ only through $\mu_{i+} = \sum_j \exp(\phi_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta})$. Thus, the condition stated in (9.1) is satisfied and the conditional MLE $\widehat{\boldsymbol{\beta}}^{(c)}$ is the same as the MLE of $\boldsymbol{\beta}$ maximizing the full likelihood.

In this example it can be shown that the Fisher information about $\boldsymbol{\beta}$ is the same from both the full and conditional models and that, in the absence of knowledge about the ϕ_i (i.e., $\phi_i \in \mathbb{R}$), the Poisson and multinomial models are equivalent for inference about $\boldsymbol{\beta}$.

In practice, if one has multinomial data then the log-likelihood in equation (9.2) needs to be maximized with respect to $\boldsymbol{\beta}$. The above algebra show that this is maximized by fitting a Poisson model conditional on treating the y_{i+} values as fixed. We can accomplish the conditioning by fitting a row effect (the minimal model) because this always fits $\widehat{\mu}_{i+} = y_{i+}$ regardless of $\boldsymbol{\beta}$. That is, as far as estimation of $\boldsymbol{\beta}$ is concerned, y_{i+} is fixed.

Example 9.2. Logistic model for binary responses. Consider the special case of the previous example with $b = 2$ and where

$$\log \mu_{i1} = \phi_i \qquad \text{and} \qquad \log \mu_{i2} = \phi_i + \boldsymbol{x}_i^T \boldsymbol{\beta}, \qquad i = 1, ..., n.$$

Conditioning on $\mathbf{y}_+$ results in a likelihood the same as that of a binomial response model with probabilities

$$p_{i2} \equiv p_i = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} .$$

Therefore, in performing logistic regression we don't need to worry about whether the $n_i \equiv y_{i+} = y_{i1} + y_{i2}$ are fixed in advance or not.

Example 9.3. Capture of fish in a trouser trawl. Assume that the numbers of fish "arriving" at the mouth of a trawl is Poisson. Let μ_{l1} be the rate of arrival of length l fish to a small mesh net that retains everything, i.e., hence μ_{l1} is also the rate of capture of length l fish. Let μ_{l2} be the rate of capture in a large mesh net with selection curve $r(l)$ = the prob that a length l fish will be retained by this net.

A trouser trawl is a modified trawl fishing a small mesh net and large mesh net simultaneously. Then it is often reasonable to assume that $\mu_{l2} = \mu_{l1} r(l)$, i.e., $\log \mu_{l2} = \log \mu_{l1} + \log(r(l))$. Here, the $\log \mu_{l1}$ are the nuisance parameters and the conditioning argument leads to analysis of a binomial response model where

$$p_{i2} \equiv p_i = \frac{r(l)}{1 + r(l)}$$

is the probability that a length l fish was captured in the large mesh net given that it was captured by the trouser trawl (Millar 1992). □

In the above examples, the equivalence of conditional and full likelihood depended on an assumed "lack of information" about the nuisance parameter. In the trouser trawl example the nuisance parameters are arrival rates of fish. If one had information on these arrival rates (e.g., from a length-based stock assessment) then there would be information on the nuisance parameters, and hence a loss of information from conditioning on them. For a look at such issues, see Scott and Wild (2001).

Example 9.4. Cox's partial likelihood for proportional-hazards model, Section 6.3.3.

Roughly speaking, we conditioned on the censoring events and failure times because they acted as sufficient statistics for the baseline hazard $\lambda(t)$, with the result that the likelihood no longer involved $\lambda(t)$. The rational for the conditioning was that, in the absence of knowledge about $\lambda(t)$, the censoring events and failure times contain no information about the parameters of interest $\boldsymbol{\beta}$.

In his original paper on this approach Cox (1972) referred to the likelihood obtained from the above conditioning as "conditional likelihood". Critics of Cox's paper noted that this was not strictly true as the sequential nature of the data required a product of different conditional likelihoods. Thus, Cox (1975) introduced the idea of "partial likelihood", which generalizes the idea of conditional likelihood to sequential models of this type.

Example 9.5. Residual Maximum Likelihood, REML

Residual maximum likelihood has been presented as an application of marginal likelihood (see next Section) in which estimation of variance components was based on residuals, or more generally, error contrasts (Harville 1977). More recently, REML has been given a more appealing justification as an application of conditional likelihood (Smyth and Verbyla 1996). Here we use the example of a linear Gaussian model.

Let $\mathbf{Y}$ be an n-dimensional multivariate Normal with $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\lambda}$, ($\boldsymbol{\lambda} \in \mathbb{R}^p$) and with variance matrix $\Sigma(\boldsymbol{\psi})$ parameterized by the unknown $\boldsymbol{\psi}$. If p increases with n then the MLE of $\boldsymbol{\psi}$ may not be consistent. This can be overcome by applying a conditional likelihood argument with $\boldsymbol{\lambda}$ as the "nuisance" parameters.

The data, $\mathbf{y}$, can be expressed as

$$\mathbf{a} = (\mathbf{I} - \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T})\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

and

$$\mathbf{u} = \mathbf{H}\mathbf{y} \ .$$

Although $\mathbf{a}$ and $\mathbf{u}$ are both vectors of length n, their "true" dimensions are $n - p$ and p, respectively, because they are degenerate multivariate normal.

$\boldsymbol{\lambda}$ does not appear in the conditional likelihood $f(\mathbf{a}|\mathbf{u})$. In fact, for this model, $\mathbf{a}$ is independent of $\mathbf{u}$ and hence $f(\mathbf{a}|\mathbf{u}) = f(\mathbf{a})$. Evaluation of $f(\mathbf{a})$ is complicated by the degeneracy of $\mathbf{a}$. In the case that $\psi \in \mathbb{R}$ and $\Sigma(\psi) = \psi\mathbf{I}$ it can be shown that the log likelihood for $\mathbf{a}$ reduces to

$$-\frac{1}{2}(n-p)\log(\psi) - \frac{\mathbf{a^T a}}{2\psi}$$

and so $\widehat{\psi}^{(m)} = \mathbf{a^T a}/(n-p)$. Note that this estimator is unbiased, whereas the full likelihood estimator has bias $-\frac{p}{n}\psi$, which will not vanish with increasing sample size if p increases in proportion with n.

9.2 Marginal Likelihood

Another way of dealing with unwanted incidental parameters is to work with functions of the data that do not depend on these parameters. Specifically, we assume that there is a non-singular transformation mapping $\boldsymbol{y} = (y_1, ..., y_n)^T$ to $(u_1, ..., u_{n-r}, a_1, ..., a_r)^T$ such that the density function factorizes in the form

$$f_Y(\boldsymbol{y}; \psi, \lambda) = f_a(a_1, ..., a_r; \psi) f_{u|a}(u_1, ..., u_{n-r}|a_1, ..., a_r; \psi, \lambda) \ .$$

Note that the distribution of $a_1, ..., a_r$ does not depend on λ. (That is, for any fixed value of ψ, $a_1, ..., a_r$ are ancillary for λ.) The marginal likelihood estimator, $\widehat{\psi}^{(m)}$, is the value of ψ maximizing $f_a(a_1, ..., a_r; \psi)$.

To minimize loss of information it is desirable that $f_{u|a}(u_1, ..., u_{n-r}|a_1, ..., a_r; \psi, \lambda)$ contains little or no information about ψ in the absence of knowledge about λ.

Example 9.6. In its simplest form, marginal likelihood may simply involve discarding some data. For example, suppose that $Y_i, i = 1, ...n_1$ are from a $N(\mu, \sigma^2)$ distribution and $Y_i, i = n_1 + 1, ..., n_1 + n_2$ are from a mixture distribution known to include $N(\mu, \sigma^2)$ as a component. To estimate μ and σ^2 it will be much easier to throw out the latter set of data and avoid the hassles of dealing with a mixture and all the associated parameters (e.g the parameters of the other component distributions and the mixing proportions). This could be a reasonable thing to do

unless n_2 is large or there is good separation between the components of the mixture distribution. □

In situations where they can be applied, conditional and marginal likelihood will work well if it can be argued that there is little loss of information. The next method, while generally not as satisfactory, can be used in all circumstances.

9.3 Profile likelihood

Let $\widehat{\lambda}(\psi)$ be the MLE of λ for fixed ψ. The partially maximized log-likelihood function

$$l^*(\psi; \boldsymbol{y}) \equiv l(\psi, \widehat{\lambda}(\psi); \boldsymbol{y}) = \max_{\lambda} \; l(\psi, \lambda; \boldsymbol{y})$$

is called the profile log-likelihood for ψ. Note that the $\widehat{\psi}$ maximizing l^* is the usual MLE.

Care has to be taken when using l^* to make inference about ψ, because it may not behave like a log-likelihood function. The chain rule gives

$$\frac{\partial l^*}{\partial \psi} \equiv \frac{\partial l(\psi, \widehat{\lambda}(\psi))}{\partial \psi} = \frac{\partial l(\psi, \widehat{\lambda})}{\partial \psi} + \frac{\partial l(\psi, \widehat{\lambda})}{\partial \widehat{\lambda}} \frac{\partial \widehat{\lambda}}{\partial \psi} = \frac{\partial l(\psi, \widehat{\lambda})}{\partial \psi} \tag{9.3}$$

because

$$\frac{\partial l(\psi, \widehat{\lambda})}{\partial \widehat{\lambda}} \equiv \frac{\partial l(\psi, \lambda)}{\partial \lambda} \Big/_{\lambda = \widehat{\lambda}(\psi)} = 0$$

since $\widehat{\lambda}(\psi)$ maximizes $l(\psi, \widehat{\lambda})$.

Expanding $l^* \equiv l(\psi, \widehat{\lambda})$ about a fixed λ_0 gives

$$\begin{aligned} l^* = l(\psi, \widehat{\lambda}) = \quad & l(\psi, \lambda_0) \quad + \frac{\partial l(\psi, \lambda_0)}{\partial \lambda}(\widehat{\lambda} - \lambda_0) \\ & + \quad \frac{1}{2}\frac{\partial^2 l(\psi, \lambda_0)}{\partial \lambda^2}(\widehat{\lambda} - \lambda_0)^2 + \frac{1}{6}\frac{\partial^3 l(\psi, \lambda_0)}{\partial \lambda^3}(\widehat{\lambda} - \lambda_0)^3 \ldots \end{aligned} \tag{9.4}$$

Substituting expansion (9.4) into (9.3) gives

$$\frac{\partial l^*}{\partial \psi} = \frac{\partial l(\psi, \lambda_0)}{\partial \psi} + \frac{\partial^2 l(\psi, \lambda_0)}{\partial \psi \partial \lambda}(\widehat{\lambda}(\psi) - \lambda_0) + \frac{1}{2}\frac{\partial^3 l(\psi, \lambda_0)}{\partial \psi \partial \lambda^2}(\widehat{\lambda}(\psi) - \lambda_0)^2 + \ldots$$

It is desirable that $E_{\psi_0}[\frac{\partial l^*(\psi_0)}{\partial \psi}]$ is not too different from zero. (Recall from the consistency of MLE's section that the asymptotic normality of MLE's was obtained

from a Taylor series expansion of $l^{'}(\widehat{\theta}_n)$ about $l^{'}(\theta_0)$. The argument required the results that $E_{\theta_0}[l^{'}(\theta_0)] = 0$, and $-E_{\theta_0}[l^{''}(\theta_0)] = I(\theta_0)$.) The term $\frac{\partial l(\psi,\lambda_0)}{\partial \psi}$ has mean 0 (when (ψ_0, λ_0) is true). It can be shown that the other two terms have means that are $O(1)$ when $\widehat{\lambda}(\psi)$ is a consistent estimator of λ_0. Since $l^{'}(\theta_0)$ is of "order" $\sqrt{n}$ in the standard case, the fact that $\frac{\partial l^*}{\partial \psi}$ may have a $O(1)$ bias should be inconsequential for large n. However, the expectations of the last two terms may be of a higher order if $\widehat{\lambda}(\psi)$ is not consistent.

Example 9.7. Finite mixture estimation with learning samples - approximate profile likelihood

Consider an m component mixture of $\mathrm{N}(\mu_k, \sigma_k^2)$ distributions, $k = 1, ..., m$. The mixing proportions $\mathbf{p} = (p_1, ..., p_{m-1})$ are the parameters of interest.

In addition to data from the mixture distribution, assume that learning samples, $\mathcal{L}_k$ are observed. The learning samples consist of data of known origin from each of the component distributions. Hence, the likelihood for this model is

$$L(p_1, ..., p_{m-1}, \mu_1, ..., \mu_m, \sigma_1^2, ..., \sigma_m^2) = \prod_{y \in mixture} f(y) \prod_{y \in \mathcal{L}_1} f_1(y), ..., \prod_{y \in \mathcal{L}_m} f_1(m) \ .$$

In practice, there may be little information about the μ_k and σ_k^2 parameters in the mixture sample. Hence, it may be reasonable to estimate these parameters solely from the learning sample and to use their MLE's in the mixture data likelihood.

Example 9.8. Growth curves. Let y_{ijr}, $r = 1, ..., n_{ij}$ be the observed lengths of age j individuals from cohort (year-class) i. (E.g, Cohort i may correspond to individuals born in year 1960+i, say. Then y_{i4r} denotes the observed length of a 4 year old individual, measured in year 1960+$i+4$). The 3-parameter von-Bertalanffy growth curve model is

$$y_{ijr} = L_{\max}(1 - a \exp(-kj)) + \epsilon_{ijr}$$

where ϵ_{ijr} are iid $N(0, \sigma^2)$ and $a, L_{\max}, k$ and σ^2 are to be estimated. Parameters $L_{\max}$ and k are of physiological relevance. Parameter a reflects only the birth date of the cohort relative to time zero.

In some populations, a may vary between cohorts, whence

$$y_{ijr} = L_{\max}(1 - a_i \exp(-kj)) + \epsilon_{ijr}$$

would be a better model. The a_i can be considered nuisance parameters. Profile likelihood is particularly convenient here because the model is linear in the a_i (given $L_{\max}$ and k) and so it is very easy to calculate $\widehat{a_i}(L_{\max}, k)$.

However, since information about a_i comes solely from the i^{th} cohort, the MLE's $\widehat{a}_i$ will not be consistent, and so the usual asymptotic theory for $L_{\max}$ and k would not apply. □

In the above example the problem is that there is one nuisance parameter per cohort, so the number of parameter increases with n as more data is added each year. Alternatively, it may be reasonable to postulate that each a_i is an unobserved random variable from some distribution. This leads into the next two approaches.....

9.4 Penalized likelihood

If the nuisance parameters, λ, are randomly distributed with density function $g(\lambda; \beta)$ then the joint density function for the data $\mathbf{y}$ and λ is

$$f(\mathbf{y}, \lambda; \psi, \beta) = f(\mathbf{y}; \psi, \lambda) g(\lambda; \beta) \ .$$

Penalized likelihood maximizes this with respect to ψ and λ and β, and in that sense one simple interprets $g(\lambda; \beta)$ as a penalty term added to the likelihood $L(\psi, \lambda) = f(\mathbf{y}; \psi, \lambda)$. Parameter β is called a hyper/super parameter.

Note that penalized likelihood is explicitly calculating estimates for all nuisance parameters, but the penalty term acts to "smooth" out the estimates of the nuisance parameters. However, in many applications, penalized likelihood is known to produce inconsistent estimators of ψ.

The degree of "smoothing" of the $\widehat{\lambda}$ estimates is controlled by the value of β. In some applications it is not possible to estimate β, and indeed, in some cases the penalized likelihood is an unbounded function of its parameters. Hence, β is often taken to be known.

If the data, $\mathbf{y}$ and λ are normally distributed then this technique is sometimes called the method of *total least squares* because the log of the above density function includes the sum of two sums-of-squares (or quadratic forms).

Example 9.4.8 continued. We could postulate that the a_i are iid $N(\nu, \tau^2)$ random variables. (Note, this may be suspect since the a_i will largely be influenced by environmental conditions which may have long term trends.) Then, the log penalized-likelihood is

$$\begin{aligned} l_{pen}(\psi, \mathbf{a}, \beta) = & - \frac{\sum_{ijr} \left(y_{ijr} - L_{\max}(1 - a_i \exp(-kj))\right)^2}{2\sigma^2} - \sum_{ijr} \log(\sigma) \\ & - \frac{\sum_i (a_i - \nu)^2}{2\tau^2} - \sum_i \log(\tau) \ , \end{aligned}$$

where $\psi = (L_{\max}, k)$, $\mathbf{a} = a_1, a_2, ...$ and $\beta = (\nu, \tau)$. Note that this function is unbounded. (Why?)

9.5 Integrated likelihood

Examples

Could mention Bin(N,p) with unknown N, as used by Berger, Liseo and Wolpert (1999), and the genetic differentiation example of **?** *(unfortunately, this is too complex to give any detail).*

Present ratio of regression coefficients as a worked example (Ghosh, Datta, Kim and Sweeting 2006).

Harville (1974) shows that REML is equivalent to integrated likelihood. This is implemented in ADMB in Chap 11 (this example is tractable).

Comment that REML more challenging outside of linear normal mixed models because of the difficulty to evaluate the density function of the error contrasts. REML via IL provides an extension, and is applied to multi-center trials in Chap 11.

Could note that integrated likelihood is an example of a joint Bayesian-frequentist approach, and note that Yuan (2009) provides a more comprehensive framework for this.

The approach of Noh and Lee (2007) looks very closely related to integrated likelihood, because it uses Laplace approximation (of 1st or 2nd order) to eliminate b and beta for purposes of estimating variance components. Their lead example is the salamander data that I worked with in 2004. Could be worth implementing in ADMB.

9.6 Mixed-effects models (aka. Mixture models, Empirical Bayes models)

In a typical mixed-effects model the nuisance parameters are assumed to be random effects (i.e., unobserved random variables). If the distribution of the random effects is assumed to be $N(\nu, \tau^2)$, say, then this formulation adds two parameters (ν, τ) to the model. (In contrast, if the nuisance parameters were treated as fixed effects then the number of parameters could increase dramatically, and increase with sample size.)

The likelihood function is given by the marginal density of the observed data. For a mixed-effects models this can be obtained by integrating the random effects out of the joint density function for data and random effects. If the data $y_i, i = 1, ..., n$ (possibly vectors) are iid then their density is given by

$$f(y_i; \psi, \beta) = \int f(y_i; \psi, \lambda) g(\lambda; \beta) d\lambda \ , \tag{9.5}$$

where λ denotes the random effect with distribution $g(\lambda; \beta)$. More generally, this is an example of a mixture distribution because the density function is a mixture (over λ) of $f(y_i; \psi, \lambda)$ densities according to the mixing weights given by $g(\lambda; \beta)$.

If the random effects distribution $g(\lambda; \beta)$ is normal and the model is linear with normal observation errors then the marginal distribution of Y is also normal and can be stated explicitly as a function of ψ and β. More generally, 9.5 will be intractable

and obtained the MLE (ψ, β) can be challenging. There is currently a lot of research effort and software development devoted to this issue.

Example 9.9. Random effects (Component of variance) models. The one-way ANOVA random effects model is given by

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where μ_i are iid $N(\mu, \tau^2)$ and ϵ_{ij} are iid $N(0, \sigma^2)$ (independent of μ_i). This model treats the μ_i as random variables rather than as fixed effects (parameters to be estimated). Thus, $E[Y_{ij}] = \mu$ and

$$Cov(Y_{ij}, Y_{bc}) = \begin{cases} 0 & \text{if } i \neq b \\ \tau^2 & \text{if } i = b \text{ and } ij \neq bc \\ Var[Y_{ij}] = \tau^2 + \sigma^2 & \text{if } ij = bc \end{cases}$$

This model is "routine" to fit using maximum likelihood or residual maximum likelihood. For example, the R language has function `lme` and SAS has `PROC MIXED`.

Example 9.6.9 continued. With the assumption that the a_i are iid $N(\nu, \tau^2)$ random variables, then

$$y_{ijr} = L_{\max}(1 - a_i \exp(-kj)) + \epsilon_{ijr}$$

are normally distributed with $E[Y_{ijr}] = L_{\max}(1 - \nu \exp(-kj))$ and variance $Var[Y_{ijr}] = (L_{\max} \exp(-kj))^2 \tau^2 + \sigma^2$. Also,

$$Cov(Y_{ijr}, Y_{bcd}) = \begin{cases} 0 & \text{if } i \neq b \\ L_{\max}^2 \exp(-kj) \exp(-kc) & \text{if } i = b \text{ and } ijr \neq bcd \\ Var[Y_{ijr}] & \text{if } ijr = bcd \end{cases}$$

Under this model the number of parameters (ψ and β) remains fixed as n increases and therefore the asymptotic MLE results will generally hold. □

When the random effects occur non-linearly in the model, or the data are non-normal then things become more challenging. Both R and SAS have functions that maximize pseudo-likelihoods obtained from linear approximations of the model. R

has function `nlme` for nonlinear mixed-effects models. SAS has macros NLINMIX and GLIMMIX (Littell, Milliken, Stroup and Wolfinger 1996) for nonlinear mixed-effects and generalized linear mixed-effects models, respectively. These pseudo-likelihood fits are known to possess undesirable properties in some situations (e.g., see Breslow and Lin 1995, Millar and Willis 1999) and must be used with care.

The recent version of SAS include a new procedure, `PROC NLMIXED` which uses numerical integration to evaluate (9.5). The next example demonstrates its application to a mixed-effects model of binomial data.

Example 9.10. Model of treatment success

The data (Beitler and Landis 1985) are from testing the effectiveness of two cream treatments (drug or control) for infection. The trials were performed at eight clinics, and the binomial outcome is the number of favourable outcomes.

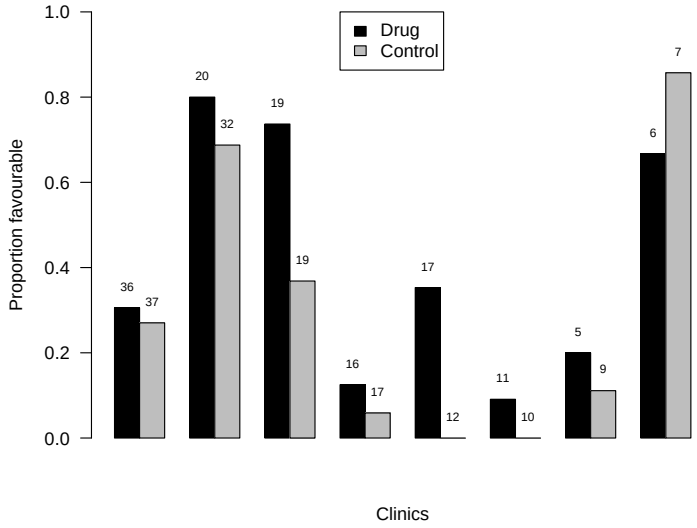
SAS Institute Inc (1999) used PROC NLMIXED to obtain the maximum likelihood fit (Table 2) of the logistic model with random clinic effect. Specifically, letting

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}$$

be the probability of a favourable outcome from treatment j in clinic i, this model is

$$\eta_{ij} = a + bt_j + u_i \ , \tag{9.6}$$

where, t_j is an indicator variable for the drug treatment and u_i are iid $N(0, \sigma_u^2)$.




```
DATA infection;
INPUT clinic t x n;
LINES;
1 1 11 36
1 0 10 37
2 1 16 20
2 0 22 32
3 1 14 19
3 0 7 19
4 1 2 16
4 0 1 17
5 1 6 17
5 0 0 12
6 1 1 11
6 0 0 10
7 1 1 5
7 0 1 9
8 1 4 6
8 0 6 7
RUN;

ODS PS FILE="NLMIXED.ps";

PROC NLMIXED DATA=infection;
PARMS beta0=-1 beta1=1 s2u=2;
eta = beta0 + beta1*t + u;
expeta = exp(eta);
p = expeta/(1+expeta);
MODEL x ~ BINOMIAL(n,p);
RANDOM u ~ NORMAL(0,s2u) SUBJECT=clinic;
RUN;

ODS PS CLOSE;
```

The SAS System 15:03 Monday, October 6, 2003 **1**

The NLMIXED Procedure

Specifications	
Data Set	WORK.INFECTION
Dependent Variable	x
Distribution for Dependent Variable	Binomial
Random Effects	u
Distribution for Random Effects	Normal
Subject Variable	clinic
Optimization Technique	Dual Quasi−Newton
Integration Method	Adaptive Gaussian Quadrature

Dimensions	
Observations Used	16
Observations Not Used	0
Total Observations	16
Subjects	8
Max Obs Per Subject	2
Parameters	3
Quadrature Points	5

Parameters			
beta0	**beta1**	**s2u**	**NegLogLike**
−1	1	2	37.5945925

Iteration History						
Iter		**Calls**	**NegLogLike**	**Diff**	**MaxGrad**	**Slope**
1		2	37.3622692	0.232323	2.882077	−19.3762
2		3	37.1460375	0.216232	0.921926	−0.82852
3		5	37.0300936	0.115944	0.315897	−0.59175
4		6	37.0223017	0.007792	0.01906	−0.01615
5		7	37.0222472	0.000054	0.001743	−0.00011
6		9	37.0222466	6.57E−7	0.000091	−1.28E−6
7		11	37.0222466	5.38E−10	2.078E−6	−1.1E−9

NOTE: GCONV convergence criterion satisfied.

The SAS System 15:03 Monday, October 6, 2003 **2**

The NLMIXED Procedure

Fit Statistics	
−2 Log Likelihood	74.0
AIC (smaller is better)	80.0
AICC (smaller is better)	82.0
BIC (smaller is better)	80.3

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
beta0	−1.1974	0.5561	7	−2.15	0.0683	0.05	−2.5123	0.1175	−3.1E−7
beta1	0.7385	0.3004	7	2.46	0.0436	0.05	0.02806	1.4488	−2.08E−6
s2u	1.9591	1.1903	7	1.65	0.1438	0.05	−0.8554	4.7736	−2.48E−7

Example 9.11. Nonparametric mixtures. No parametric form of the mixing distribution G is assumed,

$$f(y_i; \psi, G) = \int f(y_i; \psi, \lambda) dG(\lambda).$$

It can be shown that, in general, ψ can still be estimated consistently (with the usual $\sqrt{n}$ rate of convergence), and that estimates $\widehat{G}$ of G exist such that $\widehat{G} \rightarrow_D G$ in distribution.

Chapter 10

Latent variable models

10.1 Introduction

Latent variable models provide a rich and widely used collection of models that are especially suited for inference from experiments where it is not relevant or feasible to collect a sample of independently distributed observations. In many cases this will be because the data are grouped. For example, while it may be desired to test a drug treatment on 1000 patients, these patients might be selected from a much smaller number of hospitals. If there are any differences in the efficacy of the drug between the hospitals (perhaps due to staff skills, socio-economic status of patients, selection criterion for participation in the study, etc) then the 1000 patients can not be considered an iid sample. Similarly, biological experiments face constraints over the availability of equipment, so an experiment measuring the growth of a large number of fish may only be able to utilize a handful of tanks. While all reasonable attempts can be made to keep the tanks equal, this can rarely be achieved – one dead fish in a tank could affect the water quality for all other fish in that tank. In ecological field experiments, a large number of samples may necessarily have to be taken from a relatively small number of locations. These locations can be chosen to be similar, but they can never be identical.

Failure to take into account any grouping that is inherent in the data can lead to erroneous inference. Indeed, Hurlbert (1984) caused quite a clamour by revealing that about one half of all quantitative results from manipulative ecological field experiments (published between 1960 and 1981) contained questionable statistical

inference due to failure to accommodate grouping structure in the data. Hurlbert (1984) referred to this as pseudo-replication.

More generally, the class of latent variable models subsumes many other types of model, including classes of models that are variously called mixed-effects models, hierarchical models, state-space models and volatility models. Latent variable models also include semi-parametric regression models, by virtue of the mixed-effects model formulation of penalized smoothing splines (Wand 2003). They also includes mixture models, and the next section uses the binormal mixture model introduced in Chapter 2 to develop the notation for likelihood-based inference from latent variable models.

10.2 Developing the likelihood

Throughout this text, the notation $f(\boldsymbol{y};\boldsymbol{\theta})$ represents the density function of $\boldsymbol{Y}$ under repetition of the experiment from which the observed data were generated. Regarded as a function of $\boldsymbol{\theta}$, $f(\boldsymbol{y};\boldsymbol{\theta})$ is the all-important likelihood. In latent variable models the likelihood can not be expressed explicitly (except in special cases), due to the presence of unobserved additional sources of randomness that are relevant to inference about $\boldsymbol{\theta}$.

The binormal model for the waiting times of the Old Faithful geyser (Example 2.9) was presented in the form of a latent variable model. In that example, the binormal distribution was expressed as a mixture of two normal distributions where, conditional on an unobserved B_i from a Bernoulli(p) experiment, Y_i was observed from a $N(\mu, \sigma^2)$ distribution when $B_i = 1$, else from a $N(\nu, \tau^2)$ distribution. Here, the B_i are latent random variables, in the sense that they remain hidden, yet are relevant to determination of the likelihood function. In this particular example, it was straightforward to obtain the likelihood for parameters $\boldsymbol{\theta} = (p, \mu, \sigma, \nu, \tau)$ because

the latent variables are dichotomous. Specifically, for each observed y_i

$$\begin{aligned} f(y_i;\boldsymbol{\theta}) &= \sum_{b_i\in\{0,1\}} f(y_i,b_i;\boldsymbol{\theta}) && (10.1)\\ &= \sum_{b_i\in\{0,1\}} f(y_i|b_i;\boldsymbol{\theta})P(B_i=b_i;\boldsymbol{\theta}) && (10.2)\\ &= p\phi(y_i;\mu,\sigma^2)+(1-p)\phi(y_i;\nu,\tau^2) \end{aligned}$$

where $\phi(y;\mu,\sigma^2)$ denotes the $N(\mu,\sigma^2)$ density function evaluated at y.

The latent variable models considered in this chapter have a likelihood of the same form as (10.2), except that the marginalization over the latent variable involves an integral rather than a summation. In the general case, let $\boldsymbol{U}$ generically denote all unobserved random variables in the model, with density denoted $f(\boldsymbol{u};\boldsymbol{\theta})$. Letting $f(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta})$ denote the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{U}$, the joint density of $\boldsymbol{Y}$ and $\boldsymbol{U}$ is

$$f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta})f(\boldsymbol{u};\boldsymbol{\theta})\ . \tag{10.3}$$

Marginalization over $\boldsymbol{u}$ gives the likelihood function for $\boldsymbol{\theta}$ arising from observation of $\boldsymbol{y}$,

$$L(\boldsymbol{\theta}) = f(\boldsymbol{y};\boldsymbol{\theta}) = \int f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\theta})d\boldsymbol{u}\ . \tag{10.4}$$

In general, the integral in (10.4) can not be expressed in closed form.

When the conditional distribution $\boldsymbol{Y}|\boldsymbol{U}$ is a linear-normal model, then the latent variable model is termed a linear mixed-model (LMM). Similarly, the acronyms GLMM and NLMM are used for generalized linear mixed-model and nonlinear mixed-model, respectively. In the case of state-space models for time-series data, it is often the case that $f(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta})$ has the form of an autoregressive model.

This Chapter focuses on demonstration of the current capabilities of R, SAS and ADMB for maximum likelihood inference from a likelihood of the form in equation (10.4). In the examples that follow (Sections 10.4–10.7), the latent variables $\boldsymbol{U}$ are assumed to be normally distributed. The use of alternative distributions for the latent variables can be accommodated by ADMB because this software requires the user to code the equation for $f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\theta})$ within the template file, and places no explicit restrictions on its form.

10.3 Software

10.3.1 Background

The likelihood in (10.4) does possess a closed form in some popular applications. These include linear-normal mixed models, and in the time-series context, the Kalman filter provides an iterative algorithm for determining the closed form of the likelihood for linear-normal state-space models (Meinhold and Singpurwalla 1983, Wei 2006). Closed form calculation of likelihood permits standard methods of optimization to be applied. For example `PROC MIXED` uses a Newton-Raphson type algorithm to fit linear-normal mixed models.

The Expectation-Maximization (EM) algorithm was applied to the binormal mixture model in Section 5.3, with the latent variables taking the role of "missing data" in the construction of the algorithm. The EM algorithm has also been used in linear-normal mixed models (e.g., Laird and Ware 1982), but is not convenient to implement in more general latent variable models because the E and M steps are no longer available in closed form.

Likelihoods for GLMM and NLMM models can not generally be obtained in closed form. Historically, likelihood-based methods for fitting these models have used an assortment of modified likelihoods obtained through various linear approximations to the model (e.g., Stiratelli, Laird and Ware 1984, Lindstrom and Bates 1990, Schall 1991, Breslow and Clayton 1993, Wolfinger and O'Connell 1993, Littell et al. 1996). A common feature to these methods is that the fitting algorithm contains a step in which a linearized approximation to the model is fitted using linear-mixed models. For example, the pseudo-likelihood approach of Wolfinger and O'Connell (1993) fits a GLMM using alternating fits of a generalized linear model (using `PROC GENMOD`) and a linear mixed model (using `PROC MIXED`). This algorithm was implemented in SAS macro GLIMMIX, and a version for nonlinear Gaussian mixed models was implemented in SAS macro NLINMIX (Littell et al. 1996). Within R, the leading package for mixed-effects modeling is `lme4`, and earlier versions of this package used the penalized-quasi likelihood method of Breslow and Clayton (1993).

While these modified likelihoods had their uses, they did not enjoy the full ad-

vantages of working with true likelihood. For example, they have been shown to give inconsistent estimates in some situations and can be of dubious use for the purposes of model comparison and selection. McCulloch and Searle (2001) called for an end to the use of modified likelihoods, and their use has indeed waned. The NLINMIX and GLIMMIX macros of Littell et al. (1996) have now been replaced by the NLMIXED and GLIMMIX procedures (in SAS Version 9.2). Both of these procedures perform numerical evaluation of the likelihood in (10.4), although pseudo-likelihood remains an option in GLIMMIX. The NLMIXED procedure and the R package `lme4` now use only methods based on numerical evaluation of (10.4).

The next section presents the Laplace approximation for numerical evaluation of the integral in (10.4). It has been established that the Laplace approximation works extremely well in a wide variety of latent-variable models (e.g., see Skaug and Fournier 2006, Rue, Martino and Chopin 2009). Moreover, the accuracy of the approximation can be improved by use of importance sampling (Section 10.3.3). In addition, for low-dimensional integrals, Gaussian quadrature can be used to provide a higher degree of approximation.

Box 10.1.

It has been shown that penalized-quasi likelihood (Breslow and Clayton 1993) and pseudo-likelihood (Wolfinger and O'Connell 1993) have much in common with the Laplace approximation (Wolfinger 1993, Vonesh 1996, McCulloch and Searle 2001). However, the weakness in these modified-likelihood methods is that they use additional approximations to avoid the optimization and second derivative calculations required by the Laplace approximation (Section 10.3.2).

10.3.2 The Laplace approximation and Gaussian quadrature

For ease of presentation, the Laplace approximation will be motivated for the scalar case $u \in \mathbb{R}$, and it will be assumed that the domain of integration is the entire real line.

For any fixed $\boldsymbol{y}$ and $\boldsymbol{\theta}$, and with the notation $h(u) = f(\boldsymbol{y}, u; \boldsymbol{\theta})$, the integral in

(10.4) can be written

$$L(\boldsymbol{\theta}) = \int_{\mathbb{R}} h(u)du = \int_{\mathbb{R}} e^{\log(h(u))}du \ . \tag{10.5}$$

Let u^* be the value that maximizes $h(u)$, and hence also $\log(h(u))$. The first derivative of $\log(h(u))$ is zero at u^* and so a second-order Taylor's series approximation of $\log(h(u))$ around u^* yields

$$\log(h(u)) \approx \log(h(u^*)) - \frac{c(u-u^*)^2}{2} \ ,$$

where $c > 0$ is given by

$$c = - \left. \frac{\partial^2 \log(h(u))}{\partial u^2} \right|_{u=u^*} \ .$$

Substituting this expansion into (10.5) gives

$$L(\boldsymbol{\theta}) \approx h(u^*) \int_{\mathbb{R}} e^{-\frac{c(u-u^*)^2}{2}} du \ . \tag{10.6}$$

The integral in (10.6) has the form of the normalizing identity for the normal distribution

$$\int_{\mathbb{R}} e^{-\frac{(u-u^*)^2}{2\sigma^2}} = \sqrt{2\pi}\sigma \ ,$$

with $\sigma^2 = c^{-1}$. This leads immediately to the Laplace approximation

$$L(\boldsymbol{\theta}) \approx h(u^*)\sqrt{\frac{2\pi}{c}} \ . \tag{10.7}$$

In the multi-dimensional case $\boldsymbol{u} \in \mathbb{R}^q$ the above derivation extends in a natural way, and the Laplace approximation is

$$\int_{\mathbb{R}^q} h(\boldsymbol{u})d\boldsymbol{u} \approx (2\pi)^{\frac{q}{2}} h(\boldsymbol{u}^*) \det\left(-H(\boldsymbol{u}^*)\right)^{-\frac{1}{2}} \tag{10.8}$$

where $\det(-H(\boldsymbol{u}^*))$ is the determinant of the negative of the q by q Hessian matrix of second derivatives of $h(\boldsymbol{u})$, evaluated at $\boldsymbol{u}^*$. That is,

$$H(\boldsymbol{u}^*) = \left. \frac{\partial^2 \log(h(\boldsymbol{u}))}{\partial \boldsymbol{u}^2} \right|_{\boldsymbol{u}=\boldsymbol{u}^*} \ .$$

The Laplace approximation is exact when $\log(h(\boldsymbol{u}))$ is of quadratic form. Gaussian quadrature approximation provides an extension to the Laplace approximation that is exact for a wider class of integrands. In the one-dimensional case, Gaussian

quadrature is able to evaluate the integral exactly if $\log(h(u)/p(u))$ is quadratic, where $p(u)$ is a polynomial in u (McCulloch and Searle 2001). However, Gaussian quadrature is only computationally efficient for integrals of low dimension. For this reason, in the examples that follow, Gaussian quadrature was only available in SAS, R and ADMB for the models in which the likelihood in equation (10.4) could be expressed as a product of one-dimensional integrals (see Section 10.3.4).

10.3.3 Importance sampling

The Laplace approximation assumes that, to within a multiplicative constant, the integrand $h(\boldsymbol{u})$ is well approximated by the density function of a q-dimensional multivariate normal centered at the value $\boldsymbol{u}^*$. Importance sampling provides a technique for polishing this approximation.

Importance sampling is derived from the simple identity

$$\begin{aligned} \int_{\mathbf{R}^q} h(\boldsymbol{u})d\boldsymbol{u} &= \int_{\mathbf{R}^q} \frac{h(\boldsymbol{u})}{g(\boldsymbol{u})} g(\boldsymbol{u})d\boldsymbol{u} \\ &= \int_{\mathbf{R}^q} r(\boldsymbol{u})g(\boldsymbol{u})d\boldsymbol{u} \end{aligned} \tag{10.9}$$

where $r(\boldsymbol{u}) = h(\boldsymbol{u})/g(\boldsymbol{u})$, and it is assumed that the support of $g(\boldsymbol{u})$ contains the support of $h(\boldsymbol{u})$. If $g(\boldsymbol{u})$ is the density function of a q-dimensional random variable, $\boldsymbol{U}$, then (10.9) is the expected value of $r(\boldsymbol{U})$, $E[r(\boldsymbol{U})]$. Thus, provided that $r(\boldsymbol{U})$ has finite variance, the sample mean of m iid realizations of $r(\boldsymbol{U})$, with $\boldsymbol{U}$ generated according to $g(\boldsymbol{U})$, is a consistent estimator of (10.9). This estimator is simply

$$\overline{r} = \frac{1}{m}\sum_{i}^{m} r(\boldsymbol{U}_{(i)}) \ ,$$

where $\boldsymbol{U}_{(i)}$ denotes the ith realization of $\boldsymbol{U}$.

Importance sampling works best when $g(\boldsymbol{u})$ is close to proportional to $h(\boldsymbol{u})$, because then $r(\boldsymbol{U})$ will have little variability and hence $\text{var}(\overline{r})$ will be small for moderate m. The q-dimensional multivariate normal used in the Laplace approximation makes a suitable initial choice for $g(\boldsymbol{u})$. However, this approximation does sometimes result in $r(\boldsymbol{U})$ having a very large variance. This can occur if the tails of $h(\boldsymbol{u})$ are wider than those of $g(\boldsymbol{u})$, thereby permitting the occasional occurrence

of highly extreme values of $r(\boldsymbol{u})$. This problem can usually be cured by replacing the multivariate normal approximating density with a density having fatter tails. A multivariate t-density, or a mixture of multivariate normals with covariance matrices of differing scale, are suitable for this purpose.

10.3.4 Separability

In many classes of latent variable models, the computational burden of the q-dimensional marginalization in (10.4) can be reduced by obtaining the integral as a product of lower dimensional integrals. For example, suppose that the data $\boldsymbol{y}$ can be partitioned into q subsets $\boldsymbol{y}_{(1)}, ...\boldsymbol{y}_{(q)}$ where $\boldsymbol{y}_{(i)}|\boldsymbol{u}, i = 1, ..., q$ are mutually independent. Furthermore, if $\boldsymbol{y}_{(i)}|\boldsymbol{u}$ does not depend on $u_j, j \neq i$ and $U_i, i = 1, ..., q$ are independent, then the joint density $f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})$ can be expressed

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}) &= f(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\theta}) f(\boldsymbol{u}; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{q} f(y_{(i)}|\boldsymbol{u}; \boldsymbol{\theta}) \prod_{i=1}^{q} f(u_i; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{q} \left(f(y_{(i)}|u_i; \boldsymbol{\theta}) f(u_i; \boldsymbol{\theta}) \right) \ . \end{aligned}$$

This joint density has the form of a separable integrand, in the sense that its integral with respect to $\boldsymbol{u}$ can be separated into the product of q one-dimensional integrals. That is,

$$L(\boldsymbol{\theta}) = \int_{\mathbb{R}^q} f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}) d\boldsymbol{u} = \prod_{i}^{q} \int_{\mathbb{R}} f(\boldsymbol{y}_{(i)}, u_i; \boldsymbol{\theta}) \ . \tag{10.10}$$

Separating the integral into a product of low-dimensional integrals typically permits the efficient use of Gaussian quadrature approximation. The NLMIXED procedure takes advantage of separability and Gaussian quadrature is the default method of integral approximation. However, this procedure is limited by only permitting one level of grouping structure. Similarly, PROC GLIMMIX, the `lme4 package` and ADMB[1] all provide quadrature approximation for separable models, but otherwise revert to the Laplace approximation.

[1] ADMB requires the user to explicitly code the joint density in separable form – see ADMB-project (2008b)

Integrals of low dimension can accurately be evaluated using numerical integration functions provided within R. The `integrate` function evaluates integrals of one dimension, and the `adapt` package provides the `adapt` function for integrals of low dimension. Thus, for latent variable models with joint density functions that can be expressed in separable form, it will often be possible to write an R function that returns the evaluated value of $L(\boldsymbol{\theta})$. This can then be maximized using `optim`.

10.3.5 Overview of examples

Sections 10.4–10.7 present examples of linear mixed models, nonlinear mixed models, generalized linear mixed models, and Poisson state-space models, respectively. The LMM example uses data that were measured according to a multilevel hierarchical design. However, for ease of presentation. the model fitted in Section 10.4 uses only the top level of grouping. The addition of a second level of grouping is left as an exercise (see Exercises 10.1 and 10.2).

The second example finds fault in the fit of a NLMM to orange tree circumference data. The faulty model includes only a single grouping variable. Instead, a crossed-effects NLMM is postulated and fitted using ADMB. In the GLMM example, dichotomous treatment outcomes from a medical study are modeled using a binomial model with nested random effects, and it was found that the nested random effect was not required. The final example uses ADMB to fit a Poisson state-space model to a times series of disease counts.

10.4 One-way linear mixed-effects model

The data presented in Table 10.1 are from an experiment to study the reproducibility of measurements of estrone. Sixteen vials of serum were taken from each of five subjects. There were additional levels of grouping in the design of this experiment (Gail, Fears, Hoover, Chandler, Donaldson, Hyer, Pee, Ricker, Siiteri, Stanczyk, Vaught and Ziegler 1996), but only subject will be considered for the purposes of this example.

	Subject				
Vial	1	2	3	4	5
1	23	25	38	14	46
2	23	33	38	16	36
3	22	27	41	15	30
4	20	27	38	19	29
5	25	30	38	20	36
6	22	28	32	22	31
7	27	24	38	16	30
8	25	22	42	19	32
9	22	26	35	17	32
10	22	30	40	18	31
11	23	30	41	20	30
12	23	29	37	18	32
13	27	29	28	12	25
14	19	37	36	17	29
15	23	24	30	15	31
16	18	28	37	13	32

Table 10.1: Sixteen estrone measurements (pg/mL), from each of five postmenopausal women, taken from (Fears et al. 1996).

Gail et al. (1996) assumed that the data were normally distributed on the log scale. They used $\log_{10}$ (log base 10) rather than natural log, and the same is done here. The five subjects are considered to be randomly chosen from a relevant population of postmenopausal women, and so the subject effect is a random variable. The model can be written

$$y_{ij}|u_i \sim N(a+u_i, \sigma^2) \quad , i = 1, ..., 5, \ j = 1, ..., 16 \tag{10.11}$$

$$u_i \sim N(0, \sigma_u^2) \quad , i = 1, ..., 5 \tag{10.12}$$

where y_{ij} is the $\log_{10}$-transformed estrone measurement from replicate j on subject i.

There are only five subjects with which to estimate between-subject variability, and so it makes sense to use restricted maximum likelihood (REML) to take into account the loss of a degree of freedom from estimation of the intercept parameter a. The above model is a one-way random effects ANOVA, and its likelihood $L(a, \sigma^2, \sigma_u^2)$ and restricted likelihood can be written in closed form because the data $\boldsymbol{y}$ possess a

multivariate normal distribution (e.g., see Pawitan 2001).

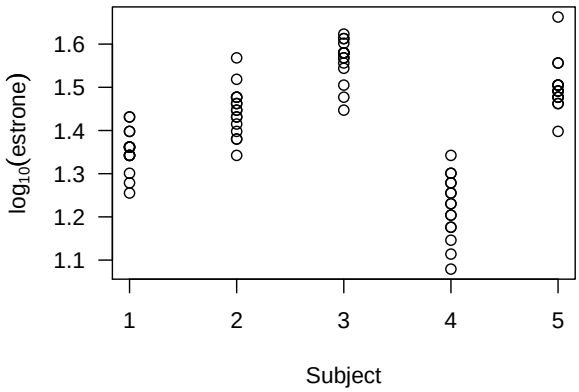


Figure 10.1: $\log_{10}$-estrone measurements

A visual inspection of the data (Fig. 10.1) shows that there is without doubt considerable between-subject variability. Nonetheless, for completeness, the null hypothesis $H_0 : \sigma_u^2 = 0$ is considered here. Under H_0 the LRT-statistic has an asymptotic distribution that is an equal mixture of a zero and a χ_1^2 random variable (Self and Liang 1987). However, Pinheiro and Bates (2000) have shown that this provides a crude approximation to the true sampling distribution of the LRT statistic when the number of groups is small. The estrone experiment has just five subjects. A better alternative would be to explore the sampling distribution of the LRT statistic bootstrapping. A second alternative is available to R users. The R-package `RLRsim` provides a function that numerically evaluates the true finite-sampling distribution of the LRT-statistic for testing hypotheses of the form $H_0 : \sigma_u^2 = 0$ in linear mixed models.

To create a bit of variety in the application of SAS, R and ADMB, each of these software was used to provide a different piece of the inferential picture. Specifically, PROC MIXED was used to produce a profile (restricted) likelihood contour plot for the variance parameters σ^2 and σ_u^2. R function `exactLRT` was used to test $H_0 : \sigma_u^2 = 0$, and the ADMB implementation demonstrates the iterative use of ADMB from within R.

10.4.1 SAS

The SAS code below assumes that dataset `estrone` contains two variables, `y` ($\log_{10}$(estrone)) and `person` (a factor variable for subject). The `MODEL` and `RANDOM` statements in PROC MIXED are used to specify the fixed and random components, respectively, of the LMM. Here, the only fixed term in the model is the intercept parameter a, and this is included by default. The `RANDOM INT / SUBJECT=person` statement is used to specify that `person` is the grouping variable, and that u_i (the random effect of `person`) is to be added to the intercept, as specified in equation (10.12).

The many features of PROC MIXED include the `PARMS` statement to calculate the (restricted) log-likelihood over a grid of values of σ_u^2 and σ^2. This facilitated the production of a contour plot of the profiled restricted log-likelihood (Fig. 10.2),

$$l_R^*(\sigma^2, \sigma_u^2) = \max_a l_R(a, \sigma^2, \sigma_u^2) \ .$$

PROC MIXED code for analysis of $\log_{10}$ estrone data

```
ODS OUTPUT ParmSearch=parms;
ODS PS FILE="EstroneREML.ps" STYLE=ALM;
ODS SELECT IterHistory CovParms SolutionF;
TITLE "REML fit to log10(estrone) data using PROC MIXED";
*Use NOPROFILE option below, to prevent partial profiling of likelihood;
PROC MIXED DATA=estrone NOPROFILE;
  MODEL y= / SOLUTION;
  RANDOM INT / SUBJECT=person;
  PARMS (0.0 TO 0.2 BY 0.0001) (0.0015 TO 0.0060 BY 0.00005);
RUN;
ODS PS CLOSE;

AXIS2 LABEL=(ANGLE=90 "Residual variance");
AXIS1 LABEL=("Between subject variance");
TITLE "REML log-likelihood contour plot";
PROC GCONTOUR DATA=parms;
  PLOT covP2*covP1=resloglike / LEVELS=(100.032 101.107 102 103)  NOLEGEND
    HREF=(0.005 TO 0.2 BY 0.005) LHREF=34 AUTOLABEL HAXIS=AXIS1 VAXIS=AXIS2;
  RUN;
QUIT;
```

The REML estimate is $(\widehat{a}, \widehat{\sigma}_u^2, \widehat{\sigma}^2) = (1.42, 0.0175, 0.00325)$. The approximate variance of $\widehat{\sigma}_u^2$ can be obtained by using the `ASYCOV` procedure option, however it is preferable that inference regarding σ_u^2 be made using likelihood ratio methods because the profile likelihood for $\widehat{\sigma}_u^2$ can be highly asymmetric (see Fig 10.3).

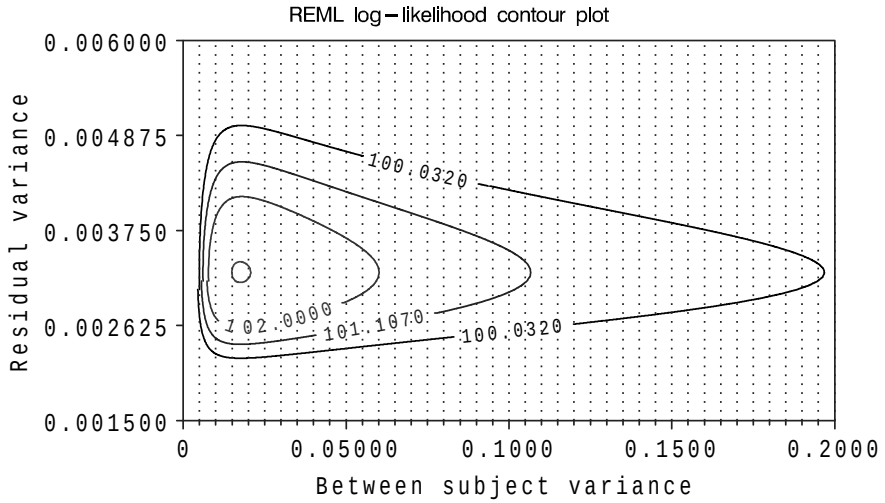


Figure 10.2: Contour plot of the profiled restricted log-likelihood from a linear mixed effects fit to $\log_{10}$ estrone data.

The maximal value of the restricted log-likelihood is 103.03. A 95% likelihood ratio confidence interval for σ_u^2 can be visually determined from Figure 10.2 using the contour corresponding to log-likelihood of $103.03 - 0.5\chi^2_{1,0.95} = 101.11$. This interval is approximately $(0.055, 0.106)$.

The waiting times are iid normal under $H_0 : \sigma_u^2 = 0$, and re-running `PROC MIXED` with the `RANDOM` statement removed calculates the maximized restricted log-likelihood under H_0, which is 45.68. The LRT-statistic for H_0 is therefore 114.70. Under the null hypothesis, the parameter lives on the boundary of the parameter space, and the usual asymptotic behaviour of the LRT-statistic does not apply. This matters little here, due to the extreme magnitude of the test statistic, and there is clearly massive evidence against H_0. For completeness, an exact test of H_0 is demonstrated in Section 10.4.2.

REML fit to log10(estrone) data using PROC MIXED

Iteration History			
Iteration	**Evaluations**	**−2 Res Log Like**	**Criterion**
1	2	−206.05467655	0.00000000

Covariance Parameter Estimates		
Cov Parm	**Subject**	**Estimate**
Intercept	person	0.01749
Residual		0.003254

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	1.4175	0.05949	4	23.83	<.0001

10.4.2 R

Here it is assumed that dataframe `estrone.df` contains two variables, `y` $= \log_{10}$(estrone) and a subject variable `Person`. Since the intercept term is fitted by default, the model expression simply requires the syntax `(1|Person)` to specify that the random effect of `Person` is added to the intercept.

```
> REMLfit=lmer(y~(1|Person),data=estrone.df)
> REMLfit
Linear mixed model fit by REML
Formula: y ~ (1 | Person)
   Data: estrone.df
    AIC    BIC logLik deviance REMLdev
 -200.1 -192.9  103.0   -209.9  -206.1
Random effects:
 Groups   Name        Variance  Std.Dev.
 Person   (Intercept) 0.0174942 0.132265
 Residual             0.0032544 0.057047
Number of obs: 80, groups: Person, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.41751    0.05949   23.83
```

The exact finite sample p-value for the likelihood ratio test of $H_0 : \sigma_u^2$ is estimated using function `exactRLRT` function in package `RLRsim` (Scheipl et al. 2008). This function numerically evaluates the formulae derived in Crainiceanu and Ruppert

(2004).

```
> library(RLRsim)
> exactRLRT(REMLfit)

simulated finite sample distribution of RLRT.  (p-value based on 10000
simulated values)

data:
RLRT = 114.6992, p-value < 2.2e-16
```

10.4.3 ADMB

The ADMB template file, `EstroneREMLProf.tpl` is given in Section 10.8.1. The log-joint density function of the observed data and latent variables, $f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})$, is programmed within the `PROCEDURE_SECTION` of the ADMB template. From the specification of the model in (10.11) and (10.12), this joint density is

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{u}; a, \sigma^2, \sigma_u^2) &= f(\boldsymbol{u}; \sigma_u^2) f(\boldsymbol{y}|\boldsymbol{u}; a, \sigma^2) \\ &= \prod_i \left(\frac{1}{\sqrt{2\pi}\sigma_u} e^{-u_i^2/2\sigma_u^2} \prod_j \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_{ij}-(a+u_i))^2/2\sigma^2} \right) \end{aligned} \quad (10.13)$$

The built-in likelihood profiling capabilities ADMB are not available for latent variable models. In the R code below, function `Profile` is created to call an ADMB executable that, for a given value of σ_u^2, calculates the profiled restricted log-likelihood

$$l_R^*(\sigma_u^2) = \max_{a,\sigma^2,} l_R(a, \sigma^2, \sigma_u^2) \ .$$

Function `plkhci` (introduced in Section 3.4.1) is then used to find a 95% LR confidence interval for σ_u^2. Note that the ADMB program uses the parameterization $(a, \log(\sigma), \log(\sigma_u))$, and so the value of `est` specified in the list object `x` is $\log(\widehat{\sigma}_u)$.

```
> #Create data file. y is a vector of log10(estrone)
> cat("#np \n",5,"\n #m \n",16,"\n #y \n",y,file="EstroneREMLProf.dat")
>
> #Define function to return profile (negative) log-likelihood
> Profile=function(logsigu) {
+   write(logsigu,"Logsigu.txt")
+   runAD("EstroneREMLProf",argvec="< Logsigu.txt > Out.txt")
```

```
+   lhood=(scan("EstroneREMLProf.rep",nmax=1))
+   return(lhood) }

>
> #Find LR interval for logsigu using plkhci function
> library(Bhat)
> x=list(label="logsigu",est=0.5*log(0.01749),low=-3,upp=-1)
> logsiguCI=plkhci(x,Profile,"logsigu")
> logsiguCI
[1] -2.599026 -1.121132

> #Back transform to get LR interval for between subject variance
> exp(2*logsiguCI)
[1] 0.005527317 0.106217707
```

Strong asymmetry in the profiled restricted log-likelihood for σ_u^2 shows the danger of assuming approximate normality of this estimator.

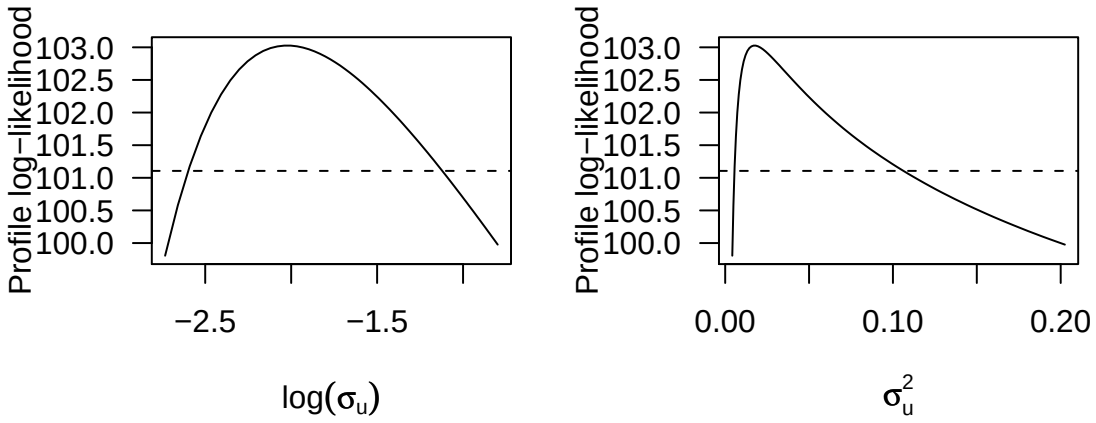


Figure 10.3: Profile restricted log-likelihoods for $\log(\sigma_u)$ and σ_u^2, for a linear mixed-model fit to $\log_{10}$ estrone data.

10.5 Nonlinear mixed-effects models

The data (Table 10.2) are measurements of the circumference (mm) of five orange trees at seven different sampling occasions (Draper and Smith 1981, p. 524). These data have been heavily used as an example of nonlinear mixed modeling, including by SAS Institute Inc (1999), and within the R-package `lme4` where they are provided as dataframe `Orange`. In these examples, the expected size of a tree is modeled as a scaled logistic function of age (measured as number of days since 31 Dec 1968). With x_j denoting the tree age on sampling occasion $j, j = 1, ..., 7$, the expected

	Tree				
Age	1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

Table 10.2: Circumference (mm) of five orange trees, from Draper and Smith (1981).

circumference can be expressed as

$$\mu_j = \frac{a}{1 + \exp(-(x_j - b)/c)} \ . \tag{10.14}$$

In this equation, parameter a corresponds to the asymptotic (as age goes to infinity) expected circumference, and b corresponds to the age as which expected circumference is half of this maximum. Parameter c controls the rate of growth.

However, it is clear from a scatter plot of the data (Fig. 10.4) that the measured circumferences are not independently distributed about the expected circumference. In particular, Figure 10.4 indicates that each tree appears to have its own underlying growth curve, These individual curves appear to differ primarily in vertical scale, but otherwise have similar shape. This suggests using a model that permits the asymptotic size to vary between tree. The trees can be assumed to be a random sample from the population of all suitable orange trees, and it would therefore be appropriate that their asymptotic sizes be modeled using random effects. Letting subscript $i, i = 1, ..., 5$ denote the tree, a potential model for the measurement y_{ij} on tree $i = 1, ..., 5$ at at time $j = 1, ..., 7$ would be

$$\begin{aligned} y_{ij}|u_i &\sim N(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \frac{a + u_i}{1 + \exp(-(x_j - b)/c)} \qquad (10.15) \\ u_i &\sim N(0, \sigma_u^2) \ . \qquad (10.16) \end{aligned}$$

Under this model, the asymptotic size of tree i is $a + u_i$. The parameters to be estimated are $\boldsymbol{\theta} = (a, b, c, \sigma^2, \sigma_u^2)$.

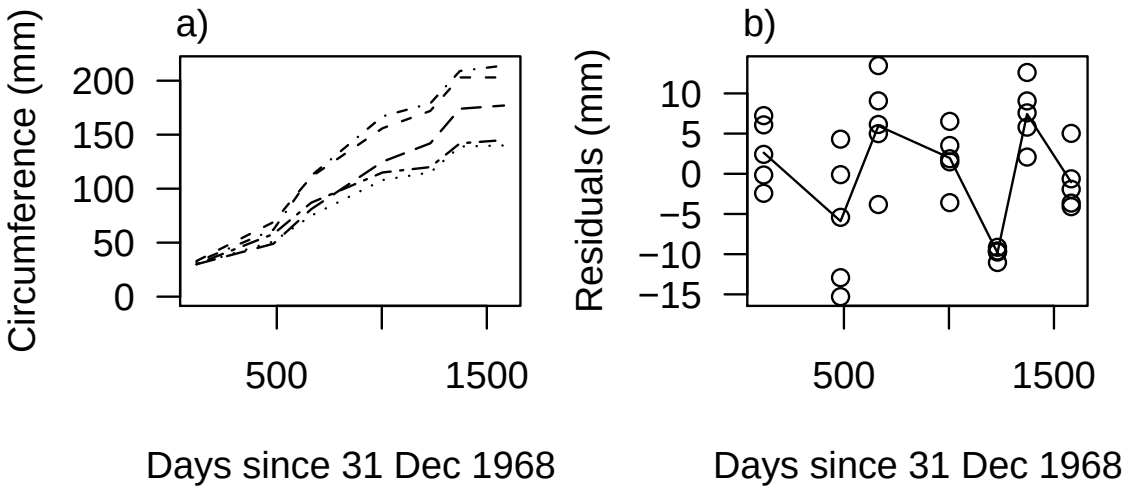


Figure 10.4: a) Orange tree circumference data with the measurements from each tree connected by a dashed line; b) Residuals from the fit of the tree-effect model.

A closer inspection of Figure 10.4 shows that the measured circumferences on the five trees appear to lack independence at each sampling time. For example, all five trees had a noticeably higher measured growth rate between sampling occasions 5 and 6 than between occasions 4 and 5. This may well be due to seasonal growth factors, but in the absence of knowledge about the conditions under which the trees were reared (but see Exercise 10.5), an alternative model for including this feature of the data would be to include a random effect of time. For example,

$$\begin{aligned} y_{ij}|\mu_{ij} &\sim N(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \frac{a + u_i + v_j}{1 + \exp(-(x_j - b)/c)} \\ u_i &\sim N(0, \sigma_u^2) \\ v_j &\sim N(0, \sigma_v^2) \ . \end{aligned}$$

This model adds an extra parameter σ_v^2, and is an example of a crossed-effects NLMM. Note that this model does not take into account the temporal order of the sampling – this could be accommodated by imposing an autoregressive structure on $v_j, j = 1, ..., 7$ (see Exercise 10.6).

The model with random tree effect was fitted using SAS procedure `NLMIXED`. This procedure uses Gaussian quadrature and can accommodate only one random effect. ADMB was used to fit the model having both random tree and day effects. The help

files in the R package `lme4` include example code for fitting the tree-effect model to these data, but it is not working correctly in the current version (`lme4_0.999375-28`) of this package. This will be corrected in later versions of the `lme4` package, but the reader is advised to check by ensuring that `lmer` returns the same fits as SAS and ADMB (see Exercise 10.3).

10.5.1 SAS

The required PROC NLMIXED code includes programming statements to calculate the expected circumference as specified by equation (10.15), and a `RANDOM` statement to specify that variable u varies by group according to a $N(0, \sigma_u^2)$ distribution.

PROC NLMIXED code for random tree effect model of orange trees

```
PROC NLMIXED DATA=tree;
  PARMS a=200 b=725 c=350 sigmasq=50 sigmasqu=1000;
  ExpSize=(a+u)/(1+exp(-(day-b)/c));
  MODEL y~NORMAL(ExpSize,sigmasq);
  RANDOM u~NORMAL(0,sigmasqu) SUBJECT=tree;
  PREDICT  y-ExpSize OUT=Residuals;
RUN;
```

The maximized value of the log-likelihood was -131.57, and the MLE $\widehat{\boldsymbol{\theta}}$ and the approximate standard errors are shown below.

Random tree−effect logistic model of orange tree circumferences

Parameter	Estimate	StandardError
a	192.06	15.6491
b	727.95	35.2561
c	348.09	27.0834
sigmasq	61.5090	15.8806
sigmasqu	1000.03	647.58

Residuals from this fit were obtained from the PREDICT statement. The residual plot (Fig. 10.4) confirms the lack of independence between sampling occasion. These residuals are calculated as $y_{ij} - \widehat{\mu}_{ij}$, where $\widehat{\mu}_{ij}$ is obtained from (10.15) by setting $\boldsymbol{u}$ and $\boldsymbol{v}$ to the values that maximize $f(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}; \widehat{\boldsymbol{\theta}})$.

	ADMB		Simulated ML	
	$\widehat{\boldsymbol{\theta}}$	std. error	$\widehat{\boldsymbol{\theta}}$	sim. error
a	196.2	19.4	195.9	0.2
b	748.4	62.3	747.6	0.3
c	352.9	33.3	352.7	0.1
σ^2	28.1	8.2	28.1	0.01
σ_u^2	1061.0	68.8	1059.8	0.5
σ_v^2	109.9	9.1	109.1	0.2

Table 10.3: ADMB and simulated ML fits of the crossed-effects NLMM to the orange tree circumference data. The Monte-Carlo simulation error of the simulated MLE is shown in the right-hand column.

10.5.2 ADMB

Letting $\boldsymbol{u} = (u_1, ..., u_5)$ denote the latent tree effects, and $\boldsymbol{v} = (u_1, ..., u_7)$ the latent day effects, the ADMB template file (Section 10.8.2) contains programming code to calculate the joint density

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) &= f(\boldsymbol{u}; \sigma_u^2) f(\boldsymbol{v}; \sigma_v^2) f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{v}; a, b, c, \sigma^2) \\ &= \left(\prod_i \frac{1}{\sqrt{2\pi}\sigma_u} e^{-u_i^2/2\sigma_u^2} \right) \left(\prod_j \frac{1}{\sqrt{2\pi}\sigma_v} e^{-v_i^2/2\sigma_v^2} \right) \left(\prod_{i,j} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_{ij}-\mu_{ij})^2/2\sigma^2} \right) , \end{aligned}$$

where $\boldsymbol{\theta} = (a, b, c, \sigma^2, \sigma_u^2, \sigma_v^2)$.

The fit of this model has previously been obtained using simulated maximum likelihood[2] (Millar 2004). The ADMB estimates, obtained using Laplace approximation, were very close to those obtained by simulated likelihood.

The maximized log-likelihood of the crossed-effects model is $l(\widehat{\boldsymbol{\theta}}) = -125.45$. The LRT-statistic for the hypothesis of no day effect, $H_0, \sigma_v^2 = 0$, is therefore $2(-125.45 + 131.57) = 12.24$. This is sufficiently large to provide very strong evidence against H_0. Bootstrap simulation of the p-value is left to Exercise 10.4.

[2] A compute intensive approach that iteratively fits a Bayesian version of the model, followed by a maximization step.

10.6 Generalized linear mixed-effects models

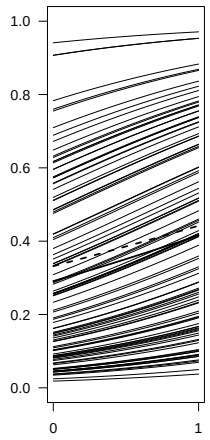
Multi-center trials. Compare to GEE. Note that deviance can not be used in GLMM context.

10.6.1 SAS

Can use either NLMIXED or GLIMMIX.

10.6.2 R

10.6.3 ADMB



10.7 State-space models

Conjunctivitus data. (ADMB only)

10.8 ADMB template files

10.8.1 One-way linear mixed effects model

Features to note:

- The `PROCEDURE_SECTION` programs the log of the joint density function given in (10.13).

- For the purposes of calculating a profile likelihood for σ_u^2, the value of $\log(\sigma_u)$ is input as a command line argument. To maximize over $\widehat{\sigma}_u^2$, the `LOCAL_CALCS` section would be removed and `log_sigma_u` would be declared as an `init_bounded_number` in the `PARAMETER_SECTION`.

- This code implements a REML fit via integrated likelihood (Section 9.5), which treats a as a latent variable. (The current version of ADMB does not include a `random_effects_number` and it was necessary to treat a as a random vector of length 1.) A ML fit would require changing the declaration of a to an `init_number`.

- The third argument in the specification of the random effects vector, $\boldsymbol{u}$, specifies that these are fitted in phase 2 of optimization. That is, a fit is achieved with $\boldsymbol{u} = \mathbf{0}$ in phase 1, and this is then used as the starting point for full optimization.

Estrone.tpl

```
DATA_SECTION
  init_int np                   //Number of people
  init_int m                    //Number of measurements on each
  init_matrix y(1,np,1,m)       // Data matrix (5 by 16)
  number log_sigma_u
 LOCAL_CALCS
  cin>>log_sigma_u;             //Read from text file named in command line
 END_CALCS

PARAMETER_SECTION
  init_bounded_number log_sigma(-10,10)  // log(residual s.e)
  random_effects_vector a(1,1)           // Vector of length 1
  random_effects_vector u(1,np,2)      // Random effects vectors
  objective_function_value neg_log_joint

PROCEDURE_SECTION
  int i,j;
  const double pi=3.14159265358979323846;
  dvariable pred, f; //f is the integrand
  dvariable sigma = exp(log_sigma);
  dvariable sigma_u=exp(log_sigma_u);
  f=0;
  for(i=1;i<=np;i++)
  {
    //Random effect contribution to joint density
    f = f-0.5*log(2*pi)-log_sigma_u-0.5*(square(u(i)/sigma_u));
    pred=a[1]+u[i];
    for(j=1;j<=m;j++)
    {
      //Data contribution to joint density
      f = f-0.5*log(2*pi)-log_sigma-0.5*square((y(i,j)-pred)/sigma);
    }
```

```
  }
  neg_log_joint=-f;

REPORT_SECTION
  report << neg_log_joint << endl;
```

10.8.2 Nonlinear mixed-effects model

OrangeTrees.tpl

```
DATA_SECTION
  init_int ntree                    //Number of trees
  init_int nday                     //Number of days
  init_matrix y(1,ntree,1,nday)     // Response vector
  init_vector t(1,nday)             // Days of measurement

PARAMETER_SECTION
  init_number a                          // Asymptotic size
  init_number b                            // Age at mid size
  init_number c                            // Growth rate paramter
  init_bounded_number log_sigma(-5,5,1)    // log(residual s.e)
  init_bounded_number log_sigma_u(-5,5,2) // log(tree effect s.e)
  init_bounded_number log_sigma_v(-5,5,2) // log(day effect s.e)
  random_effects_vector u(1,ntree,2)   // Random tree effects
  random_effects_vector v(1,nday,2)        // Random day effect
  objective_function_value neg_log_joint
  sdreport_number sigmasq
  sdreport_number sigmasq_u
  sdreport_number sigmasq_v

PROCEDURE_SECTION
  int i,j;
  const double pi=3.14159265358979323846;
  dvariable pred, f; // f is the integrand
  sigmasq = exp(2*log_sigma);
  sigmasq_u = exp(2*log_sigma_u);
  sigmasq_v = exp(2*log_sigma_v);
  f = 0.0;
  for(i=1;i<=ntree;i++)
  {
    for(j=1;j<=nday;j++)
    {
      pred = (a+u(i)+v(j))/(1+exp(-(t(j)-b)/c));
      f = f-0.5*log(2*pi)-log_sigma-0.5*square(y(i,j)-pred)/sigmasq;
    }
  }
// Random effect contribution from u
  for(i=1;i<=ntree;i++)
  {
    f = f-0.5*log(2*pi)-log_sigma_u-0.5*square(u(i))/sigmasq_u;
  }
// Random effects contribution from v
  for(j=1;j<=nday;j++)
  {
    f = f-0.5*log(2*pi)-log_sigma_v-0.5*square(v(j))/sigmasq_v;
  }
  neg_log_joint=-f;
```

10.9 Exercises

10.1 The estrone data in Table 10.1 in batches of four vials (Gail et al. 1996). For each subject, the first four vials comprise batch 1, the next four comprise batch 2, and so on.

1. Using restricted maximum likelihood, fit a nested LMM to the estrone data, with grouping factors subject and batch (nested within subject).
2. Which model do you prefer, the model with nested random effects or the model with only the subject random effect?

10.2 For the nested LMM in Exercise 10.1, conduct a bootstrap simulation to evaluate the sampling distribution of the LRT statistic for testing the hypothesis of no batch effect, and compare to the actual-LRT statistic.

10.3 Check whether the `nlme` function from R-package `lme4` is giving correct results by applying it to the orange tree data used in Section 10.4.

1. Fit a model with tree effect only, and compare the results to those from SAS (Section 10.4.1).
2. Create an indicator variable for the age variable, and fit a model with tree and age effects. Compare the results to those from ADMB (Section 10.4.3).

10.4 For the orange tree data, conduct a bootstrap simulation to evaluate the sampling distribution of the LRT-statistic for the hypothesis of no day effect, and compare to the actual LRT-statistic.

10.5 Instead of adding a random day effect to the tree-effect model, fit a seasonal model to the orange tree data. Assume a northern hemisphere climate which causes the rate of growth to be at its minimum on 31 December, and at its greatest on 30 June. Specifically, let the growth rate parameter vary with calendar day $k = 1, ..., 365$, according to the sinusoidal curve

$$c_k = c_{min} + \delta(1 - \cos(2\pi k/365) \ .$$

Here the parameters to be estimated are $\boldsymbol{\theta} = (a, b, c_{min}, \delta, \sigma^2, \sigma_u^2)$.

10.6 Fit an AR(1) temporal model to orange tree data.

10.7 Taking advantage of separability, use the `integrate` function to calculate the likelihood for clinic only model... eg 3.

10.8 Write this model in separable form to use GQ in ADMB.

10.9 Use REML in eg 3.

Part Three: Theoretical foundations

Chapter 11

Cramér-Rao inequality and Fisher information

11.1 Introduction

The Cramér-Rao inequality specifies a performance limit on unbiased estimators (should they exist) of $\theta \in \mathbb{R}$, or of scalar functions $g(\boldsymbol{\theta})$ when $\boldsymbol{\theta} \in \mathbb{R}^s, s \geq 1$. It does this by providing an explicit lower bound on the variance of all such estimators. In general practice, the CR inequality is of limited utility because the construction of unbiased estimators that achieve the lower bound is only possible within the class of exponential family models (Chapter **??**) and subject to $g(\boldsymbol{\theta})$ being of a specific form (Pawitan 2001, p. 223), notwithstanding that this class of models does include the normal-linear models. More generally, all unbiased estimators of $g(\boldsymbol{\theta})$ could have variance that exceeds the lower bound of the CR inequality. It might then be possible to apply the Rao-Blackwell method (Pawitan 2001, p. 225 for details) to find one having the smallest variance.

However, a theory based on unbiased estimators is far too restrictive to be a tool for general-purpose statistical inference. Indeed, in many applications it will be the case that no unbiased estimators exists (e.g. see Example 11.2). MLEs are not generally unbiased, but typically they are approximately unbiased for sufficiently large sample size. Moreover, as seen in Chapter 12, the CR lower bound is attained by MLEs in an asymptotic sense, as the variance of a limiting distribution. For this reason, MLEs are said to be "asymptotically efficient".

11.1.1 Notation

The notation $E_{\boldsymbol{\theta}}$ denotes expectation (over repeat observation of $\boldsymbol{Y}$) under the distribution with density $f(\boldsymbol{y};\boldsymbol{\theta})$, and similarly for variance $Var_{\boldsymbol{\theta}}$. That is, $E_{\boldsymbol{\theta}}$ can be considered an abbreviation for the more explicit but cumbersome notation $E_{\boldsymbol{Y};\boldsymbol{\theta}}$.

It will be assumed that the model is a collection of distributions that are absolutely continuous with respect to some sigma-finite measure. However, familiarity with measure theory is not required, and the notation will omit these details for simplicity. The notation employed will be for that of a continuous distribution, and so, for example, the expected value of a real-valued function $h(\boldsymbol{Y})$ will be written *Ref?*

$$E_{\boldsymbol{\theta}}[h(\boldsymbol{Y})] = \int h(\boldsymbol{y}) f(\boldsymbol{y};\boldsymbol{\theta}) d\boldsymbol{y} \ .$$

If $\boldsymbol{Y}$ have a discrete distribution, then $f(\boldsymbol{y};\boldsymbol{\theta})$ denotes a probability mass function and it is implicitly understood that

$$\int h(\boldsymbol{y}) d\boldsymbol{y} \equiv \sum_{\boldsymbol{y}} h(\boldsymbol{y}) f(\boldsymbol{y};\boldsymbol{\theta}) \ .$$

11.2 The Cramér-Rao inequality for $\theta \in \mathbb{R}$

The Cramér-Rao inequality makes use of the *expected Fisher information*. This quantity is also a central element of the asymptotic theory of MLEs.

Definition 11.1 Expected Fisher information, $\theta \in \mathbb{R}$. *Let $\mathbf{Y}$ have distribution with density function $f(\mathbf{y};\theta), \theta \in \mathbb{R}$. Assuming that $\frac{\partial f(y;\theta)}{\partial \theta}$ exists, the expected Fisher information about θ contained in $\mathbf{Y}$ is*

$$I(\theta) = E_\theta \left[\left(\frac{\partial l(\theta;\, \boldsymbol{Y})}{\partial \theta} \right)^2 \right] \ . \tag{11.1}$$

Expected Fisher information can appear to be an unusual beast at first sight. It may help to write it more explicitly as

$$I(\theta_0) \equiv E_{\theta_0} \left[\left(\frac{\partial l(\theta;\, \boldsymbol{Y})}{\partial \theta} \Big|_{\theta=\theta_0} \right)^2 \right] \ , \tag{11.2}$$

where $I(\theta_0)$ is the expected Fisher information evaluated at the fixed parameter value θ_0. Examples of the calculation of $I(\theta)$ are given after the statement of the Cramér-Rao inequality.

In expression (11.2), the derivative of the log-likelihood

$$s(\theta, \boldsymbol{Y}) = \frac{\partial l(\theta; \boldsymbol{Y})}{\partial \theta} , \tag{11.3}$$

is commonly known as the *score function* when regarded as a function of θ (for fixed $\boldsymbol{Y}$), or the *score statistic* when regarded as function of $\boldsymbol{Y}$ (for fixed θ). The score statistic can be regarded as a transformation of $\boldsymbol{Y}$ from $\mathbb{R}^n$ to $\mathbb{R}$ and hence is simply a random variable.

Before stating the various forms of the Cramér-Rao inequality, a formal definition of unbiased estimator is needed.

Definition 11.2 Unbiased estimator *The estimator $T(\mathbf{Y})$ of $g(\boldsymbol{\theta}) \in \mathbb{R}$ is an* unbiased estimator *of $g(\boldsymbol{\theta})$ if $E_{\boldsymbol{\theta}}[T(\mathbf{Y})] = g(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$.*

Theorem 11.1 Cramér-Rao inequality for estimation of $\theta \in \Theta \subset \mathbb{R}$.
Let $\mathbf{Y} = (Y_1, ..., Y_n)$ be a sample from a distribution with joint density function $f(\mathbf{y}; \theta)$ and let $T(\mathbf{Y})$ be any unbiased estimator of $\theta \in \Theta \subset \mathbb{R}$. If $Var_\theta(T) < \infty$ and $\int T(\mathbf{y}) f(\mathbf{y}; \theta) d\mathbf{y}$ and $\int f(\mathbf{y}; \theta) d\mathbf{y}$ can be differentiated with respect to θ under the integral sign then,

$$Var_\theta(T) \geq \frac{1}{I(\theta)} \tag{11.4}$$

provided that $0 < I(\theta) < \infty$. The right-hand side of (11.4) is called the Cramér-Rao lower bound

Proof of CR inequality. We want to show that $Var_\theta(T) I(\theta) \geq 1$. Letting f denote $f(\boldsymbol{y}, \theta)$,

$$\begin{aligned} Var_\theta(T) I(\theta) &= \int (T(\boldsymbol{y}) - \theta)^2 f d\boldsymbol{y} \int \left(\frac{\partial \log f}{\partial \theta} \right)^2 f d\boldsymbol{y} \\ &\geq \left(\int (T(\boldsymbol{y}) - \theta) \left(\frac{\partial \log f}{\partial \theta} \right) f d\boldsymbol{y} \right)^2 \end{aligned}$$

by the Cauchy-Schwarz inequality. The right hand side equals

$$\left(\int (T(\boldsymbol{y})-\theta)\frac{f'}{f} f d\boldsymbol{y}\right)^2 = \left(\int (T(\boldsymbol{y})-\theta) f' d\boldsymbol{y}\right)^2 \tag{11.5}$$

where $f' \equiv f'(\boldsymbol{y};\theta)$ is the derivative of $f(\boldsymbol{y};\theta)$ with respect to θ. Now, $\int f(\boldsymbol{y};\theta)d\boldsymbol{y}$ can be differentiated under the integral sign, and so

$$\begin{aligned}\int \theta f' d\boldsymbol{y} &= \theta \int f' d\boldsymbol{y} \\ &= \theta \, \frac{\partial \int f d\boldsymbol{y}}{\partial\theta} = \theta \, \frac{\partial 1}{\partial\theta} = 0,\end{aligned} \tag{11.6}$$

where we used the fact that

$$\int f(\boldsymbol{y};\theta)d\boldsymbol{y} = 1 \text{ for all } \theta \in \Theta.$$

Similarly

$$\int T(\boldsymbol{y}) f' d\boldsymbol{y} = \frac{\partial \int T(\boldsymbol{y}) f d\boldsymbol{y}}{\partial\theta} = \frac{\partial\theta}{\partial\theta} = 1 \tag{11.7}$$

since T is an unbiased estimator of θ. Substituting (11.6) and (11.7) into (11.5) gives the required result. □

Example 11.1. Binomial. Let Y be distributed Binomial(n,p). To within a constant the likelihood is $L(p;y) = p^y(1-p)^{n-y}$ and the log-likelihood is

$$l(p;y) = y\log(p) + (n-y)\log(1-p)\ , \tag{11.8}$$

so

$$\frac{\partial l(p;y)}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p} = \frac{y-np}{p(1-p)}.$$

The expected Fisher information for p provided by observing Y is therefore

$$\begin{aligned} I(p) &= E_p\left[\left(\frac{Y-np}{p(1-p)}\right)^2\right] \\ &= \frac{E_p[(Y-np)^2]}{p^2(1-p)^2} \\ &= \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}. \end{aligned}$$

That is, for any unbiased estimator $T(Y)$ of p,

$$Var_p(T(Y)) \geq \frac{1}{I(p)} = \frac{p(1-p)}{n}$$

and in this simple example the unbiased estimator $\widehat{p} = Y/n$ attains the bound. □

The Cramér-Rao inequality does *not* imply the existence of an unbiased estimator of θ that will achieve the lower bound, nor even the existence of *any* unbiased estimator of θ.

The Cramér-Rao inequality is next extended to estimators of functions of $\theta \in \mathbb{R}$ (Section 11.3) and later to functions of $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$ (Section 11.6) .

11.3 CR inequality for functions of θ

Here, it is desired to obtain a version of the Cramér-Rao inequality for unbiased estimators of $\zeta = g(\theta)$ where $g : \Theta \to \mathbb{R}$ is assumed differentiable with derivative denoted $g'(\theta) = \frac{\partial \zeta}{\partial \theta}$.

If $g(\theta)$ is invertible then the model can be parameterized as a function of ζ. To distinguish models parameterized by θ from those parameterized by ζ, a temporary modification of notation will be employed by introducing subscripts to the log-likelihood, so that l_θ and l_ζ denote the log-likelihood functions for θ and ζ respectively. That is, $l_\zeta(\zeta; \boldsymbol{y}) = l_\theta(g^{-1}(\zeta); \boldsymbol{y}) = l_\theta(\theta; \boldsymbol{y})$. From the chain rule of differentiation

$$\begin{aligned} \frac{\partial l_\zeta(\zeta; \boldsymbol{y})}{\partial \zeta} &= \frac{\partial l_\theta(\theta; \boldsymbol{y})}{\partial \theta} \frac{\partial \theta}{\partial \zeta} \\ &= \frac{\partial l_\theta(\theta; \boldsymbol{y})}{\partial \theta} \left(\frac{\partial \zeta}{\partial \theta} \right)^{-1} = \frac{1}{g'(\theta)} \frac{\partial l_\theta(\theta; \boldsymbol{y})}{\partial \theta} \ . \end{aligned}$$

The expected Fisher information about ζ from observing $\boldsymbol{y}$ can therefore be obtained as

$$\begin{aligned} I(\zeta) &= E_\zeta \left[\left(\frac{\partial l_\zeta(\zeta; \mathbf{Y})}{\partial \zeta} \right)^2 \right] \\ &= \frac{1}{g'(\theta)^2} E_\theta \left[\left(\frac{\partial l_\theta(\theta; \mathbf{Y})}{\partial \theta} \right)^2 \right] = \frac{I(\theta)}{g'(\theta)^2} \end{aligned} \tag{11.9}$$

Example 11.2. Log-odds Suppose that Y is distributed Binomial$(n, p), 0 < p < 1$, and that the quantity of interest is the log-odds, $\zeta = g(p) = \text{logit}(p) = \log(p/(1-p))$. The inverse of g is

$$p = g^{-1}(\zeta) = \frac{e^{\zeta}}{1+e^{\zeta}}$$

and $I(\zeta)$ can be obtained from working with the binomial likelihood (11.8) expressed as a function of ζ (see Exercise 11.3). Alternatively, using (11.9), the expected Fisher information about ζ from Y is

$$\begin{aligned} I(\zeta) &= \left(\frac{\partial \log(p/(1-p))}{\partial p}\right)^{-2} I(\theta) \\ &= \left(\frac{1}{p(1-p)}\right)^{-2} \frac{n}{p(1-p)} \\ &= np(1-p) \end{aligned}$$

The CR inequality for unbiased estimators $T(Y)$ of $g(p)$ is therefore

$$Var_p(T) \geq \frac{1}{I(\zeta)} = \frac{1}{np(1-p)}.$$

Here, it can be shown that no unbiased estimator of $g(p)$ exists - see Box 11.1 □

Box 11.1.

For the estimation problem in Example 11.2, if $T(Y)$ is an unbiased estimator of the log-odds then

$$E_p(T(Y)) = \sum_{y=0}^{n} \binom{n}{y} T(y) p^y (1-p)^{n-y} = \log(p/(1-p)) \qquad \forall p,\ 0 < p < 1.$$

Note that $E_p(T(Y)) \leq \max_y T(y)$, yet $g(p) = \log(p/(1-p))$ is unbounded as $p \to 1$. Hence no unbiased estimator of the log-odds can exist.

It is immediate from (11.9) that the Cramér-Rao lower bound for unbiased estimators $T(\boldsymbol{Y})$ of $\zeta = g(\theta)$ is $g'(\theta)^2 I(\theta)^{-1}$. This lower bound also holds more generally for differentiable functions $g(\theta)$ that are not necessarily invertible, and is stated below.

Theorem 11.2 *If $T(\mathbf{Y})$ is an unbiased estimator of the differentiable function $g(\theta)$, and the conditions of Theorem 11.1 hold, then*

$$Var_\theta(T) \geq \frac{(g'(\theta))^2}{I(\theta)} \tag{11.10}$$

where $g'(\theta) = \frac{\partial g}{\partial \theta}$ evaluated at θ.

Theorem 11.2 can be proved by minor modification to the proof of Theorem 11.1 (Exercise 11.1).

11.4 Alternative formulae for $I(\theta)$

The next lemma provides alternative formulae for calculation of $I(\theta)$. It is often the case that calculating Fisher information using the form (11.2) can be cumbersome compared to use of the alternative (11.13).

Lemma 11.1 *If $\frac{\partial f(y;\theta)}{\partial \theta}$ exists and $\int f(\mathbf{y};\theta)d\mathbf{y}$ can be differentiated (with respect to θ) under the integral sign then*

$$E_\theta\left[\frac{\partial l(\theta;\mathbf{Y})}{\partial \theta}\right] = 0 \tag{11.11}$$

from which it immediately follows that

$$I(\theta) = Var_\theta\left[\frac{\partial l(\theta;\boldsymbol{Y})}{\partial \theta}\right] . \tag{11.12}$$

Furthermore, if $\frac{\partial^2 f(y;\theta)}{\partial \theta^2}$ exists and $\int f(\mathbf{y};\theta)d\mathbf{y}$ can be twice differentiated under the integral sign, then

$$I(\theta) = -E_\theta\left[\frac{\partial^2 l(\theta;\boldsymbol{Y})}{\partial \theta^2}\right] . \tag{11.13}$$

Proof. Note that

$$\begin{aligned} E_\theta\left[\frac{\partial l(\theta;\boldsymbol{Y})}{\partial \theta}\right] &\equiv E_\theta\left[\frac{\partial \log f(\boldsymbol{Y};\theta)}{\partial \theta}\right] \\ &= \int \frac{f'(\boldsymbol{y};\theta)}{f(\boldsymbol{y};\theta)} f(\boldsymbol{y};\theta)d\boldsymbol{y} \\ &= \int f'(\boldsymbol{y};\theta)d\boldsymbol{y} \\ &= \frac{\partial \int f(\boldsymbol{y};\theta)d\boldsymbol{y}}{\partial \theta} = \frac{\partial 1}{\partial \theta} = 0 \end{aligned}$$

To show (11.13), from the chain rule of differentiation

$$\begin{aligned}\frac{\partial^2 l(\theta;\boldsymbol{y})}{\partial\theta^2} &\equiv \frac{\partial}{\partial\theta}\frac{f'(\boldsymbol{y};\theta)}{f(\boldsymbol{y};\theta)} \\ &= \frac{f''(\boldsymbol{y};\theta)}{f(\boldsymbol{y};\theta)} - \left(\frac{f'(\boldsymbol{y};\theta)}{f(\boldsymbol{y};\theta)}\right)^2 .\end{aligned}$$

The first term on the right hand side vanishes when we take its expectation (since it is the double derivative (with respect to θ) of $\int f(\boldsymbol{y};\theta)d\boldsymbol{y}$), giving the required result.

Example 11.1 ctd. Binomial. We have that

$$\frac{\partial l(p;y)}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p}$$

and so the second derivative of the log-likelihood is

$$\frac{\partial^2 l(p;y)}{\partial p^2} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2}. \tag{11.14}$$

Taking expectations gives

$$\begin{aligned}I(p) &= E_p\left[\frac{Y}{p^2} + \frac{n-Y}{(1-p)^2}\right] \\ &= \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)} .\end{aligned}$$

□

11.5 The iid data case

In many experiments the data $Y_1, ..., Y_n$ will be iid. In this case it would seem natural that the information from the sample will be n times the information from a single data point. If Y_i are iid, then

$$\frac{\partial l(\theta;\,\boldsymbol{Y})}{\partial\theta} = \sum_{i=1}^{n}\frac{\partial \log f(Y_i;\theta)}{\partial\theta}$$

where $\frac{\partial \log f(Y_i;\theta)}{\partial\theta}, i = 1, ..., n$ are also iid random variables. Hence

$$\begin{aligned}I(\theta) &= Var_\theta\left[\frac{\partial l(\theta;\,\boldsymbol{Y})}{\partial\theta}\right] \\ &= \sum_{i=1}^{n} Var_\theta\left[\frac{\partial \log f(Y_i;\theta)}{\partial\theta}\right] = nI_1(\theta)\end{aligned}$$

where $I_1(\theta)$ is the expected Fisher information from one data point.

Example 11.1 ctd. Binomial. A binomial experiment is equivalent to observing n iid Bernoulli trials. $I_1(p)$, the information from a single Bernoulli trial, is $1/(p(1-p))$.

11.6 The multi-parameter case, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$

The notation, concepts and results of Sections 11.2–11.5 extend in a natural way to the multi-parameter case $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$. In particular, assuming differentiability of $f(\boldsymbol{y}; \boldsymbol{\theta})$ with respect to each $\theta_i, i = 1, ..., s$, the score statistic is now a s-dimensional random vector

$$s(\boldsymbol{\theta};\, \boldsymbol{Y}) = \frac{\partial l(\boldsymbol{\theta};\, \boldsymbol{Y})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial l(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \theta_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial l(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \theta_s} \end{pmatrix} .$$

Definition 11.3 Expected Fisher information, $\boldsymbol{\theta} \in \mathbb{R}^s$. *The expected Fisher information matrix, $I(\boldsymbol{\theta})$, is defined to be the $s \times s$ matrix with i, j^{th} element*

$$I(\boldsymbol{\theta})_{ij} = E_{\boldsymbol{\theta}} \left[\frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right] . \tag{11.15}$$

That is, $I(\boldsymbol{\theta})$ is the expectation of the outer product of the score statistic with itself

$$I(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left[\frac{\partial l}{\partial \boldsymbol{\theta}} \frac{\partial l}{\partial \boldsymbol{\theta}}^T \right] . \tag{11.16}$$

Ref???

Box 11.2.

From 11.16, $I(\boldsymbol{\theta})$ is necessarily a non-negative definite matrix, and for well-posed models it will be positive definite, and therefore non-singular (i.e., invertible). In practice, if $I(\boldsymbol{\theta})$ is singular then it will be due to the model being non-identifiable, as can arise in linear regression (Example 11.6) when the design matrix is not full rank.

The multi-parameter version of the Cramér Rao inequality is stated analogously to Theorem 11.2, but with the lower bound expressed using a quadratic form.

Theorem 11.3 Multiparameter Cramér Rao inequality $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$.
If $T(\boldsymbol{Y})$ *is an unbiased estimator of the differentiable function* $g(\boldsymbol{\theta}) : \Theta \to \mathbb{R}$ *then*

$$var_{\boldsymbol{\theta}}(T) \geq (g^{'})^T I^{-1}(\boldsymbol{\theta}) g^{'} \ , \tag{11.17}$$

where $g^{'} \equiv g^{'}(\boldsymbol{\theta}) = \frac{\partial g}{\partial \boldsymbol{\theta}}$ *is the column vector with* i^{th} *element*

$$g_i^{'} = \frac{\partial g}{\partial \theta_i} \ .$$

The proof of Theorem 11.3 requires an extension of the Cauchy-Schwarz inequality and is not provided here. It can be found in (Lehmann 1983, Section 2.7).

Example 11.3. Bound for θ_i If $g(\boldsymbol{\theta}) = \theta_i$, then $g^{'}(\boldsymbol{\theta})$ is the s-dimensional vector having unity as its ith element, and zero elsewhere. Hence, the lower bound on the variance of unbiased estimators of θ_i is $[I^{-1}(\boldsymbol{\theta})]_{ii}$, the ith diagonal element of the inverse Fisher information matrix. □

11.6.1 Alternative formulae for $I(\boldsymbol{\theta})$

Analogously to Section 11.5, if the data are iid then $I(\boldsymbol{\theta}) = nI_1(\boldsymbol{\theta})$ where $I_1(\boldsymbol{\theta})$ is the information matrix for a single observation. The alternative formulae for calculation of expected Fisher information extend to the multi-parameter case in a natural way.

Lemma 11.2 *If the first derivatives (with respect to* θ_i*) of* $f(y; \boldsymbol{\theta})$ *exist and* $\int f(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y}$ *can be differentiated (with respect to each* θ_i*) under the integral sign then*

$$E_{\boldsymbol{\theta}} \left[\frac{\partial l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}} \right] = \mathbf{0} \tag{11.18}$$

from which it immediately follows that $I(\boldsymbol{\theta})$ *is the* $s \times s$ *variance matrix of the score statistic. That is,*

$$I(\boldsymbol{\theta})_{ij} = Cov_{\boldsymbol{\theta}} \left(\frac{\partial l}{\partial \theta_i}, \frac{\partial l}{\partial \theta_j} \right) \ . \tag{11.19}$$

Furthermore, if all second derivatives of $f(y; \boldsymbol{\theta})$ *exist and* $\int f(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y}$ *can be twice differentiated under the integral sign, then*

$$I(\boldsymbol{\theta})_{ij} = -E \left[\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \theta_i \partial \theta_j} \right] \ . \tag{11.20}$$

That is, $I(\boldsymbol{\theta})$ *is the negative of the expected value of the* $s \times s$ *Hessian matrix of second derivatives of the log-likelihood.*

11.6.2 Fisher information for re-parameterized model

Suppose that the model is re-parameterized using $\boldsymbol{\zeta} = g(\boldsymbol{\theta})$, where $g : \Theta \to \mathbb{R}^s$ is differentiable and invertible. Denote

$$g(\boldsymbol{\theta}) = \begin{pmatrix} g_1(\boldsymbol{\theta}) \\ . \\ . \\ . \\ g_s(\boldsymbol{\theta}) \end{pmatrix}$$

where each co-ordinate $g_i : \Theta \to \mathbb{R}$ has derivative $g_i^{'} = (\frac{\partial g_i}{\partial \theta_1}, ..., \frac{\partial g_i}{\partial \theta_s})^T$. Then the Fisher information matrix $I(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta}$ can be obtained from the Fisher information matrix $I(\boldsymbol{\theta})$ using the natural extension of formula 11.9. That is,

$$I(\boldsymbol{\zeta}) = [G(\boldsymbol{\theta})^{-1}]^T I(\boldsymbol{\theta}) G(\boldsymbol{\theta})^{-1} \ , \tag{11.21}$$

where $G(\boldsymbol{\theta})$ is the s by s Jacobian matrix of derivatives

$$G(\boldsymbol{\theta}) = \begin{pmatrix} g_1^{'}(\boldsymbol{\theta})^T \\ . \\ . \\ . \\ g_s^{'}(\boldsymbol{\theta})^T \end{pmatrix} .$$

11.6.3 Examples

Example 11.4. Normal with both μ and σ^2 unknown. Let $Y_1, ..., Y_n$ be iid $\mathrm{N}(\mu, \sigma^2)$. Here, note that the information calculations are made with respect to parameters $(\theta_1, \theta_2) = (\mu, \sigma^2)$. Application of formula (11.21) is then demonstrated to obtain the information matrix for parameters $(\zeta_1, \zeta_2) = (\mu, \sigma)$.

Now, the contribution to the log-likelihood from a single data point y is

$$l(\boldsymbol{\theta}; y) = l(\mu, \sigma^2; y) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-\mu)^2$$

and so

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}(y-\mu) \\ \frac{\partial l}{\partial \sigma^2} &= \frac{-1}{2\sigma^2} + \frac{1}{2\sigma^4}(y-\mu)^2 \end{aligned}$$

The elements of $I_1(\boldsymbol{\theta}) = I_1(\mu, \sigma^2)$ are

$$\begin{aligned}
[I_1]_{11} &= E_{\boldsymbol{\theta}}\left[\left(\frac{\partial l}{\partial \mu}\right)^2\right] \\
&= \frac{1}{\sigma^4} E_{\boldsymbol{\theta}}\left[(Y-\mu)^2\right] = \frac{1}{\sigma^2} \\
[I_1]_{12} &= E_{\boldsymbol{\theta}}\left[\frac{\partial l}{\partial \mu}\frac{\partial l}{\partial \sigma^2}\right] \\
&= \frac{1}{2\sigma^4} E_{\boldsymbol{\theta}}\left[-(Y-\mu) + \frac{1}{\sigma^2}(Y-\mu)^3\right] = 0 \\
[I_1]_{22} &= E_{\boldsymbol{\theta}}\left[\left(\frac{\partial l}{\partial \sigma^2}\right)^2\right] \\
&= \frac{1}{4\sigma^4} E_{\boldsymbol{\theta}}\left[1 - \frac{2}{\sigma^2}(Y-\mu)^2 + \frac{1}{\sigma^4}(Y-\mu)^4\right] \\
&= \frac{1}{4\sigma^4}\left[1 - \frac{2\sigma^2}{\sigma^2} + \frac{3\sigma^4}{\sigma^4}\right] = \frac{1}{2\sigma^4}
\end{aligned}$$

Alternatively, using formula (11.20) based on second derivative calculations,

$$\begin{aligned}
[I_1]_{11} &= -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 l}{\partial \mu^2}\right] \\
&= -E_{\boldsymbol{\theta}}\left[\frac{-1}{\sigma^2}\right] = \frac{1}{\sigma^2} \\
\mathrm{I}_1]_{12} &= -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 l}{\partial \mu \partial \sigma^2}\right] \\
&= \frac{1}{\sigma^4} E_{\boldsymbol{\theta}}\left[(Y-\mu)\right] = 0 \\
[I_1]_{22} &= -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 l}{\partial (\sigma^2)^2}\right] \\
&= \frac{-1}{2\sigma^4} + \frac{1}{\sigma^6} E_{\boldsymbol{\theta}}\left[(Y-\mu)^2\right] = \frac{1}{2\sigma^4}\ .
\end{aligned}$$

Thus, the information matrix for a single observation is

$$I_1(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$$

and so

$$I(\mu, \sigma^2) = n \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix} \qquad \text{and} \qquad I^{-1}(\mu, \sigma^2) = \frac{1}{n}\begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\ .$$

That is, for a sample of size n, the CR lower bounds for unbiased estimators of μ and σ^2 are $[I^{-1}]_{11} = \sigma^2/n$ and $[I^{-1}]_{22} = 2\sigma^4/n$ respectively. The MLE of μ is $\overline{Y}$ which is unbiased and has variance that equals the CR lower bound.

It can be shown that no unbiased estimator of σ^2 can achieve the CR lower bound (e.g., see Pawitan 2001, p. 223). The MLE of σ^2 is $\widehat{\sigma}^2 = \frac{n-1}{n}S^2$ where S^2 is the familiar unbiased sample variance. Th estimator S^2 has distribution (Seber and Lee 2003, p. 48)

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} \sim \frac{\sigma^2\chi^2_{n-1}}{n-1}.$$

Since $E[\chi^2_{n-1}] = n-1$, S^2 is unbiased for σ^2 and $\widehat{\sigma}^2$ has bias $\frac{-\sigma^2}{n}$ Also,

$$\begin{aligned} \text{var}(S^2) &= \left(\frac{\sigma^2}{n-1}\right)^2 \text{var}(\chi^2_{n-1}) \\ &= \left(\frac{\sigma^2}{n-1}\right)^2 2(n-1) = \frac{2\sigma^4}{n-1}, \end{aligned}$$

and

$$\begin{aligned} \text{var}(\widehat{\sigma}^2) &= \left(\frac{n-1}{n}\right)^2 \text{var}(S^2) \\ &= \left(\frac{n-1}{n}\right)\frac{2\sigma^4}{n}. \end{aligned}$$

For large n, the bias of $\widehat{\sigma}^2$ tends to zero and it is said to be asymptotically unbiased. The limiting variance (as $n \to \infty$) of both S^2 and $\widehat{\sigma}^2$ is the CR lower bound, and hence they are both asymptotically efficient.

Finally, if this model were re-parameterized using $\boldsymbol{\zeta} = (\mu, \sigma) = (\theta_1, \sqrt{\theta_2})$, then $I(\boldsymbol{\zeta})$ could be obtained from re-doing the information calculations, or, obtained from equation (11.21) as

$$\begin{aligned} I_1(\boldsymbol{\zeta}) = I_1(\mu, \sigma) &= \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2\sqrt{\theta_2}} \end{pmatrix}^{-1} I(\boldsymbol{\theta}) \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2\sqrt{\theta_2}} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} \\ &= \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix} \end{aligned}$$

□

In the above example the information matrix $I(\boldsymbol{\theta})$ is diagonal and consequently the CR lower bound for each parameter does not depend on whether the other parameter is known. For example, if σ^2 were known, one would simply calculate

$I^{-1}(\mu) = \sigma^2/n$. If μ were known, then $I^{-1}(\sigma^2) = 2\sigma^4/n$ and in this case the MLE of σ^2 would be $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \mu)^2$. This is an unbiased estimator of σ^2 and has variance that equals the CR lower bound.

Example 11.5. Exponential family distributions. Let $Y_i, i = 1, ..., n$ be independent with distribution belonging to the exponential family such that the log-density can be written in the form

$$\log f(y_i; \psi_i, \phi_i) = \frac{y\psi_i - b(\psi_i)}{a(\phi_i)} + c(y, \phi_i) \ . \tag{11.22}$$

For each i, it will assumed that $a(\phi_i) = \phi/w_i$ for some $\phi > 0$ and known values w_i. In general, ϕ is a parameter to be estimated, but in some cases may be known (e.g., see Example 7.2). Furthermore, ψ_i is assumed to be a differentiable function of parameters $\beta_j \in \mathbb{R}, j = 1, ..., p$. The parameters to be estimated are $\boldsymbol{\theta} = (\beta_1, ..., \beta_p, \phi)$, and $I(\boldsymbol{\beta})$ will be used to denote the upper-left $p \times p$ sub-matrix of $I(\boldsymbol{\theta})$ corresponding to information calculations with respect to $\boldsymbol{\beta}$ only.

Denoting $l_i = l(\boldsymbol{\beta}, \phi; y_i)$,

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, \phi; \mathbf{y})}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i}{\partial \beta_j} &= \sum_{i=1}^n \frac{\partial l_i}{\partial \psi_i}\frac{\partial \psi_i}{\partial \beta_j} \\ &= \sum_{i=1}^n \frac{(y_i - b'(\psi_i))}{a(\phi_i)}\frac{\partial \psi_i}{\partial \beta_j} \ . \end{aligned} \tag{11.23}$$

It is more intuitive to re-express (11.23) in terms of the mean and variance of Y_i. This can be accomplished by noting that, from Exercise 11.8,

$$\mu_i = E[Y_i] = b'(\psi_i) \ , \tag{11.24}$$

and

$$\text{var}(Y_i) = b''(\psi_i)a(\phi) \ . \tag{11.25}$$

Thus,

$$\frac{\partial \psi_i}{\partial \beta_j} = \frac{\partial \psi_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} = \frac{1}{b''(\psi_i)}\frac{\partial \mu_i}{\partial \beta_j} = \frac{a(\phi_i)}{\text{var}(Y_i)}\frac{\partial \mu_i}{\partial \beta_j} \ .$$

Substituting into formula (11.23) gives

$$\frac{\partial l(\boldsymbol{\beta}, \phi; \mathbf{y})}{\partial \beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i)}{\text{var}(Y_i)}\frac{\partial \mu_i}{\partial \beta_j} \ . \tag{11.26}$$

The (j,k)'th element of $I(\boldsymbol{\beta})$ is

$$\begin{aligned}
[I(\boldsymbol{\beta})]_{jk} &= E_{\boldsymbol{\theta}}\left[\frac{\partial l(\boldsymbol{\beta},\phi;\mathbf{y})}{\partial \beta_j}\frac{\partial l(\boldsymbol{\beta},\phi;\mathbf{y})}{\partial \beta_k}\right] \\
&= \sum_{i=1}^{n}\frac{E_{\boldsymbol{\theta}}\left[(Y_i-\mu_i)^2\right]}{\mathrm{var}(Y_i)^2}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k} \\
&= \sum_{i=1}^{n}\frac{\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}}{\mathrm{var}(Y_i)}
\end{aligned} \tag{11.27}$$

In matrix notation,

$$I(\boldsymbol{\beta}) = \mathbf{G}^T V^{-1}\mathbf{G} \tag{11.28}$$

where V is the $n\times n$ diagonal matrix with diagonal entry i equal to $\mathrm{var}(Y_i)$, and G is the $n\times p$ Jacobian matrix with (i,j) element equal to $\partial\mu_i/\partial\beta_j$.

Differentiation of (11.23) with respect to ϕ results in a summation of terms in which y_i appears only through the multiplicative term $(y_i - b'(\psi_i))$ in the numerator. These have expectation of zero, and it follows that $I(\boldsymbol{\theta})$ has block diagonal form

$$I(\boldsymbol{\theta}) = \begin{pmatrix} I(\boldsymbol{\beta}) & \mathbf{0} \\ \mathbf{0} & [I(\boldsymbol{\theta}]_{\phi,\phi} \end{pmatrix}, \tag{11.29}$$

where $[I(\boldsymbol{\theta}]_{\phi,\phi}$ denotes Fisher information with respect to ϕ.

Example 11.6. Linear regression model. Let each $Y_i, i = 1,...,n$ be independently distributed as $\mathrm{N}(\boldsymbol{x}_i^T\boldsymbol{\beta},\sigma^2)$ where $\boldsymbol{x}_i$ is a length p vector of known covariates and $\boldsymbol{\beta}\in\mathbb{R}^p$ and σ^2 are to be estimated. The normal distribution is exponential family (Example 7.1) with parameters $\boldsymbol{\theta} = (\beta_1,...,\beta_p,\sigma^2)$, and hence from (11.28),

$$I(\boldsymbol{\beta}) = \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}$$

where $\mathbf{X}$ is the $n\times p$ design matrix having $\boldsymbol{x}_i^T$ as row i.

Analogously to the iid $N(\mu,\sigma^2)$ case in Example (11.4), $[I(\boldsymbol{\theta})]_{\sigma^2,\sigma^2} = n/2\sigma^4$. From (11.29), the Fisher information for $\boldsymbol{\theta}$ from observation of $Y_i, i = 1,...,n$ is the $(p+1)\times(p+1)$ matrix

$$I(\boldsymbol{\beta},\sigma^2) = \begin{pmatrix} \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{pmatrix}. \tag{11.30}$$

The likelihood equation for $\widehat{\boldsymbol{\beta}}$ can be obtained from (11.26). Using matrix notation, this is

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}(\mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\mathbf{X})\boldsymbol{\beta} = 0 \ . \tag{11.31}$$

Assuming that the design matrix is full rank, the least squares estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} \tag{11.32}$$

is the unique MLE solving this equation. This estimator has variance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ and therefore for any linear function $g(\boldsymbol{\beta})$ the unbiased estimator $g(\widehat{\boldsymbol{\beta}})$ attains the CR lower bound. □

Example 11.7. Nonlinear regression model. Let each $Y_i, i = 1, ..., n$ be independently distributed as $\mathrm{N}(\mu_i(\boldsymbol{\beta}), \sigma^2)$, where each μ_i is differentiable with respect to $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$. It follows immediately from (11.29) that

$$I(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{pmatrix} \ . \tag{11.33}$$

□

Example 11.8. Generalized linear model with canonical link function. In the notation of Section 7.1.2, the canonical link function $g(\mu)$ is the inverse of b'. Since $\mu = b'(\psi)$, if follows that

$$\psi = (b')^{-1}(\mu) = g(\mu) = \eta \ .$$

That is, ψ is the linear predictor. Therefore $\partial\psi_i/\partial\beta_j = x_{ij}$, which does not depend on $\boldsymbol{\beta}$. From (11.23), it follows that the Hessian matrix, $\partial^2 l/\partial\boldsymbol{\beta}^2$, does not depend on y_i. This establishes that the expected and observed Fisher information are identical when the canonical link is used in a GLM. □

The next example is one where $I(\boldsymbol{\theta})$ is not diagonal and the CR lower bound for estimation of one of the parameters is affected by whether the other parameter is known or not.

Information orthogonality

Example 11.9. Gamma. Let $Y_1, ..., Y_n$ be iid from a Gamma(α, β) distribution, i.e.,

$$f(y; \alpha, \beta) = \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} \quad , y > 0 \ .$$

Then (Exercise 11.7),

$$I_1(\alpha, \beta) = \begin{pmatrix} \Psi'(\alpha) & 1/\beta \\ 1/\beta & \alpha/\beta^2 \end{pmatrix} \tag{11.34}$$

where $\Psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ (the digamma function). Thus,

$$I_1^{-1}(\alpha, \beta) = \frac{\beta^2}{\alpha\Psi'(\alpha) - 1} \begin{pmatrix} \alpha/\beta^2 & -1/\beta \\ -1/\beta & \Psi'(\alpha) \end{pmatrix} .$$

and for a sample of size n the CR lower bounds for unbiased estimators of α and β are therefore $\alpha/n(\alpha\Psi'(\alpha) - 1)$ and $\beta^2\Psi'(\alpha)/n(\alpha\Psi'(\alpha) - 1)$ respectively. Since the CR lower bound is the inverse of information, it can be said that $n(\alpha\Psi'(\alpha) - 1)/\alpha$ is the information for α when β is unknown, and similarly, that $n(\alpha\Psi'(\alpha) - 1)/\beta^2\Psi'(\alpha)$ is the information for β when α is unknown.

Compare the above to the case where α (the shape parameter) is known. Then the CR lower bound for β is $\beta^2/n\alpha$. (In this case the MLE is $\widehat{\beta} = \overline{Y}/\alpha$ and it is unbiased with variance equal to the CR lower bound.) Note that $\beta^2\Psi'(\alpha)/n(\alpha\Psi'(\alpha) - 1) \geq \beta^2/n\alpha$. This is suggesting that β is harder to estimate when α is not known. □

In the above example, it was seen that the CR lower bound for estimation of one parameter was reduced when it was assumed that the value of the other parameter was known. It can be shown that this phenomenon holds more generally (Exercise 11.10).

11.7 Exercises

11.1 Prove (11.10) via slight modification to the proof of the CR inequality.

11.2 Let $Y_1, ..., Y_n$ be iid Poisson(λ).

1. Verify that $I(\lambda) = n/\lambda$.
2. Suppose that we are interested in estimating $g(\lambda) = P(Y = 0) = \exp(-\lambda)$. Verify that the CR lower bound for unbiased estimation of $g(\lambda)$ is $\lambda \exp(-2\lambda)/n$.

11.3 Using 11.1 or 11.13, calculate $I(\zeta)$ for a binomial experiment with n trials and log-odds ζ. That is, using the log-likelihood

$$l(\zeta; y) = y(\zeta - \log(1 + e^\zeta)) - (n - y)\log(1 + e^\zeta) \ .$$

11.4 Let Y be from a geometric distribution with density function

$$f(y) = p^y(1-p) \quad , y = 0, 1, 2, ...$$

where $0 < p < 1$. Show that $I_1(p) = \frac{1}{p(1-p)^2}$.
You may find it helpful to utilize the knowledge that $E[Y] = \frac{p}{1-p}$.

11.5 Let Y be from the distribution with density

$$f(y; \beta) = \frac{\beta}{(y+\beta)^2}, \ x \geq 0.$$

Show that $I_1(\beta) = 1/(3\beta^2)$.

11.6 Let Y have a logistic(θ) distribution with density

$$f(y, \theta) = \frac{\exp(y-\theta)}{(1+\exp(y-\theta))^2}, \quad y \in \mathbb{R}, \quad \theta \in \mathbb{R} \ .$$

Show that $I_1(\theta) = 1/3$.

11.7 Verify the Fisher information matrix, $I_1(\alpha, \beta)$, in (11.34) for Y distributed Gamma(α, β).

11.8 Let Y have an exponential family distribution $f(y; \psi, \phi)$ with density given by equation (11.22).

1. Using the identities $E_\theta[\frac{\partial l}{\partial \theta}] = 0$ and $E_\theta[\frac{\partial^2 l}{\partial \theta^2}] + E_\theta[(\frac{\partial l}{\partial \theta})^2] = 0$, show that

$$E[Y] = b^{'}(\psi) \quad \text{and} \quad \text{var}(Y) = b^{''}(\psi)a(\phi) \ .$$

2. The binomial proportion $Y \sim \text{Bin}(n, p)/n$ has exponential family distribution (Example 7.2). Use the above result to obtain the mean and variance of Y.

11.9 Repeat the information calculations in Example 11.5 using $[I(\boldsymbol{\beta})]_{jk} = -E[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}]$.

11.10 Suppose that $\boldsymbol{\theta} = (\psi, \lambda)$ where $\boldsymbol{\theta} \in \mathbb{R}^s$, $\psi \in \mathbb{R}^r$ and $\lambda \in \mathbb{R}^{s-r}$. Denote information matrices

$$I(\boldsymbol{\theta}) = I(\psi, \lambda) = \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix} \quad \text{and} \quad I^{-1}(\boldsymbol{\theta}) = I^{-1}(\psi, \lambda) = \begin{pmatrix} I^{\psi\psi} & I^{\psi\lambda} \\ I^{\lambda\psi} & I^{\lambda\lambda} \end{pmatrix} .$$

where $I_{\psi\psi}$ and $I^{\psi\psi}$ are $r \times r$ and $I_{\lambda\lambda}$ and $I^{\lambda\lambda}$ are $(s-r) \times (s-r)$. You may assume that $I(\boldsymbol{\theta})$ is positive definite.

a. By expanding $I^{-1}(\psi, \lambda)I(\psi, \lambda) = I_s$ (the $s \times s$ identity matrix), show that

$$I^{\psi\psi} = (I_{\psi\psi} - I_{\psi\lambda}I_{\lambda\lambda}^{-1}I_{\lambda\psi})^{-1}$$

b. Hence, show that for any vector $g^{'} \in \mathbb{R}^r$,

$$(g^{'})^T I^{\psi\psi} g^{'} \geq (g^{'})^T I_{\psi\psi}^{-1} g^{'} .$$

Hint: If A and B are $p \times p$ matrices that are symmetric and positive definite, and

$$\mathbf{h}^T A \mathbf{h} \geq \mathbf{h}^T B \mathbf{h} \quad \text{for all} \quad \mathbf{h} \in \mathbb{R}^p$$

then

$$\mathbf{h}^T A^{-1} \mathbf{h} \leq \mathbf{h}^T B^{-1} \mathbf{h} \quad \text{for all} \quad \mathbf{h} \in \mathbb{R}^p \ .$$

Ref???

Chapter 12

Asymptotic theory and approximate normality

12.1 Introduction

Section 12.2 establishes the central-limit theorem for maximum likelihood estimators as sample size, n, tends to infinity, for the case of iid data and $\theta \in \mathbb{R}$. This is obtained in three steps, via Lemmas 12.1, 12.2 and Theorem 12.1. In plain language, the first lemma shows that the likelihood will be greater when evaluated at the true unknown parameter value θ_0 then at any other value you choose to specify, for sufficiently large sample size. The second lemma uses this result to establish that MLEs are *consistent* estimators, at least in the case where the MLE is unique. Finally, Theorem 12.1 uses a Taylor series expansion of the derivative of the log-likelihood function (i.e., the score statistic) to obtain the asymptotic normality of a $\sqrt{n}$-standardized sequence of maximum likelihood estimators. The result is stated for multi-parameter models with $\boldsymbol{\theta} \in \mathbb{R}^s$ and non-iid data in sub-sections 12.2.2 and 12.2.5, respectively.

Section 12.3 provides a translation of the central-limit result into a more pragmatic interpretation based on approximate normality, and shows how this can be put to good use. Examples are used to caution that approximate normality of the MLE does not in any way guarantee that the MLE inherits the usual properties of the normal distribution. For example, although the MLE may be well approximated by a normal, it may not have an expected value (Example 12.9), or not exist with some small positive probability (Example 12.2).

Section 12.4 looks at the construction of asymptotically correct hypothesis tests

and confidence intervals/regions. These are asymptotically correct in the sense that they have approximately the desired level or coverage for sufficiently large sample size. Section 12.4.3 extends this to functions $g(\boldsymbol{\theta}) : \mathbb{R}^s \to \mathbb{R}^p$. Section 12.5 presents the use of likelihood ratio tests and confidence intervals, and Section 12.6 provides brief coverage of the lesser-used Rao-score statistic.

Throughout this chapter, n will be used to denote sample size. For example, T_n in Definition 12.1 denotes an estimator based on a sample of size n. Similarly, $\widehat{\theta}_n$ in Lemma 12.2 and Theorem 12.1 denotes the estimator obtained as a root of the likelihood equation from a size n sample.

12.2 Consistency and asymptotic normality

The central-limit theorem for maximum likelihood estimators requires the MLE to be *consistent*, that is, that the MLE converges to the true parameter value in probability. The formal definition of a consistent estimator is given below. In this definition and in Lemma 12.1, the parameter is assumed to be in $\mathbb{R}^s$. Restriction to the scalar-parameter case is required for Lemma 12.2 and Theorem 12.1.

Definition 12.1 Consistent estimator

A sequence of estimators $T_n \equiv T_n(Y_1, ..., Y_n)$ of $g(\boldsymbol{\theta}) \in \mathbb{R}$ is said to be consistent if for every $\boldsymbol{\theta}$ in Θ

$$T_n \to_{P_{\boldsymbol{\theta}}} g(\boldsymbol{\theta})$$

where $\to_{P_{\boldsymbol{\theta}}}$ denotes convergence in probability when $\boldsymbol{\theta}$ is the true parameter, i.e., for every $\epsilon > 0$, $P_{\boldsymbol{\theta}}(|T_n - g(\boldsymbol{\theta})| > \epsilon) \to 0$ as $n \to \infty$.

The following regularity conditions are required in the statement of Lemma 12.1.

Regularity conditions

R1: The observations $Y_i, i = 1, ..., n$ are iid from the distribution $P_{\boldsymbol{\theta}}$ having density function $f(y; \boldsymbol{\theta})$, within the parametric statistical model indexed by $\boldsymbol{\theta}$.

R2: The distributions $P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta$ are identifiable, that is, they are distinct.

R3: The distributions $P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta$ have common support, i.e., the set $\mathbf{A} = \{y : f(y;\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.

Lemma 12.1 $\boldsymbol{\theta}_0$ is best.

If the true value of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$, and regularity conditions R1-R3 hold, then

$$P_{\boldsymbol{\theta}_0}\left(\prod_{i=1}^{n} f(Y_i;\boldsymbol{\theta}_0) > \prod_{i=1}^{n} f(Y_i;\boldsymbol{\theta})\right) \to 1 \quad \text{as} \quad n \to \infty$$

for any fixed $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

That is, for large sample size, the likelihood of the true parameter value $\boldsymbol{\theta}_0$ is greater than that of any other fixed $\boldsymbol{\theta}$ (with high probability).

Proof. Note that the conclusion of this lemma is equivalent to

$$P_{\boldsymbol{\theta}_0}\left(\sum_{i=1}^{n} \log\left(\frac{f(Y_i;\boldsymbol{\theta})}{f(Y_i;\boldsymbol{\theta}_0)}\right) < 0\right) \to 1$$

or

$$P_{\boldsymbol{\theta}_0}\left(\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(Y_i;\boldsymbol{\theta})}{f(Y_i;\boldsymbol{\theta}_0)}\right) < 0\right) \to 1\ . \tag{12.1}$$

Now, by the weak law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(Y_i;\boldsymbol{\theta})}{f(Y_i;\boldsymbol{\theta}_0)}\right) \to_p E_{\boldsymbol{\theta}_0}\left[\log\left(\frac{f(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta}_0)}\right)\right]$$

and by Jensen's inequality applied to the convex function -log, we have

$$E_{\boldsymbol{\theta}_0}\left[\log\left(\frac{f(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta}_0)}\right)\right] < \log E_{\boldsymbol{\theta}_0}\left[\frac{f(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta}_0)}\right] = \log(1) = 0.$$

Thus, we have shown that $\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(Y_i;\boldsymbol{\theta})}{f(Y_i;\boldsymbol{\theta}_0)}\right)$ converges in probability to a negative quantity, from which the result follows (see Exercise 13.3). □

The next lemma will assume the following condition:

R4: The parameter space $\Theta \subset \mathbb{R}$ and $\theta_0 \in \Omega$ where Ω is an open subset of Θ.

Lemma 12.2 Consistency of $\widehat{\theta}_n$

Assume R1 - R4, and that $f(y;\theta)$ is differentiable with respect to θ in Ω, with derivative $f'(y;\theta)$. Then with probability tending to 1 as $n \to \infty$, the likelihood equation

$$\frac{\partial L(\theta; y_1, ..., y_n)}{\partial \theta} \equiv \frac{\partial L(\theta; \mathbf{y})}{\partial \theta} \equiv \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(y_i; \theta) = 0$$

or equivalently, the equation

$$\frac{\partial \log L(\theta; \mathbf{y})}{\partial \theta} \equiv \frac{\partial l(\theta; \mathbf{y})}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log f(y_i; \theta) = \sum_{i=1}^{n} \frac{f'(y_i; \theta)}{f(y_i; \theta)} = 0$$

has a root $\widehat{\theta}_n \equiv \widehat{\theta}_n(\mathbf{y})$ such that $\widehat{\theta}_n(\mathbf{Y}) \equiv \widehat{\theta}_n(Y_1, ..., Y_n)$ tends to the true value θ_0 in probability.

Proof. Since θ_0 lies within the open subset $\Omega \subset \Theta$, the interval $(\theta_0 - a, \theta_0 + a)$ is entirely contained within Θ for sufficiently small $a > 0$. By virtue of the differentiability of the likelihood, it is necessarily the case that the likelihood has a turning point (i.e., a root of the likelihood equation) in the interval $(\theta_0 - a, \theta_0 + a)$ if the likelihood is higher at θ_0 then at both $\theta_0 - a$ and $\theta_0 + a$. By the previous lemma, the probability of this tends to one because it is the intersection of the event $\{L(\theta_0; \boldsymbol{Y}) > L(\theta_0 - a; \boldsymbol{Y})\}$ and the event $\{L(\theta_0; \boldsymbol{Y}) > L(\theta_0 + a; \boldsymbol{Y})\}$, both of which have probability tending to one as $n \to \infty$ (see Exercise 12.1)

If the likelihood equation has multiple roots in $(\theta_0 - a, \theta_0 + a)$ then take $\widehat{\theta}_n$ to be the root closest to θ_0 (this removes the dependence of $\widehat{\theta}_n$ on a.) Hence, we have shown that for arbitrarily small a, with probability $\to 1$ there exists a root $\widehat{\theta}_n$ such that $|\widehat{\theta}_n - \theta_0| < a$. That is, $P(|\widehat{\theta}_n - \theta_0| > a) \to 0$, which establishes that $\widehat{\theta}_n \underset{p}{\to} \theta_0$. □

Lemma 12.2 comes very close to establishing the existence of a consistent sequence of roots $\widehat{\theta}_n$ of the likelihood equation, but stops just short of this. Note that a root of the likelihood equation need not exist for every possible outcome $\boldsymbol{y}$. However, a root does exist with probability tending to unity, and so a consistent sequence of estimators $\widehat{\theta}_n^*$ can be obtained as

$$\widehat{\theta}_n^* = \begin{cases} \widehat{\theta}_n & \text{if a root exists} \\ \theta_a & \text{otherwise} \end{cases} \tag{12.2}$$

where θ_a is any arbitrary value in the parameter space Θ.

As in the proof of Lemma 12.2, if the likelihood equation has multiple roots then $\widehat{\theta}_n$ in (12.2) is taken to be the root closest to θ_0. If the likelihood equation has unique roots then (12.2) provides a description to formally construct a consistent sequence of estimators of θ_0. If the likelihood equation can have multiple roots, but with the root being unique with probability tending to unity as $n \to \infty$, then consistency of $\widehat{\theta}_n^*$ will be assured for any choice of root of the likelihood equation.

The next two examples demonstrate the strange behaviour that the likelihood may exhibit on subsets of the sample space having zero probability or probability tending to zero.

Example 12.1. No root of iid $N(0, \sigma^2)$ likelihood equation. Let $Y_1, ..., Y_n$ be iid $N(0, \sigma^2)$ with parameter space given by the open interval defined by $\sigma > 0$. If all y_i are identically zero then there is no root of the likelihood equation in Θ and the likelihood is unbounded as σ is made arbitrarily small. This mis-behaviour of the likelihood iid is vacuous in practice, because the outcome $y_i = 0, i = 1, ..., n$ is a zero-probability event. □

Example 12.2. No root of binomial likelihood equation for the log-odds. Let $B_1, ..., B_n$ be iid Bernoulli with $P(B_i = 1) = p,\ 0 < p < 1$, with the model parameterized by the log-odds $\zeta = \text{logit}(p) = \log(p/(1-p)) \in \Theta = \mathbb{R}$. Letting $Y_n = \sum B_i$ denote the binomially distributed sum of the n Bernoulli observations, the likelihood for ζ given the observation of y_n is

$$\begin{aligned} L(\zeta; y_n) &\propto p^{y_n}(1-p)^{n-y_n} \\ &= \frac{e^{\zeta y_n}}{(1+e^{\zeta})^n} . \end{aligned}$$

If y_n is other than 0 or n then the likelihood has a root at $\widehat{\zeta}_n = \text{logit}(\widehat{p}_n)$ where $\widehat{p}_n = y_n/n$. However, if y_n is 0 or n then this likelihood does not possess a root in $\mathbb{R}$. For example, if $y_n = 0$ then $L(\zeta; y_n) = (1+e^{\zeta})^{-n}$ is monotone decreasing on $\mathbb{R}$. Note that $P(1 \le Y_n \le n)$ has probability tending to one as $n \to \infty$. From the weak

law of large numbers, Y_n/n converges in probability to the true unknown p_0. With probability tending to 1, the likelihood equation has a root $\widehat{\zeta}_n = \text{logit}(Y_n/n)$, and (from property P3 in Section 13.3) this sequence of roots converges in probability to the true value ζ_0. □

12.2.1 Asymptotic normality: $\theta \in \mathbb{R}$

Establishing the asymptotic normality of maximum likelihood estimators requires further regularity conditions on the statistical model. The last of these conditions, R7, places a very specific bounding condition on the third derivative of the log-likelihood. The proof of Theorem 12.1 employs a Taylor-series expansion of the score statistic in (12.4), and this condition is sufficient to ensure that the remainder term in this expansion can be ignored.

R5: The integral $\int f(y;\theta)dy$ can be twice differentiated (w.r.to θ) under the integral sign.

R6: $0 < I(\theta) < \infty$.

R7: The third derivative of f (w.r.to θ) exists and

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(y;\theta)\right| \leq M(y)$$

for all $y \in \mathbf{A}$ (the support set) and $\theta_0 - c < \theta < \theta_0 + c$ where $E_{\theta_0}[M(Y)] < \infty$.

Theorem 12.1 Asymptotic normality and efficiency of $\widehat{\theta}_n \in \mathbb{R}$

If regularity conditions A0-A6 hold, then any consistent sequence $\widehat{\theta}_n \equiv \widehat{\theta}_n(Y_1, ..., Y_n)$ of roots of the likelihood equation satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightarrow_D N\left(0, \frac{1}{I_1(\theta_0)}\right). \tag{12.3}$$

where $I_1(\theta_0)$ is the expected Fisher information about θ_0 from one observation.

Proof.

Expand the score function $l'(\theta) = \frac{\partial l(\theta;\boldsymbol{Y})}{\partial\theta}$ about θ_0 using the mean-value form of the Taylor-series expansion, and evaluate at $\widehat{\theta}_n$. This gives Ref?

$$\begin{aligned} 0 &= l'(\widehat{\theta}_n) \\ &= l'(\theta_0) + (\widehat{\theta}_n - \theta_0) l''(\theta_0) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^2 l'''(\theta_n^*) \end{aligned} \tag{12.4}$$

where θ_n^* lies between θ_0 and $\widehat{\theta}_n$. Rearranging gives

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{(1/\sqrt{n}) l'(\theta_0)}{-(1/n) l''(\theta_0) - (1/2n)(\widehat{\theta}_n - \theta_0) l'''(\theta_n^*)}. \tag{12.5}$$

The required result is obtained after showing that

$$\frac{1}{\sqrt{n}} l'(\theta_0) \to_D N(0, I_1(\theta_0)) \ , \tag{12.6}$$

and

$$-\frac{1}{n} l''(\theta_0) \to_p I_1(\theta_0) \ , \tag{12.7}$$

and

$$\frac{1}{n}(\widehat{\theta}_n - \theta_0) l'''(\theta_n^*) \to_p 0 \ . \tag{12.8}$$

From (12.7) and (12.8), application of Slutsky's theorem establishes that the denominator in (12.5) converges in probability to $I_1(\theta_0)$. That is,

$$-(1/n) l''(\theta_0) - (1/2n)(\widehat{\theta}_n - \theta_0) l'''(\theta_n^*) \to_p I_1(\theta_0) \ .$$

Thus, (12.5) can be expressed as A_n/B_n where A_n converges in distribution to $N(0, I_1(\theta_0))$ and $1/B_n$ converges in probability to $1/I_1(\theta_0)$. A second application of Slutsky's theorem then gives

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_D \frac{1}{I_1(\theta_0)} N(0, I_1(\theta_0)) = N\left(0, \frac{1}{I_1(\theta_0)}\right)$$

as required.

The following arguments establish (12.6)–(12.8).

- The left-hand side of (**??**) can be written

$$\frac{1}{\sqrt{n}} l'(\theta_0) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} l'(\theta_0, Y_i)\right)$$

  where the terms in the summation are iid with mean 0 (from Lemma 11.1). It follows from the central limit theorem that

$$\begin{aligned} \frac{1}{\sqrt{n}} l'(\theta_0) &\to_D N\left(0, \mathrm{var}(l'(\theta_0, Y))\right) \\ &= N(0, I_1(\theta_0)) \ . \end{aligned} \tag{12.9}$$

- The left-hand side of (12.7) is

$$-\frac{1}{n}l''(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n} l''(\theta_0, Y_i)$$

where each term in the summation is iid. It follows from the weak law of large numbers that

$$\begin{aligned} -\frac{1}{n}l''(\theta_0) &\to_p -E_{\theta_0}[l''(\theta_0, Y)] \\ &= I_1(\theta_0) \ , \end{aligned} \tag{12.10}$$

from Lemma 11.1.

- Finally, to show (12.8) note that $\theta_0 - c < \theta_n^* < \theta_0 + c$ with probability tending to 1 and hence from condition A6

$$|\frac{1}{n}l'''(\theta_n^*)| = |\frac{1}{n}\sum_{i=1}^{n} l'''(\theta_n^*, Y_i)| \leq \frac{1}{n}\sum_{i=1}^{n} M(Y_i)$$

with probability tending to 1 also. By the weak law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} M(Y_i) \to_p E_{\theta_0}[M(Y)] < \infty$$

and hence $l'''(\theta_n^*)/n$ is bounded in probability. That is, there exists $B < \infty$ (take B to be any value greater than $E_{\theta_0}[M(Y)]$) such that

$$P(|\frac{1}{n}l'''(\theta_n^*)| > B) \to 0 \quad \text{as} \quad n \to \infty \ . \tag{12.11}$$

Since $\widehat{\theta}_n - \theta_0 \to_p 0$, equation (12.8) follows (see Exercise 13.4). □

Example 12.3. Asymptotic distribution of the binomial proportion. Let $B_1, ..., B_n$ be iid Bernoulli with $P(B_i = 1) = p$, $0 < p < 1$, and let $\widehat{p}_n = \frac{1}{n}\sum B_i$ denote the ML estimator of p. The expected Fisher information from a single Bernoulli observation is $I_1(p) = 1/(p(1-p))$ (Example 11.1), and so, with p_0 denoting the true unknown value of p,

$$\sqrt{n}(\widehat{p}_n - p_0) \to_D N\left(0, \frac{1}{I_1(p_0)}\right) = N(0, p_0(1-p_0)) \ .$$

Example 12.2 ctd. Binomial model indexed by log-odds. From Example 11.2, $I_1(\zeta_0) = p_0(1-p_0)$ is the expected Fisher information about $\zeta_0 = \text{logit}(p_0)$ from a single Bernoulli observation. From Theorem 12.1,

$$\sqrt{n}(\widehat{\zeta}_n - \zeta_0) \to_D N\left(0, \frac{1}{p_0(1-p_0)}\right) . \tag{12.12}$$

where $\widehat{\zeta}_n = \text{logit}(\widehat{p}_n)$ is the MLE, provided that $\widehat{p}_n$ is not equal to 0 or 1. For $\widehat{\zeta}_n$ to be a properly defined estimator it is enough to set $\widehat{\zeta}_n$ to any arbitrary real number when $\widehat{p}_n \in \{0, 1\}$. These events have probability tending to zero as $n \to \infty$ (since $0 < p_0 < 1$), and so the estimator $\widehat{\zeta}_n$ so defined is a consistent estimator of ζ_0 and (12.12) holds.

12.2.2 Asymptotic normality: $\boldsymbol{\theta} \in \mathbb{R}^s$

Extending the consistency and asymptotic normality results to the multiparameter case, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$ is reasonably straightforward (but extremely tedious) and so only the results are stated below.

When $\boldsymbol{\theta} = (\theta_1, ..., \theta_s) \in \Theta \subset \mathbb{R}^s$, recall (Section 11.6) that the $s \times s$ expected Fisher information matrix $I(\boldsymbol{\theta})$ has i, j^{th} element

$$I(\boldsymbol{\theta})_{ij} = E_{\boldsymbol{\theta}}\left[\frac{\partial l(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \theta_i}\frac{\partial l(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \theta_j}\right] = Cov_{\boldsymbol{\theta}}\left(\frac{\partial l(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \theta_i}, \frac{\partial l(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \theta_j}\right) \tag{12.13}$$

and, provided the likelihood can be twice differentiated under the integral sign, it can also be obtained as

$$I(\boldsymbol{\theta})_{ij} = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \theta_i \partial \theta_j}\right] . \tag{12.14}$$

Theorem 12.2 Asymptotic normality and efficiency of $\widehat{\boldsymbol{\theta}}_n \in \mathbb{R}^s$ *Under appropriate conditions (similar to those of the one-parameter case), any consistent sequence $\widehat{\boldsymbol{\theta}}_n \equiv \widehat{\boldsymbol{\theta}}_n(Y_1, ..., Y_n)$ of roots of the likelihood equations satisfies*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_D N_s\left(\mathbf{0}, I_1(\boldsymbol{\theta}_0)^{-1}\right) , \tag{12.15}$$

where $I_1(\boldsymbol{\theta})$ is the expected Fisher information (matrix) about $\boldsymbol{\theta}_0$ from one observation.

Example 12.4. The expected Fisher information matrix about (μ, σ^2) from a single observation $Y \sim N(\mu, \sigma^2)$ was obtained in Example 11.4. Letting $(\overline{Y}_n, \widehat{\sigma}_n^2)$ denote the ML estimator from an iid sample of size n distributed $N(\mu_0, \sigma_0^2)$, it follows that

$$\sqrt{n}\left(\begin{pmatrix} \overline{Y}_n \\ \widehat{\sigma}_n^2 \end{pmatrix} - \begin{pmatrix} \mu_0 \\ \sigma_0^2 \end{pmatrix}\right) \to_D N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & 2\sigma_0^4 \end{pmatrix}\right).$$

For this example, the exact distributions of $\overline{Y}_n$ and $\widehat{\sigma}_n^2$ are known to be $\overline{Y}_n \sim N(\mu_0, \sigma_0^2/n)$ and $\widehat{\sigma}_n^2 \sim \sigma_0^2 \chi_{n-1}^2/n$. That is,

$$\begin{aligned} \sqrt{n}(\overline{Y}_n - \mu_0) &\sim N(0, \sigma_0^2) \\ \sqrt{n}(\widehat{\sigma}_n^2 - \sigma_0^2) &\sim \sqrt{n}\sigma_0^2 \left[\frac{\chi_{n-1}^2}{n} - 1\right] . \end{aligned} \tag{12.16}$$

Using (12.16), the convergence in distribution of $\sqrt{n}(\widehat{\sigma}_n^2 - \sigma_0^2)$ to $N(0, 2\sigma_0^4)$ can be established from first principles (Exercise 13.7). □

12.2.3 Asymptotic normality under model mis-specification

Suppose that the statistical model is mis-specified by working with the density function $\widetilde{f}(\boldsymbol{y}; \boldsymbol{\theta})$ rather than the correct density function $f(\boldsymbol{y}; \boldsymbol{\theta})$, Define $\widetilde{\boldsymbol{\theta}}_0 \in \Theta$ to be the parameter value such that the difference between density functions $\{\widetilde{f}(\boldsymbol{y}; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ and $f(\boldsymbol{y}; \boldsymbol{\theta}_0)$ is minimized.[1] That is

$$\Delta(\widetilde{f}(\boldsymbol{y}; \widetilde{\boldsymbol{\theta}}_0), f(\boldsymbol{y}; \boldsymbol{\theta}_0)) = \min_{\boldsymbol{\theta} \in \Theta} \Delta((\widetilde{f}(\boldsymbol{y}; \boldsymbol{\theta}), f(\boldsymbol{y}; \boldsymbol{\theta}_0)) ,$$

where it is assumed that $\widetilde{\boldsymbol{\theta}}_0$ exists and is unique. In some cases it may be that $\widetilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$ but this is not so in general.

Given observations $\boldsymbol{y}$, the likelihood function being maximized, $\widetilde{l}(\boldsymbol{\theta}) = \log \widetilde{f}(\boldsymbol{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$, is not the true likelihood function. However, by virtue of the definition of $\widetilde{\boldsymbol{\theta}}_0$, and subject to appropriate regularity conditions, it follows (see Pawitan 2001, pp. 370-374 for details) that

$$E_{\boldsymbol{\theta}_0}\left[\left.\frac{\partial \widetilde{l}(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}_0}\right] = \mathbf{0} \tag{12.17}$$

[1] The difference is quantified using the Kullback-Leibler divergence (Kullback 1959) $\Delta(f_1, f_2) = \int f_1(\boldsymbol{y}) \log\left(\frac{f_1(\boldsymbol{y})}{f_2(\boldsymbol{y})}\right) d\boldsymbol{y}$.

where $E_{\boldsymbol{\theta}_0}$ continues to denote expectation under the true model. Letting $\widetilde{\boldsymbol{\theta}}_n$ denote the MLE under maximization of $\widetilde{l}$, it can be shown that $\widetilde{\boldsymbol{\theta}}_n$ is a consistent estimator of $\widetilde{\boldsymbol{\theta}}_0$ (Pawitan 2001), and asymptotically normal.

For $\theta \in \mathbb{R}$, the asymptotic normality of $\widetilde{\boldsymbol{\theta}}_n$ can be established by repeating the steps in the proof of Theorem 12.1, but now using a Taylor series expansion of $\widetilde{l}'(\widetilde{\theta}_n)$ around $\widetilde{\theta}_0$. It is not the case that $\text{var}(\widetilde{l}'(\widetilde{\theta}_0, Y))$ and $-E_{\theta_0}[\widetilde{l}''(\widetilde{\theta}_0, Y)]$ are necessarily equal, and so using (12.9) and (12.10) in place of (12.6) and (12.7) gives

$$\sqrt{n}(\widetilde{\theta}_n - \widetilde{\theta}_0) \to_D N\left(0, \frac{\text{var}(\widetilde{l}'(\widetilde{\theta}_0, Y))}{E_{\theta_0}[\widetilde{l}''(\widetilde{\theta}_0, Y)]^2}\right) . \tag{12.18}$$

In the multi-parameter case, $\boldsymbol{\theta} \in \mathbb{R}^s$, the convergence result becomes

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_0) \to_D N_s\left(\mathbf{0}, A^{-1}BA^{-1}\right) . \tag{12.19}$$

where A and B are the $s \times s$ matrices

$$A = -E_{\boldsymbol{\theta}_0}[\widetilde{l}''(\widetilde{\boldsymbol{\theta}}_0, Y)] , \tag{12.20}$$

and

$$B = \text{var}(\widetilde{l}'(\widetilde{\boldsymbol{\theta}}_0, Y)) . \tag{12.21}$$

12.2.4 Asymptotic normality of M-estimators

The previous section introduced the notion that asymptotically normal estimators can be obtained when the model is incorrectly specified. This can be used to our advantage, especially in situations where there is difficulty in specifying the correct likelihood. Useful estimators can be obtained by solving equations that possess similar properties to a likelihood equation. The name *M-estimator* alludes to the fact that these estimators are similar in construction to maximum likelihood estimators.

Specifically, suppose $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$, and let $U_1(\boldsymbol{\theta}; y) : \mathbb{R}^s \to \mathbb{R}^s$ be such that it satisfies the zero-mean property

$$E_{\boldsymbol{\theta}_0}[U_1(\boldsymbol{\theta}; Y)] = \mathbf{0}. \tag{12.22}$$

For an iid sample $\boldsymbol{y} = y_i, i = 1, ..., n$, an estimating function can be defined as

$$U_n(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} U_1(\boldsymbol{\theta}; y_i) .$$

Then, subject to appropriate regularity conditions (e.g., see Pawitan 2001, p. 405), for sufficiently large n, a solution $\widetilde{\boldsymbol{\theta}}_n$ to the estimating equation

$$U_n(\boldsymbol{\theta}) = \mathbf{0} \tag{12.23}$$

exists such that

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightarrow_D N_s\left(\mathbf{0}, \mathbf{A}_1^{-1}\mathbf{B}_1\mathbf{A}_1^{-T}\right) \ . \tag{12.24}$$

where $\mathbf{A}_1^{-T}$ denotes the inverse of $\mathbf{A}_1^T$, and $\mathbf{A}_1$ and $\mathbf{B}_1$ are the $s \times s$ matrices

$$\mathbf{A}_1 = -E_{\boldsymbol{\theta}_0}[U_1'(\boldsymbol{\theta}_0, Y)] \ , \tag{12.25}$$

and

$$\mathbf{B}_1 = \mathrm{var}(U_1(\boldsymbol{\theta}_0, Y)) \ . \tag{12.26}$$

Example 12.5. Suppose $\theta \in \mathbb{R}$ and y_i are iid with mean μ and variance σ^2. It is assumed that μ is a differentiable and invertible function of θ. The estimating function

$$U_1(\theta; y) = (y - \mu)$$

satisfies the zero-mean property, and the estimating equation is

$$U_n(\theta, \boldsymbol{y}) = \sum_{i=1}^{n}(y_i - \mu) = 0 \ .$$

The estimating equation is satisfied when $\mu = \overline{y}_n$, and hence the M-estimator of θ is the value $\widetilde{\theta}_n$ such that $\mu(\widetilde{\theta}_n) = \overline{y}_n$.

Now, $\mathbf{A}_1$ and $\mathbf{B}_1$ are scalar quantities given by

$$\mathbf{A}_1 = -\frac{\partial \mu}{\partial \theta}, \quad \text{and} \quad \mathbf{B}_1 = \sigma^2 \ ,$$

and so it follows from (12.24) that

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \rightarrow_D N\left(0, \left(\frac{\partial \theta}{\partial \mu}\right)^2 \sigma^2\right) \ .$$

□

Heyde (1997) gives a rigorous presentation of M-estimators and their properties, and considers the optimality of classes of estimators. In particular, it can be shown

that the log-likelihood is the optimal M-estimator (Theorem 12.3). For convenience, the proof of this theorem assumes that the estimating function is of standardized form, whereby $\mathbf{A}_1$ and $\mathbf{B}_1$ are equal. This can be assumed without loss of generality (Exercise 12.3). In standardized form, the asymptotic variance in (12.24) reduces to the inverse of $\text{var}(U_1)$.

Example 12.5 ctd. The standardized form of U_1 is

$$U_1^{(s)} = \frac{\partial \mu}{\partial \theta} \frac{(y-\mu)}{\sigma^2} \ ,$$

which has variance

$$\text{var}\left(U_1^{(s)}\right) = \left(\frac{\partial \mu}{\partial \theta}\right)^2 \frac{1}{\sigma^2} \ .$$

□

Theorem 12.3 *Suppose $\theta \in \mathbb{R}$ and y_i are iid, and let $U_1(\theta; \mathbf{y})$ be a zero-mean estimating function. Then, the asymptotic variance of $\sqrt{n}(\widetilde{\theta}_n - \theta_0)$ is minimized when $U_1(\theta; y_i)$ is the score function $\frac{\partial l(\theta; y_i)}{\partial \theta}$.*

Proof: Without loss of generality it can be assumed that U_1 is in standardized form. Then, it is required to show that $\text{var}(U_1) \leq I_1(\theta) = \text{var}\left(\frac{\partial l(\theta; Y_i)}{\partial \theta}\right)$. Since

$$E[U_1] = \int U_1(\theta; y) f(y; \theta) dy = 0 \quad \text{for all } \theta \in \Theta \ ,$$

then, assuming differentiation under the integral sign is valid, it follows that

$$\frac{\partial}{\partial \theta} \int U_1 f dy = \int \frac{\partial U_1}{\partial \theta} f dy + \int U_1 \frac{\partial f}{\partial \theta} dy = 0 \ .$$

That is,

$$\begin{aligned} \text{var}(U_1) = -E\left[\frac{\partial U_1}{\partial \theta}\right] &= \int U_1 \frac{\partial f}{\partial \theta} dy \\ &= \int U_1 \frac{\partial l}{\partial \theta} f dy \\ &= \text{cov}\left(U_1, \frac{\partial l}{\partial \theta}\right) \leq \sqrt{\text{var}(U_1) \text{var}\left(\frac{\partial l}{\partial \theta}\right)} \end{aligned}$$

from which the required result is immediate. □

12.2.5 The non-iid case

The asymptotic normality results can be generalized to cases where the data are not iid. For example, in the case of independent but not identically distributed data, asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$ will be maintained provided that the log-likelihood $l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l(\boldsymbol{\theta}; y_i)$ is not overly influenced by individual $l(\boldsymbol{\theta}; y_i)$ terms. More formally, the expected Fisher information from observing $\boldsymbol{Y}$ is

$$I(\boldsymbol{\theta}) = \sum_{i=1}^{n} I_i(\boldsymbol{\theta})$$

where $I_i(\boldsymbol{\theta})$ is the information from observation of Y_i. Under the condition that the averaged Fisher information matrix converges element-wise, i.e.,

$$\overline{I}_n(\boldsymbol{\theta}) = \frac{I(\boldsymbol{\theta})}{n} \to \overline{I}(\boldsymbol{\theta}) \tag{12.27}$$

for some positive definite matrix $\overline{I}(\boldsymbol{\theta})$, it can be shown (Hoadley 1971) that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_D N_s\left(\mathbf{0}, \overline{I}(\boldsymbol{\theta}_0)^{-1}\right). \tag{12.28}$$

Note that in the iid case, $I_1(\boldsymbol{\theta}_0) = \overline{I}(\boldsymbol{\theta}_0)$.

12.3 Approximate normality

The convergence result of (12.28) can be written in normalized form by multiplying both sides by $\overline{I}^{\frac{1}{2}}(\boldsymbol{\theta}_0)$, which is the square-root matrix of the positive definite matrix $\overline{I}(\boldsymbol{\theta}_0)$ in the multi-parameter case $\boldsymbol{\theta} \in \mathbb{R}^s$. That is,

$$\sqrt{n}\overline{I}^{\frac{1}{2}}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_D N_s(\mathbf{0}, \mathbf{I_s}) \ ,$$

where $\mathbf{I_s}$ is the s-dimensional identity matrix. From (12.27) it follows that

$$I(\boldsymbol{\theta}_0)^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_D N_s(\mathbf{0}, \mathbf{I_s}) \ . \tag{12.29}$$

That is, for n sufficiently large,

$$\sqrt{I(\boldsymbol{\theta}_0)}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \dot{\sim} N_s(\mathbf{0}, \mathbf{I_s}) \ , \tag{12.30}$$

where $\sim$ denotes "approximately distributed". Equivalently,

$$\widehat{\boldsymbol{\theta}}_n \sim N_s\left(\boldsymbol{\theta}_0, I(\boldsymbol{\theta}_0)^{-1}\right) \ . \tag{12.31}$$

Equation **??** is the pragmatic statement of the asymptotic normality of MLE's. Specifically, that the distribution of $\widehat{\boldsymbol{\theta}}_n$ (under repetition of the experiment) is approximately equal to that of a multivariate normal with mean $\boldsymbol{\theta}_0$ and variance matrix given by the inverse of the expected Fisher information $I(\boldsymbol{\theta}_0)$.

Example 12.4 ctd. From (**??**), it follows that for $Y_i \sim N(\mu_0, \sigma_0^2), i = 1, ..., n$, and n sufficiently large

$$\begin{pmatrix} \overline{Y}_n \\ \widehat{\sigma}_n^2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_0 \\ \sigma_0^2 \end{pmatrix}, \frac{1}{n}\begin{pmatrix} \sigma_0^2 & 0 \\ 0 & 2\sigma_0^4 \end{pmatrix}\right).$$

□

Loosely speaking, it can be said that, for sufficiently large n, $\widehat{\boldsymbol{\theta}}_n$ is approximately an unbiased estimator of $\boldsymbol{\theta}_0$ with variance matrix $I(\boldsymbol{\theta}_0)^{-1}$. By virtue of the delta theorem, any differentiable function $g(\boldsymbol{\theta}) : \Theta \to \mathbb{R}$ is also an approximately unbiased estimator of $g(\boldsymbol{\theta}_0)$, with approximate variance equal to the Cramér-Rao lower bound in (11.17). Consequently maximum likelihood estimators are said to be asymptotically efficient.

Remember that we have to be very careful to remember that "approximately" is in the sense of having a distribution that is approximately normal. There is no absolute guarantee that the expected value or variance of $\widehat{\theta}_n$ even exist (Example 12.9), or that a root of the likelihood exists (Example 12.2) So, for example, saying that $\widehat{\theta}_n$ is approximately unbiased has to be recognized as a statement about its

limiting distribution.

Box 12.1.

In practice, ML estimators typically will have finite expected values and variance if the model is sensibly parameterized, notwithstanding the possible need to correct for possible ill-behaviour that may occur with probability tending to zero as sample size increases (e.g., equation 12.2). They typically have order $1/n$ bias, denoted $O(n^{-1})$. That is $b(\widehat{\theta}_n) = |E[\widehat{\theta}_n] - \theta_0| \leq \frac{M}{n}$ for some finite M. From (12.31), the variance is also order $O(n^{-1})$. The mean-squared error of $\widehat{\theta}_n$ is

$$MSE = b(\widehat{\theta}_n)^2 + \mathrm{Var}(\widehat{\theta}_n) = O(n^{-2}) + O(n^{-1}) \ .$$

This means that the bias of MLE's is dominated by their variance for large n.

12.3.1 Estimation of approximate variance

In practice, calculation of expected Fisher information may be difficult or intractable and a more convenient approach is required. Since the expected Fisher information is the negative of the expected Hessian matrix of the log-likelihood (equation 12.14), it is natural to approximate it using the negative of the Hessian evaluated using the observed data $\boldsymbol{y}$. This approximation is called the observed Fisher information and will be denoted by $O(\boldsymbol{\theta})$, having i, j element

$$O_{ij}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial\theta_i \partial\theta_j} \ . \tag{12.32}$$

An alternative approximation can be based on the representation of expected Fisher information as the variance of the score function (see Exercise 12.9, but this is not widely used.

In general, it is usually most convenient to evaluate the observed Fisher information at $\widehat{\boldsymbol{\theta}}_n$, but it could also be evaluated at a hypothesized value $\boldsymbol{\theta}_0$ specified under a null hypothesis. Then, the approximate normality of the MLE can be expressed as

$$\widehat{\boldsymbol{\theta}}_n \sim N_s\left(\boldsymbol{\theta}_0, O(\boldsymbol{\theta}_0)^{-1}\right) \ . \tag{12.33}$$

or

$$\widehat{\boldsymbol{\theta}}_n \sim N_s\left(\boldsymbol{\theta}_0, O(\widehat{\boldsymbol{\theta}}_n)^{-1}\right) \ . \tag{12.34}$$

Not only is the approximation in (12.34) convenient, it has been shown that this form of the normal approximation is generally more accurate than the alternative which uses expected Fisher information (Efron and Hinkley 1978).

The approximation in (12.34) is by far the most widely used, and is the form used throughout Part 2 of this text, with the exception of Section 8.2. There, an estimate of the asymptotic variance in (12.19) is employed. In particular, this requires estimation of matrix B (in equation 12.21) using the formula given in Box **??**. The estimate of variance is commonly called the sandwich estimator of variance because matrix B is sandwiched between matrix A^{-1} on both the left and right. The sandwich estimate of variance is not used routinely for inference, because it can have considerably greater sampling variability than the observed Fisher information, thereby leading to confidence intervals with inferior coverage probability (Kauermann and Carroll 2001).

12.3.2 Approximate normality of M-estimators

For fixed n, let $\boldsymbol{y}$ be observed from the distribution with density $f(\boldsymbol{y}; \boldsymbol{\theta}_0)$, and let $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^s$ denote the M-estimator obtained as the solution to the s-dimensional system of equations

$$U(\boldsymbol{\theta}; \boldsymbol{y}) = \mathbf{0} \ .$$

The estimating function U is assumed to satisfy the zero-mean property

$$E_{\boldsymbol{\theta}}[U(\boldsymbol{\theta};\, \boldsymbol{Y})] = \mathbf{0} \quad \text{for all} \quad \boldsymbol{\theta} \in \Theta \ . \tag{12.35}$$

For sufficiently large n, and subject to appropriate regularity conditions, the finite sample version of the asymptotic normality result in Section 12.2.4 is

$$\widetilde{\boldsymbol{\theta}} \sim N_s(\boldsymbol{\theta}_0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-T}) \ , \tag{12.36}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the $s \times s$ matrices

$$\mathbf{A} = -E_{\boldsymbol{\theta}_0}[U^{'}(\boldsymbol{\theta}_0,\, \boldsymbol{Y})] \ , \tag{12.37}$$

and

$$\mathbf{B} = \text{var}(U(\boldsymbol{\theta}_0,\, \boldsymbol{Y})) \ . \tag{12.38}$$

The following example features an estimation equation that is of great practical relevance. It is widely used to extend generalized linear models to grouped data, and its application is demonstrated in Section 8.2.

Example 12.6. The Wedderburn and GEE estimating function. The form of estimating equation used throughout Chapter 8 can be written

$$U(\boldsymbol{\theta}; \boldsymbol{y}) = \dot{\boldsymbol{\mu}}^T \mathbf{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0} \ , \tag{12.39}$$

where $\boldsymbol{\mu} = E[\boldsymbol{Y}] = (\mu_1, ..., \mu_n)^T$ depends on parameters $\boldsymbol{\theta} \in \mathbb{R}^s$ and $\dot{\boldsymbol{\mu}}$ denotes the $n \times s$ Jacobian matrix with i, j element

$$[\dot{\boldsymbol{\mu}}]_{ij} = \frac{\partial \mu_i}{\partial \theta_j} \ .$$

Matrix $\mathbf{V}$ is $n \times n$ and positive definite, and may depend on $\boldsymbol{\theta}$, and on correlation parameters $\boldsymbol{\alpha}$ in the GEE context. Ideally, $\mathbf{V}$ is equal, or proportional, to var($\boldsymbol{Y}$). However, the approximate normality result in (12.36) holds if $\mathbf{V}$ is a mis-specification of var($\boldsymbol{Y}$), and hence $\mathbf{V}$ has a general interpretation as a "working" variance matrix (see the application in Section s:GEEs).

When $Y_i, i = 1, ..., n$ are independent then the estimating equation has the form used by Wedderburn (1974), as demonstrated in Section 8.1,

$$\sum_i^n \frac{\frac{\partial \mu_i}{\partial \theta_j}(y_i - \mu_i)}{v_i} = 0 \ , \qquad j = 1, ..., s \ , \tag{12.40}$$

where v_i is the assumed variance of Y_i.

In the generalized estimating equation context of Liang and Zeger (1986) the data have the form of independent multivariate observations $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})^T, i = 1, ..., m$ (Section 8.2) and the estimating equation can be written in the form

$$\sum_i^m \dot{\boldsymbol{\mu}}_i{}^T \mathbf{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0 \ , \qquad j = 1, ..., s \ .$$

where $\boldsymbol{\mu}_i = (E[Y_{i1}], ..., E[Y_{in_i}])^T$ and $\mathbf{V}_i$ is the assumed variance matrix of $\boldsymbol{Y} = (Y_{i1}, ..., Y_{in_i})^T$.

Since $\dot{\boldsymbol{\mu}}$ and $\mathbf{V}$ are functions of $\boldsymbol{\theta}$ alone, it follows immediately that the estimating function $U(\boldsymbol{\theta}; \boldsymbol{y})$ in (12.39) satisfies the zero-mean property (12.35). Moreover, matrices $\mathbf{A}$ and $\mathbf{B}$ are

$$\mathbf{A} = \dot{\boldsymbol{\mu}}^T \mathbf{V}^{-1} \dot{\boldsymbol{\mu}} \ , \tag{12.41}$$

and

$$\mathbf{B} = \dot{\boldsymbol{\mu}}^T \mathbf{V}^{-1} \widehat{\text{var}}(\boldsymbol{Y}) \mathbf{V}^{-1} \dot{\boldsymbol{\mu}} \ , \tag{12.42}$$

and so the approximate variance matrix of the M-estimator $\widetilde{\boldsymbol{\theta}}$ is

$$\begin{aligned} \widehat{\text{var}}(\widetilde{\boldsymbol{\theta}}) &= \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-T} \\ &= \left(\dot{\boldsymbol{\mu}}^T \mathbf{V}^{-1} \dot{\boldsymbol{\mu}} \right)^{-1} \dot{\boldsymbol{\mu}}^T \mathbf{V}^{-1} \text{var}(\boldsymbol{Y}) \mathbf{V}^{-1} \dot{\boldsymbol{\mu}} \left(\dot{\boldsymbol{\mu}}^T \mathbf{V}^{-1} \dot{\boldsymbol{\mu}} \right)^{-1} \end{aligned} \tag{12.43}$$

When $\mathbf{V} = \text{var}(\boldsymbol{Y})$ then this simplifies to

$$\widehat{\text{var}}(\widetilde{\boldsymbol{\theta}}) = \dot{\boldsymbol{\mu}}^T \widehat{\text{var}}(\boldsymbol{Y})^{-1} \dot{\boldsymbol{\mu}} \ . \tag{12.44}$$

In the GEE context, if it is believed that $\mathbf{V}(\boldsymbol{\theta}) = \text{var}(\boldsymbol{Y})$ then $\mathbf{V}(\widehat{\boldsymbol{\theta}})$ can be used as the estimate of $\text{var}(\boldsymbol{Y})$ in (12.44). However, if $\mathbf{V}$ is regarded as a working approximation to $\text{var}(\boldsymbol{Y})$ then equation (12.43) can be employed. This requires using an empirical estimate of $\text{var}(\boldsymbol{Y})$, obtained from appropriately standardized residuals (Hardin and Hilbe 2003). This estimator of $\text{var}(\widetilde{\boldsymbol{\theta}})$ is variously known as the robust-, or empirical-, or sandwich estimator of variance.

It can be shown that the variance matrix (12.44) is smaller (with respect to Löwner order) than (12.43) (e.g., see Exercise 12.4). That is, $\text{var}(\boldsymbol{Y})$ is the optimal choice for $\mathbf{V}$. □

12.4 Wald tests and confidence regions

12.4.1 Wald test statistics

Wald test statistics are obtained directly from the approximate normality of $\widehat{\boldsymbol{\theta}}_n$. Here W_1, W_2, W_3, and W_4 will be used to denote four variants of the Wald test statistic, differing only in the form of Fisher information employed in the statement

of approximate normality. For ease of notation, their dependence on sample size n will not be explicitly denoted.

In the scalar parameter case, $\theta \in \Theta \subset \mathbb{R}$, squaring both sides of the convergence result in (12.29) gives (by property P3, Section 13.3),

$$W_1 = I(\theta_0)(\widehat{\theta}_n - \theta_0)^2 \rightarrow_D N(0,1)^2 = \chi^2_1 \,, \tag{12.45}$$

where W_1 denotes a Wald test statistic for the null hypothesis $H_0 : \theta = \theta_0$.

From the definition of convergence in distribution, (12.45) establishes that $P_{\theta_0}(W_1 \leq w)$ converges to $F_{\chi^2_1}(w)$, where P_{θ_0} denotes probability under $H_0 : \theta = \theta_0$, and $F_{\chi^2_1}(w)$ denotes the distribution function of a χ^2_1 random variable. In particular, letting $\chi^2_{1,1-\alpha}$ denote the $(1-\alpha)$ quantile of the χ^2_1 distribution (i.e., if $X \sim \chi^2_1$ then $P(X > \chi^2_{1,1-\alpha}) = \alpha$), $P_{\theta_0}(W_1 > \chi^2_{1,1-\alpha})$ is approximately equal to α for large n under H_0. Hence, the critical region

$$\left\{ \boldsymbol{Y} : W_1 > \chi^2_{1,1-\alpha} \right\}$$

defines an approximate level α test of $H_0 : \theta = \theta_0$.

In the multi-dimensional case, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s (s \geq 1)$, the corresponding Wald test statistic for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is a quadratic form, and it follows from equation (15.2) that

$$W_1 = (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightarrow_D \chi^2_s \,. \tag{12.46}$$

Other forms of the Wald test statistic are obtained by replacing $I(\boldsymbol{\theta}_0)$ with an asymptotically equivalent[2] alternative. By virtue of the consistency of $\widehat{\boldsymbol{\theta}}_n$ (i.e., $\widehat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}_0$) one alternative is to use the expected Fisher information evaluated at $\widehat{\boldsymbol{\theta}}_n$, $I(\widehat{\boldsymbol{\theta}}_n)$. Other alternatives include using the observed Fisher information evaluated at either $\widehat{\boldsymbol{\theta}}_n$ or $\boldsymbol{\theta}_0$. That is,

$$W_2 = (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T I(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{12.47}$$

$$W_3 = -(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T l''(\theta_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{12.48}$$

and

$$W_4 = -(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T l''(\widehat{\theta}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{12.49}$$

[2]Two sequences of random variables are said to be asymptotically equivalent if the difference between them converges in probability to 0.

can all be used in place of W_1.

Example 12.7. Binomial Let Y be distributed Binomial(n, p_0) and let $\widehat{p} = y/n$ denote the MLE of p_0. From Example 11.1 we have that

$$\begin{aligned} l''(p) &= -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} \\ I(p) &= \frac{n}{p(1-p)} \,. \end{aligned}$$

Hence,

$$\begin{aligned} W_1 &= I(p_0)(\widehat{p} - p_0)^2 = \frac{n(\widehat{p} - p_0)^2}{p_0(1-p_0)} \\ W_2 &= I(\widehat{p})(\widehat{p} - p_0)^2 = \frac{n(\widehat{p} - p_0)^2}{\widehat{p}(1-\widehat{p})} \\ W_3 &= -l''(p_0)(\widehat{p} - p_0)^2 = \left(\frac{y}{p_0^2} + \frac{n-y}{(1-p_0)^2}\right)(\widehat{p} - p_0)^2 \\ W_4 &= -l''(\widehat{p})(\widehat{p} - p_0)^2 = \frac{n(\widehat{p} - p_0)^2}{\widehat{p}(1-\widehat{p})} \end{aligned}$$

The four Wald test statistics all have an asymptotic χ^2_1 distribution under $H_0 : p = p_0$, equivalently, their square-root has asymptotically a standard normal distribution. Here, W_2 and W_4 are identical, by virtue of the fact that $I(\widehat{p}) = -l''(\widehat{p})$. (This equivalence of the expected and observed Fisher information does not hold in general, but in Example 11.8 it is seen to hold for generalized linear models that use the canonical link function.)

Statistical texts differ in their statement of the Wald test for the binomial probability, with some using test statistic W_1 (e.g., Collett 1991), and others using the form of test statistics W_2 and W_4 (e.g., Wild and Seber 2000). The justification for statistics W_2 and W_4 is that they are easier to invert for purposes of obtained confidence intervals (see Section 12.4.2), however, it is test statistic W_1 that is to be preferred (Agresti and Coull 1998, Brown et al. 2001, Brown et al. 2002) because it makes use of the value of p_0 that is specified under the null hypothesis. □

12.4.2 Wald confidence intervals and regions

In Section 13.1 it was shown that the $(1-\alpha)100\%$ confidence region for $\boldsymbol{\theta}_0 \in \mathbb{R}^s$ is the collection of all values $\boldsymbol{\theta}^*$ that are not rejected by the size α test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^*$. The Wald statistic therefore leads to the approximate $(1-\alpha)100\%$ confidence region given by all values such that

$$W \leq \chi^2_{s,1-\alpha}$$

where W denotes any of the choices of Wald test statistic in (12.46)–(12.49).

The computation of the confidence region is generally more tractable for statistics W_2 and W_4 because these use forms of the Fisher information that are evaluated at the MLE $\widehat{\boldsymbol{\theta}}_n$.

Example 12.7 ctd. Although the Wald test statistic W_1 is preferred for testing $H_0 : p = p_0$, constructing a $(1-\alpha)100\%$ confidence interval using this test statistic requires determination of the collection of values p^* such that

$$\left\{ p^* : \frac{n(\widehat{p} - p_*)^2}{p_*(1-p_*)} < \chi^2_{1,1-\alpha} \right\} .$$

This can be done explicitly, by finding the roots of the appropriate quadratic (Exercise 12.6), and is sometimes called the the Wilson interval because its first use appears to be Wilson (1927).

The most common interval is obtained from expedient inversion of test statistics W_2 (or W_4). This choice give the desired interval to be the collection of p^* values such that

$$\left\{ p^* : \frac{n(\widehat{p} - p_*)^2}{\widehat{p}(1-\widehat{p})} < \chi^2_{1,1-\alpha} \right\}$$

which is the familiar interval

$$\widehat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} . \tag{12.50}$$

□

12.4.3 Wald tests and regions for $g(\boldsymbol{\theta}) \in \mathbb{R}^p$

Consider inference about $g(\boldsymbol{\theta}) \in \mathbb{R}^p$ where $g(\boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}_0 \in \mathbb{R}^s$ with $p \times s$ Jacobian matrix $G(\boldsymbol{\theta}_0)$. Then, applying the delta theorem to (12.15) gives the convergence result

$$\begin{aligned} \sqrt{n}(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_0)) &\rightarrow_D G(\boldsymbol{\theta}_0) N_s(\mathbf{0}, I_1(\boldsymbol{\theta}_0)^{-1}) \\ &= N_p(\mathbf{0}, G(\boldsymbol{\theta}_0) I_1(\boldsymbol{\theta}_0)^{-1} G(\boldsymbol{\theta}_0)^T) \ . \end{aligned} \tag{12.51}$$

from which approximate Wald tests and confidence regions for $g(\boldsymbol{\theta}_0)$ can be obtained following the methodology of the previous section, as demonstrated in the following example.

Example 12.8. Suppose that it is desired to test a hypothesis that specifies the values of the first r elements of parameter vector $\boldsymbol{\theta} \in \mathbb{R}^s$, i.e., $H_0 : (\theta_1, ..., \theta_r) = (\theta_{01}, ..., \theta_{0r})$. Here, $g(\boldsymbol{\theta}) = (\theta_1, ..., \theta_r)$ and so the first r columns of the $r \times s$ Jacobian matrix $G(\boldsymbol{\theta}_0)$ form the $r \times r$ identity matrix, and the remaining $s-r$ columns contain only zeroes. Let $\boldsymbol{\psi} = (\theta_1, ..., \theta_r)$, $\widehat{\boldsymbol{\psi}}_n = (\widehat{\theta}_1, ..., \widehat{\theta}_r)$ and $\boldsymbol{\psi}_0 = (\theta_{01}, ..., \theta_{0r})$ denote the first r elements of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$, respectively. Then, from application of (12.51),

$$\sqrt{n}(\widehat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0) \rightarrow_D N_r\left(\mathbf{0}, [I_1(\boldsymbol{\theta}_0)^{-1}]_{[rr]}\right)$$

where $[I_1(\boldsymbol{\theta}_0)^{-1}]_{[rr]}$ is the upper $r \times r$ submatrix of $I_1(\boldsymbol{\theta}_0)^{-1}$.

Consequently, from equation(15.2), it follows that a Wald statistic for testing the hypothesis $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ is

$$W_1 = (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T [(I(\boldsymbol{\theta}_0)^{-1}]_{[r,r]})^{-1} (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \rightarrow \chi^2_r \ .$$

Alternatively, replacing expected information with observed information evaluated at $\widehat{\boldsymbol{\theta}}_n$,

$$W_4 = (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T [(O(\widehat{\boldsymbol{\theta}}_n)^{-1}]_{[r,r]})^{-1} (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \rightarrow \chi^2_r \ .$$

□

Example 12.9. For Y_i distributed iid $N(\mu, \sigma^2)$, the asymptotic convergence of the MLEs was given in Example 12.4. Now, suppose that the asymptotic distribution of

the ML estimator of the "coefficient of variation" σ/μ is of interest. Here $g(\mu, \sigma^2) = \sigma/\mu = \sqrt{\sigma^2}/\mu$ and so $g'(\mu, \sigma^2) = (-\sigma/\mu^2, 1/(2\sigma\mu))$. Therefore

$$\sqrt{n}\left(\frac{\widehat{\sigma}}{\overline{Y}_n} - \frac{\sigma}{\mu}\right) \to_D N\left(0, \frac{\sigma^2}{\mu^2}\left(\frac{\sigma^2}{\mu^2} + \frac{1}{2}\right)\right).$$

Although the statistic $\sqrt{n}\left(\frac{\widehat{\sigma}}{\overline{Y}_n} - \frac{\sigma}{\mu}\right)$ converges in distribution, it can be shown that its expectation and variance do not exist. □

12.5 Likelihood ratio statistic

The asymptotic convergence of the likelihood ratio test statistic is stated and proved in Section 12.5.1 for the scalar parameter case. Extension to the multidimensional case is natural, and is stated in Section 12.5.2.

12.5.1 Likelihood ratio test: $\theta \in \mathbb{R}$

Theorem 12.4 Asymptotic distribution of the likelihood ratio statistic.

In the scalar parameter case, $\theta \in \Theta \subset \mathbb{R}$, if the assumptions of theorem 12.1 hold then

$$2[l(\widehat{\theta}_n; \mathbf{Y}) - l(\theta_0; \mathbf{Y})] = 2\log\frac{L(\widehat{\theta}_n; \mathbf{Y})}{L(\theta_0; \mathbf{Y})} \to_D \chi^2_1. \tag{12.52}$$

Proof. Expand $l(\theta; \boldsymbol{Y})$ about $\widehat{\theta}_n$ and evaluate at the true parameter value θ_0,

$$l(\theta_0) = l(\widehat{\theta}_n) + l'(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n) + \frac{1}{2}l''(\theta^*_n)(\theta_0 - \widehat{\theta}_n)^2 \tag{12.53}$$

where θ^*_n lies between θ_0 and $\widehat{\theta}_n$. Since $l'(\widehat{\theta}_n) = 0$ this gives

$$2[l(\widehat{\theta}_n) - l(\theta_0)] = -l''(\theta^*_n)(\theta_0 - \widehat{\theta}_n)^2 \tag{12.54}$$

The result follows from Slutsky's theorem since

$$(\sqrt{n}(\theta_0 - \widehat{\theta}_n))^2 \to_D (N(0, \frac{1}{I_1(\theta_0)}))^2 = \frac{1}{I_1(\theta_0)}\chi^2_1$$

and

$$\frac{1}{n}l''(\theta^*_n) \to_p -I_1(\theta_0)$$

by virtue of the fact that $\theta_n^* \to_p \theta_0$. □

Thus, for n sufficiently large, an approximate size α hypothesis test of $H_0 : \theta = \theta_0$ is given by

$$\text{reject } H_0 \text{ if } \quad 2[l(\widehat{\theta}_n;\, \boldsymbol{Y}) - l(\theta_0;\, \boldsymbol{Y})] > \chi^2_{1,1-\alpha} \ .$$

12.5.2 Likelihood ratio test for $\boldsymbol{\theta} \in \mathbb{R}^s$ and $g(\boldsymbol{\theta}) \in \mathbb{R}^p$

The convergence result in Theorem 12.4 generalizes to the multidimensional case, $\boldsymbol{\theta} \in \mathbb{R}^s$, with the degrees of freedom of the limiting chi-square being s. That is,

$$2[l(\widehat{\boldsymbol{\theta}}_n;\, \boldsymbol{Y}) - l(\boldsymbol{\theta}_0;\, \boldsymbol{Y})] = 2 \log \frac{L(\widehat{\boldsymbol{\theta}}_n;\, \boldsymbol{Y})}{L(\boldsymbol{\theta}_0;\, \boldsymbol{Y})} \to_D \chi^2_s$$

The convergence result for composite hypotheses is analogous, with the limiting degrees of freedom equal to the reduction in dimension under the null hypothesis. These results can be generalized by the following theorem, for purposes of testing the hull hypothesis $H_0 : g(\boldsymbol{\theta}) = g_0 \in \mathbb{R}^p$ where $1 \le p \le s$.

Theorem 12.5 *Given $g(\boldsymbol{\theta}) : \Theta \to \mathbb{R}^p$, and subject to appropriate regularity conditions,*

$$2[l(\widehat{\boldsymbol{\theta}}_n; \mathbf{Y}) - \max_{\boldsymbol{\theta}: g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_0)} l(\boldsymbol{\theta}; \mathbf{Y})] \to_D \chi^2_p \ . \tag{12.55}$$

If $g_0 = g(\theta_0)$ then the null hypothesis $H_0 : g(\boldsymbol{\theta}) = g_0$ is true, and the convergence result of Theorem 12.5 then leads to a test having asymptotically correct size.

The proof of Theorem 12.5 can be obtained from Wilks (1938) by assuming that the model $\{f(\boldsymbol{y}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s\}$ can be re-parameterized using $\boldsymbol{\zeta} = (g(\boldsymbol{\theta}), \lambda(\boldsymbol{\theta}))$ for some suitable transformation $\lambda(\theta) \in \mathbb{R}^{s-p}$.

12.6 Rao-score test statistic†

The Rao-score test is seldom used in practice and is included here only for completeness.

Another asymptotically equivalent test statistic is obtained by noting from equation (12.5) that $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is asymptotically equivalent to $\frac{l'(\theta_0)}{\sqrt{n} I_1(\theta_0)}$. The asymptotic

convergence of the square of this statistic is therefore

$$\frac{l'(\theta_0)^2}{nI_1(\theta_0)^2} \rightarrow_D N(0, \frac{1}{I_1(\theta_0)})^2 \equiv \frac{1}{I_1(\theta_0)}\chi_1^2$$

which leads to the Rao-score statistic

$$\frac{l'(\theta_0)^2}{I(\theta_0)} \rightarrow_D \chi_1^2 \,. \tag{12.56}$$

The convergence result in equation (12.56) can be obtained more directly from equation (12.6). However, the approach taken above establishes the asymptotic equivalence of the Wald and Rao-score statistics.

Using this statistic to test $H_0 : \theta = \theta_0$ is the Rao (or score) test. An advantage of this test is that it does not require calculation of the MLE $\widehat{\theta}_n$. However, this test is rarely seen in practice, because it requires both the score and information functions to be determined and it does not have the intuitive appeal of the Wald or likelihood ratio tests.

In the vector parameter case,

$$l'(\boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)^{-1} l'(\boldsymbol{\theta}_0) \rightarrow_D \chi_s^2$$

is the Rao-score test statistic of the simple hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. In the composite hypothesis case, with $\widehat{\boldsymbol{\theta}}_{0n}$ denoting the MLE under H_0, the Rao-score statistic (also known as the Lagrange multiplier statistic) is

$$l'(\widehat{\boldsymbol{\theta}}_{0n})^T I^{-1}(\widehat{\boldsymbol{\theta}}_{0n}) l'(\widehat{\boldsymbol{\theta}}_{0n}) \rightarrow_D \chi_r^2 \,.$$

Again, note that these test statistics do not require maximization of the likelihood with respect to the full model.

Example 12.7 ctd. Let Y be distributed Binomial(n, p) and let $\widehat{p} = y/n$ denote the MLE of p. Now,

$$\begin{aligned} l'(p) &= \frac{y - np}{p(1-p)} \\ I(p) &= \frac{n}{p(1-p)} \,, \end{aligned}$$

and hence the Rao-score test statistic is

$$\frac{l'(p_0)^2}{I(p_0)} = \frac{n(\widehat{p} - p_0)^2}{p_0(1-p_0)} \ . \tag{12.57}$$

In this particular case, the Rao-score test is identical to the Wald test statistic W_1.

□

12.7 Exercises

12.1 Let $A_1, ..., A_n$ and $B_1, ..., B_n$ denote sequences of events such that $P(A_n) \to 1$ and $P(B_n) \to 1$ as $n \to \infty$. Show that $P(A_n \cap B_n) \to 1$ as $n \to \infty$.

12.2 Let $Y_1, ..., Y_n$ be iid Poisson(λ) and consider the problem of estimating $p = P(Y = 0) = \exp(-\lambda)$. The MLE of λ is the sample mean $\overline{Y}_n$ and so the ML estimator of $\exp(-\lambda)$ is $\widehat{p} = \exp(-\overline{Y}_n)$. In Exercise 11.2 we saw that $I_1(\lambda) = 1/\lambda$.

1. The asymptotic convergence result (12.3) gives

$$\sqrt{n}(\overline{Y}_n - \lambda) \to_D N(0, \lambda).$$

   Use the delta theorem to determine the convergence (in distribution) of $\sqrt{n}(\widehat{p} - p)$.

2. Instead of using the ML estimator of p one could simply use the proportion of Y_i which take the value zero, which will be denoted $T(\boldsymbol{Y})$. This is a binomial experiment and so by the central limit theorem

$$\sqrt{n}(T(\boldsymbol{Y}) - p) \to_D N(0, p(1-p)).$$

   Show the MLE $\widehat{p}$ is asymptotically more efficient than $T(\boldsymbol{Y})$. That is, show that $p(1-p)$, is greater than the limiting variance of $\sqrt{n}(\widehat{p} - p)$ (calculated in part 1 above).

12.3 Let U be an estimating function satisfying the zero-mean property (12.35), and with $\mathbf{A}$ and $\mathbf{B}$ defined in (12.37) and (12.38), respectively. Then, the estimating function $U^{(s)} = \mathbf{B}^{-1}\mathbf{A}U$ has the same solution as the estimating function U.

1. Show that $U^{(s)}$ satisfies the zero-mean property.
2. Show that $U^{(s)}$ is standardized, that is, $-E_{\boldsymbol{\theta}_0}[U^{(s)'}(\boldsymbol{\theta}_0, \boldsymbol{Y})] = \text{var}(U^{(s)}(\boldsymbol{\theta}_0, \boldsymbol{Y}))$.

12.4 Consider the (weighted) linear regression model

$$\boldsymbol{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \ ,$$

where $\mathbf{V}$ is a diagonal matrix with diagonal elements $v_{ii} = w_i^{-1}\sigma^2$ for known weights w_i. For convenience, attention here is restricted to estimation of $\boldsymbol{\beta} \in \mathbb{R}^p$, and without loss of generality it can be assumed that σ^2 is known because of the information orthogonality of $\boldsymbol{\beta}$ and σ^2.

If $\boldsymbol{\beta}$ is estimated using an ordinary (unweighted) linear regression then the least-squares estimator $\tilde{\boldsymbol{\beta}}$ is the MLE from a mis-specified model in which all w_i are taken to be unity.

1. The variance of $\tilde{\boldsymbol{\beta}}$ is

$$\text{var}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \ .$$

   Verify that $\text{var}(\tilde{\boldsymbol{\beta}})$ is of the sandwich form, $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, where

$$\mathbf{A} = -E_{\boldsymbol{\beta}_0}[\tilde{l}''(\boldsymbol{\beta}_0, \boldsymbol{Y})] \ , \quad \text{and} \quad \mathbf{B} = \text{var}(\tilde{l}'(\boldsymbol{\beta}_0, \boldsymbol{Y})) \ ,$$

   where $\tilde{l}$ denotes log-likelihood under the ordinary least squares model.

2. The MLE $\widehat{\boldsymbol{\beta}}$ is the weighted least-squares estimator and $\text{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}\mathbf{V}^{-1}\mathbf{X})^{-1}$. Show that $\text{var}(\tilde{\boldsymbol{\beta}}) \geq \text{var}(\widehat{\boldsymbol{\beta}})$, where $\geq$ is defined using the Löwner ordering. That is, show that

$$\boldsymbol{b}^T\text{var}(\tilde{\boldsymbol{\beta}})\boldsymbol{b} \geq \boldsymbol{b}^T\text{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{b} \ ,$$

   for any $\boldsymbol{b} \in \mathbb{R}^p$.

12.5 The method-of-moments estimator for the parameters of the zero-inflated Poisson was derived using the estimating equation in Exercise 8.2. For the ZIP data of Exercise 3.7, calculate the approximate variance of $(\widetilde{\lambda}, \widetilde{p})$ given in (12.36). For the purpose of calculating $\mathbf{B}$, you may use the empirical estimate.

12.6 Determine the general form of the Wilson confidence interval for the binomial probability, for y successes from n trials. Calculate this interval with $y = 10$ successes observed from $n = 100$ trials, and compare to the more expedient interval obtained from (12.50).

12.7 Nonlinear regression: Suppose that $Y_i \sim N(g_i(\beta), \sigma^2)$, $i = 1, ..., n$, where $\beta \in \mathbb{R}$ and each g_i is a differentiable function of β (and may depend on covariates associated with observation i). For simplicity, you may assume that σ^2 is known. Let $\widehat{\beta}_n$ be the MLE of β. Determine the form of the Wald statistics W_1 and W_3.

12.8 Let A_i, $i = 1, ..., n$ be independent and identically distributed Bernoulli random variables with $P(A_i = 0) = P(A_i = 1) = 1/2$. If $A_i = 0$ then Y_i is observed from a Normal(μ, σ_0^2) distribution, otherwise Y_i is observed from a Normal(μ, σ_1^2) distribution. The values of σ_0 and σ_1 are known and are unequal. Both A_i and Y_i are observed, so the data are of the form $(a_1, y_1), ..., (a_n, y_n)$, or equivalently $(\boldsymbol{y}, \boldsymbol{a})$ where $\boldsymbol{y} = (y_1, ..., y_n)$ and $\boldsymbol{a} = (a_1, ..., a_n)$.

1. Write down the likelihood function $L(\mu) = f(\boldsymbol{y}, \boldsymbol{a}; \mu)$. (*Hint:* It may simplify notation to denote the distribution of $Y_i|A_i$ as Normal$(\mu, \sigma_{a_i}^2)$.)
2. Show that the maximum likelihood estimator of μ is the weighted mean

$$\widehat{\mu} = \frac{\sum_{i=1}^n y_i/\sigma_{a_i}^2}{\sum_{i=1}^n 1/\sigma_{a_i}^2}$$

3. Calculate the observed Fisher information $O(\mu)$ for the data.
4. Calculate the expected Fisher information $I(\mu)$ (for a size n sample $(A_1, Y_1), ..., (A_n, Y_n)$).
5. $I(\mu)$ and $O(\mu)$ calculated above are unequal except when n is even and $\sum_{i=1}^n a_i = n/2$. Which do you prefer as an estimator of the variance of $\widehat{\mu}$, $1/I(\widehat{\mu})$ or $1/O(\widehat{\mu})$? Justify your choice.

12.9 The representation of expected Fisher information as the variance matrix of the score function gives rise to an alternative estimate of $\mathbf{V}(\boldsymbol{\theta}_0)$ that can be used when Y_i are independent. Evaluated at $\widehat{\boldsymbol{\theta}}$, this estimate is

$$B(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \left(\frac{\partial l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}} \right)^2 \Bigg|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}} .$$

This estimate of $\mathbf{V}(\boldsymbol{\theta}_0)$ is not much used in practice, but in SAS it is available using the `COV=5` procedure option in `PROC NLP`.

1. If $Y_i, i = 1, ...n$ are iid Bernoulli(p), determine the forms of $O(\widehat{p})$ and $B(\widehat{p})$, and show that they are identical.
2. Suppose that Y_i are iid from a mixture of two known distributions (i.e., from a distribution with density $f(y_i; p) = pf_1(y_i) + (1-p)f_2(y_i)$), leaving only p to estimate.) For any $p, 0 < p < 1$, determine the forms of $O_(p)$ and $B_(p)$, and show that they are identical.

Chapter 13

Theoretical Tools

13.1 Equivalence of tests and confidence intervals

Test statistics can be used to construct confidence intervals for the true unknown $\boldsymbol{\theta}_0$, and vice-versa. Indeed, the following theorem formalizes the equivalency between hypothesis tests and confidence intervals.

Theorem 13.1

I) A $(1-\alpha)100\%$ *confidence region for* $\boldsymbol{\theta}_0$ *is given by the collection of all values* $\boldsymbol{\theta}^*$ *for which the size* α *hypothesis test of* $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^*$ *is not rejected.*

II) A size α *hypothesis test is given by rejecting* $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ *if and only if the* $(1-\alpha)100\%$ *confidence region does not include* $\boldsymbol{\theta}_0$*.*

Proof.

I) The confidence region contains $\boldsymbol{\theta}_0$ iff $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is not rejected, which has probability $(1-\alpha)$ as required.

II) The hypothesis test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is rejected iff the confidence region fails to contain $\boldsymbol{\theta}_0$, which has probability α, as required. □

13.2 Transformation of variables

Theorem 13.2 *Let* Y *be a continuous random variable with density function* f_Y *and sample space* S_Y*. Let* $Z = u(Y)$ *where the function* u *is a monotone (increasing*

or decreasing) function from S_Y to S_Z (the sample space of Z) with inverse $v(z) = y$ having derivative $v'(z)$. Then, Z has density function

$$f_Z(z) = \begin{cases} f_Y(v(z))|v'(z)|, & z \in S_Z \\ 0, & \textit{otherwise} \end{cases}$$

Proof. Note that u and v are either both monotone increasing or both monotone decreasing. If u and v are monotone increasing, then, for any $y \in \mathbb{R}$,

$$F_Z(z) = P(Z \leq z) = P(Z \in (-\infty, z]) = P(Y \in (-\infty, v(z)]) = \int_{-\infty}^{v(z)} f(t)dt$$

and substituting $s = u(t)$ gives

$$F_Z(z) = \int_{(-\infty,z]\cap S_Z} f(v(s))v'(s)ds = \int_{-\infty}^{z} f_Z(s)ds$$

where f_Z is defined above. Hence, f_Z is the density function of Z. The proof is similar in the monotone decreasing case. □

The above theorem extends in a natural way to the multidimensional case where $\boldsymbol{Y}$ and $\boldsymbol{Z} = u(\boldsymbol{Y})$ are both in $\mathbb{R}^n$. In this case, u is assumed to be one-to-one with differentiable inverse $v : \mathbb{R}^n \to \mathbb{R}^n$. Then,

$$f_Z(\boldsymbol{z}) = \begin{cases} f_Y(v(\boldsymbol{z}))|\det(v'(\boldsymbol{z}))|, & \boldsymbol{z} \in S_{\boldsymbol{Z}} \\ 0, & \text{otherwise} \end{cases} \tag{13.1}$$

where $v'(\boldsymbol{z})$ is the $n \times n$ dimensional derivative matrix (Jacobian) of v.

Box 13.1.

It follows from (13.1) that maximum likelihood inference is invariant to transformation of the data $\boldsymbol{y}$. This assumes that the transformation $\boldsymbol{z} = u(\boldsymbol{z})$ is fixed (i.e., does not depend on any parameters), because it is then the case that $|\det(v'(\boldsymbol{z}))|$ is a constant, and hence $f_Z(\boldsymbol{z};\boldsymbol{\theta}) \propto f_Y(\boldsymbol{y};\boldsymbol{\theta})$.

13.3 Relevant probability theory

This section focuses on the concept of convergence for sequences of random variables. The context in which this is relevant to maximum likelihood estimation is that the ML estimator is a random variable under repetition of the experiment, with

distribution that depends on the sample size n. This section provides the tools to understand the behaviour of this sequence of random variables (after suitable standardization,) as n tends to infinity (Chapter 12).

Definition 13.1 Convergence in distribution *Let $X_1, X_2, ...$ be a sequence of random variables with distributions functions denoted $F_n(x) = P(X_n \leq x), n = 1, 2, ...,$ and let X be a random variable with distribution function $F(x)$. Then the sequence X_n is said to* converge in distribution *(or* in law*) ($X_n \to_D X$) if*

$$F_n(x) \to F(x) \tag{13.2}$$

for all points x at which F is continuous.

In practice, notation such as $X_n \to_D N(0,1)$ (say), is used to denote $X_n \to_D X$ where X is distributed according to a standard normal distribution.

Some of the more theoretical texts will use the terminology that F_n converges *weakly* to F, and for this reason convergence in distribution is often called *weak convergence*.

Convergence in distribution extends naturally to the random vector case, $\boldsymbol{X} \in \mathbb{R}^s$, with (13.2) being replaced by the requirement that $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ for all points $\boldsymbol{x} \in \mathbb{R}^s$ at which F is continuous. In this case, $F_n(\boldsymbol{x}) = P(X_{n1} \leq x_1, ..., X_{ns} \leq x_s)$ where $\boldsymbol{X}_n = (X_{n1}, ..., X_{ns})$ and $\boldsymbol{x} = (x_1, ..., x_s)$.

Example 13.1. Let X be distributed $N(0,1)$, and define X_n as

$$X_n = (-1)^n X + Y_n$$

where Y_n are independently distributed $N(0, \frac{1}{n})$. Note that X_n is distributed $N(0, 1+\frac{1}{n})$ and hence the distribution function of X_n is $F_n(x) = P(X_n \leq x) = P(X_n/\sqrt{1+1/n} \leq x/\sqrt{1+1/n}) = \Phi(x/\sqrt{1+1/n})$ where Φ is the distribution function of the standard normal. From the continuity of Φ it follows that $F_n(x)$ converges to $\Phi(x)$ as $n \to \infty$, for all $x \in \mathbb{R}$. Since Φ is the distribution function of X, this establishes that $X_n \to_D X$. □

Convergence in distribution is often used to express the asymptotic behaviour of some suitably standardized estimator, as sample size becomes large. This is demonstrated below by the Central Limit Theorem for the sample mean (see Billingsley 1979, ,for proof). This theorem is extended to ML estimators in Chapter 12. *Newer Ref?*

Example 13.2. Central Limit Theorem: If $Y_1, Y_2, ...$ are iid random variables with finite mean and variance, μ and σ^2, and $\overline{Y}_n$ is the mean of $Y_1, ..., Y_n$ then

$$X_n = \sqrt{n}(\overline{Y}_n - \mu) \rightarrow_D N(0, \sigma^2).$$

□

Example 13.3. Normal approximation to the binomial: A binomial random variable $Y \sim Bin(n, p)$ is the sum of n iid Bernoulli(p) random variables, each with mean $\mu = P(Y_i = 1) = p$ and variance $\sigma^2 = p(1-p)$. The sample mean of these Bernoulli is of course $\widehat{p} = Y/n$, and so from the central limit theorem,

$$\sqrt{n}(\widehat{p} - p) \rightarrow_D N(0, p(1-p)).$$

Figure 13.1 shows the pointwise convergence of the distribution function of $\sqrt{n}(\widehat{p} - p)$ for increasing n. □

Definition 13.2 Convergence in probability

Let $X_1, X_2, ...$ and X be random variables on some probability space. Then the sequence X_n is said to converge in probability ($X_n \rightarrow_p X$) if for every $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty. \tag{13.3}$$

Convergence in distribution simply requires the distribution function of X_n to converge to the distribution function of X, but does not make any other assumptions. In particular, it does not assume that X_n takes similar values to X. In contrast, convergence in probability requires that X_n and X be defined on the same probability space (so that the random variable $X_n - X$ is defined) and that X_n takes values close to X with high probability.

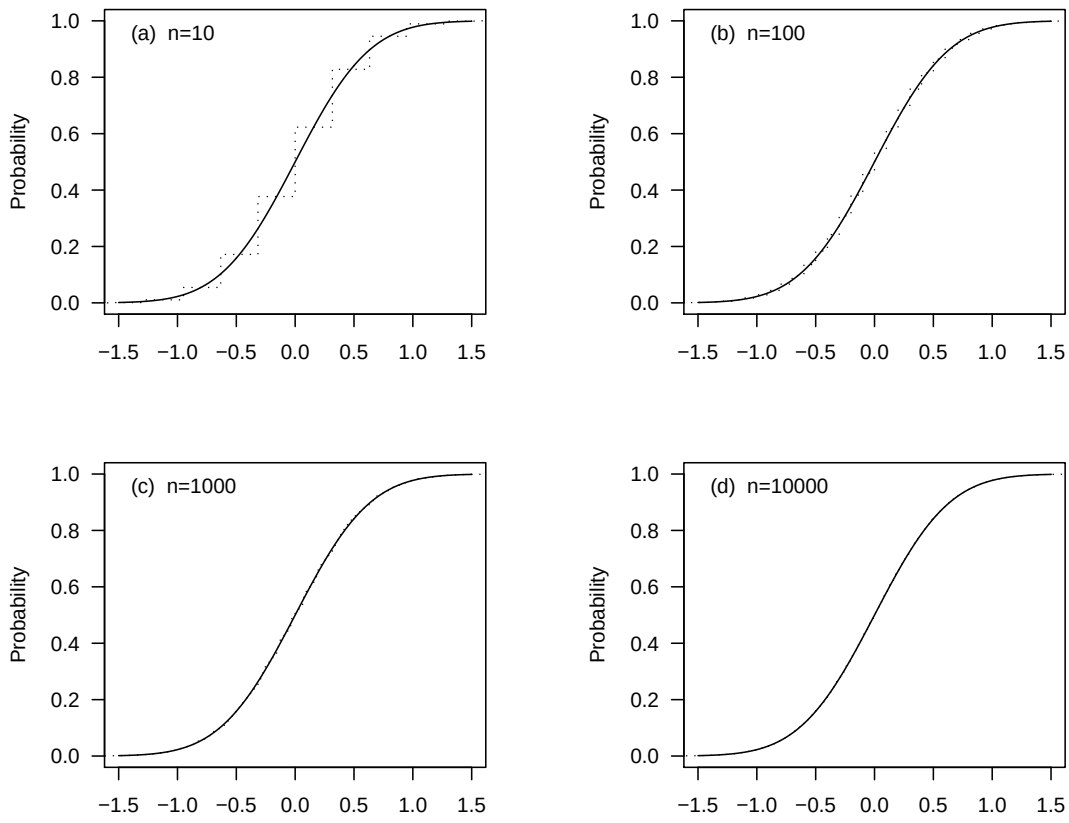


Figure 13.1: Distribution function (dashed) of $\sqrt{n}(\widehat{p}-p)$ where $\widehat{p}$ is the observed proportion from a binomial$(n, 0.5)$ experiment, for $n = 10, 100, 1000, 10000$. The limiting $N(0, 0.25)$ distribution is overlaid (solid).

Example 13.4. Weak law of large numbers: Let $Y_1, Y_2, ...$ be iid and let $\overline{Y}_n$ denote the mean of $Y_1, ..., Y_n$. Then,

$$\overline{Y}_n \rightarrow_p \mu$$

for some constant μ iff

$$nP(|Y_1| > n) \rightarrow 0 \quad \text{and} \quad \int_{[-n,n]} ydF(y) \rightarrow \mu \text{ as } n \rightarrow \infty.$$

The above condition is weaker than existence of $E[Y_1]$. If $E[Y_1]$ exists then convergence is almost sure (Billingsley 1979), and $\mu = E[Y_1]$ (this is the strong law of large numbers). □

Properties:

P1: $X_n \to_p X$ implies $X_n \to_D X$. The reverse does not hold in general (but see P2 below).

P2: $X_n \to_D c$, where c is a constant, implies $X_n \to_p c$. That is, convergence in probability and convergence in distribution are equivalent when convergence is to a constant (Exercise 13.5).

P3: If g is continuous and $X_n \to_p X$ then $g(X_n) \to_p g(X)$. The same is true for convergence in distribution, i.e., $X_n \to_D X$ implies $g(X_n) \to_D g(X)$. If X is a constant, c, then g need only be continuous at c to conclude that $g(X_n) \to_D g(c)$.

The following continuation of Example 13.1 shows that the reverse of property P1 does not hold in general.

Example 13.1 ctd. It was seen earlier that X_n converges in distribution to the standard normal distribution X. However, X_n does not converge in probability to X because when n is odd then $X_n = -X + Y_n$, and so $X_n - X = -2X + Y_n$ which has a $N(0, 4 + \frac{1}{n})$ distribution. □

Example 13.5. Under the conditions of the central limit theorem (Example 13.2)

$$\frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma} \to_D N(0,1) \ .$$

From property P3, using $g(x) = x^2$, if follows that

$$\frac{n(\overline{Y}_n - \mu)^2}{\sigma^2} \to_D \chi^2_1$$

since the square of a standard normal random variable has a χ^2_1 distribution. □

Slutsky's Theorem:

If $X_n \to_D X$ and $A_n \to_p a$, $B_n \to_p b$ (a, b constants) then

$$A_n + B_n X_n \to_D a + bX$$

Example 13.6. Asymptotic normality of t-statistic: Let $Y_1, ..., Y_n$ be iid from a distribution with mean μ and variance σ^2 and let $S_n^2 = \sum_{i=1}^n (Y_i - \overline{Y}_n)^2/(n-1)$ denote the sample standard deviation. It can be shown (Exercise 13.6) that $S_n \to_p \sigma$, and hence that $B_n = \sigma/S_n \to_p 1$ by property P3 above. From the central limit theorem (Example 13.2), $X_n \equiv \sqrt{n}(\overline{Y}_n - \mu)/\sigma \to_D N(0,1)$, and so application of Slutsky's theorem (with A_n omitted) gives

$$\frac{\sqrt{n}(\overline{Y}_n - \mu)}{S_n} = \frac{\sigma}{S_N} \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma} \to_D N(0,1) \ .$$

□

The next theorem is of great practical importance because, given convergence in distribution of (suitably standardized) X_n, it gives the asymptotic behaviour of (suitably standardized) functions of X_n. The practical version of this, the delta method, is implemented as an R function and in several SAS procedures.

Delta Theorem, $X_n \in \mathbb{R}$.

Suppose $\sqrt{n}(X_n - b) \to_D X$. If $g : \mathbb{R} \to \mathbb{R}$ is differentiable and g' is continuous at b then

$$\sqrt{n}(g(X_n) - g(b)) \to_D g'(b)X.$$

Proof:

The mean value theorem of differentiation gives

$$\sqrt{n}(g(X_n) - g(b)) = \sqrt{n}(g'(X_n^*)(X_n - b)) = g'(X_n^*)(\sqrt{n}(X_n - b)) \qquad (*)$$

where $|X_n^* - b| \leq |X_n - b|$. By Slutsky's theorem

$$X_n - b = \frac{1}{\sqrt{n}}(\sqrt{n}(X_n - b)) \to_D 0.X = 0$$

since $1/\sqrt{n} \to_p 0$. Hence $X_n - b \to_p 0$ (Fact II). Since $X_n - b \to_p 0$ then $X_n^* - b \to_p 0$ also. This is equivalent to $X_n^* \to_p b$. Now, g' is continuous at b, so $g'(X_n^*) \to_p g'(b)$

by Fact III. Applying Slutsky's theorem to $(*)$ gives the required result.

Example 13.7. If $\sqrt{n}(\overline{Y}_n - \mu) \to_D N(0, \sigma^2)$ and $g(y) = y^2$ then

$$\sqrt{n}(\overline{Y}_n^2 - \mu^2) \to_D 2\mu N(0, \sigma^2) = N(0, 4\mu^2\sigma^2).$$

□

Delta Theorem, $X_n \in \mathbb{R}^s$**.**

The delta theorem extends naturally to the random vector case. Suppose that

$$\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{b}) \to_D \boldsymbol{X}\,.$$

Let $g : \mathbb{R}^s \to \mathbb{R}^p$ where

$$g(\boldsymbol{x}) = \begin{pmatrix} g_1(\boldsymbol{x}) \\ . \\ . \\ . \\ g_p(\boldsymbol{x}) \end{pmatrix}$$

and each co-ordinate $g_i : \mathbb{R}^s \to \mathbb{R}$ has derivative $g_i' = (\frac{\partial g_i}{\partial x_1}, ..., \frac{\partial g_i}{\partial x_s})^T$ that is continuous at $\boldsymbol{b}$. Then,

$$\sqrt{n}(g(\boldsymbol{X}_n) - g(\boldsymbol{b})) \to_D G(\boldsymbol{b})\boldsymbol{X}$$

where $G(\boldsymbol{b})$ is the p by s matrix of derivatives

$$G(\boldsymbol{b}) = \begin{pmatrix} g_1'(\boldsymbol{b})^T \\ . \\ . \\ . \\ g_p'(\boldsymbol{b})^T \end{pmatrix}.$$

13.4 Relevant inequalities

Jensen's inequality for convex functions:

Let D be an interval in $\mathbb{R}$. A function $\phi : D \to \mathbb{R}$ is convex if the graph of $(y, \phi(y))$ is such the chord between any two points is completely on or above the curve (Fig 13.2). That is,

$$\phi(ay_1 + (1-a)y_2) \leq a\phi(y_1) + (1-a)\phi(y_2)$$

for all $y_1, y_2 \in D$ and $0 \leq a \leq 1$.

Jensen's inequality states that if ϕ is convex and Y is any random variable on D with finite expectation, then

$$\phi(E[Y]) \leq E[\phi(Y)] \ .$$

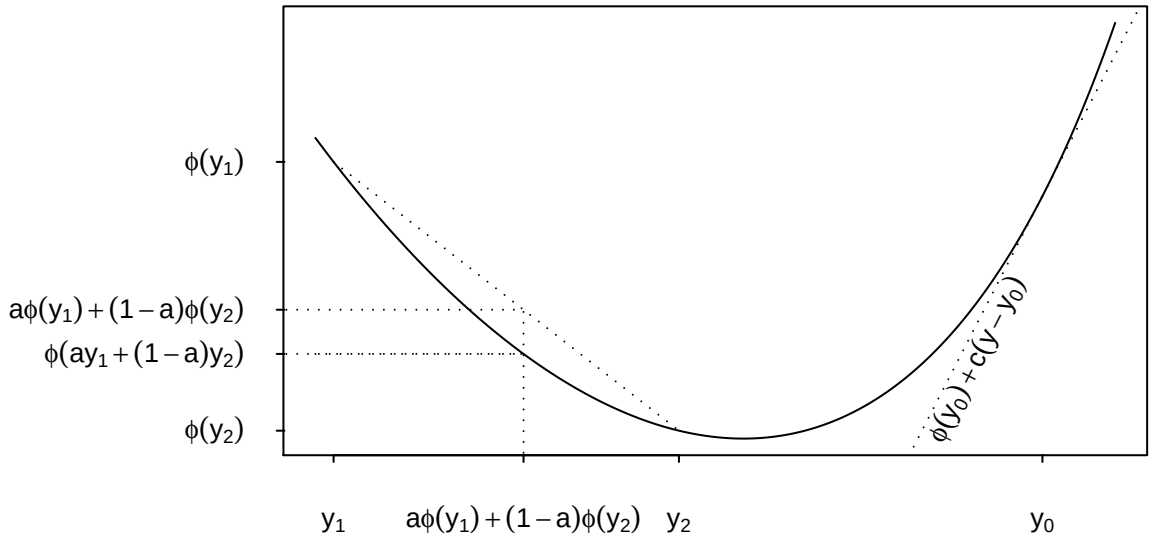


Figure 13.2: Graph of a convex function. The chord between points $(y_1, \phi(y_1))$ and $(y_2, \phi(y_2))$ lies completely on or above the curve $(y, \phi(y))$. The tangent at point y_0 lies completely on or below the curve.

Proof. An equivalent definition of convexity is, for any fixed $y_0 \in D, \exists c$ such that

$$\phi(y_0) + c(y - y_0) \leq \phi(y) \text{ for all } y \in D.$$

Jensen's inequality results from setting $y_0 = E[Y]$ and taking expectations. □

It can be shown that if a function $\phi(y)$ is twice differentiable then it is convex if and only if the second derivative is non-negative on D. This can be used to quickly establish that $\phi(y) = \exp(y)$ and $\phi(y) = y^2$ are convex functions on $\mathbb{R}$. The functions $\phi(y) = -\log(y)$ and $\phi(y) = -\sqrt{y}$ are convex on $\mathbb{R}^+$.

Example 13.8. Let Y be distributed $N(\mu, \sigma)$. Then $X = \exp(Y)$ has a lognormal distribution. The exponential function is convex, and application of Jensen's

inequality gives

$$E[X] = E[\exp(Y)] \geq \exp(E[Y]) = \exp(\mu) \ .$$

In fact, $E[X] = \exp(\mu + \sigma^2/2)$. □

Example 13.9. Negative bias of sample standard deviation Let $Y_1, ..., Y_n$ be iid from some distribution with finite mean and variance μ and σ^2. Consider the random variable given by the sample standard deviation

$$S(\boldsymbol{Y}) = \left(\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2 \right)^{1/2} .$$

Then, using Jensen's inequality with $\phi(S) \equiv S^2$ we have

$$\phi(E[S]) = (E[S])^2 \leq E[\phi(S)] = E[S^2] = \sigma^2,$$

That is, we have shown that $E[S] \leq \sigma$. This inequality will be strict provided that the distribution of Y_i is not degenerate, because the square function is strictly convex. □

Cauchy-Schwarz inequality:

For any two random variables X, Y such that $E[X^2] < \infty$ and $E[Y^2] < \infty$,

$$(E[XY])^2 \leq E[X^2]E[Y^2] \ .$$

Proof. For any $t \in \mathbb{R}$ it is the case that

$$E[(X - tY)^2] \geq 0 \ .$$

That is,

$$E[X^2] - 2tE[XY] + t^2 E[Y^2] \geq 0 \ . \tag{13.4}$$

Equation (13.4) is a quadratic in t of the form $at^2 + bt + c$, where $a = E[Y^2], b = -2E[XY]$, and $c = E[X^2]$. This quadratic takes only non-negative values and so it is necessarily the case that $b^2 - 4ac \leq 0$ (Stewart 1999). That is,

$$4(E[XY])^2 - 4E[Y^2]E[X^2] \leq 0 \ ,$$

from which the Cauchy-Schwarz inequality follows.

Corollary. Let (X, Y) be a discrete bivariate random variable having point mass $\frac{1}{n}$ at values $(x_i, y_i), i = 1, ..., n$. The pairs (x_i, y_i) are not required to be distinct. From application of the Cauchy-Schwarz inequality, it follows that

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 \ . \tag{13.5}$$

Example 13.10. Covariance inequality: Let X and Y be random variables with finite mean and variance μ_x, σ_x^2, and μ_y, σ_y^2, respectively. Then $\text{Cov}[X, Y]^2 = (E[(X - \mu_x)(Y - \mu_y)])^2 \leq E[(X - \mu_x)^2]E[(Y - \mu_y)^2] = \sigma_x^2\sigma_y^2$. □

13.4.1 Useful identities

Mean and variance identities:

The mean and variance identities below can be very useful if the distribution of Y is complicated, but the distributions of X and $Y|X$ are of convenient form (e.g., Exercise ex:MeanAndVarOfBinormal).

If X and Y are any two random variables then

$$E[Y] = E_X[E[Y|X]] \tag{13.6}$$

and

$$\text{var}(Y) = E_X[\text{var}(Y|X)] + Var_X(E[Y|X]) \ , \tag{13.7}$$

where E_X and Var_X denote expectation and variance with respect to X, respectively.

The proof of 13.6 is quick and is given below. See (Casella and Berger 1990, Section 4.4) for proof of 13.7.

Proof:

$$\begin{aligned} E[Y] &= \int\int yf(x,y)dxdy \\ &= \int\int yf(y|x)f(x)dxdy \\ &= \int\int yf(y|x)dyf(x)dx \\ &= \int E[Y|X]f(x)dx \\ &= E_X[E[Y|X]] \end{aligned}$$

Example 13.11. Let $N \sim \text{Poisson}(\lambda)$ and let $Y|N \sim \text{binomial}(N,p)$. From (13.6) and (13.7) the mean and variance of Y are

$$E[Y] = E_N[E[Y|N]] = E_N[Np] = pE_N[N] = p\lambda$$

and

$$\begin{aligned} \text{var}(Y) &= E_N[\text{var}(Y|N)] + Var_N(E[Y|N]) \\ &= E_N[Np(1-p)] + Var_N(Np) \\ &= p(1-p)\lambda + p^2\lambda = p\lambda \ . \end{aligned}$$

In fact, it can be shown that $Y \sim \text{Poisson}(p\lambda)$ – one way to establish this is from the identity

$$f(y) = \sum_{N \geq y} f(y|N)f(N) = \sum_{N \geq y} \frac{N!}{y!(N-y)!} p^y (1-p)^{N-y} \frac{e^{-\lambda}\lambda^N}{N!} \ .$$

This is left as an Exercise. □

13.5 Exercises

13.1 Provide a counter example to show that $X_n \to_D X$ and $Y_n \to_D Y$ does *not* imply that $X_n + Y_n \to_D X + Y$.

13.2 The definition of convergence in probability (Definition 13.2) used a strict inequality. Show that this inequality does not need to be strict, that is, show that the condition

$$P(|X_n - X| > \epsilon) \to 0 \text{ as } n \to \infty \quad \forall \epsilon > 0 \tag{13.8}$$

is equivalent to the condition that

$$P(|X_n - X| \geq \varepsilon) \to 0 \text{ as } n \to \infty \quad \forall \varepsilon > 0 \ . \tag{13.9}$$

13.3 Let $k < 0$ be a constant and let $X_1, X_2, ...$ be a sequence of random variables. Prove that $X_n \to_p k$ implies $P(X_n < 0) \to 1$.

13.4 Suppose that $X_n \to_p 0$ and that B_n is bounded in probability, that is, there exists a $b > 0$ such that $P(|B_n| > b) \to 0$. Show that $X_n B_n \to_p 0$.

13.5 Let c be a constant and let $X_1, X_2, ...$ be a sequence of random variables. Prove that

a) $X_n \to_p c$ implies $X_n \to_D c$.

b) $X_n \to_D c$ implies $X_n \to_p c$.

13.6 Let $Y_1, ..., Y_n$ be iid with mean μ and variance σ^2. Show that the sample variance S_n^2 converges in probability to σ^2. Hint: Show that the sample variance can be expressed as

$$S_n^2 = \sum (Y_i - \overline{Y}_n)^2/(n-1) = \frac{n}{n-1}\left[\sum (Y_i - \mu)^2/n + K_n\right]$$

where $K_n \to_p 0$.

13.7 Let $X_n \sim \chi^2_n$. Using the fact that X_n has distribution equal to that of the sum of n iid χ^2_1 random variables, show that

a) Show that $\sqrt{n}\left(\frac{X_n}{n} - 1\right) \to_D N(0, 2)$,

b) and hence (see Example 12.4) that $\sqrt{n}\sigma^2\left(\frac{X_{n-1}}{n} - 1\right) \to_D N(0, 2\sigma^4)$.

13.8 The following counter examples show that convergence in probability (and hence also convergence in distribution) do not imply convergence of moments.

a) Construct an example in which $X_n \to_p X$ where $E[X_n]$ exist for all n, and $\lim_{n\to\infty} E[X_n] = \mu$, but $E[X] = \mu_X \neq \mu$.

b) Construct an example in which $X_n \to_p X$ where $E[X]$ exists and $E[X_n]$ exist for all n, but $\lim_{n\to\infty} E[X_n]$ does not exist.

13.9 Let Y be distributed according to the five-parameter binormal mixture model of Example 2.9. By conditioning on the unobserved Bernoulli random variable B, use (13.6) and (13.7) to find $E[Y]$ and $\text{var}(Y)$.

13.10 For Y distributed according to the zero-inflated Poisson (see Section 7.5.2), use (13.6) and (13.7) to show that $E[Y] = (1-p)\lambda$ and $\text{var}(Y) = (1 + p\lambda)E[Y]$.

13.11 Let $\lambda \sim \text{Gamma}(m, m^{-1}\mu)$ and let $Y|\lambda \sim \text{Poisson}(\lambda)$. Using (13.6) and (13.7), show that Y has mean μ and variance $\mu(1 + \mu/m)$. (Note: It can be shown that Y has a negative binomial distribution, $\text{NB}(m, \mu)$.)

Chapter 14

Fundamental paradigms and principles of inference

The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation — Sir Ronald Fisher

14.1 Introduction

Not all statisticians think alike. In fact, there are several schools of thought regarding the appropriate framework for statistical inference. These include, frequentist, fiducial (Fisher 1933) and Bayesian[1].

The frequentist approach is also known as the "classical" or "traditional" approach. It is (at the time of writing) the form of statistical inference that is most widely taught at schools and universities, and with the widespread use of familiar frequentist-based methods (e.g., linear regression, ANOVA, chi-squared tests, generalized linear models) it continues to be the most widely used of the statistical paradigms. The frequentist approach is built around the idea of repeat experiments. Indeed, frequentists define the probability of an event to be the proportion of times it occurs out of a large number of independent repeat experiments, and the tools of frequentist inference are based on this notion. For example, the method used to calculate a 95% confidence interval is such that if the experiment is repeated a

[1]See Section 15.2 for a quick self-test of frequentist versus Bayesian thinking.

large number of times then the calculated 95% confidence intervals will contain the unknown parameter about 95% of the time.

The fiducial approach essentially advocates basing inference on the likelihood function alone, without recourse to prior knowledge or the concept of repeat sampling. However, this approach appears to have clouded rather than clarified the debate, and has gone into statistical obscurity.

The Bayesian approach also advocates using the likelihood function, but in conjunction with a prior distribution on the parameters. Moveover, the Bayesian approach departs markedly from frequentist and fiducial approaches by viewing probability as an expression of belief.

In this section we scratch the surface on some of the considerations behind the search for the correct paradigm for statistics. The presentation takes a look at three principles, namely, the sufficiency, conditionality and likelihood principles. Birnbaum (1962, and discussion thereof) caused a statistical brawl between frequentist and Bayesian statisticians with his ground-breaking work on the interplay between these relationships. It is the flavour of this controversy that the presentation here attempts to capture.

In the following sections, notation of the form $(E, \boldsymbol{y})$ the "statistical evidence" (Birnbaum 1962), where $\boldsymbol{y}$ represents "what was observed" and E represents "under what experimental plan and conditions". Conceptually, it may help the reader to (temporarily) unlearn their previous statistical training. So, for example, when conducting an experiment to make inference about the probability that a passenger car has more than one occupant, suspend your knowledge that an appropriately conducted experiment will yield a binomial random variable. Instead, consider what it means to repeat the experiment (e.g., see Example 14.5), and what it is that the observer actually measures (e.g., see Example 14.2).

14.2 Sufficiency principle

Speaking very crudely, the notion behind the sufficiency principle is that the data $\boldsymbol{y} = (y_1, ..., y_n)$ can be partitioned into a useful part and an irrelevant part. It is

only the useful part that should be utilized in making inference, and the irrelevant part should be discarded because it will simply add noise (i.e., increase statistical variability). This is seen in the case of iid $\mathrm{N}(\mu, \sigma^2)$ data where it is only the sample mean, $\overline{y}$, and the sample variance $s^2 = \sum_i^n (y_i - \overline{y})^2/(n-1)$ that are used for inference about μ or σ. The data, $\boldsymbol{y}$, can be thrown away once $\overline{y}$ and s^2 have been calculated.

Somewhat more formally, the concept of sufficiency is used to "reduce" the data in such a way that no relevant information about the parameters is lost. In the definition of sufficiency below, the statistic $T(\boldsymbol{Y}) : \mathbb{R}^n \to \mathbb{R}^m$ where $m \leq n$ can be considered a "reduction" of the data $\boldsymbol{Y}$ if $m < n$. The essence of sufficiency is that the data $\boldsymbol{Y}$ can, loosely speaking, be partitioned into the pieces $T(\boldsymbol{Y})$ and $\boldsymbol{Y}|T(\boldsymbol{Y})$, where the distribution of the latter does not depend on $\boldsymbol{\theta}$.

Definition 14.1 Sufficient statistic

$T(\boldsymbol{Y}) : \mathbb{R}^n \to \mathbb{R}^m$ is a *sufficient statistic* for $\boldsymbol{\theta}$ if the distribution of $\boldsymbol{Y}$ given $T(\boldsymbol{Y})$ does not depend on $\boldsymbol{\theta}$. □

The argument can be made that if $\boldsymbol{Y}|T(\boldsymbol{Y})$ does not depend on $\boldsymbol{\theta}$ then it can not provide information about $\boldsymbol{\theta}$. Hence, all of the information in the data about $\boldsymbol{\theta}$ is contained solely in $T(\boldsymbol{Y})$. This logic is formally embodied in the sufficiency principle.

The sufficiency principle: Let E_1 be a specified experiment and let $T(\boldsymbol{Y})$ be any sufficient statistic. If $\boldsymbol{y}$ is observed from E_1, and if E_2 is the experiment derived from E_1 by observing only $T(\boldsymbol{y})$, then $(E_1, \boldsymbol{y})$ and $(E_2, T(\boldsymbol{y}))$ have the same "information content". □

One can establish that a statistic is sufficient from the above definition of a sufficient statistic. This can be cumbersome because it requires the conditional distribution of $\boldsymbol{Y}$ given $T(\boldsymbol{Y})$.

Example 14.1. Let Y_1 and Y_2 be iid Poisson(λ). The statistic $T = Y_1 + Y_2$ is sufficient for λ. That is, the distribution of (Y_1, Y_2) given T does not depend on λ.

Proof: Now,

$$\begin{aligned} f(\boldsymbol{y}|T(\boldsymbol{Y})=t;\lambda) &= f(y_1,y_2|t;\lambda) \\ &= \frac{f(y_1,y_2,t;\lambda)}{f(t;\lambda)} = \begin{cases} \frac{f(y_1,y_2;\lambda)}{f(t;\lambda)} & ,y_1+y_2=t \\ 0 & ,y_1+y_2\neq t \end{cases} \end{aligned}$$

where $f(y_1,y_2;\lambda) = f(y_1;\lambda)f(y_2;\lambda)$ because Y_1 and Y_2 are independent. Using moment generating functions, or from direct calculation, it can be shown that T has a Poisson(2λ) distribution, and so

$$\frac{f(y_1,y_2;\lambda)}{f(t;\lambda)} = \frac{\frac{e^{-\lambda}\lambda^{y_1}}{y_1!}\frac{e^{-\lambda}\lambda^{y_2}}{y_2!}}{\frac{e^{-2\lambda}(2\lambda)^t}{t!}} = \frac{t!}{y_1!y_2!}\left(\frac{1}{2}\right)^t$$

which does not depend on λ.

Note that since $y_2 = t - y_1$, we have

$$f(y_1|t) = \frac{t!}{y_1!(t-y_1)!}\left(\frac{1}{2}\right)^t .$$

That is, conditional on t, y_1 is Bin$(t,\frac{1}{2})$. □

It can be extremely tedious to verify sufficiency of $T(\boldsymbol{Y})$ from its definition, it is much easier to use the Factorization theorem.

Theorem: Factorization Theorem/Fisher-Neyman criterion
$T(\boldsymbol{Y})$ is sufficient for $\boldsymbol{\theta}$ iff there exist non-negative functions $g(T(\boldsymbol{y});\boldsymbol{\theta})$ and $h(\boldsymbol{y})$ such that $f(\boldsymbol{y};\boldsymbol{\theta})$ can be written

$$f(\boldsymbol{y};\boldsymbol{\theta}) = g(T(\boldsymbol{y});\boldsymbol{\theta})h(\boldsymbol{y}) \tag{14.1}$$

where $h(\boldsymbol{y})$ does not depend on $\boldsymbol{\theta}$.

Proof. To show that sufficiency ⇒ Fisher-Neyman criterion, assume that $T(\boldsymbol{Y})$ is sufficient for $\boldsymbol{\theta}$. Then,

$$\begin{aligned} f(\boldsymbol{y};\boldsymbol{\theta}) &= f(\boldsymbol{y}|T(\boldsymbol{y});\boldsymbol{\theta})f(T(\boldsymbol{y});\boldsymbol{\theta}) \\ &= f(\boldsymbol{y}|T(\boldsymbol{y}))f(T(\boldsymbol{y})|\boldsymbol{\theta}) \;\equiv\; h(\boldsymbol{y})g(T(\boldsymbol{y});\boldsymbol{\theta}) \end{aligned}$$

where $h(\boldsymbol{y}) = f(\boldsymbol{y}|T(\boldsymbol{y}))$ does not depend on $\boldsymbol{\theta}$ because $T(\boldsymbol{Y})$ is sufficient for $\boldsymbol{\theta}$.

To show that Fisher-Neyman criterion $\Rightarrow$ sufficiency, note that the density of $T(\boldsymbol{Y})$ is

$$\begin{aligned} f(t;\boldsymbol{\theta}) &= \int_{\boldsymbol{y}:T(\boldsymbol{y})=t} f(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\ &= \int_{\boldsymbol{y}:T(\boldsymbol{y})=t} g(t;\boldsymbol{\theta})h(\boldsymbol{y})d\boldsymbol{y} \\ &= g(t;\boldsymbol{\theta})\int_{\boldsymbol{y}:T(\boldsymbol{y})=t} h(\boldsymbol{y})d\boldsymbol{y} \;=\; k_t g(t;\boldsymbol{\theta}) \end{aligned} \tag{14.2}$$

where k_t depends only on t. Thus, the conditional density of $\boldsymbol{Y}$ given $T(\boldsymbol{y}) = t$ is

$$f(\boldsymbol{y}|T(\boldsymbol{y}) = t) \;=\; \frac{f(\boldsymbol{y};\boldsymbol{\theta})}{f(t;\boldsymbol{\theta})} \;=\; \frac{g(t;\boldsymbol{\theta})h(\boldsymbol{y})}{k_t g(t;\boldsymbol{\theta})} \;=\; \frac{h(\boldsymbol{y})}{k_t}$$

which does not depend on $\boldsymbol{\theta}$.

Example 14.1 ctd. We have

$$f(y_1, y_2; \lambda) = \frac{e^{-2\lambda}\lambda^{y_1+y_2}}{y_1!y_2!}$$

which can be written in the form of (14.1) with $T(\boldsymbol{y}) = y_1 + y_2$, $g(T(\boldsymbol{y}); \lambda) = e^{-2\lambda}\lambda^{y_1+y_2}$ and $h(\boldsymbol{y}) = 1/(y_1!y_2!)$. Hence $T(\boldsymbol{Y}) = Y_1 + Y_2$ is sufficient for λ by the Factorization Theorem. □

Example 14.2. Bernoulli trials. Suppose that we observe a sequence of n iid Bernoulli(p) random variables, $Y_1, ..., Y_n$. Then, by the factorization theorem, $t = \sum_{i=1}^n y_i$ is sufficient for p because $f(y_1, ..., y_n; p) = p^t(1-p)^{n-t}$. The sufficiency principle says that for purposes of making inference about p it is enough to work with just the sufficient statistic $T = \sum_{i=1}^n Y_i$ because the binomial experiment in which only T is observed has the same information content (about p) as the experiment in which the Bernoulli sequence $Y_1, ..., Y_n$ is observed. □

Example 14.3. Normal. For iid data from a N(μ, σ^2) distribution, the factorization theorem quickly establishes that $(\overline{Y}, \sum Y_i^2)$ is sufficient for $\boldsymbol{\theta} = (\mu, \sigma)$. The

sufficiency principle can therefore be used to say that the experiment E_2 in which only the bivariate sufficient statistic $(\overline{y}, \sum y_i^2)$ is observed has the same information content about μ and σ as the experiment E_1 in which the entire data $\boldsymbol{y} = y_1, ..., y_n$ are observed. Note that observing $(\overline{y}, \sum y_i^2)$ is equivalent to observing $(\overline{y}, \sum (y_i - \overline{y})^2)$ or $(\overline{y}, s^2)$ because each can be calculated from any other. □

In frequentist statistics there is a standard body of theory based on the use of sufficient statistics for the construction of unbiased estimators with minimum variance (MVUE's). At the core of this theory is the Rao-Blackwell theorem which states that if a MVUE exists then it must be a function of only a minimal[2] sufficient statistic.

It is immediate from the factorization theorem that maximum likelihood estimators are functions of sufficient statistics, because the maximization of $L(\boldsymbol{\theta}; \boldsymbol{y})$ with respect to $\boldsymbol{\theta}$ depends on $\boldsymbol{y}$ only through the sufficient statistic $T(\boldsymbol{y})$.

14.3 Conditionality principle

The idea behind the conditionality principle is to condition upon (i.e. treat as fixed) aspects of the experiment which contain no information about $\boldsymbol{\theta}$. Some form of conditioning is vital in frequentist statistics, because it is required to determine what exactly is meant by "repetition of the experiment".

The statement of the principle that is given below uses a mixture of just two experiments (see Example 14.4), but it applies more generally to a mixture of an arbitrary number of experiments (see Example 14.5).

Let E_1 and E_2 be two experiments with the same parameter space Θ and with densities $f_1(\boldsymbol{y}_1; \boldsymbol{\theta})$ and $f_2(\boldsymbol{y}_2; \boldsymbol{\theta})$ respectively, where the unknown $\boldsymbol{\theta} \in \Theta$ is the same in both experiments.

The conditionality principle: Let the (mixture) experiment, E, consist of $(E_1, \boldsymbol{y}_1)$ with probability p, or $(E_2, \boldsymbol{y}_2)$ with probability $1 - p$. If $(E_i, \boldsymbol{y}_i)$ is the experiment actually observed then the "information content" from $(E, (E_i, \boldsymbol{y}_i))$ equals the "in-

[2] A sufficient statistic $T(\boldsymbol{Y})$ is said to be minimal sufficient if for any other sufficient statistic $S(\boldsymbol{Y})$ there exists a function h_S such that $T(\boldsymbol{Y}) = h_S(S(\boldsymbol{Y}))$.

formation content" from $(E_i, \boldsymbol{y}_i)$. □

In the statement of the conditionality principle, probability p does not depend on $\boldsymbol{\theta}$, and hence there is no information about p from the choice of whether it is experiment E_1 or E_2 that is observed. The implication is that it is appropriate to condition on the experiment actually observed.

Example 14.4. Random experiment. Suppose that the experiment consists of a laboratory making four iid measurements of the concentration of a chemical in a supplied sample. The laboratory has two spectrometers, an older machine, and a new one that is an order of magnitude more precise than its predecessor. For ease of exposition, it will be assumed that the measurements (under repeat use) from the two machines are normally distributed about the true concentration, μ, with standard deviations of 1.0 and 0.1 micro-molg^{-1}, respectively. Unfortunately, on the day of measuring, the new machine was already being used and it was necessary to use the old machine.

Having taken a sample of four measurements using the old machine, inference would proceed by calculating the mean of these measurements and using knowledge about the distribution of the sample mean under replication of the experiment. In this case, that the sample mean had a normal distribution with mean equal to the true unknown concentration, and standard deviation of $1.0/\sqrt{4} = 0.5$ micro-molg^{-1}. Or does it?

The experiment described above is a mixture experiment, in the sense that the measured concentrations are obtained from either using the new spectrometer (experiment E_1, say) or the old one (experiment E_2), depending on the availability of the new spectrometer. In the above scenario the old machine was used and it is totally irrelevant that there was a more precise machine that could have been used on another occasion. By calculating a standard deviation of 0.5 micro-molg^{-1} for the sample mean of the four measurements, the experiment that is being hypothetically repeated is observation of measurements $\boldsymbol{y}_2$ from experiment E_2.

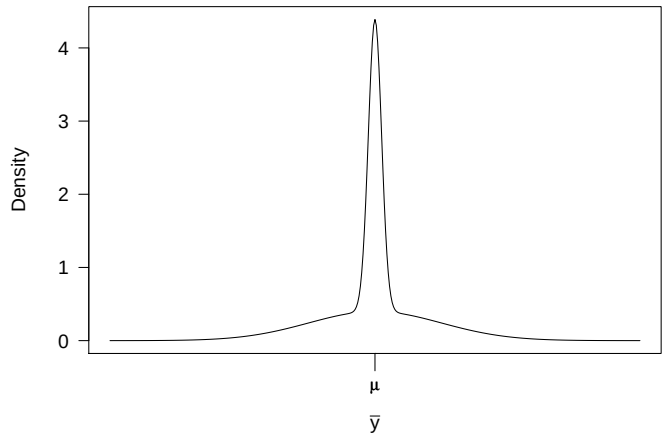


Figure 14.1: Binormal density for equal mixture of $N(\mu, 0.5^2)$ and $N(\mu, 0.05^2)$

If the conditionality principle was not used here, then the distribution of the sample mean under repetition of the experiment would be binormal (see Example 2.9). Figure 14.1 shows the shape of this binormal density assuming that the more accurate spectrometer is available 50% of the time (at random). The mean and variance of this binormal distribution are μ and 0.12625, respectively (see Exercise 13.9). □

Example 14.5. Random number of trials. In practice, the number of trials in a binomial experiment is often random. (This is equivalent to random sample size in a sequence of Bernoulli trials.) For example, to estimate the probability that a passenger car has more than one occupant, you might record the number of occupants in all passenger cars passing through an intersection for an hour. The sample size is therefore random because if you were to repeat this hour-long experiment at a different time then you would most likely have a different sample size. This is an example of a mixture experiment consisting of many different binomial experiments (having different number of trials).

To make inference about the probability of multi-occupant passenger cars you would not use the mixture experiment in which the sample size is random. Rather, you would use the binomial model with the total number of trials *fixed* and equal to

the the number of passenger cars that you observed during the hour of observation. (Also, see Exercise 14.6.) □

Box 14.1.

Chapter 9 presents further use of conditioning, in the context of coping with nuisance parameters. There, it is seen that conditioning can be appropriate even when the probabilities of observing from the component experiments (e.g., E_1 and E_2 in the statement of the conditionality principle) do depend on $\boldsymbol{\theta}$. This is because the nuisance parameters may prevent the knowledge of which experiment was performed from providing any relevant information about the parameters of interest.

14.4 The likelihood principle

This is the third and final principle. In essence, it states that if two experiments result in the same likelihood function then those two experiments provide the same information content about $\boldsymbol{\theta}$. The likelihood principle provides a rationale for discarding measurements that provide no relevant information about $\boldsymbol{\theta}$ (see the continuation of Example 14.5 below).

The likelihood principle: If $\boldsymbol{y}_1$ observed from E_1 and $\boldsymbol{y}_2$ observed from E_2 have the same likelihood functions (to within a constant), i.e $f_1(\boldsymbol{y}_1;\boldsymbol{\theta}) = cf_2(\boldsymbol{y}_2;\boldsymbol{\theta})$, $\forall \boldsymbol{\theta} \in \Theta$ then the "information content" concerning inference about $\boldsymbol{\theta}$ is the same from both $(E_1, \boldsymbol{y}_1)$ and $(E_2, \boldsymbol{y}_2)$. □

Example 14.5 ctd. In Example 14.5 it could be that additional observations were made, say, the colour of the car, gender of the driver, etc. If these observations have distribution that does not depend on the probability of multiple occupancy then the likelihood of these extra observations is subsumed by the constant c in the statement of the likelihood principle, because it does not depend on $\boldsymbol{\theta}$. □

However, it has been argued that the likelihood principle is contradictory with frequentist inference, as demonstrated by the following example.

Example 14.6. Negative binomial vs binomial. A market researcher is interested in households that have gross income within a certain range. She needs to question four such households, and a total of 12 are contacted to get the desired four. What proportion p of households are in the desired income range?

This experiment is negative binomial because it is the number of trials required to get four "successes" that is random. The likelihood function from this negative binomial experiment is

$$\binom{11}{3} p^4 (1-p)^8.$$

To within a constant, this is the same as the likelihood for a binomial where 4 successes were observed in 12 trials, and therefore the likelihood principle says the negative binomial experiment has the same information content as a binomial(12,p) experiment in which 4 successes are observed. The MLE of p is 1/3 for both of these experiments.

The frequentist would prefer to use the minimum variance unbiased estimator, should it exist. For a negative binomial requiring $k(\geq 2)$ successes, each with probability p, the MVUE of p does exist and is $\widehat{p}_u = \frac{k-1}{N-1}$ (Lehmann 1983, p.134) where N is the total number of trials. Here, $\widehat{p}_u = 3/11$. In contrast, had the experiment been binomial with $y = 4$ observed from $n = 12$ trials then the minimum variance unbiased estimator (and also MLE in this case) would be $\widehat{p} = 1/3$. □

14.4.1 Relationship with sufficiency and conditionality

It is relatively easy to show that the likelihood principle implies both the sufficiency and conditionality principles, and this is given below in Theorem 14.1. Birnbaum (1962) showed that the converse also holds, that is, the sufficiency and conditionality principles jointly also imply the likelihood principle. (Also of note, Evans, Fraser and Monette (1986) presented a modified version of the conditionality principle that by itself implies the likelihood principle.) This shattered the peace, because the sufficiency and conditionality principles were generally considered reasonable by frequentists. Indeed, frequentists make heavy use of sufficient statistics, and conditionality is fundamental to defining the experiment that (hypothetically) would be

repeated. However, the likelihood principle is less palatable to frequentists due to its potential conflict with their approach to inference, as demonstrated in Example 14.6. More generally, note that the likelihood principle is based solely on the likelihood function that results from the data that were actually observed. It makes no use of the concept of "repeatable experiments".

Theorem 14.1 The likelihood principle implies the sufficiency principle and conditionality principle.

Proof: Likelihood principle ⇒ sufficiency principle: If $T(\boldsymbol{Y})$ is a sufficient statistic then, from the Factorization Theorem and equation (14.2), the density function of $\boldsymbol{y}$ can be written

$$f(\boldsymbol{y};\boldsymbol{\theta}) = g(T(\boldsymbol{y});\boldsymbol{\theta})h(\boldsymbol{y}) = \frac{f(T(\boldsymbol{y});\boldsymbol{\theta})}{k_t}h(\boldsymbol{y}) = c(\boldsymbol{y})f(T(\boldsymbol{y});\boldsymbol{\theta})$$

where $f(T(\boldsymbol{y});\boldsymbol{\theta})$ is the density function of $T(\boldsymbol{y})$, and $c(\boldsymbol{y})$ does not depend on $\boldsymbol{\theta}$. Applying the likelihood principle, with E_1 being the experiment where $\boldsymbol{y}$ is observed from the distribution with density function $f_1 = f(\boldsymbol{y};\boldsymbol{\theta})$, and E_2 being the experiment where $T(\boldsymbol{y})$ is observed from the distribution with density function $f_2 = f(T(\boldsymbol{y});\boldsymbol{\theta})$, we conclude that the information content of $(E_1, \boldsymbol{y})$ and $(E_2, T(\boldsymbol{y}))$ is the same. This is the sufficiency principle.

Likelihood principle ⇒ conditionality principle: The mixture experiment assigns probabilities p and $1-p$ to experiments 1 and 2, respectively. Hence, under the mixture experiment, the density function for observing $\boldsymbol{y}_i$ from experiment $i = 1, 2$ is given by the probability that experiment i is performed, multiplied by the density function for $\boldsymbol{y}_i$ under experiment i. That is, the observation $(E_i, \boldsymbol{y}_i)$ has probability density function

$$p_i f_i(\boldsymbol{y}_i;\boldsymbol{\theta}) \quad , i = 1, 2,$$

where $p_1 = p$ and $p_2 = 1-p$ do not depend on $\boldsymbol{\theta}$. This density function is proportional to the density function for observing $\boldsymbol{y}_i$ from experiment E_i, $f_i(\boldsymbol{y}_i;\boldsymbol{\theta})$. The likelihood principle therefore says that the information content of $(E, (E_i, \boldsymbol{y}_i))$ is the same as that of $(E_i, \boldsymbol{y}_i)$. This is the conditionality principle. □

Birnbaum (1962) and others (e.g., Pawitan 2001) comment that it is not clear that the likelihood principle really is contrary to frequentist inference. The likelihood principle is saying that the information content is the same from two experiments if the likelihoods are equal (to within a constant). However, this does not necessarily imply that these experiments should result in the same *inference*. It may be that the manner in which the experiment could be repeated is relevant to inference, but this knowledge is not captured in the likelihood. In Example 14.6, the observed proportion is the parameter value that is most supported by the likelihood under both the negative binomial and binomial models. However, under the negative binomial model this estimator is biased, but it is unbiased under the binomial model.

The monograph of Berger and Wolpert (1988) argues in favour of the likelihood principle, and moreover, that the Bayesian paradigm is compatible with this principle. Of course, it must be noted that the Bayesian approach does not lead to identical inference from identical likelihoods if different priors are used.

Box 14.2.

At present, Bayesian inference is enjoying a resurgence of interest. This is largely due to recent computational advances rather than fundamentalism.

14.5 Statistical significance versus statistical evidence †

To conclude this chapter, the first example below is a well-known contrived example which is often used to question the common interpretation that p-values provide measures of statistical evidence. However, this is rather moot, because the real lesson from the example is that hypothesis testing should not have been performed in the first place since there is no motivation for preferring to falsify one model (Popper 1959) rather than the other. Indeed, Exercise 14.8 provides a more meaningful approach to choosing between two simple hypotheses. The second example is a continuation of Example 2.4.

Royall (1997) provides a more in-depth coverage of the relationship between statistical significance and statistical evidence, and Lehmann (2006) considers modified forms of the Neyman-Pearson test.

Neyman-Pearson theorem for simple hypotheses.

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_a : \boldsymbol{\theta} = \boldsymbol{\theta}_a$. Let $L(\boldsymbol{\theta}; \boldsymbol{y}) = f_n(\boldsymbol{y}; \boldsymbol{\theta})$ denote the likelihood function. The hypothesis test with critical region

$$C = \left\{ \boldsymbol{y} : \frac{L(\boldsymbol{\theta}_a; \boldsymbol{y})}{L(\boldsymbol{\theta}_0; \boldsymbol{y})} > k \right\} , \tag{14.3}$$

is the most powerful size α test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_a : \boldsymbol{\theta} = \boldsymbol{\theta}_a$, where k is such that $P_{\boldsymbol{\theta}_0}(\boldsymbol{Y} \in C) = \alpha$.

Example 14.7. Let $Y \in \{1, 2, 3\}$ and test $H_0 : \theta = 0$ versus $H_a : \theta = 1$ where $P_\theta(Y = y)$ is given by

	y		
	1	2	3
P_0	.009	.001	.99
P_1	.001	.989	.01

The test with critical region $\{1, 2\}$ is a most powerful test of size 0.01, and so, if $y \in \{1, 2\}$ then the p-value can be no bigger than 0.01. However, the likelihood ratio is 9 to 1 in favour of H_0 when $y = 1$ is observed!!! □

Example 2.4 ctd. Recall, that the observation Y was either from a $N(0, 1)$ distribution or $N(0, 100^2)$ distribution, but it is unknown which. To test $H_0 : Y \sim N(0, 1)$ versus $H_a : Y \sim N(0, 100^2)$ the critical region in (14.3) has the form

$$C = \{ \boldsymbol{y} : |\boldsymbol{y}| > k \} .$$

Therefore, the most powerful 5% level test rejects for all y *outside* of the interval $(-z_{0.025}, z_{0.975} \approx (-1.96, 1.96)$. Conversely, the most powerful test of $H_0 : Y \sim N(0, 100^2)$ versus $H_a : Y \sim N(0, 1)$ is rejected at the 5% level for all y *inside*

the interval $(-6.2707, 6.2707)$ (Exercise 14.7). Thus, the value of $y = 2$ would be rejected by both tests.

Using the likelihood ratio as a measure of support (Fig. 14.2), the $N(0, 1)$ model has greater support than the $N(0, 100^2)$ model for all y in $(-3.0351, 3.0351)$. □

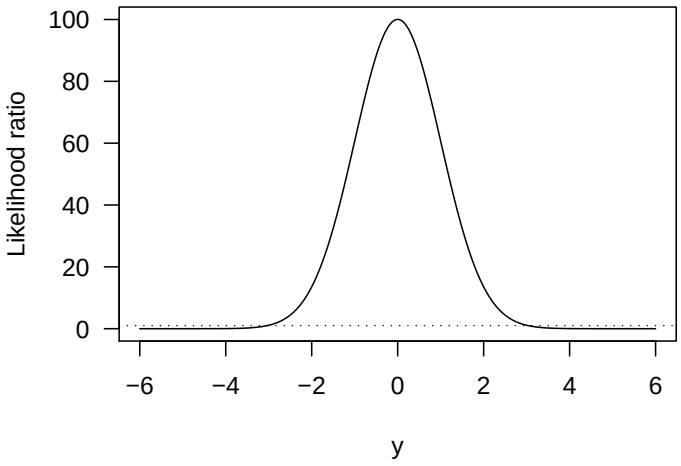


Figure 14.2: Likelihood ratio of $N(0, 1)$ and $N(0, 100^2)$ models, with a dashed line at unity. Note: Over the range of y values plotted, the likelihood of the $N(0, 100^2)$ model is sufficiently flat that the shape of the likelihood ratio is very much like that of the bell-shaped curve of the $N(0, 1)$ model.

14.6 Exercises

14.1 Let $\mathbf{Y} = Y_1, ..., Y_n$ be iid from a chi-square distribution with ν degrees of freedom. This distribution has density

$$f(y; \nu) = \frac{y^{\nu/2-1} e^{-y/2}}{\Gamma(\nu/2) 2^{\nu/2}} \quad , y > 0 \ .$$

Use the factorization theorem to obtain a univariate sufficient statistic for ν.

14.2 Let $Y_i, i = 1, ..., n$ be iid from a Pareto(α, M) distribution, for some $\alpha > 0, M > 0$ (see Exercise 2.8). Find a bivariate sufficient statistic for (α, M).

14.3 Let Y (possibly vector valued) be iid with exponential family density of the form

$$f(y; \boldsymbol{\theta}) = c(y) k(\boldsymbol{\theta}) \exp \left\{ \sum_{j=1}^{s} \theta_j t_j(y) \right\} \ . \tag{14.4}$$

For an iid sample, $Y_i, i = 1, ..., n$, from this distribution, use the factorization theorem to establish that $(T_1(\boldsymbol{y}), ..., T_s(\boldsymbol{y}))$ is a sufficient statistic for $\boldsymbol{\theta} = (\theta_1, ..., \theta_s)$ where $T_j(\boldsymbol{y}) = \sum_{i=1}^{n} t_j(y_i)$.

14.4 Under Hardy-Weinberg equilibrium, alleles A and B occur independently and thus the probabilities of genotypes AA, AB and BB are $p^2, 2p(1-p)$ and $(1-p)^2$, respectively, where p is the probability of allele type A. Let n be the number of genotypes measured and let (N_1, N_2, N_3) (where $N_3 = n - N_1 - N_2$) be the observed number of AA, AB and BB alleles, respectively. Use the factorization theorem to show that the number of A alleles observed, $T = 2N_1 + N_2$, is sufficient for p.

14.5 In Example 14.2 the factorization theorem was used to show that $T = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for p when Y_i are iid Bernoulli(p) random variables. What is the distribution of $(Y_1, Y_2, ..., Y_n)|T$?

14.6 Let Y_1 and Y_2 be independent Poisson random variables, with means parameterized as $p\lambda$ and $(1-p)\lambda$ respectively for $0 < p < 1$ and $\lambda \in \mathbb{R}^+$. Show that $Y_1 + Y_2 \sim$ Poisson(λ) and hence that $Y_1|(Y_1 + Y_2 = n) \sim$ binomial(n, p). If interest lies only in parameter p, argue that the information content from observing Y_1, Y_2 is the same as that from observing $Y_1|Y_1 + Y_2$.

14.7 For the most powerful test of $H_0 : Y \sim N(0, 100^2)$ versus $H_a : Y \sim N(0, 1)$ show that $C = (-100z_{0.475}, 100z_{0.475})$.

14.8 Given the observation of $\boldsymbol{y}$, it is required to choose between the two hypotheses $\boldsymbol{Y} \sim f(\boldsymbol{Y}; \theta_0)$ and $\boldsymbol{Y} \sim f(\boldsymbol{Y}; \theta_1)$, by specifying a region, C, such that H_0 is chosen if $\boldsymbol{y} \in C$, else H_1 is chosen. Let $p_i, i = 0, 1$ denote the probability of falsely choosing H_i when H_{1-i} is true.

1. Show that the region

$$C = \{\boldsymbol{y} : f(\boldsymbol{y}; \theta_0) > f(\boldsymbol{y}; \theta_1)\} \tag{14.5}$$

minimizes $p_0 + p_1$.

2. Show that the region obtained by replacing the strict inequality in (14.5) by a non-strict inequality also minimized $p_0 + p_1$.

That is, the sum of error probabilities is minimized by choosing the hypothesis that has greater support from the likelihood.

Chapter 15

Miscellaneous

15.1 Notation

Θ	Parameter space
Θ_0	Reduced parameter space under a null hypothesis
$L(\theta)$	Likelihood function
$l(\theta)$	Log-likelihood function
$\mathbb{R}$	Real numbers
$\mathbb{R}^+$	Positive reals, i.e., $(0, \infty)$.
$\sim$	Distributed as
$\stackrel{\sim}{}$	Approximately distributed as
$\stackrel{=}{}_D$	Equal in distribution
$\approx_D$	Approximately equal in distribution
$\stackrel{\rightarrow}{}_D$	Converges in distribution
$\stackrel{\rightarrow}{}_p$	Converges in probability
E_θ	Abbreviation for $E_{\boldsymbol{Y}\mid\theta}$
$E_{\boldsymbol{Y}\mid\theta}$	Expectation with respect to $f(\boldsymbol{y};\theta)$
Var_θ	Abbreviation for $Var_{\boldsymbol{Y}\mid\theta}$
$Var_{\boldsymbol{Y}\mid\theta}$	Variance with respect to $f(\boldsymbol{y};\theta)$
$\chi^2_{r,q}$	The q quantile of a χ^2_r distribution, i.e., if $X \sim \chi^2_r$ then $P(X \leq \chi^2_{r,q}) = q$.
z_q	The q quantile of a $N(0,1)$ distribution, e.g., $z_{0.975} \approx 1.96$.
$\Gamma(\alpha)$	Gamma function (see Fig. 6.5).

The following table shows the Greek symbols used within this text.

α	alpha	β	beta	γ	gamma	δ	delta
ϵ	epsilon	ε	epsilon (variation)	ζ	zeta	η	eta
θ	theta	λ	lambda	μ	mu	ν	nu
π	pi	ρ	rho	σ	sigma	τ	tau
ϕ	phi	χ	chi	ψ	psi	ω	omega
Θ	Theta	Φ	Phi	Ψ	Psi	Ω	Omega

15.2 Do you think like a frequentist or a Bayesian?

The customary way to toss a coin is to flip it into the air and catch it with one hand, and to place that hand palm down over the other hand with the coin hidden. The required levels of suspense and theatrics are obtained by the delay, and flair, of lifting the catching hand to reveal the upturned face of the coin.

It will assumed that the coin is balanced and tossed in a fair way, so that heads and tails have equal probability of 0.5.

Question: The coin has been tossed and caught, and is being held in the hidden position between the two hands. What is the probability that the coin shows heads when the catching hand is lifted?

See the Solutions Chapter for diagnosis.

15.3 Useful distributions

Most of the distributions used in this text are listed below, with a very brief explanation of their typical use. An authoritative reference on a vast assortment of distributions is provided by Johnson, Kotz and Kemp (1992), Johnson, Kotz and Balakrishnan (1994) and Johnson, Kotz and Balakrishnan (1995).

15.3.1 Discrete distributions

Bernoulli. Bernoulli(p), $0 \leq p \leq 1$:

A random variable taking the value 1 or 0, with probabilities p and $(1-p)$ respectively. Used to model a single trial from an experiment with only two possible outcomes such as Yes/No, Success/Fail, Heads/Tails, etc.

Density: $f(y;p) = p^y(1-p)^{1-y}, \quad y = 0, 1$

Mean: $E[Y] = p$

Variance: $\text{var}(Y) = p(1-p)$

Binomial. $\text{Bin}(n, p),\ 0 \le p \le 1$:

A random variable taking an integer value between 0 and n, corresponding to the number of 1's observed from n iid Bernoulli trials.

Density: $f(y; p) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y}, \quad y = 0, 1, ..., n$

Mean: $E[Y] = np$

Variance: $\text{var}(Y) = np(1-p)$

Multinomial process. $\text{MultProc}(p_1, ..., p_s),\ 0 \le p_i \le 1,\ \sum_{i=1}^{s} p_i = 1$:

A generalization of the Bernoulli to $s(\ge 3)$ possible outcomes. The multinomial process random variable Y takes an integer value between 1 and s with probabilities $p_1, ..., p_s$, respectively. Used to model a single trial from an experiment with s possible outcomes (e.g., Yes, No, or Undecided).

Density: $f(y; p_1, ..., p_s) = p_y, \quad y = 1, ..., s$

Mean[1]: $E[Y] = \sum_{i=1}^{s} i p_i$

Variance[1]: $\text{var}(Y) = \sum_{i=2}^{s} i(i-1) p_i$

Multinomial. $\text{Mult}(n, p_1, ..., p_s),\ 0 \le p_i \le 1,\ \sum_{i=1}^{s} p_i = 1$:

A random vector, $\boldsymbol{Y} = (Y_1, ..., Y_s)$ of length s, generated from n iid trials of a Multinomial process. The value of Y_i is the number of trials that produced outcome i.

Density: $f(y_1, ..., y_s; p_1, ..., p_s) = \frac{n!}{y_1! ... y_s!} p_1^{y_1} ... p_s^{y_s}, \quad y_i = 0, ..., n, \quad \sum_{i=1}^{s} y_i = n$

Mean: $E[Y_i] = np_i$

Variance: $\text{var}(Y_i) = np_i(1-p_i),\ \text{cov}(Y_i, Y_j) = -np_i p_j, i \ne j$

[1] In many situations, the multinomial-process random variable will be a factor variable. That is, the values 1,...,s are simply labels for the possible outcomes (e.g. 1 ≡ Yes, 2 ≡ No, 3 ≡ Undecided). In that case, the mean and variance are not meaningful.

Poisson. Pois(λ), $\lambda > 0$:

A random variable taking a non-negative integer value. The Poisson distribution is commonly used to model count data. However, it is often the case that the Poisson model will be inadequate, in which case a more flexible alternative (e.g., negative binomial) may be required.

Density: $f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, ...$
Mean: $E[Y] = \lambda$
Variance: $\text{var}(Y) = \lambda$

Negative binomial. NB(m, p), $m > 0$, $0 \leq p \leq 1$:

A random variable taking a non-negative integer value. Often used to model over-dispersed count data.

Density: $f(y;m,p) = \frac{\Gamma(y+m)}{\Gamma(m)y!}p^m(1-p)^y, \quad y = 0, 1, ...$
Mean: $E[Y] = \frac{m(1-p)}{p}$
Variance: $\text{var}(Y) = \frac{m(1-p)}{p^2} = E[Y]\left(1 + \frac{E[Y]}{m}\right)$

The Poisson is obtained in the limit as $m \to \infty$ and $p \to 0$ such that $\mu = E[Y] = m(1-p)/p$ converges to a fixed limit.

When m is a positive integer, then Y is distributed as the number of Bernoulli(p) trials that result in a 0 before m 1's are observed. The geometric distribution is the special case of the negative binomial where $m = 1$.

The negative binomial density is sometimes parameterized using m and μ. Since $p = m/(\mu + m)$, this gives

$$f(y;m,\mu) = \frac{\Gamma(y+m)}{\Gamma(m)y!}\left(\frac{m}{\mu+m}\right)^m\left(\frac{\mu}{\mu+m}\right)^y . \tag{15.1}$$

For a constant value of m, negative binomial data, $y_1, ..., y_n$, can be modeled using generalized linear models (e.g., see Section 7.6). With constant m, note that var(Y) is a quadratic function of μ.

An alternative is to allow m to vary for each i, such that $p = m_i/(\mu_i + m_i)$ is constant. This parameterization results in a linear mean-variance relationship $\text{var}(Y) = \mu/p$. However, this is a more challenging model to fit.

Hypergeometric. $\mathrm{H}(r, n, m)$:

The $\mathrm{H}(r, n, m)$ random variable is distributed as the number of white balls removed from an urn containing n white and m black balls, when $r(\leq n+m)$ balls are removed by random sampling *without* replacement. Clearly, Y can not exceed the smaller of r and n. Also, if $r > m$ then at least $r - m$ balls must be white.

Density: $f(y; r, n, m) = \dfrac{\binom{n}{y}\binom{m}{r-y}}{\binom{n+m}{r}}, \quad \max(0, r-m) \leq y \leq \min(r, n)$

Mean: $E[Y] = rp$

Variance: $\mathrm{var}(Y) = rp(1-p)(n+m-r)/(n+m-1)$

where $p = n/(n+m)$ is the proportion of white balls in the urn.

15.3.2 Continuous distributions

Uniform. $\mathrm{U}(a, b),\ b > a$:

A random variable with density that is constant between a and b, and zero otherwise.

Density: $f(y; a, b) = \frac{1}{b-a}, \quad a \leq y \leq b$

Mean: $E[Y] = (a+b)/2$

Variance: $\mathrm{var}(Y) = (b-a)^2/12$

Normal. $\mathrm{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma > 0$:

Also know as the Gaussian distribution. In many ways, this is the "default" distribution for continuous data because of the well developed statistical methods for normally distributed data (e.g., regression, ANOVA, etc.) Moreover, this is the limiting distribution of the central limit theorem, which guarantees approximate normality of calculated statistics (e.g., sample means, MLEs) under appropriate regularity conditions.

Density: $f(y; \mu, \sigma^2) = \frac{e^{-(y-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}, \quad y \in \mathbb{R}$

Mean: $E[Y] = \mu$

Variance: $\text{var}(Y) = \sigma^2$

Lognormal. $\text{LN}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma > 0$:

A lognormal random variable is so named because its log is normal. That is, the $\text{LN}(\mu, \sigma^2)$ random variable Y is such that $\log(Y)$ is distributed $\text{N}(\mu, \sigma^2)$. The lognormal is convenient for the modeling of right-skew data $Y_i \in \mathbb{R}^+$, because this is equivalent to modeling $\log(Y_i)$ as normally distributed.

Density: $f(y; \mu, \sigma^2) = \frac{e^{-(\log(y)-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma y}, \quad y \in \mathbb{R}$

Mean: $E[Y] = e^{\mu+\sigma^2/2}$

Variance: $\text{var}(Y) = (e^{\sigma^2} - 1)E[Y]^2$

Exponential. $\text{Exp}(\mu),\ \mu > 0$:

Commonly used to model the duration of time until an event occurs, e.g., often used in survival analysis to model the time until failure.

Density: $f(y; \mu) = \frac{e^{-y/\mu}}{\mu}, \quad 0 \leq y$

Mean: $E[Y] = \mu$

Variance: $\text{var}(Y) = \mu^2$

The exponential density is often re-parameterized using the rate parameter $\lambda = 1/\mu$.

Chi-square. $\chi^2_r,\ r > 0$:

The chi-square distribution arises from quadratic forms of normally distributed random variables, such as residual sums-of-squares. This is the property that results in many familiar test statistics having a (possibly approximate) chi-square distribution.

Density: $f(y; r) = \frac{y^{r/2-1}e^{-y/2}}{2^{r/2}\Gamma(r/2)}, \quad 0 \leq y$

Mean: $E[Y] = r$

Variance: $\text{var}(Y) = 2r$

If $Z \sim N(\mu, \sigma^2)$ then

$$\frac{(Z-\mu)^2}{\sigma^2} \sim \chi^2_1 \ .$$

In the multidimensional case, if $\boldsymbol{Z} \sim N_r(\boldsymbol{\mu}, \Sigma)$ then (Seber and Lee 2003, p. 30)

$$(\boldsymbol{Z}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{Z}-\boldsymbol{\mu}) \sim \chi^2_r \ . \tag{15.2}$$

Gamma. Gamma(α, β), $\alpha > 0$, $\beta > 0$:

The gamma family of distributions provides a lesser-used alternative to the lognormal for the modeling of right-skewed data on $\mathbb{R}^+$. It's main relevance is that the family of gamma distributions includes exponential and χ^2 distributions.

In the parameterization of the Gamma(α, β) density used below, α is the shape parameter and β is the scale parameter (i.e., the mean and standard deviation are proportional to β). The gamma is also sometimes parameterized using α and $\beta^* = 1/\beta$, where β^* is the rate parameter.

Density: $f(y; \alpha, \beta) = \frac{y^{\alpha-1} e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)}, \quad 0 \le y$
Mean: $E[Y] = \alpha\beta$
Variance: $\text{var}(Y) = \alpha\beta^2$

An Exp(μ) distribution corresponds to a Gamma$(1, \mu)$ distribution, and a χ^2_p distribution corresponds to a Gamma$(p/2, 2)$ distribution.

F-distribution. F_{r_1,r_2}, $r_1 > 0$, $r_2 > 0$:

The F-distribution arises as the ratio of chi-squares that have been divided by their degrees of freedom. Specifically, if Y_1 and Y_2 are $\chi^2_{r_1}$ and $\chi^2_{r_2}$ distributed then

$$F = \frac{Y_1/r_1}{Y_2/r_2} \sim F_{r_1,r_2} \ .$$

As r_2 tends to infinity, F_{r_1,r_2} converges to the distribution of a $\chi^2_{r_1}/r_1$ random variable. The F_{1,r_2}-distribution is equivalent to the square of a t_{r_2}-distribution.

Density: $f(y; r_1, r_2) = \frac{(r_1/r_2)^{r_1/2}}{B(r_1/2, r_2/2)} \frac{y^{r_1/2-1}}{(1+r_1 y/r_2)^{(r_1+r_2)/2}}$

Mean: $E[Y] = r_2/(r_2-2)\ , \quad r_2 > 2$

Variance: $\text{var}(Y) = \frac{2r_2^2(r_1+r_2-2)}{r_1(r_2-2)^2(r_2-4)}$

where $B()$ denotes the beta function.

Multivariate normal. $N_s(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^s$, $\boldsymbol{\Sigma} \in \mathbb{R}^{s \times s}$ is positive definite:

The multivariate normal is the approximating distribution (subject to regularity conditions) of vector-valued MLEs $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^s, s \geq 2$. Here, $\boldsymbol{\mu} = \mu_1, ..., \mu_s$ and $\boldsymbol{\Sigma}$ is the $s \times s$ variance matrix with i, j element denoted σ_{ij}^2.

Density: $f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi}^s |\boldsymbol{\Sigma}|^{1/2}} e^{-(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}-\boldsymbol{\mu})}$

Mean: $E[Y_i] = \mu_i$

Variance: $\text{var}(Y_i) = \sigma_{ii}^2, \ \text{cov}(Y_i, Y_j) = \sigma_{ij}^2$

15.4 Software extras

15.4.1 R function `Plkhci` for likelihood ratio confidence intervals

At the time of writing, the current version of the `plkhci` function in library `Bhat` does not accept additional arguments to be passed to the negative log-likelihood function. This is useful functionality because the additional arguments will typically include the data. Where this flexibility is required, this text uses function `Plkhci`.

Function `Plkhci` is defined in the text file `Plkhci.R`, which can be downloaded from www.stat.auckland.ac.nz/~millar. It is obtained from a very minor modification to the `plkhci` function. Specifically, the first two lines of `plkhci` are

```
function (x, nlogf, label, prob=0.95, eps=0.001, nmax=10, nfcn=0)
{
```

and in `Plkhci` these are replaced by

```
function (x, nllhood, label, prob=0.95, eps=0.001, nmax=10, nfcn=0,...)
{
    nlogf=function(x) nllhood(x,...)
```

15.4.2 R function `Profile` for calculation of profile likelihoods

The `Profile` function calculates the profile log-likelihood

$$l^*(\boldsymbol{\psi}; \boldsymbol{y}) \equiv \max_{\boldsymbol{\lambda}} \; l(\boldsymbol{\psi}, \boldsymbol{\lambda}; \boldsymbol{y}) \; . \tag{15.3}$$

The profile parameter, $\boldsymbol{\psi}$, can be any subset of θ.

By way of example, in the binormal mixture model with parameters $\boldsymbol{\theta} = (p, \mu, \sigma, \nu, \tau)$ labeled as `parnames = c("p","mu","sigma","nu","tau")`, and negative log-likelihood function `nllhood` (Section 3.3.3), suppose that $\psi = (\sigma, \tau)$. The profile log-likelihood at $\boldsymbol{\psi} = (5, 7)$ is

```
> Profile(parnames,nllhood,label=c("sigma","tau"),psi=c(5,7),
+          lambda=c(0.5,55,80),y=waiting)$value
[1] -1037.363
```

where argument `label` specifies the subset of parameter names in `parnames` that correspond to $\boldsymbol{\psi}$, Argument `lambda` gives the start values for the maximization in (15.3), in the order in which they appear in `parnames`. Here, $\boldsymbol{\lambda} = (p, \mu, \nu)$.

The code

```
> Profile(parnames,nllhood,label=c("p","mu","sigma"),psi=c(1/3,55,5),
+          lambda=c(80,5),y=waiting)$value
[1] -1036.829
```

provides easier implementation of the likelihood ratio test in Section 3.4.1.

15.4.3 SAS macro `Plkhci` for profile likelihood confidence intervals

For calculation of likelihood-ratio confidence intervals. Useful for users who do not have access to `PROC NLP`, and/or require the functionality of `PROC NLMIXED`. This macro repeatedly calls a second user-specified macro in which the model is fitted in `PROC NLMIXED` with the parameter of interest passed as an argument to that macro. Macro `Plkhci` uses the bisection method to calculate the likelihood ratio CI at the desired level.

An example of general use is provided in Section 3.4.1. The first argument to `Plkhci` is the name of the user-specified macro. It searches within the range specified by its next two arguments. The fourth argument is the maximal value of the

log-likelihood, $l(\widehat{\boldsymbol{\theta}})$. The fifth argument, is used to specify whether it the left side (`side="L"`) or right side (`side="R"`) bound that is desired. Optional argument `alpha` specifies that a $(1-\alpha)100\%$ confidence interval is desired, and optional argument `tol` specifies the desired convergence tolerance of the confidence interval bound.

15.4.4 SAS macro `Profile` for calculation of profile likelihoods

The `Profile` macro can be used to calculate the profile log-likelihood $l^*(\psi)$ for individual model parameters. That is, where $\psi = \theta_k$ for some k. It uses the same user-defined macro mentioned in the description of the `Plkhci` macro. An example of use is provided in Section 4.5.1.

15.4.5 SAS macro `DeltaMethod` for application of the delta method

For application of the delta method to general transformation of the parameters, $g(\boldsymbol{\theta}) : \mathbb{R}^s \rightarrow \mathbb{R}^p$ for $p \leq s \leq 2$. The `DeltaMethod` macro constructs a call to `PROC NLMIXED` in which the `ESTIMATE` statement is used to obtain the variance of $g(\boldsymbol{\theta})$.

Bibliography

ADMB-project: 2008a, *An introduction to AD Model Builder version 9.0.0 for use in nonlinear modeling and statistics*, admb-project.org/documentation.

ADMB-project: 2008b, *Random effects in AD model builder: ADMB-RE user guide*, admb-project.org/documentation.

Agresti, A. and Coull, B. A.: 1998, Approximate is better than "exact" for interval estimation of binomial proportions, *Amer. Statist.* **52**, 119–126.

Akaike, H.: 1974, A new look at the statistical model identification, *IEEE Trans. Automat. Contr.* **15**, 716–723.

Anonymous: 2003, *R: A language and environment for statistical computing*, R Development Core Team, Vienna.

Beitler, P. J. and Landis, J. R.: 1985, A mixed-effects model for categorical data, *Biometrics* **41**, 991–1000.

Berger, J. O.: 1985, *Statistical decision theory and Bayesian analysis*, 2nd edn, Springer-Verlag, New York.

Berger, J. O., Liseo, B. and Wolpert, R. L.: 1999, Integrated likelihood methods for eliminating nuisance parameters, *Stat. Sci.* **14**, 1–28.

Berger, J. O. and Wolpert, R. L.: 1988, *The likelihood principle, 2nd edition*, Vol. 6, IMS Lecture Notes - Monograph Series.

Billingsley, P.: 1979, *Probability and measure*, Wiley, New York.

Birnbaum, A.: 1962, On the foundations of statististical inference (with discussion), *J. Amer. Stat. Assoc.* **57**, 269–326.

Bjørnstad, J. F.: 1996, On the generalization of the likelihood function and the likelihood principle, *J. Amer. Stat. Assoc.* **91**, 791–806.

Bøhning, D., Dietz, E. and Schlattmann, P.: 1999, The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology, *AmerStat.* **46**, 327–333.

Borchers, D. L., Buckland, S. T. and Zucchini, W.: 2002, *Esimating Animal Abundance: Closed Populations*, Springer-Verlag, London.

Box, G. E. P. and Cox, D. R.: 1964, An analysis of transformations (with discussion), *J. Roy. Stat. Soc. Series B* **26**, 211–252.

Breslow, N. E. and Clayton, D. G.: 1993, Approximate inference in generalized linear mixed models, *J. Amer. Stat. Assoc.* **88**, 9–25.

Breslow, N. E. and Lin, X.: 1995, Bias correction in generalised linear mixed models with a single component of dispersion, *Biometrika* **82**, 81–91.

Brown, L. D., Cai, T. T. and DasGupta, A.: 2001, Interval estimation for a binomial proportion (with discussion), *Statistical Science* **2**, 101–133.

Brown, L. D., Cai, T. T. and DasGupta, A.: 2002, Confidence intervals for a binomial proportion and asymptotic expansions, *Ann. Statist.* **30**, 160–201.

Burnham, K. P. and Anderson, D. R.: 2002, *Model Selection and Multimodel Inference - A Practical Information-Theoretic Approach*, 2nd edn, Springer-Verlag, New York.

Carey, V., Zeger, S. L. and Diggle, P.: 1993, Modelling multivariate binary data with alternating logistic regressions, *Biometrika* **80**, 517–526.

Casella, G. and Berger, R. L.: 1990, *Statistical inference*, Duxbury Press, California.

Chambers, E. A. and Cox, D. R.: 1967, Discrimination between alternative binary response models, *Biometrika* **54**, 573–578.

Chernick, M. R.: 2008, *Bootstrap methods: a guide for practitioners and researchers*, 2nd edn, Wiley, Hoboken, New Jersey.

Clark, J. R.: 1957, Effect of length of haul on cod end escapement, *Technical Report Paper S25*, ICNAF/ICES/FAO Workshop on Selectivity, Lisbon.

Collett, D.: 1991, *Modelling binary data*, Chapman and Hall, London.

Cormack, R. M.: 1992, Interval estimation for mark-recapture studies of closed populations, *Biometrics* **48**, 567–576.

Cox, D. R.: 1972, Regression models and life tables (with discussion), *J. Roy. Stat. Soc. Series B* **34**, 187–220.

Cox, D. R.: 1975, Partial likelihood, *Biometrika* **62**, 269–276.

Crainiceanu, C. M. and Ruppert, D.: 2004, Likelihood ratio tests in linear mixed models with one variance component, *J. R. Statist. Soc. B* **66**, 165–185.

Davison, A. C. and Hinkley, D. V.: 1997, *Bootstrap methods and their applications*, Cambridge University Press, New York.

Delwiche, L. D. and Slaughter, S. J.: 2003, *The little SAS book: A primer*, 3rd edn, SAS Institute Inc., Cary, N.C.

Dempster, A. P., Laird, N. M. and Rubin, D. B.: 1977, Maximum likelihood from incomplete data via the em algorithm, *J. Roy. Stat. Soc. Series B* **39**, 1–38.

Der, G. and Everitt, B. S.: 2009, *A handbook of statistical analyses using SAS*, 3rd edn, Chapman and Hall/CRC, Boca Raton, FL.

Draper, N. R. and Smith, H.: 1981, *Applied regression analysis*, 2nd edn, Wiley, New York.

Edwards, A. W. F.: 1972, *Likelihood*, Cambridge University Press, Cambridge.

Efron, B.: 1979, Bootstrap methods: another look at the jackknife, *Ann. Statist.* **7**, 1–26.

Efron, B.: 1987, Better bootstrap confidence intervals (with discussion), *J. Amer. Stat. Assoc.* **82**, 171–200.

Efron, B. and Hinkley, D. V.: 1978, Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information, *Biometrika* **65**, 457–487.

Efron, B. and Tibshirani, R. J.: 1993, *An introduction to the bootstrap*, Chapman and Hall, New York.

Evans, M. A., Kim, H. and O'Brien, T. E.: 1996, An application of profile-likelihood based confidence interval to capture-recapture estimators, *J. Agric. Biol. Env. Stat.* **1**, 131–140.

Evans, M. J., Fraser, D. A. S. and Monette, G.: 1986, On principles and arguments to likelihood, *Can. J. Statist.* **14**, 181–199.

Everitt, B. S. and Hothorn, T.: 2006, *A handbook of statistical analyses using R*, Chapman and Hall/CRC, Boca Raton, FL.

Faraggi, D., Izikson, P. and Reiser, B.: 2003, Confidence intervals for the 50 per cent response dose, *Statist. Med.* **22**, 1977–1988.

Fears, T. R., Benichou, J. and Gail, M. H.: 1996, A reminder of the fallibility of the wald statistic, *Amer. Stat.* **50**, 226–227.

Felsenstein, J.: 1981, Evolutionary trees from dna sequences: A maximum likelihood approach, *Journal of Molecular Evolution* **17**, 368–376.

Fieller, E. C.: 1954, Some problems in interval estimation, *J. Roy. Stat. Soc. Series B* **16**, 175–185.

Finney, D. J.: 1971, *Probit analysis*, 3rd edn, Cambridge University Press, London.

Fisher, R. A.: 1912, On an absolute criterion for fitting frequency curves, *Messeng. Math.* **41**, 155–160.

Fisher, R. A.: 1921, On the 'probable error' of a coefficient of correlation deduced from a small sample, *Metron* **1**, 3–32.

Fisher, R. A.: 1933, The concepts of inverse probability and fiducial probability referring to unknown parameters, *Proc. Roy. Soc. A* **139**, 343–348.

Freireich, E. J. et al.: 1963, The effect of 6-mercaptopurine on the duration of steroid-induced remissions of acute leukemia: A model for evaluation of other potentially useful therapy, *Blood* **21**, 699–716.

Fushiki, T., Komaki, F. and Aihara, K.: 2004, On parametric bootstrapping and bayesian prediction, *Scand. J. Stat.* **31**, 403–416.

Gail, M. H., Fears, T. R., Hoover, R. N., Chandler, D. W., Donaldson, J. L., Hyer, M. B., Pee, D., Ricker, W. V., Siiteri, P. K., Stanczyk, F. Z., Vaught, J. B. and Ziegler, R. G.: 1996, Reproducibility studies and interlaboratory concordance for assays of serum hormone levels: estrone, estradiol, estrone sulfate, and progesterone, *Cancer Epidemiology, Biomarkers & Prevention.* .

Gelman, A. and Hill, J.: 2007, Data analysis using regression and multilevel/hierarchical models, p. 625.

Ghosh, M., Datta, G. S., Kim, D. and Sweeting, T. J.: 2006, Likelihood-based inference for the ratios of regression coefficients in linear models, *Annals Inst. Stat. Math.* **58**, 457–473.

Golub, G. H. and Pereyra, V.: 1973, The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, *SIAM Journal on Numerical Analysis* **10**, 413–432.

Goodman, L. A.: 1960, On the exact variance of products, *J. Amer. Stat. Assoc.* **55**, 708–713.

Gribben, P. E., Helson, J. and Millar, R. B.: 2004, Population abundance estimates of the new zealand geoduck clam, *panopea zelandica*, using north american methodology: is the technology transferable, *Shellfish Res.* **23**, 683–692.

Hall, P., Peng, L. and Tajvidi, N.: 1999, On prediction intervals based on predictive likelihood or bootstrap methods, *Biometrika* **86**, 871–880.

Hardin, J. W. and Hilbe, J. M.: 2003, *Generalized estimating equations*, Chapman and Hall, New York.

Harrell, F. E.: 2001, *Regression modeling strategies with applications to linear models logistic regression, and survival analysis*, Springer-Verlag, New York.

Harris, I. R.: 1989, Predictive fit for natural exponential families, *Biometrika* **76**, 675–684.

Harville, D. A.: 1974, Bayesian inference for variance components using only error contrasts, *Biometrika* **61**, 383–385.

Harville, D. A.: 1977, Maximum likelihood approaches to variance component estimation and to related problems, *J. Amer. Stat. Assoc.* **72**, 320–340.

Hathaway, R. J.: 1985, A constrained formulation of maximum-likelihood estimation for normal mixture distributions, *Ann. Statist.* **13**, 795–800.

Heyde, C. C.: 1997, *Quasi-likelihood and its application: A general approach to optimal parameter estimation*, Springer-Verlag, New York.

Hoadley, B.: 1971, Asymptotic properties of maximum likelihood estimators for the independent not indentically distributed case, *Ann. Math. Statist.* **42**, 1977–1991.

Hurlbert, S. H.: 1984, Pseudoreplication and the design of ecological field experiments, *Ecological Monographs* **54**, 187–211.

Hutton, J. L. and Monaghan, P. F.: 2002, Choice of parametric accelerated life and proportional hazards models for survival data: asymptotic results, *Lifetime Data Analysis* **8**, 375–393.

ICES: 1979, *Reports of the ICES Advisory Committee on fisheries management, 1978. ICES Co-operative Research Report 85*, International Council for the Exploration of the Sea, Charlottenlund, Denmark.

Ihaka, R. and Gentleman, R.: 1996, R: A language for data analysis and graphics, *J. Comput. Graphical Stat.* **5**, 299–314.

Jamshidian, M. and Jennrich, R. I.: 2000, Standard errors for em estimation, *J. Roy. Stat. Soc. Series B* **62**, 257–270.

Joe, H. and Xu, J. J.: 1996, The estimation method of inference functions for margins for multivariate models, *Technical Report 166*, Department of Statistics, University of British Columbia. 21 pp.

Johnson, N. L., Kotz, S. and Balakrishnan, N.: 1994, *Continuous univariate distributions*, Vol. 1, 2nd edn, Wiley, New York.

Johnson, N. L., Kotz, S. and Balakrishnan, N.: 1995, *Continuous univariate distributions*, Vol. 2, 2nd edn, Wiley, New York.

Johnson, N. L., Kotz, S. and Kemp, A. W.: 1992, *Univariate discrete distributions*, 2nd edn, Wiley, New York.

Jørgensen, B.: 1993, A review of conditional inference: is there a universal definition of nonformation?, *Bull. Intl. Stat. Inst.* **55**, 323–340.

Kauermann, G. and Carroll, R. J.: 2001, A note on the efficiency of sandwich covariance matrix estimation, *J. Amer. Stat. Assoc.* **96**, 1387–1396.

Kleinbaum, D. G. and Klein, M.: 2005, *Survival analysis: A self-learning text*, 2nd edn, Springer, New York.

Kullback, S.: 1959, *Information theory and statistics*, Wiley, New York.

Laird, N. M. and Ware, J. H.: 1982, Random-effects models for longitudinal data, *Biometrics,* **38**, 963–974.

Lange, K.: 2002, *Mathematical and statistical methods for genetic analysis*, 2nd edn, Springer, New York.

Lehmann, E. L.: 1983, *Theory of Point Estimation*, Wiley, New York.

Lehmann, E. L.: 2006, *On likelihood ratio tests*, IMS Lecture Notes, Vol 49, Institute of Mathematical Statistics, Haywood, CA, pp. 1–8.

Liang, K. Y. and Zeger, S. L.: 1986, Longitudinal data analysis using generalized linear models, *Biometrika.* **73**, 13–22.

Lindstrom, M. J. and Bates, D. M.: 1990, Nonlinear mixed effects models for repeated measures data, *Biometrics* **46**, 673–687.

Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D.: 1996, *SAS System for Mixed Models*, SAS Institute Inc., Cary, NC.

Lo, Y.: 2005, Likelihood ratio tests of the number of components in a normal mixture with unequal variances, *Stat. & Prob. Letters* **71**, 225–235.

Louis, T. A.: 1982, Finding the observed information matrix using the *em* algorithm, *J. Roy. Stat. Soc. Series B* **44**, 226–233.

Ludwig, D.: 1996, Uncertainty and the assessment of extinction probabilities, *Ecol. Appl.* **6**, 1067–1076.

Manly, B. F. J.: 1997, *Randomization, bootstrap and Monte Carlo methods in biology*, 2nd edn, Chapman and Hall, London.

Marin, J. A., Jones, O. P. and Hadlow, W. C. C.: 1993, Micropropogation of columnar apple trees, *The journal of horticultural science* **68**, 289–297.

McCullagh, P. and Nelder, J. A.: 1989, *Generalized linear models*, 2nd edn, Chapman and Hall, New York.

McCulloch, C. E. and Searle, S. R.: 2001, *Generalized, linear, and mixed models*, John Wiley, New York.

McDonald, J. H., Verrelli, B. C. and Geyer, L. B.: 1996, Lack of geographic variation in anonymous nuclear polymorphisms in the american oyster, *crassostrea virginica*, *Molecular Biol. Evol.* **13**, 1114–1118.

McLachlan, G. J.: 1987, On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, *Appl. Statist.* **36**, 318–324.

McLachlan, G. J. and Krishnan, T.: 2008, *The EM algorithm and extensions*, 2nd edn, Wiley, New Jersey.

Meinhold, R. J. and Singpurwalla, N. D.: 1983, Understanding the kalman filter, *Amer. Stat.* **37**, 123–127.

Millar, R. B.: 1992, Estimating the size-selectivity of fishing gear by conditioning on thetotal catch, *J. Amer. Stat. Assoc.* **87**, 962–968.

Millar, R. B.: 2004, Simulated maximum likelihood applied to non-gaussian and nonlinear mixed effects and state-space models, *Aust. N. Z. J. Stat.* **46**, 543–554.

Millar, R. B. and Willis, T. J.: 1999, Estimating the relative density of snapper in and around a marine reserve using a log-linear mixed effects model, *Aust. NZ. J. Stat.* **41**, 383–394.

Navidi, W.: 1997, A graphical illustration of the EM algorithm, *The American Statistician* **51**, 29–31.

Nocedal, J. and Wright, S. J.: 2006, *Numerical optimization*, 2nd edn, Springer, New York.

Noh, M. and Lee, Y.: 2007, Reml estimation for binary data in glmms, *J. Multivariate Analysis,* **98**, 896–915.

Pace, L. and Salvan, A.: 1997, *Principles of Statistical Inference from a Neo-Fisherian Perspective*, World Scientific, River Edge, NJ.

Pawitan, Y.: 2001, *In all likelihood: statistical modelling and inference using likelihood*, Oxford University Press, Oxford.

Petersen, C. G. J.: 1896, The yearly immigration of young plaice into the limfjord from the german sea, *Report of the Danish Biological Station* **6**, 1–48.

Pinheiro, J. C. and Bates, D. M.: 2000, *Mixed-effects models in S and S-PLUS*, Springer, New York.

Pledger, S.: 2000, Unified maximum likelihood estimates for closed capture-recapture models using mixtures, *Biometrics* **56**, 434–442.

Pollock, K. H.: 1975, A k-sample tag-recapture model allowing for unequal survival and catchability, *Biometrika* **62**, 577–583.

Popper, K.: 1959, *The logic of scientific discovery*, Routledge, London.

Press, W. H., Teulolsky, S. A., Vetterling, W. T. and Flannery, B. P.: 2007, *Numerical recipes: the art of scientific computing*, 3rd edn, Cambridge University Press, New York.

Proschan, F.: 1963, Theoretical explanation of observed decreasing failure rate, *Technometrics* **5**, 375–383.

SAS Institute Inc: 1999, *SAS/STAT User's Guide, Version 8*, SAS Institute Inc., Cary, NC.

Royall, R. M.: 1997, *Statistical evidence: a likelihood paradigm*, Chapman and Hall, London.

Royston, P. and Parmar, M. K. B.: 2002, Flexible parametric proportional-hazards and propotional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects, *Statistics in Medicine* **21**, 2175–2197.

Rue, H., Martino, S. and Chopin, N.: 2009, Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations, *J. R. Statist. Soc. B* **71: In press**.

Schafer, J. L.: 1997, *Analysis of incomplete multivariate data*, Chapman and Hall, London.

Schall, R.: 1991, Estimation in generalized linear models with random effects, *Biometrika* **78**, 719–727.

Scheipl, F., Greven, S. and Küchenhof, H.: 2008, Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models, *Comp. Stat. Data. Anal.* **52**, 3283–3299.

Schwarz, G.: 1978, Estimating the dimension of a model, *Ann. Stat.* **6**, 461–464.

Scott, A. J. and Wild, C. J.: 2001, Maximum likelihood for generalised case-control studies, *J. Stat. Plan. Inf.* **96**, 3–27.

Seber, G. A. F.: 1982, *Estimation of Animal Abundance and Related Parameters, 2nd edn*, Blackburn Press, New Jersey.

Seber, G. A. F. and Lee, A. J.: 2003, *Linear Regression Analysis, 2nd edn*, Wiley, New Jersey.

Self, S. G. and Liang, K. Y.: 1987, Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *J. Amer. Stat. Assoc.* **82**, 605–610.

Severini, T. A.: 1994, Approximately bayesian inference, *J. Amer. Stat. Assoc.* **89**, 242–249.

Simpson, J. A. and Weiner, E. S. C. (eds): 1989, *The Oxford English Dictionary*, 2nd edn, Oxford University Press, Oxford, U.K.

Skaug, H. J. and Fournier, D. A.: 2006, Automatic approximation of the marginal likelihood in non-gaussian hierarchical models, *Comp. Stat. Data Anal.* **51**, 699–709.

Smyth, G. K. and Verbyla, A. P.: 1996, A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models, *J. Roy. Stat. Soc. Series A* **58**, 565–572.

Sprott, D. A.: 1975, Marginal and conditional sufficiency, *Biometrika* **62**, 599–605.

Stewart, J.: 1999, *Calculus, 4th edn*, Brooks Cole, Pacific Grove, CA.

Stiratelli, R., Laird, N. M. and Ware, J. H.: 1984, Random-effects models for serial observations with binary response, *Biometrics,* **40**, 961–971.

Trivedi, P. K. and Zimmer, D. M.: 2007, Copula modeling: An introduction for practitioners, *Foundations and Trends in Econometrics* **1**, 1–111.

Tuyl, F., Gerlach, R. and Mengersen, K.: 2009, Posterior predictive arguments in favor of the bayes-laplace prior as the consensus prior for binomial and multinomial parameters, *Bayesian Analysis* **4**, 151–158.

Van Deusen, P. C.: 2002, An EM algorithm for capture-recapture estimation, *Environmental and Ecological Statistics* **9**, 151–167.

Venzon, D. J. and Moolgavkar, S. H.: 1988, A method of computing profile-likelihood based confidence intervals, *Appl. Statist.* **37**, 87–94.

Vonesh, E. F.: 1996, A note on the use of laplace's approximation for nonlinear mixed-effects models, *Biometrika* **83**, 447–452.

Wald, A.: 1943, Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Tran. Amer. Math. Soc.* **54**, 426–482.

Wand, M. P.: 2003, Smoothing and mixed models, *Computational Statistics* **18**, 223–249.

Warton, D. I.: 2005, Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data, *Environmetrics* **16**, 275–289.

Wedderburn, R. W. M.: 1974, Quasi-likelihood functions, generalized linear models and the gauss -newton method, *Biometrika* **61**, 439–447.

Wedderburn, R. W. M.: 1976, On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models, *Biometrika* **63**, 27–32.

Wei, W. W. S.: 2006, *Time series analysis: univariate and multivariate methods*, 2nd edn, Addison Wesley, New York.

Wild, C. J. and Seber, G. A. F.: 2000, *Chance encounters: a first course in data analysis and inference*, John Wiley, New York.

Wilks, S. S.: 1938, The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Annals Math. Stat.* **9**, 60–62.

Williams, D. A.: 1986, Interval estimation of the median lethal dose, *Biometrics* **42**, 641–645.

Willis, T. J. and Millar, R. B.: 2005, Using marine reserves to estimate fishing mortality, *Ecol. Let.* **8**, 47–52.

Wilson, E. B.: 1927, Probable inference, the law of succession, and statistical inference, *J. Amer. Stat. Assoc.* **22**, 209–212.

Wolfinger, R.: 1993, Laplace's approximation for nonlinear mixed models, *Biometrika* **80**, 791–795.

Wolfinger, R. and O'Connell, M.: 1993, Generalized linear mixed models: a pseudo-likelihood approach, *J. Statist. Comput. Simul.,* **48**, 233–243.

Wu, C. F. J.: 1983, On the convergence properties of the em algorithm, *Ann. Statist.* **11**, 95–103.

Yan, J.: 2007, Enjoy the joy of copulas: With a package `copula`., *Journal of Statistical Software* **21**, 1–21.

Yuan, A.: 2009.

Zheng, X. and Loh, W.-Y.: 1995, Consistent variable selection in linear models, *J. Amer. Stat. Assoc.,* **90**, 151–156.

Index