Foundations of Medical Statistics

Statistical Inference 9: Inference with multiple parameters (1)

Aims

The aim of this section is to generalise the ideas of the previous sections to models with more than one parameter.

Objectives

The objective is that, at the end of the session, you should undestand how to use the likelihood ratio, Wald and score tests for inference with more than one parameter, and how conditional likelihoods can sometimes be used to remove unwanted parameters from a model.

9.1 Introduction

Sections 3–7 have introduced the concepts of likelihood, estimation and hypothesis testing in situations where the model is defined in terms of only one parameter.

Usually, there will be at least two parameters in the model. For example:

- In a clinical trial comparing treatments A and B, a continuous measurement is made on each of n_A patients randomised to receive treatment A, and n_B patients randomised to receive treatment B. These measurements are assumed to be observations of random variables $X_1, \dots, X_{n_A} \overset{iid}{\sim} N(\mu_A, \sigma^2)$, $Y_1, \dots, Y_{n_B} \overset{iid}{\sim} N(\mu_B, \sigma^2)$ – parameters μ_A, μ_B, σ^2 are unknown; interest is in $\mu_A - \mu_B$, the difference between the means in the two groups.
- $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ – parameters μ and σ^2 are both unknown.
- In a cross sectional survey, the random variable Y_i, systolic blood pressure of the i^{th} individual, is assumed to be distributed $N(\mu_i, \sigma^2)$, and μ_i is a function of age and body mass index ($\mu_i = \beta_0 + \beta_1 age_i + \beta_2 BMI_i$) – parameters $\beta_0, \beta_1, \beta_2, \sigma^2$ are unknown, interest is in β_1, β_2.

Sections 9 and 10 cover how the likelihood methodology is extended to situations where there is more than one parameter.

9.2 Multiple parameters – general situation

9.2.1 Likelihood

If data $\underline{x} = (x_1, \dots, x_n)$ are assumed to be independent observations generated from a model defined in terms of k parameters $\theta_1, \dots, \theta_k$, the likelihood is given by

$$L(\theta_1, \dots, \theta_k | \underline{x}) = f(\underline{x} | \theta_1, \dots, \theta_k) = \prod_{i=1}^{n} f(x_i | \theta_1, \dots, \theta_k)$$

(see section 3.7) and the log-likelihood by

$$l(\theta_1, \dots, \theta_k | \underline{x}) = \sum_{i=1}^{n} \log f(x_i | \theta_1, \dots, \theta_k)$$

The MLEs of $\theta_1, \dots, \theta_k$ are obtained by solving the k simultaneous equations:

$$\left.\begin{aligned} \frac{\partial l}{\partial \theta_1} &= l'(\theta_1) = 0 \\ \frac{\partial l}{\partial \theta_2} &= l'(\theta_2) = 0 \\ &\vdots \\ \frac{\partial l}{\partial \theta_k} &= l'(\theta_k) = 0 \end{aligned}\right\} k \text{ equations for } k \text{ unknown parameters}$$

Notes

- These equations are sometimes referred to as **score equations** (the partial derivative of the log likelihood being the **score function**). Score equations are one of a more general class of **general estimating equations**, which may involve modifications to the likelihood.
- The invariance property of an MLE, introduced previously for one-parameter regular functions, extends to regular transformations of multiple parameters (see for example Casella & Berger p320, 321).

The variance-covariance matrix of the k MLEs is a $k \times k$ symmetric matrix, with variances of each MLE along the diagonal, and covariances between pairs of MLEs in the off-diagonal positions. Thus, if $\underline{\hat{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$, then

$$Var(\underline{\hat{\theta}}) \approx -\begin{pmatrix} \dfrac{\partial^2 l}{\partial \theta_1^2} & \dfrac{\partial^2 l}{\partial \theta_2 \partial \theta_1} & \cdots & \dfrac{\partial^2 l}{\partial \theta_k \partial \theta_1} \\ \dfrac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \dfrac{\partial^2 l}{\partial \theta_2^2} & \cdots & \dfrac{\partial^2 l}{\partial \theta_k \partial \theta_2} \\ & & \ddots & \\ \dfrac{\partial^2 l}{\partial \theta_1 \partial \theta_k} & \dfrac{\partial^2 l}{\partial \theta_2 \partial \theta_k} & \cdots & \dfrac{\partial^2 l}{\partial \theta_k^2} \end{pmatrix}^{-1}_{\underline{\theta} = \underline{\hat{\theta}}}$$

analogous to $-1/l''(\theta)|_{\theta=\hat{\theta}}$ for one parameter.

Notes

- The matrix of second derivatives, $\underline{l''(\theta)}$, is known as the Hessian matrix. The negative of this matrix, after substituting the relevant MLEs, $-\underline{l''(\hat{\theta})}$, is called the **observed information matrix** (OIM). You will see the term 'OIM' mentioned in the Stata output for **generalized linear models** which you will meet in the spring term. OIM corresponds to the observed Fisher information in the one-parameter case. The expected information matrix is the corresponding matrix obtained by taking the expectation of each element in the observed information matrix, and this corresponds to the expected Fisher information in the one-parameter case.

- Standard errors for each MLE can thus be obtained by taking the square roots of the diagonal elements of the matrix shown above, the inverse of the OIM.

9.2.1 Hypothesis testing

A null hypothesis may now be a statement about several parameters. For example if there are two parameters,

$H_0: \theta_1 = 0, \theta_2 = 5$ vs $H_1: \theta_1 \neq 0$ or $\theta_2 \neq 5$ (or both).

The log-likelihood ratio test readily generalises to this situation:

$$-2llr(\underline{\theta}_0) = -2\{l(\underline{\theta}_0) - l(\hat{\underline{\theta}})\} \dot{\sim} \chi^2_r$$

under $H_0: \underline{\theta} = \underline{\theta}_0$ where r is the number of parameters restricted under the null hypothesis (in the example above, $r = 2$).

Often we will only want to test hypotheses concerning a subset of the parameters. We come back to this in section 9.5 and section 10.

The Wald and score tests also have multivariate extensions, defined in terms of vectors and matrices. We now turn to these.

EXERCISE 9.2.1

The univariate Wald test for parameter ϕ is W_ϕ, where

$$H_0: \phi = \phi_0 \quad \Rightarrow \quad W_\phi = \left(\frac{\phi_0 - \hat{\phi}}{S}\right)^2 \dot{\sim} \chi^2_1$$

where $1/S^2 = -l''(\hat{\phi})$.

Suppose θ, λ are two *independent* parameters (e.g. from two independent datasets). Can you think of a way of constructing a test statistic, giving its distribution, to test $H_0: \theta = \theta_0, \lambda = \lambda_0$ vs $H_1: \theta \neq \theta_0$ and/or $\lambda \neq \lambda_0$?

9.3 Multivariate Wald test

Recall for a univariate MLE $\hat{\theta}$, with standard error S, the Wald test statistic for $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$ is

$$\left(\frac{\theta_0 - \hat{\theta}}{S}\right)^2$$

The multivariate version of this, for a joint test of multiple parameters, involves matrix multiplication. For simplicity we only present the case of two parameters, but this makes the general case clear.

Suppose we have two parameters, λ and ψ, and we wish to test $H_0: \lambda = \lambda_0, \psi = \psi_0$. Then the log-likelihood is $l(\lambda, \psi)$ and the derivative of the log-likelihood is a 2×1 column-vector:

$$\underline{l'} = \begin{pmatrix} \dfrac{\partial l}{\partial \lambda} \\ \dfrac{\partial l}{\partial \psi} \end{pmatrix},$$

and the second derivative is now a 2 × 2 matrix

$$\underline{l''} = \begin{pmatrix} \dfrac{\partial^2 l}{\partial \lambda^2} & \dfrac{\partial^2 l}{\partial \lambda \partial \psi} \\ \dfrac{\partial^2 l}{\partial \psi \partial \lambda} & \dfrac{\partial^2 l}{\partial \psi^2} \end{pmatrix}$$

where

$$\frac{\partial^2 l}{\partial \psi \partial \lambda} = \frac{\partial^2 l}{\partial \lambda \partial \psi}$$

Then the two parameter version of $(\theta_0 - \hat{\theta})$ is

$$\begin{pmatrix} \lambda_0 - \hat{\lambda} \\ \psi_0 - \hat{\psi} \end{pmatrix}.$$

The Wald test statistic multiplies $\left(\theta_0 - \hat{\theta}\right)^2$ by $1/S^2$ i.e. $-\underline{l''}\left(\hat{\lambda}, \hat{\psi}\right)$ in the multivariate case. The two parameter Wald test statistic is thus

$$\left(\lambda_0 - \hat{\lambda}, \psi_0 - \hat{\psi}\right)\left(-\underline{l''}(\hat{\lambda}, \hat{\psi})\right)\begin{pmatrix} \lambda_0 - \hat{\lambda} \\ \psi_0 - \hat{\psi} \end{pmatrix}.$$

Although it looks complicated, this is actually just a number. This is because it is the product of a 1 × 2 vector, a 2 × 2 matrix ($-\underline{l''}\left(\hat{\lambda}, \hat{\psi}\right)$) and a 2 × 1 vector, which is a 1 × 1 'matrix', i.e. a number.

As with the multivariate log-likelihood ratio test, we compare this to a χ^2_r distribution, where r is the number of restricted parameters. In this example $r = 2$.

EXERCISE 9.3
Suppose θ, λ are two *independent* parameters (e.g. in connection with two independent datasets). Construct the multivariate Wald test statistic (giving its distribution) to test $H_0: \theta = \theta_0, \lambda = \lambda_0$ vs $H_1: \theta \neq \theta_0$ or $\lambda \neq \lambda_0$, showing that it is the same as that derived in Exercise 9.2.1.

9.4 Multivariate score test

Again, we illustrate the case when we have two parameters and we wish to test $H_0: \lambda = \lambda_0, \psi = \psi_0$.

Recall the univariate score test is U^2/V, where $U = l'(\theta_0)$ and $V = E[-l''(\theta_0)]$.

For two parameters, therefore,

$$\underline{U} = \begin{pmatrix} \dfrac{\partial l}{\partial \lambda} \\ \dfrac{\partial l}{\partial \psi} \end{pmatrix}_{\lambda=\lambda_0,\psi=\psi_0}$$

and

$$\underline{V} = E\left[-\underline{H}\right] \text{ where } \underline{H} = \begin{pmatrix} \dfrac{\partial^2 l}{\partial \lambda^2} & \dfrac{\partial^2 l}{\partial \lambda \partial \psi} \\ \dfrac{\partial^2 l}{\partial \psi \partial \lambda} & \dfrac{\partial^2 l}{\partial \psi^2} \end{pmatrix}_{\lambda=\lambda_0,\psi=\psi_0}$$

Thus $\underline{V}^{-1} = \left(E\left[-\underline{H}\right]\right)^{-1}$, and the score statistic is then

$$\underline{U}^T\underline{V}^{-1}\underline{U} = \left(\frac{\partial l}{\partial \lambda}, \frac{\partial l}{\partial \psi}\right)_{\lambda_0,\psi_0} \left(E\left[-\begin{pmatrix} \dfrac{\partial^2 l}{\partial \lambda^2} & \dfrac{\partial^2 l}{\partial \lambda \partial \psi} \\ \dfrac{\partial^2 l}{\partial \psi \partial \lambda} & \dfrac{\partial^2 l}{\partial \psi^2} \end{pmatrix}_{\lambda_0,\psi_0}\right]\right)^{-1} \begin{pmatrix} \dfrac{\partial l}{\partial \lambda} \\ \dfrac{\partial l}{\partial \psi} \end{pmatrix}_{\lambda_0,\psi_0}$$

which again is a number. We compare this statistic to a χ_r^2 distribution, where r is the number of restricted parameters, and in this case again $r = 2$.
Often $E\left[\underline{H}\right] = \underline{H}$, so $\underline{V}^{-1} = -\underline{H}^{-1}$, and then

$$\underline{V}^{-1} = -\frac{1}{\left(\dfrac{\partial^2 l}{\partial \lambda^2}\dfrac{\partial^2 l}{\partial \psi^2} - \left(\dfrac{\partial^2 l}{\partial \psi \partial \lambda}\right)^2\right)_{\lambda_0,\psi_0}} \begin{pmatrix} \dfrac{\partial^2 l}{\partial \psi^2} & -\dfrac{\partial^2 l}{\partial \lambda \partial \psi} \\ -\dfrac{\partial^2 l}{\partial \psi \partial \lambda} & \dfrac{\partial^2 l}{\partial \lambda^2} \end{pmatrix}_{\lambda_0,\psi_0}$$

All the comments of section 7 apply on the relative merits of the three approximate tests: the Wald and Score approximations are now no longer quadratic curves but *quadratic forms* which approximate the likelihood ratio *surface.*

9.5 Conditional likelihood

In many situations there are only a few parameters of interest out of the (possibly hundreds) in the model. One way to focus on the parameters of interest is to use profile log-likelihood, which is discussed in Section 10.

However, an alternative in certain settings is to use a conditional likelihood. In conditional likelihood, some aspect of the study results containing no information about the parameter of interest is considered fixed. Then a new probability model for the data is defined in terms of conditional probabilities. This new probability model in effect considers only the relevant subset of the sample space.

Consider the following example. Suppose k_1 and k_2 events are observed in p_1 and p_2 person-years of follow-up respectively in two population groups; interest is in the rate ratio $\theta = \lambda_2/\lambda_1$ and we can assume a Poisson model for the rates.

Now consider $k = k_1 + k_2$, the total number of events. This provides no information about how the events divide between the two groups, so let's condition on this total k. As we will see below, this will allow us to move from the Poisson probability model, where we are interested in the probability of a certain number of events occurring, to a binomial model, where we are interested in the probability of events belonging to one group. In the process we throw away information, but not information which is relevant to θ. We discard information on the absolute number of events observed, but we are only interested in the relative numbers in the two groups.

Let us consider the probability of observing k_1 events in group 1, given that there are k events overall:

$$\text{Prob}(k_1 \text{ events in group 1}|\ k \text{ events in total}) = \frac{\text{Prob}(k_1 \text{ events in group 1 and } k - k_1 \text{ events in group 2})}{\text{Prob}(k \text{ events in total})} \quad (1)$$

Since the two groups are independent, the total number of events is Poisson with mean $(\lambda_1 p_1 + \lambda_2 p_2)$, so

$$\text{Prob}(k \text{ events in total}) = e^{-(\lambda_1 p_1 + \lambda_2 p_2)} \frac{(\lambda_1 p_1 + \lambda_2 p_2)^k}{k!}$$

Similarly, because the groups are independent,

$$\begin{aligned}&\text{Prob}(k_1 \text{ events in group 1 and } k - k_1 \text{ events in group 2})\\ &\quad = \frac{e^{-\lambda_1 p_1}(\lambda_1 p_1)^{k_1}}{k_1!} \times \frac{e^{-\lambda_2 p_2}(\lambda_2 p_2)^{k-k_1}}{(k-k_1)!}.\end{aligned}$$

So the conditional probability in equation (1) is

$$\begin{aligned}&\frac{e^{-\lambda_1 p_1} e^{-\lambda_2 p_2} (\lambda_1 p_1)^{k_1} (\lambda_2 p_2)^{k-k_1} k!}{e^{-(\lambda_1 p_1 + \lambda_2 p_2)} (\lambda_1 p_1 + \lambda_2 p_2)^k k_1!\,(k-k_1)!}\\ &\quad = \left(\frac{\lambda_1 p_1}{\lambda_1 p_1 + \lambda_2 p_2}\right)^{k_1} \left(\frac{\lambda_2 p_2}{\lambda_1 p_1 + \lambda_2 p_2}\right)^{k-k_1} \frac{k!}{k_1!\,(k-k_1)!}\\ &\quad = \left(\frac{\lambda_1 p_1}{\lambda_1 p_1 + \lambda_2 p_2}\right)^{k_1} \left(1 - \frac{\lambda_1 p_1}{\lambda_1 p_1 + \lambda_2 p_2}\right)^{k-k_1} \frac{k!}{k_1!\,(k-k_1)!},\end{aligned}$$

which is the probability for a binomial distribution with $\pi = \lambda_1 p_1/(\lambda_1 p_1 + \lambda_2 p_2)$, total number of trials k, and number of successes k_1.

Very importantly:

$$\pi = \frac{\lambda_1 p_1}{\lambda_1 p_1 + \lambda_2 p_2} = \frac{p_1}{p_1 + \frac{\lambda_2}{\lambda_1} p_2} = \frac{p_1}{p_1 + \theta\, p_2},$$

so by conditioning on k we have obtained a probability (and hence a likelihood) with one parameter only – and it is the one we want. Thus the conditional log-likelihood is

$$l_c(\theta) = k_1 \log\left(\frac{p_1}{p_1 + \theta p_2}\right) + (k - k_1) \log\left(1 - \frac{p_1}{p_1 + \theta p_2}\right)$$

(using the form for the binomial log-likelihood, with k_1 successes after k trials, and with π given above). After simplification (using $k = k_1 + k_2$) and ignoring terms not involving θ, the log-likelihood becomes

$$l_c(\theta) = k_2 \log(\theta) - k \log(p_1 + \theta p_2) \, ,$$

which is the **conditional log-likelihood** for θ.

9.5.1 Notes on conditional log-likelihood

1. The conditional log-likelihood is a **true** log-likelihood, based on a (conditional) probability of observed data.
2. A conditional approach relies on our ability to find suitable quantities on which to condition, so that the conditional log-likelihood depends only on the parameter of interest. For example, a method which leads to a conditional log-likelihood for the rate difference $\lambda_2 - \lambda_1$ has not been found.
3. In contrast, the profile log-likelihood, described in Section 10, can always be constructed (numerically if not algebraically).

Note that conditional log-likelihoods play a key role in survival analysis, as the basis for the Cox proportional hazards model; and conditional log-likelihoods form the basis also of conditional logistic regression (both studied in Term 2).