ECON4136, Lcture 6: Likelihood

Tore Schweder

September 27, 2012

1 Introduction

The following is basically taken from the forthcoming book *Confidence, Likelihood, Probability* by Nils Lid Hjort and myself.

In 1922 R.A. Fisher (of the Fisher distribution) introduced the likelihood function and a whole theory of statistical inference based on likelihood. The likelihood function is defined in parametric statistical model, such as the linear regression model with independent and normally distributed errors with the same variance σ^2. The parameter of this model is then the vector (β, σ) where β is the vector of regression coefficients. Fisher suggested to estimate the parameter, including β, by maximizing the likelihood function, i.e. by the maximum likelihood estimator (MLE), and he found that that for large samples its distribution is approximately normally distributed around the parameter with variance-covariance matrix equal to the inverse of the matrix of second derivatives of the likelihood function at its top.

In this lecture we will take a rather informal look at this theory, with emphasis on its intuition and logic, but not attempting to give formal proofs. Most effort will be to appreciate the large sample properties of the maximum likelihood estimator and the deviance function $D = -2\log(L(\theta)/L(\widehat{\theta})$ where L is the likelihood function and $L(\widehat{\theta})$ is its maximum. This log likelihood ratio is an important tool in hypothesis testing and in calculating confidence regions.

Large sample properties are investigated as various limits as the sample size $n \to \infty$. The two most important limit concepts in statistics are convergence in probability and convergence in distribution or law. A sequence of stochastic variables Y_n $n = 1, \cdots, \infty$ converges in probability to a if $P(|Y_n - a| < \varepsilon) \to 0$ for all $\varepsilon > 0$. This is denoted $Y_n \to_p a$. The other type of convergence is denoted $Y_n \to_d Y$ where Y is the stochastic variable having the limit distribution for Y_n. The simple Central limit theorem is then $\sqrt{n}(\bar{X}_n - \mu) \to_d Z$ when $\bar{X}_n$ is the mean of n independent observations from a distribution with mean (expected value) μ and variance $\sigma^2 = 1$. See Wooldrige (2010).

2 Likelihood methods

Suppose data Y stem from some parametric model with joint density $f(y, \theta)$, with $\theta = (\theta_1, \ldots, \theta_p)$ an unknown parameter vector belonging to an appropriate parameter set Ω in $\mathbb{R}^p$. Thhis is the simultaneous density for the full data set. Quite often the data are of the type $Y_1, \ldots, Y_n$, with these being independent, in which case

$$f(y, \theta) = f_1(y_1, \theta) \cdots f_n(y_n, \theta)$$

in terms of density f_i for Y_i. These may also encompass covariate information, say vector x_i associated with Y_i, in which case the notation fruitfully may be modified to $f(y_i \mid x_i, \theta)$, with the interpretation that this is the conditional density of Y_i given x_i.

The *likelihood function* $L(\theta)$ is simply the joint density, but now viewed as a function of the parameter vector for given data values, say for the observed $Y = y$. It is in several respects more convenient to work with the *log-likelihood function*

$$\ell(\theta) = \log L(\theta) = \log f(y, \theta) \tag{1}$$

rather than directly with the likelihood function itself. Sometimes we use $\ell_n(\theta) = \log L_n(\theta)$ to emphasise in the notation that the functions are defined in terms of the first n data points in a sequence. Technically speaking these definitions allow even artificial models and likelihood functions but usually one rules out such situations by insisting on at least a mild amount of regularity, viewed as smoothness in θ, but not necessarily in y. Thus declaring that Y is $N(\theta, 1)$ when θ is rational but a $N(\theta, 2)$ when θ is irrational, for example, arguably does define a statistical model, but it would fall outside what we would be willing to seriously consider.

The maximum likelihood estimator is the value $\widehat{\theta}$ of the parameter vector that maximises the likelihood function (or, equivalently, the log-likelihood function), for the observed data. When required we make the distinction between the estimator, i.e. the random function $\widehat{\theta} = \widehat{\theta}(Y)$, and the concrete estimate, i.e. $\widehat{\theta}_{obs} = \widehat{\theta}(y_{obs})$, where y_{obs} is the observed outcome of Y. As a general and numerically convenient estimation recipe the maximum likelihood principle enjoys various good properties, see Lehmann(1983, ch. 6)and Claeskens and Hjort (2008, ch. 2), for example, in addition to what is briefly reviewed below. Among these good and convenient properties is invariance – with respect to both data transformation and parameter transformation. Thus if $Y^* = T(Y)$ is a one-one data transformation, such as taking the logarithm of positive data, leading to a likelihood function $L^*(\theta)$ for the Y^* data, then $L^*(\theta) = L(\theta)$ and thus the maximum likelihood estimator of θ remains the same (see Exercise **??**. Secondly, if $\gamma = g(\theta)$ is a one-one parameter transformation (featuring component transformations $\gamma_j = g_j(\theta)$ for $j = 1, \ldots, p$), then the maximum likelihood estimator of γ is simply $\widehat{\gamma}_{ML} = g(\widehat{\theta}_{ML})$ (see again the exercise just pointed to).

The main theorem about the maximum likelihood estimator is that it converges in distribution as the sample size increases to a multinormal and with

an identifiable variance matrix agreeing with that dictated by the Cramér–Rao lower bound for unbiased estimators, which is the inverse of J in **??** below, see Pawitan (2001). To put up the required formal statement let us first consider the so-called *i.i.d. situation*, where observations $Y_1, Y_2, \ldots$ are independent and identically distributed, stemming from some common density $f(y, \theta)$, assumed to have two derivatives with respect to the p components of θ in at least a neighbourhood around the prospective true parameter value θ_0, explicitly assumed here to be an inner point of the parameter space. Consider $u(y, \theta) = \partial \log f(y, \theta)/\partial\theta$, the so-called score function with components $u_j(y, \theta)$ for $j = 1, \ldots, p$. These have zero means, when inserting a random Y_i from the $f(y, \theta)$ distribution, cf. Exercise **??**, and we assume that the score function has a finite variance matrix

$$J(\theta) = \mathrm{Var}_\theta\ u(Y, \theta) = -\,\mathrm{E}_\theta\, \frac{\partial^2 \log f(Y, \theta)}{\partial\theta\partial\theta^t} \tag{2}$$

of full rank under the true value θ_0 (to see why the two matrices here are identical, see the exercise just mentioned).

Under mild regularity conditions, basically that the true parameter value is an inner point in the parameter space, the log-likelihood surface is approximately quadratic near its optimum. This property drives various further important properties pertaining to the maximum likelihood estimator, profile-versions of the log-likelihood function, the deviance statistic, etc. As such the following result may be seen as the canonical 'master lemma'. Its use lies also in seeing how similar results may be reached along the same line of arguments in more general situations.

Lemma 1 (the canonical quadratic approximation) *In the i.i.d. situation described above, let θ_0 denote the true parameter, assumed to be an inner point of the parameter region, with $J = J(\theta_0)$ of (**??**) having full rank. Consider the random function*

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) \quad \textit{with } s \in \mathbb{R}^p, \tag{3}$$

with ℓ_n denoting the log-likelihood function based on the first n observations. Then, under mild further regularity conditions, we have

$$A_n(s) \to_d A(s) = s^t U - \frac{1}{2} s^t J s, \quad \textit{where} \quad U \sim N_p(0, J). \tag{4}$$

Figure **??** gives an illustration, depicting five realisations in a binomial situation with $n = 100$ and $p_0 = 0.333$. Note the near quadratic shape of each realized A_n, and thus locally the near quadratic shape of each realized log-likelihood. Such near quadratic functions provides quite accurate approximations by a two-term Taylor expansion. The so-called first order asymptotic likelihood theory, which we shall take a look on, is actually a study of this Taylor approximation of the log-likelihood, and the distribution of its first term.

We note that $A_n(s)$ is only defined as long as $\theta_0 + s/\sqrt{n}$ is inside the parameter range, which in this binomial example means $p_0 + s/\sqrt{n}$ inside $(0, 1)$;

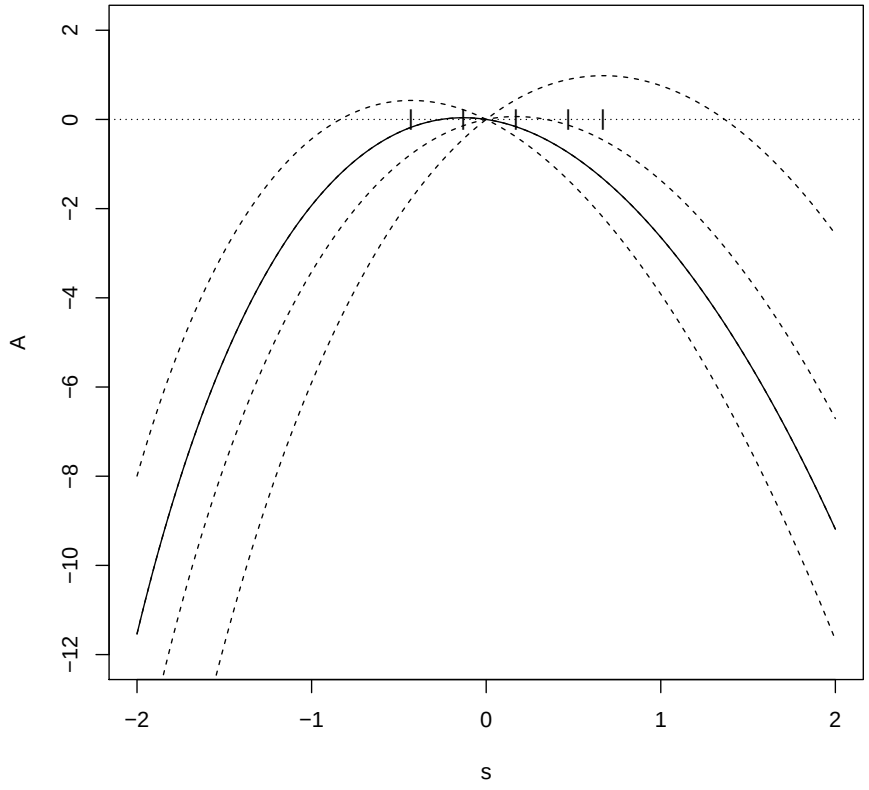

Figure 1: *Five realisations of the random $A_n(\cdot)$ function, in a binomial situation with $n = 100$ and $p_0 = 0.333$. The five maximisers $s_{0,j}$, tagged in the figure, corresponds to five versions of $\sqrt{(}\widehat{p}_j - p_0)$, with $\widehat{p}_j$ the five maximum likelihood estimates.*

but there is no real trouble since $A_n(s)$ for a given s exists for all large enough n. In particular the limit process $A(s)$ is really defined over all of $\mathbb{R}^p$. In the lemma we have merely specified that there is convergence in distribution for each fixed s. This might be strengthened to so-called functional convergence, but that takes us too far afield.

We note that different sets of precise regularity conditions manage to secure conclusion (**??**), but we choose not to go into these details here. The proof we give now is meant to give the essential ideas rather than the full story. For sets of sufficient conditions and full proofs, along with general guidelines for reaching analogous statements in similar problems, see e.g. Hjort and Pollard (1993) Lehmann and Romano (2005)

Proof 1 *A two-term Taylor expansion leads to*

$$A_n(s) = s^t U_n - \frac{1}{2} s^t J_n s + r_n(s),$$

in which

$$U_n = \ell_n'(\theta_0)/\sqrt{n} = n^{-1/2} \sum_{i=1}^n u(Y_i, \theta_0), J_n = -\ell_n''(\theta_0)/n = -n^{-1} \sum_{i=1}^n \frac{\partial^2 \log f(Y_i, \theta_0)}{\partial\theta \partial\theta^t} \tag{5}$$

*and $r_n(s)$ a remainder term. Recall that $u(Y_i, \theta_0) = \frac{\partial}{\partial\theta} \log f(Y_i, \theta_0)$ is the individual score function. Here $U_n \to_d U \sim N_p(0, J)$ by the (multi-dimensional) central limit theorem and $J_n \to_p J$ by the (multi-dimensional) law or large numbers, see e.g. Lehmann (1999); that the matrices involved are the same is precisely identity (**??**). The task of the implied mild extra regularity conditions is to ensure that the $r_n(s) \to_p 0$. For relevant details concerning such functional convergence see e.g. Hjort and Pollard (1993)*

Theorem 1 (normal approximation for the maximum likelihood estimator) *In the i.i.d. situation worked with above, let $\widehat{\theta}_n$ be the maximum likelihood estimator based on the first n observations. If the model holds, with θ_0 the true parameter, being an inner point of the parameter space, and with $J(\theta_0)$ being of full rank, then under mild further regularity assumptions we have*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d N_p(0, J(\theta_0)^{-1}) \tag{6}$$

as the sample size n tends to infinity.

There are again different sets of precise regularity assumptions that guarantee result (**??**); see e.g. Hjort and Pollard (1993). The present point is to see how easily the conclusion flows from the 'master lemma', with a mild modicum of regularity.

Proof 2 *The crux of the proof is that*

$$A_n \to_d A \quad \textit{ought to imply} \quad \arg\max_s(A_n) \to_d \arg\max_s(A);$$

securing this step is indeed the business of the mild extra regularity alluded to. For relevant details see e.g. Hjort and Pollard (1993). With $s = \sqrt{n}(\widehat{\theta} - \theta_0)$ $\arg\max(A_n) = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ *and* $\arg\max(A) = J^{-1}U$*, which is zero-mean multinormal with variance matrix* $J^{-1}JJ^{-1} = J^{-1}$*.*

Rather importantly, results (**??**)–(**??**) continue to hold also when the Y_i are independent but stemming from non-identical distributions, e.g. in a regression context where Y_i has density of the form $f(y \mid x_i, \theta)$ for some covariate vector x_i. In that case, we may define

$$J_n(\theta) = n^{-1} \sum_{i=1}^{n} J_i^*(\theta), \quad \text{with} \quad J_i^*(\theta) = \mathrm{Var}_\theta \; u_i(Y_i, \theta) = -\,\mathrm{E}_\theta \, \frac{\partial^2 \log f(Y_i \mid x_i \theta)}{\partial \theta \partial \theta^t},$$

and (**??**)–(**??**) hold with $J = J(\theta_0)$ the limit of $J_n^*(\theta_0)$ and under some mild extra regularity conditions. Essentially, the line of argument used for Theorem **??** above goes through in this more general setting, with appropriate assumptions of the Lindeberg kind to secure

$$U_n = \ell_n'(\theta_0)/\sqrt{n} = n^{-1/2} \sum_{i=1}^{n} \frac{\partial \log f(Y_i \mid x_i, \theta_0)}{\partial \theta} \to_d U \sim \mathrm{N}_p(0, J)$$

and $J_n \to_p J$; see e.g. Hjort and Pollard (1993) or Hjort and Claeskens (2003a).

There are three favourable aspects to the consequent approximation to the distribution of $\widehat{\theta}$, for a given large or moderately large sample size:

(a) The estimator is approximately unbiased – to be more pedantically correct, its exact distribution is close to a distribution for which the bias $\mathrm{E}_\theta(\widehat{\theta} - \theta)$ is smaller in size than $1/\sqrt{n}$; it is actually of order $O(1/n)$, under mild conditions.

(b) Its distribution is approximately multinormal – hence distributions of single components $\widehat{\theta}_j$ and of linear combinations are approximately normal; this makes it relatively easy to construct confidence intervals and tests with coverage and significance levels close to any intended values.

(c) Its variance matrix (again, the variance matrix of a distribution close to the exact one) is approximately equal to $J(\theta_0)^{-1}/n$ – which is identical to the guaranteed lower bound for unbiased estimators provided by the vector version of the Cramér–Rao theorem (see e.g. Lehmann (1983); hence one cannot hope for other estimation strategies to perform better than this, asymptotically.

There are various caveats here, one of which is that the convergence towards the limit distribution may be slow, and also that the implied approximation $J(\theta_0)/n$ to the variance matrix of $\widehat{\theta}$ may need modifications and improvements, in situations with many parameters. Various remarks and examples pertaining to the occasionally not so sound behaviour of the maximum likelihood estimator

for small or moderate sample sizes are offered in e.g. Lehmann (1983). Overall, though, result (**??**) remains an impressively versatile result, with a string of useful consequences for day-to-day practice of modern statistics.

For using result (**??**) in practice it is important to be able to combine it with a consistent estimator of the limit distribution variance matrix. Two such, both indeed consistent under mild regularity conditions, are

$$\widehat{J}_n = -n^{-1}\frac{\partial^2 \ell_n(\widehat{\theta})}{\partial\theta\partial\theta^t} = -n^{-1}\sum_{i=1}^n \frac{\partial^2 \log f_i(Y_i,\widehat{\theta})}{\partial\theta\partial\theta^t} \quad \text{and} \quad \widetilde{J}_n = J(\widehat{\theta}).$$

It follows that

$$B_n(\theta_0) = n(\widehat{\theta}-\theta_0)^t \widehat{J}_n(\widehat{\theta}-\theta_0) \to_d \chi^2_p, \tag{7}$$

under model conditions (with the same result being true if $\widehat{J}_n$ is replaced by $\widetilde{J}_n$, or indeed with any other consistent estimator of $J(\theta_0)$), giving us both a possibility to test point-hypotheses of the type $H_0\colon \theta = \theta_0$ against $\theta \neq \theta_0$ (one rejects if $B_n(\theta_0)$ is larger than the appropriate χ^2_p quantile), and a confidence set enveloping the true parameter with a coverage probability converging to any desired α level:

$$E_n = \{\theta\colon B_n(\theta) \leq \Gamma_p^{-1}(\alpha)\}.$$

Here and later Γ_ν with inverse Γ_ν^{-1} are used for the cumulative distribution and quantile function of the χ^2 distribution. The confidence set is an ellipsoid centred at $\widehat{\theta}$ and with a radius going to zero with speed $1/\sqrt{n}$.

For constructing such confidence regions, an alternative to $B_n(\theta_0)$ of (**??**), which uses the Hessian matrix $\widehat{J}_n$ associated with the log-likelihood function, is to use say $B'_n(\theta_0) = n(\widehat{\theta}-\theta_0)^t J(\theta_0)(\widehat{\theta}-\theta_0)$, with the explicit Fisher information matrix. Again we have $B'_n(\theta_0) \to_d \chi^2_p$ at the true value, so $E'_n = \{\theta\colon B'_n(\theta) \leq \Gamma_p^{-1}(\alpha)\}$ has the same first-order asymptotic property as has E_n. As an easy illustration of what these two approaches may mean, take $Y \sim \text{Bin}(n,p)$ (the binomial distribution); here the E_n and E'_n methods amount to respectively

$$\frac{\sqrt{n}|\widehat{p}-p|}{\{\widehat{p}(1-\widehat{p})\}^{1/2}} \leq \Gamma_1^{-1}(\alpha)^{1/2} \quad \text{and} \quad \frac{\sqrt{n}|\widehat{p}-p|}{\{p(1-p)\}^{1/2}} \leq \Gamma_1^{-1}(\alpha)^{1/2}.$$

There are arguments supporting the view that using the observed rather than the expected information tends to be the better choice; see e.g. Efron and Morris (1978).

Example 1 Normal linear regression

The linear regression model is among the most successful and widely used tools of applied statistics. Its prototypical version is that of $y = a + bx + \text{noise}$, associated with a scatterplot of (x_i, y_i) points. Here we go straight to the more general case of linear multiple regression, where observation Y_i is linked to a covariate vector $x_i = (x_{i,1},\ldots,x_{i,p})^t$ in the fashion of

$$Y_i = x_i^t\beta + \varepsilon_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \varepsilon_i \quad \textit{for} \quad i = 1,\ldots,n.$$

The error terms ε_i are taken as i.i.d. and $\mathrm{N}(0,\sigma^2)$. In compact matrix form we may write

$$Y \sim \mathrm{N}_n(X\beta, \sigma^2 I_n),$$

where the observation vector Y is $n \times 1$, the covariate matrix X is $n \times p$, the regression coefficient vector β is $p \times 1$, and I_n denotes the $n \times n$ identity matrix. The model has $p+1$ unknown parameters, and the log-likelihood function is seen to be

$$\ell_n(\beta,\sigma) = -n\log\sigma - \frac{1}{2}(1/\sigma^2)Q(\beta) - \frac{1}{2}n\log(2\pi)$$

in terms of

$$Q(\beta) = \sum_{i=1}^n (Y_i - x_i^t\beta)^2 = \|Y - X\beta\|^2.$$

The maximum likelihood estimator of β is identical to the least sum of squares estimator

$$\widehat\beta = \arg\min(Q) = (X^tX)^{-1}X^tY, \text{ leading to } Q_0 = \min_\beta Q(\beta) = \sum_{i=1}^n (Y_i - x_i^t\widehat\beta)^2,$$

where we assume that X has full rank p. The maximum likelihood estimator of σ is seen to be $\widehat\sigma = (Q_0/n)^{1/2}$. For this well-behaved model one knows the exact distributions of all relevant quantities, without the need of Theorem **??** *and other approximations; thus $\widehat\beta \sim \mathrm{N}_p(\beta, \sigma^2\Sigma_n^{-1}/n)$, where $\Sigma_n = n^{-1}\sum_{i=1}^n x_i x_i^t$, and $\widehat\beta$ is stochastically independent of $Q_0 \sim \sigma^2\chi^2_{n-p}$, see e.g. It is nevertheless of interest to see how the normal approximations apply here, and taking two derivatives of the log-likelihood functions leads to*

$$J_n(\beta,\sigma) = \frac{1}{\sigma^2}\begin{pmatrix}\Sigma_n & 0\\ 0 & 2\end{pmatrix},$$

Thus the large-sample spirit of Theorem **??** *gives a perfect match for the distribution of $\widehat\beta$ and the approximation $\mathrm{N}(0,\frac12)$ to that of $\sqrt{n}\{(\chi^2_{n-p}/n)^{1/2} - 1\}$.*

3 Focus parameters and profile likelihoods

"How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service." Indeed there is quite often a one-dimensional *focus parameter* of particular interest, as dictated by the experiment conducted and the relevant phenomena studied, and hence conforming with this particular view of Charles Darwin's. Let

$$\psi = a(\theta_1,\ldots,\theta_p)$$

be such a parameter in focus, e.g. one of the θ_j component parameters but more generally a function of the full model. It follows from the comments above that

its maximum likelihood estimator is $\widehat{\psi}_{ML} = a(\widehat{\theta}_{ML})$. Its limit distribution is easily found via the so-called delta method; in general,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d Z \text{ implies } \sqrt{n}\{a(\widehat{\theta}) - a(\theta_0)\} \to_d c^t Z = \sum_{j=1}^{p} c_j Z_j,$$

in which

$$c = \partial a(\theta_0)/\partial\theta, \quad \text{i.e. } c_j = \partial a(\theta_0)/\partial\theta_j \text{ for } j = 1, \ldots, p. \tag{8}$$

This holds provided merely that $a(\theta)$ has smooth first order derivatives in the p parameters at θ_0. Since Z in this case is multinormal, we have

$$\sqrt{n}(\widehat{\psi} - \psi_0) \to_d c^t Z \sim \mathrm{N}(0, \kappa^2) \quad \text{where} \quad \kappa^2 = c^t J(\theta_0)^{-1} c, \tag{9}$$

say. Hence we have an easy to use general large-sample recipe for any focus parameter $\psi = a(\theta)$ – it may be estimated as $\widehat{\psi} = a(\widehat{\theta})$, and confidence intervals and tests for one- or two-sided hypotheses may be drawn from

$$\sqrt{n}(\widehat{\psi} - \psi)/\widehat{\kappa} \to_d \mathrm{N}(0, 1), \tag{10}$$

with $\widehat{\kappa}$ any consistent estimator of κ.

The $\widehat{\psi}$ estimator also maximises the *profile likelihood*

$$L_{prof}(\psi) = \max\{L(\theta) \colon a(\theta) = \psi\},$$

often most conveniently studied and computed via the log-profile-likelihood, say

$$\ell_{n,prof}(\psi) = \max\{\ell_n(\theta) \colon a(\theta) = \psi\}. \tag{11}$$

Conceptually and operationally it is often convenient to carry out a suitable reparametrisation, if necessary, say from $(\theta_1, \ldots, \theta_p)$ to $(\psi, \chi_1, \ldots, \chi_{p-1})$, making the focus parameter the first component of the new parameter vector. Then

$$\ell_{n,prof}(\psi) = \max_{\chi} \ell_n(\psi, \chi) = \ell_n(\psi, \widehat{\chi}(\psi)),$$

with $\widehat{\chi}(\psi)$ the maximum likelihood estimator in the the $(p-1)$-dimensional model that fixes ψ.

Example 2 The normal model for i.i.d. data

Let $Y_1, \ldots, Y_n$ be independent and identically distributed (i.i.d.) from the normal distribution with mean and standard deviation parameters μ and σ, which we write as $\mathrm{N}(\mu, \sigma^2)$. Then from the Gaußian density formula and with a little algebra the log-likelihood function is

$$\ell_n(\mu, \sigma) = -n \log \sigma - \frac{1}{2}(1/\sigma^2)\{Q_0 + n(\bar{Y} - \mu)^2\} - \frac{1}{2} n \log(2\pi)$$

in terms of average $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ *and sum of squares* $Q_0 = \sum_{i=1}^n (Y_i - \bar{Y})^2$. *This is seen to be maximised for*

$$\widehat{\mu}_{ML} = \bar{Y} \quad \textit{and} \quad \widehat{\sigma}_{ML} = (Q_0/n)^{1/2}. \tag{12}$$

The log-profile likelihood $\ell_{n,prof,2}(\sigma)$ *for* σ *is*

$$\max_{\text{all } \mu} \ell_n(\mu,\sigma) = \ell_n(\widehat{\mu}(\sigma),\sigma) = -n\log\sigma - \frac{1}{2}(1/\sigma^2)Q_0 - \frac{1}{2}n\log(2\pi), \tag{13}$$

where $\widehat{\mu}(\sigma)$ *is the maximiser of* $\ell_n(\mu,\sigma)$ *for given* σ, *but where one actually finds that* $\widehat{\mu}(\sigma) = \bar{Y}$, *irrespective of* σ. *The situation is different when it comes to the log-profile likelihood for* μ, *where* $\ell_n(\mu,\sigma)$ *for given* μ *is maximised for*

$$\widehat{\sigma}(\mu)^2 = Q_0/n + (\bar{Y}-\mu)^2 = \widehat{\sigma}^2_{ML}\{1 + (\bar{Y}-\mu)^2/\widehat{\sigma}^2_{ML}\}.$$

This leads to log-profile likelihood

$$\ell_{n,prof,1}(\mu) = -n\log\widehat{\sigma}(\mu) - \frac{1}{2}n - \frac{1}{2}n\log(2\pi) = -\frac{1}{2}n\log\{1+(\bar{Y}-\mu)^2/\widehat{\sigma}^2_{ML}\} + K., \tag{14}$$

where K *indicates terms not depending on the model parameters.*

For a third example, let the focus parameter be the normalised mean $\psi = \mu/\sigma$. *Its log-profile likelihood becomes*

$$\ell_{n,prof,3}(\psi) = \max_\sigma \ell_n(\sigma\psi,\sigma) = \ell_n(\widehat{\sigma}(\psi)\psi, \widehat{\sigma}(\psi)),$$

where $\widehat{\sigma}(\psi)$ *is the minimiser of*

$$n\log\sigma + \frac{1}{2}(1/\sigma^2)\{Q_0 + n(\bar{Y}-\sigma\psi)^2\}$$

for given ψ; *see Exercise* **??**.

The main theorem about the log-profile likelihood is as follows. Its content is most expediently expressed as a statement about the random quantity

$$D_n(\psi) = 2\{\ell_{n,prof}(\widehat{\psi}) - \ell_{n,prof}(\psi)\}. \tag{15}$$

We term it *the profile deviance*, for the focus parameter ψ in question, and may view it both as a random curve, which we often also wish to display in a diagram, and as a random variable for a given ψ value. Note that D_n is the 'twice log-likelihood-ratio statistic' for testing $H_0\colon \psi = \psi_0$ (under which the parameter dimension is $p-1$) against $\psi \neq \psi_0$ (when the parameter dimension is p), in that

$$D_n(\psi_0) = 2\log\frac{\max_{\text{all } \theta} L_n(\theta)}{\max_{\theta\colon a(\theta)=\psi_0} L_n(\theta)} = 2\{\ell_n(\widehat{\psi},\widehat{\chi}) - \ell_n(\psi_0,\widehat{\chi}(\psi_0))\}.$$

Theorem 2 (chi-squared approximation for the deviance) *Under conditions of the model and those described for Theorem* **??**, *and under the true parameter* θ_0 *(so that the true value of the one-dimensional parametr* ψ *is* $\psi_0 = a(\theta_0)$*), assumed to be an inner point in the parameter space, we have*

$$D_n(\psi_0) = 2\{\ell_{n,prof}(\widehat{\psi}) - \ell_{n,prof}(\psi_0)\} \to_d \chi^2_1. \tag{16}$$

Results of this type are sometimes referred to generically as 'Wilks theorems'.

For proofs of variations of statement (**??**), with appropriate regularity conditions, see e.g. Pawitan (2001). The following sketch, which may be completed using finer details from Pawitan, will hopefully be instructive.

Proof 3 *The essence of the proof is that the log-likelihood function is approximately a quadratic function near its optimum, as formalised via the* $A_n \to_d A$ *convergence result of Lemma* **??**, *and that we are examining the maximum of this function under a linear type side constraint. This is since any smooth* $\psi = a(\theta)$ *may be locally linearised to the required degree of approximation, cf. the discussion around (***??***) involving* $c = \partial a(\theta_0)/\partial\theta$ *of (***??***). It is actually sufficient to examine the case of* $\psi = c^t\theta$*, for suitable* $c = (c_1, \ldots, c_p)^t$ *(we may even reparametrise to have* $c = (1, 0, \ldots, 0)^t$ *to have* ψ *appearing as the first component in the parameter vector).*

To see how this works out we start with a second order Taylor expansion of the log-likelihood function, say

$$\ell_n(\theta) = \ell_n(\widehat{\theta}_n) - \frac{1}{2}n(\theta - \widehat{\theta}_n)^t\widehat{J}_n(\theta - \widehat{\theta}_n) + \varepsilon_n(\theta),$$

with $\varepsilon_n(\theta)$ *the remainder term, typically of order* $n^{-3/2}$*. A separate investigation reveals that a quadratic form* $Q(x) = (x-a)^tB(x-a)$*, where* B *is symmetric and positive definite, when examined under the side condition that* $c^tx = d$*, is minimised for* $x_0 = a + \{(d - c^ta)/(c^tB^{-1}c)\}B^{-1}c$*, with consequent minimum value equal to* $(d - c^ta)^2/c^tB^{-1}c$*; see Exercise* **??**. *It follows that*

$$\ell_{n,prof}(\psi_0) = \ell_n(\widehat{\theta}_n) - \frac{1}{2}n\min\{(\theta-\widehat{\theta}_n)^t\widehat{J}_n(\theta-\widehat{\theta}_n)\colon c^t\theta = \psi_0\} + \delta_n = \ell_{n,prof}(\widehat{\psi}) - \frac{1}{2}n\frac{(\psi_0 - c^t\widehat{\theta}_n)^2}{c^t\widehat{J}_n^{-1}c} + \delta_n, \tag{17}$$

with δ_n *the implied remainder term. This leads to*

$$D_n(\psi_0) = \frac{n(\widehat{\psi} - \psi_0)^2}{c^t\widehat{J}_n^{-1}c} - 2\delta_n \quad \textit{converging to} \quad Z^2 = \frac{(c^tJ^{-1}U)^2}{c^tJ^{-1}c}$$

provided sufficient regularity is in place for δ_n *to tend to zero in probability; see again the references pointed to above for instances of such sets of conditions. This proves statement (***??***) in that* $c^tJ^{-1}U$ *is zero-mean normal with variance* $c^tJ^{-1}c$*; cf. also result (***??***).*

4 Sufficiency and the likelihood principle

The likelihood function and a whole theory of statistical inference based on it was introduced by Fisher (1922). Sufficiency and ancillarity are two central concepts.

A statistics $S = S(Y)$ based on data Y is sufficient when it holds all the information there is in the data regarding the model parameter θ. In formal terms the conditional distribution of Y given S is then free of θ,

$$f_{joint}(y, \theta) = f^{Y|S}(y|s) f^{S}(s, \theta) \tag{18}$$

where the conditional distribution $f^{Y|S}$ is the same for all parameter values. In view of the model, the data provides no extra information regarding θ on top of a sufficient statistic S. It is therefore obvious that any inference must be based on a sufficient statistic. This is the *sufficiency principle*.

The whole data is sufficient, but it might be reduced to a lower dimensional sufficient statistic. If say Y is a sample from the $N(\mu, \sigma^2)$, the mean and the empirical variance, $S = (\bar{Y}, V)$, make up a sufficient statistic. Any one-to-one transformation of S is also sufficient. But $S = (\bar{Y}, V)$ cannot be reduced further without loosing its property. The mean alone is for example not sufficient. A sufficient statistic S is minimal sufficient if it is a function of any other sufficient statistic. In that case a statistic $g(S)$ is not sufficient if the function g is not one-to-one.

The likelihood is sufficient. From (**??**) it is clear that the likelihood based on the whole data is proportional to that based on any sufficient statistic. It is thus minimal sufficient. Inference in a parametric model must consequently be based on the likelihood function. This is the *weak likelihood principle*.

Ancillarity is the opposite concept of sufficiency. A statistic T is ancillary when its distribution does not depend on the model parameter θ. In that case

$$f_{joint}(y, \theta) = f^{Y|T}(y|t, \theta) f^{T}(t). \tag{19}$$

The likelihood is thus proportional to the conditional likelihood given the ancillary statistic, and inference should be conditional on T. This is the *conditionality principle*. A statistic could also be ancillary for a specific parameter, say σ. Then

$$f_{joint}(y, \theta) = f^{Y|T}(y|t, \psi) f^{T}(t, \sigma),$$

when $\theta = (\psi, \sigma)$. It is then often good reasons to carry out the inference in the conditional model given T. Whether this always should be done is debatable.

The sufficiency principle and the conditionality principle imply the *strong likelihood principle* (Birnbaum 1962). It sais that all the evidence the observed data provides is embodied in the conditional likelihood given an ancillary statistic. If for example two experiments, both with model parameter θ lead to the same likelihood function, the inference should then be the same in the two cases. One could actually regard the two experiments A and B as being parts of a larger mixture experiment where a fair coin is tossed to determine whether

A or B should be carried out. The outcome of the toss is clearly an ancillary statistic. By the strong likelihood principle inference should be based on the conditional likelihood given this ancillary statistic. Since the two experiments yielded the same likelihood function, this is the conditional likelihood function in the mixture experiment.

According to the strong likelihood principle, the inference is to be based on the observed likelihood function, regardless of what the model or the data generating process is, provided it leads to the likelihood. This is a controversial principle.

Example 3 *In a binomial experiment of* $n = 20$ *Bernoulli trials all with success probability* p *the number of successes is* $y = 13$*. In another experiment Bernoulli trials are carried out until* $y = 13$ *successes are obtained. It so happens that in the second experiment the total number of trials came out* $n = 20$*. The second experiment is called negative binomial since the number of failures* $N - 15$ *is negatively binomial with parameter* $1 - p$*. Both experiments yield the likelihood function* $p^{13}(1-p)^7$ *when constant terms are removed. Suppose the purpose of both experiments was to test whether* $p \leq 1/2$ *against* $p > 1/2$*. With half-correction, the p-value is 0.095 in the binomial experiment and 0.108 in the negative binomial experiment. They are different despite the likelihoods being identical.*

The Bayesian build on the strong likelihood principle. It is only the likelihood that enters their posterior distribution. Contextual information, such as how sampling is carried out, is of no consequence to the Bayesian on top of the observed likelihood function. The frequentist builds his inference both on the actual outcome of the experiment, as expressed in the likelihood function, and on the contextual evidence available. His p-value is the probability under the null hypothesis of obtaining at least as radical results as observed. His hypothetical series of repeated experiments depends on the protocol of the experiment. As p-values, confidence intervals and confidence distributions will depend on both the observed data and the contextual evidence in the protocol of the experiment, both pieces of evidence is taken into account. Is the frequentist statistician breaching the strong likelihood principle? We would argue not, since the observed data provides us with one piece of information as expressed in the observed likelihood, while the contextual evidence comes in addition. The contextual evidence must obviously be consistent with the likelihood, but is not entirely contained in the observed likelihood. The likelihood principle is discussed in Berger and Wolpert (1988).

5 Exercises

1. *Invariance properties for maximum likelihood:* Suppose a data vector Y has joint density function $f_{joint}(y, \theta)$ in terms of a parameter vector θ.

(a) Let $Y^* = T(Y)$ be a one-one transformation of data (such as taking the logarithm of all data points). Show that Y^* has density

$$f^*_{joint}(y^*, \theta) = f_{joint}(T^{-1}(y^*), \theta) \Big| \frac{\partial T^{-1}(y^*)}{\partial y^*} \Big|$$

and that the maximum likelihood estimator $\widehat{\theta}$ remains the same, whether calculated on the basis of Y or Y^*.

(b) Show invariance with respect to parameter transformation, and that $\widehat{\tau}_{ML} = t(\widehat{\theta}_{ML})$ when $\tau = t(\theta)$ for a function t.

2. *The Fisher information matrix:* Let the dimension of the parameter be $p = 1$. Show that the score function$u(Y, \theta) = \frac{\partial}{\partial \theta} \ell(\theta, Y)$ has mean zero. The trick is to differentiate $1 = \int f(y, \theta) dy$ with respect to θ, inside the integral. Smoothness makes this ok. This yields

$$\int u(y, \theta) f(y, \theta) dy = 0$$

. Differentiating this last equation under the integral yields $\int \frac{\partial^2}{\partial \theta^2} f(y, \theta) dy = 0$, and thus

$$-\int \frac{\partial^2}{\partial \theta^2} \ell(\theta, y) f(y; \theta) dy = \int u(y; \theta)^2 f(y, \theta) dy = var(u(Y, \theta)).$$

3. *Profile likelihood for normalised mean parameter:* Consider the normal model $N(\mu, \sigma^2)$ for i.i.d. data $Y_1, \ldots, Y_n$, cf. Example **??**, where we exhibited the log-profile likelihood functions for parameters μ and σ.

(a) Let the focus parameter be $\psi = \mu/\sigma$, and find the log-profile likelihood.

(b) Then consider the case of focus parameter $\psi = \Pr\{Y_i \leq y_0\} = \Phi((y_0 - \mu)/\sigma)$. Find the log-profile likelihood function.

4. *Log-profile likelihood asymptotics:* Verify directly the χ^2_1 limits of the (profile) deviance of both μ and σ in the case of Example **??**.

5. *Profile likelihoods for logistic regression:* Consider data for n mothers and babies where each mother is characterized by a covariate vector x and the babies are categorized to be large, weighting more than 2.5 kg, or

small. The covariates are mother's weight, whether she is black or not, and whether she is a smoker or not. For a given focus parameter of the type $p_0 = Psmallbaby \mid x_0 = H(x_0^t\beta)$, with x_0 a given vector of covariates, make a programme that computes and displays the log-profile likelihood function

$$\ell_n(p_0) = \max\{\ell_n(\beta)\colon x_0^t\beta = H^{-1}(p_0)\}.$$

Show that $H^{-1}(p_0) = \log\{p_0/(1-p_0)\}$ when $H(z) = \frac{\exp z}{1+\exp z}$ is the logistic function. Simulate a data set of size $n = 200$ and apply your program.

6. *Minimisation exercises for profile log-likelihoods:* Consider a quadratic form $Q(x) = (x-a)^tB(x-a)$, where $a \in \mathbb{R}^p$ and B is a symmetric positive definite $p \times p$ matrix.

   (a) Use e.g. Lagrange multipliers to show that the x minimising $Q(x)$ under the linear side constraint $c^tx = d$ is equal to

$$x_0 = a + \frac{d - c^ta}{c^tB^{-1}c}B^{-1}c.$$

   (b) Deduce that indeed

$$\min\{(x-a)^tB(x-a)\colon c^tx = d\} = \frac{(d-c^ta)^2}{c^tB^{-1}c}.$$

References

[1] Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London.*

[2] Hjort, N. L. Pollard, D. B. 1993 Asymptotics for minimisers of convex processes.Statistical Research Report, Department of Mathematics, University of Oslo, Oslo. Unpublished

[3] Lehmann, E.L. 1983 *Theory of Point Estimation.* Wiley, New York

[4] Lehmann, E.L. and Romano, J.P. 2005 *Testing Statistical Hypotheses [3rd ed.].* Wiley, New York

[5] Pawitan Y. 2001

In all likelihood; statistical modelling and inference using likelihood. Clarendon Press, Offord