What a model with data says about theta

D.A.S. Fraser, N. Reid and A. Wong

November 29, 2006

Abstract

Recent likelihood theory gives complete inference for scalar parameters in continuous statistical models; the methodology involves conditioning at an initial stage, and marginalization at a subsequent stage. Using three simple examples we outline this general route from model with data to likelihood and p-value function for assessing an arbitrary scalar parameter. In wide generality the results are unique in the context of combining general components, but not of course for combining normal components with their multiple inherent symmetries. This gives an inference presentation, from which standard tests, estimates and confidence intervals are immediately available. Extensions for the vector parameter case are discussed.

1 Introduction

A simple example with an observed likelihood function and a special reparameterization illustrates the route to accurate p-values; the needed general case formulas are then described: see Sections 2, 3, and 4. A second example in Section 5 then illustrates the elimination of nuisance parameter effect and the related theory is described in Section 6. A third example in Section 7 then illustrates the elimination of data information that pertains to known error patterns; the corresponding general case formulas are then described in Section 8. These three examples with theory thus describe a standardized route to p-values for arbitrary scalar parameters. Section 9 then indicates how the methodology applies to the familiar Box & Cox (1964) problem, with the related theory outlined in Section 10. Section 11 then gives a brief overview and indicates the extension to the vector interest parameter case.

2 A simple example: the needed ingredients

Consider a scalar variable y providing measurement information on an unknown θ with an extreme-value error distribution; and suppose there is a single observed value. The model is

$$f(y;\theta) = \exp\{-(y-\theta) - e^{-(y-\theta)}\}. \tag{1}$$

and the observed data value is $y^0 = 21.5$.

Our position is that the total inference information concerning θ is given by two functions $L(\theta)$ and $p(\theta)$ which record probability AT the observed data point and probability LEFT of the observed data point. For the first it is often convenient to record it in logarithmic form $\ell(\theta)$. Thus for the example we have

$$\begin{aligned} \ell(\theta) &= a + \theta - e^{\theta - 21.5}, \\ p(\theta) &= \exp(-e^{\theta-21.5}) \end{aligned} \tag{2}$$

where the constant a can be viewed as arbitrary. In Figure 1a we record the probability $L(\theta)$ at the data point 21.5, scaled so the maximum value is 1; and in Figure 1b we record the probability LEFT of the data point 21.5. We view the probability LEFT of the data given by $p(\theta)$ as recording where the data point lies in the distribution with parameter value θ, that is, as recording the PERCENTILE POSITION of the data point in the θ distribution.

In general contexts $\ell(\theta)$ is typically available immediately while $p(\theta)$ is often analytically intractable in various ways not evident in the example. Fortunately highly accurate approximations to $p(\theta)$ are available in wide generality from recent likelihood theory. The approximations are computationally routinely and mechanically available from the log-likelihood function $\ell(\theta)$ together with a typically easily accessible canonical reparameterization $\varphi(\theta)$. For this example the reparameterization is

$$\varphi(\theta) = \exp(\theta - 21.5) - 1. \tag{3}$$

The corresponding exact p-value function $p(\theta)$ and the approximate p-value function say $\tilde{p}(\theta)$ are plotted as the solid and dotted curves in Figure 1b. The high accuracy of the approximation is the familiar and surprising result in applications.

The reparameterization (3) in the case of a scalar variable is obtained as the gradient of likelihood at the observed data:

$$\varphi(\theta) = \frac{\partial}{\partial y}\ell(\theta; y)\Big|_{y^0}, \tag{4}$$

and in the particular case of a full exponential model is given by any version of the canonical parameter. We will see that the basic ingredients $\ell(\theta)$ and $\varphi(\theta)$ lead quite generally to p-values for scalar interest parameter.

3 Familiar inference summaries

A model with data and an interest parameter say θ lead quite generally to the log-likelihood $\ell(\theta)$ describing log-probability AT the data and to the p-value $p(\theta)$ describing probability LEFT of the data. We view these as recording the total inference information concerning the interest parameter as provided by the model with data. Conventionally however, simple inference summaries are wanted, and these are available immediately.

For example consider the total inference given by (2). A central 95% confidence interval is given as

$$(\hat{\theta}_L, \hat{\theta}_U) = \{p^{-1}(0.975), p^{-1}(0.025)\}$$

which has the value $(17.824, 22.805)$. A median-type estimate is given as

$$\tilde{\theta} = p^{-1}(0.5)$$

which has the value 21.133. If a particular parameter value $\theta = \theta_0$ is to be assessed, the corresponding p-value is given as

$$p(\theta_0);$$

if $\theta_0 = 22$ the percentile position of the data point is 19.2%. An upper 99% bound for θ is given as

$$p^{-1}(0.01),$$

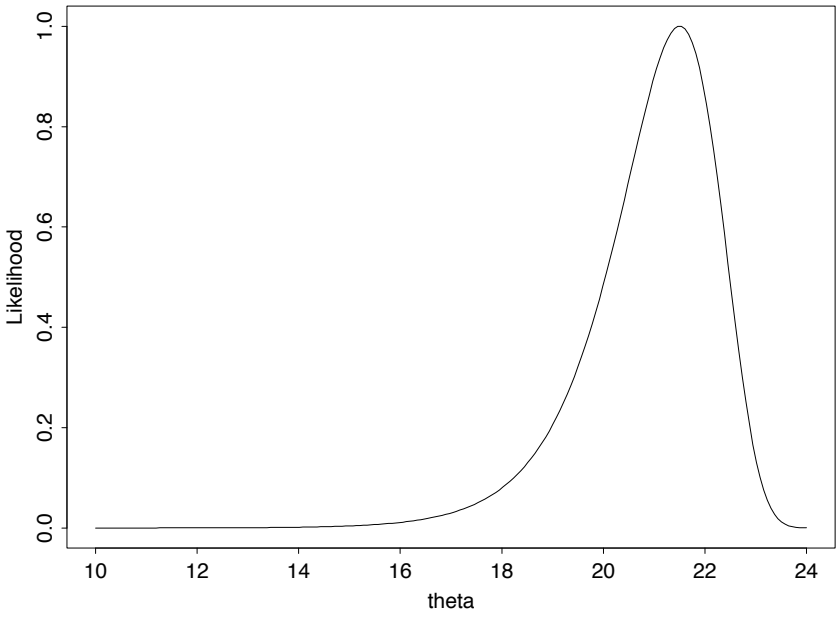
which for the numerical example has the value 23.027.

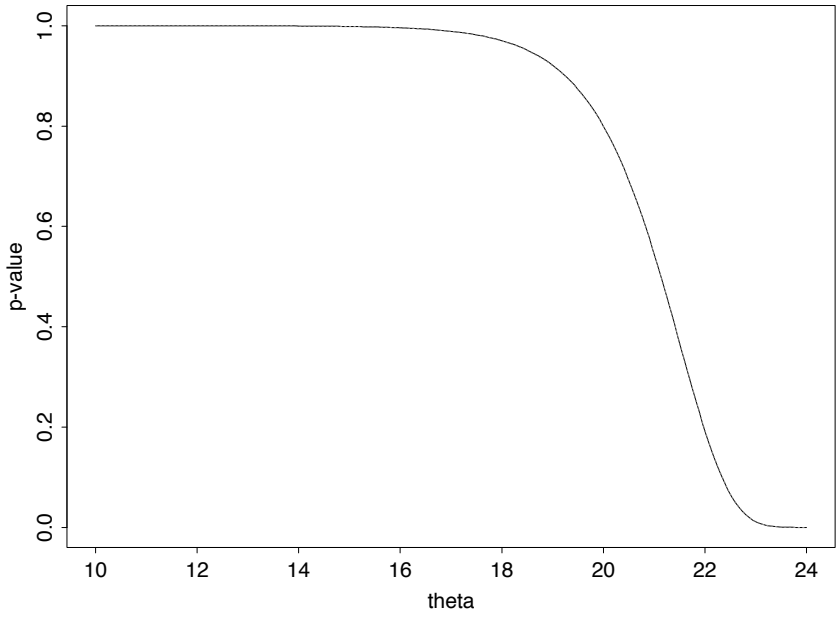
4 From the basic ingredients to p-values

For the case of a scalar parameter θ with an available log likelihood $\ell(\theta)$ and an available canonical parameter $\varphi(\theta)$, the p-value function $p(\theta)$ is obtained as

$$\begin{aligned} p(\theta) &= \Phi(r) + \left(\frac{1}{r} - \frac{1}{q}\right)\phi(r) && (5) \\ &= \Phi(r - r^{-1}\log\frac{r}{q}). && (6) \end{aligned}$$

Figure 1: Data 21.5 from the extreme value model (1): (a) probability AT the data point 21.5 scaled so the maximum value is 1; (b) probability LEFT of the data point 21.5 with exact value $p(\theta)$ as solid line and approximation $\tilde{p}(\theta)$ as dotted line (these are essentially superimposed on each other).





where r is the signed likelihood ratio quantity and q is the Wald quantity in the φ parameterization,

$$r = \text{sign}(\hat{\theta}-\theta)[2\{\ell(\hat{\theta})-\ell(\theta)\}]^{1/2} \quad (7)$$
$$q = \text{sign}(\hat{\theta}-\theta)|\hat{\varphi}-\varphi|\hat{j}_{\varphi\varphi}^{1/2}; \quad (8)$$

for this, $\hat{\theta}$ and $\hat{\varphi}$, are maximum likelihood values for θ and $\varphi(\theta)$, and $\hat{j}_{\varphi\varphi} = (-\partial^2/\partial\varphi^2)\ell(\theta)|_{\hat{\theta}}$ available as $(-\partial^2/\partial\theta^2)\ell(\theta)|_{\hat{\theta}}(\partial\varphi/\partial\theta)^{-2}|_{\hat{\theta}}$ is the observed information. Formulas (5) and (6) are third order accurate and third order equivalent.

For an exponential family model, the reparameterization $\varphi(\theta)$ is just the regular canonical parameter. The third order calculation of $p(\theta)$ by numerical integration of a density is available from Daniels (1954) and the explicit formula (5) is given in Lugannani and Rice (1980); the alternate formula (6) that avoids the risk of values outside $[0,1]$ is given in Barndorff-Nielsen (1986, 1991). In this exponential model context, these formulas correspond to saddlepoint approximations..

For much more general statistical models with asymptotic properties, the same high accuracy is available widely extending the saddlepoint point type approximation approach. This uses formula (4) for $\varphi(\theta)$ and is discussed in Fraser (1990) with various examples of its application. Its use can be based on Taylor series expansions; see Cakmak et al. (1998). And it can be derived also from Barndorff-Nielsen (1986); see for example Fraser and Reid (1995).

5 A simple example with nuisance parameter

Consider the gamma model with mean μ and shape parameter β

$$f(y;\beta,\mu) = \Gamma^{-1}(\beta)\left(\frac{\beta}{\mu}\right)^{\beta} y^{\beta-1}\exp(-\frac{\beta}{\mu}y);$$

and suppose we have an extremely small sample $n=2$ with data $(y_1^0, y_2^0) = (1,4)$; see Fraser, Reid and Wong (1997).

The observed log-likelihood and reparameterization are given in the following array:

$$\ell(\beta,\mu) = 1.38629(\beta-1) - 5\frac{\beta}{\mu} + 2\beta\log\frac{\beta}{\mu} - 2\log\Gamma(\beta), \quad (9)$$
$$\varphi'(\beta,\mu) = (\beta, \beta/\mu); \quad (10)$$

the reparameterization is the canonical parameter of the model, which is exponential. The two parameter functions $\ell(\theta)$ and $\varphi(\theta)$ lead quite generally to highly accurate p-values for any parameter of interest.

Table 1: Third order $\tilde{p}(\mu)$ and exact $p(\mu)$ from ℓ and φ in (9) and (10).

μ	1	3	5	7	9
$\tilde{p}$	.910	.466	.291	.230	.200
p	.901	.464	.318	.256	.225

For assessing a parameter say μ, the recent third order likelihood methods use formula (5) or (6) together with modified versions of r and q that are recorded in the next section. Exact values for the p-value function are available by numerical integration. Table 1 records the approximate and exact values for the following five values 1,3, 5, 7 and 9 for μ.

The actual distributions underlying these p-values for this very small sample $n = 2$ example have a U or bathtub shaped density, as reported for various values of μ in Fraser, Reid and Wong (1997). What is quite remarkable is that the use of the standard normal distribution function with (5) or (6) can give the substantial accuracy indicated in the table. The moderate departures are not particularly serious but we do note that the exact values were obtained by numerical integration that was not fine-tuned to the unusual bathtub density shape of the distribution for this VERY small sample case. Experience with the third order procedures suggests that the third order values may be more reliable here than the nominal exact values; and simulations with $N = 100,000$ provide approximations to the exact with standard error less than .0016 and confirm the accuracy.

The example has exponential model form. More generally with a variable and parameter of dimension p and with asymptotic properties, we obtain the reparameterization $\varphi(\theta)$ as the gradient of likelihood at the observed data

$$\varphi'(\theta) = \nabla \ell(\theta; y)|_{y^0} = \frac{\partial}{\partial y'} \ell(\theta; y)|_{y^0}. \tag{11}$$

In particular for an exponential model this just recovers a version of the canonical parameter, an affine function of the obvious version; either works equally well for computations.

6 From $\ell(\theta)$ and $\varphi(\theta)$ to p-value; with nuisance parameter

For the vector parameter case with an available log-likelihood $\ell(\theta)$ and canonical parameter $\psi(\theta)$, the p-value function $p(\psi)$ for assessing a scalar interest parameter $\psi(\theta)$ can be obtained using (5) or (6) together with a signed likelihood ratio r and a modified Wald type quantity q,

$$r(\psi) \;=\; \text{sign}(\hat{\psi} - \psi)[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2} \tag{12}$$

$$q(\psi) \;=\; \text{sign}(\hat{\psi}-\psi)|\hat{\chi}-\hat{\chi}_\psi|\left\{\frac{|\hat{j}_{\varphi\varphi}|}{|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|}\right\}^{1/2}; \tag{13}$$

for this $\theta=(\lambda,\psi)$ has been written in terms of a nuisance parameter λ that complements the interest parameter ψ. The scalar parameter $\chi(\theta)$ is a rotated coordinate of $\varphi(\theta)$ that acts as a surrogate for $\psi(\theta)$ and has linearity in terms of $\varphi(\theta)$. Explicit formulas for the components in (12) and (13) are recorded in the Appendix Sections 12.1 and 12.2.

The basis for the analysis just described derives from that for a full exponential model

$$f(y;\theta)=\exp\{\varphi'(\theta)t(y)-k(\theta)\}h(y)$$

where φ and t are p-dimensional and $\varphi(\theta)$ and θ are assumed to be one-one equivalent. Early analyses of this model considered the case where the interest parameter $\psi(\theta)$ was a scalar coordinate of $\varphi(\theta)$; in this case the conditional distribution of the coordinate of $t(y)$ corresponding to ψ is examined conditional on the coordinates corresponding to λ and this conditional distribution is free of λ and can be analyzed as in the example in Section 2 using the results summarized in Section 3.

This conditional approach would seem however to require for implementation an explicit form for the conditional distribution; however a third order approximation for the corresponding likelihood is available as

$$\ell(\psi)=\ell_p(\psi)+\frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)| \tag{14}$$

which then permits the direct use of the theory in Section 3. It can also be adjusted to handle the case where $\psi(\theta)$ is a ratio of canonical parameters, as is the case with the gamma example in Section 5; we do not give details here for this pattern of analysis.

For quite general cases, where $\psi(\theta)$ does not have linearity in terms of $\varphi(\theta)$, a marginal approach replaces the conditional approach. A p-value obtained from the conditional approach when available is of course a marginal p-value and does agree with that obtained from the marginal approach. More generally just the marginal version is available.

For the case of an interest parameter $\psi(\theta)$ and a nuisance parameter $\lambda(\theta)$ of dimensions d and $p-d$, a marginal distribution for assessing $\psi(\theta)$ is available for an appropriate variable of dimension d. This distribution is obtained in effect by integrating out a variable corresponding to the nuisance parameter. If the interest parameter is scalar, then the needed r and the q are given by (12, 13); for details see Fraser and Reid (1995, 2001).

7 Example with redundant variables

In the preceding examples we examined cases where the variable and the parameter have the same dimension. Of course for the case of sampling from the gamma model in Section

5 a sufficiency reduction could have provided the reduction to the two dimensions of the gamma parameters. Now we consider the more typical case where each variable is giving information on more than one parameter coordinate but overall with multiple variables there is redundancy, more variables than parameters.

For an example consider the regression model $y = X\beta+\sigma z$ where $z = (z_1, \ldots, z_n)'$ is a sample from the Student(6) distribution, a longer tailed distribution often viewed as providing a better pattern for the error as revealed by large data sets. We suppose that the highest order regression coefficient β_r is of interest.

For this example the parameter has dimension $p = r + 1$ and the variable y has dimension n which we assume to be larger than p. This is an example of a transformation model; see for example Fraser (1979) and references therein.

The likelihood function can be written as a sum $\ell(\theta) = \Sigma_1^n \ell_i(\theta)$ where

$$\ell_i(\theta) = -\log\sigma + \ell\{(y_i - X_i\beta)\sigma^{-1}\}, \tag{15}$$

$\ell(z) = \log f_6(z)$ is the logarithm of the Student(6) density function, and X_i is the ith row of the design matrix X. The canonical parameter can also be written as a sum $\varphi(\theta) = \Sigma_1^n \varphi_i(\theta)$ where

$$\varphi_i'(\theta) = \sigma^{-1} s\{(y_i - X_i\beta)\sigma^{-1}\}(X_i, \hat{z}_i) \tag{16}$$

where $\hat{z}_i$ is the ith coordinate of the standardized residual $\hat{z} = (y - X\hat{\beta})\hat{\sigma}^{-1}$ and $s(z) = d\ell(z)/dz$. The analysis then proceeds as discussed in Sections 5 and 6. For some discussion see Fraser, Wong and Wu (1999).

Simulations have been used to assess this p-value procedure (Fraser, Wong and Wu, ibid). In one case an extremely small data size was examined with $n = 2$ and $X = (1)$, and thus with a single regression coefficient which is the mean μ. In each instance a p-value was calculated for assessing the true μ. This was then repeated a total of $N = 100,000$ times and the distribution of the p-values was examined for uniformity on the interval $(0, 1)$. The proportions in the intervals formed by $(0, 0.005, 0.025, 0.5, 0.975, 0.999, 1.0)$ were then examined relative to the nominal target values of 0.5%, 2.0%, 47.5%, 47.5%, 2.0%, 0.5%. See Table 2 which also records the proportions for the familiar first order signed likelihood ratio (slr) p-value $\Phi(r(\mu))$. In each case the two standard error limit for the $N = 100,000$ simulations is recorded relative to the target probability.

The likelihood ratio values clearly differ substantially from the target values. The third order likelihood values are adequate for practical purposes and are of course calculated here for an EXTREMELY small data size $n = 2$.

Table 2: Proportions of p-values in 6 intervals on $(0, 1)$ using the signed likelihood ratio (slr), and the third order (3rd) procedures.

Interval	$(0, .005)$	$(.005, .025)$	$(.025, .500)$	$(.500, .975)$	$(.975, .995)$	$(.995, 1)$
slr	.05583	.05919	.38677	.38266	.05831	.05724
3rd	.00707	.02463	.47018	.46502	.02522	.00788
$2SE$	.0004	.0009	.003	.003	.0009	.0004
Nominal	.005	.02	.475	.475	.02	.005

8 With redundant variables

The regression model example is a special case of a transformation model which has an exact ancillary say a. Correspondingly the model then has the form $f(y;\theta) = g(s|a;\theta)h(a)$ where s and a have dimensions p and $n - p$. Now consider the more general case where a is an approximate ancillary and $f(y;\theta) = g(s|a;\theta)h(a)$ holds to the appropriate order.

For the same dimension case discussed in Sections 5 and 6 we saw that the canonical parameter $\varphi(\theta)$ was obtained (11) as the gradient of likelihood; we also noted that it could be replaced by an affine equivalent without altering the calculations of a p-value. For the present case if we accept the use of a conditional model given the ancillary a we would then have

$$\varphi'(\theta) = \frac{\partial}{\partial s'}\ell(\theta; s|a)|_{y^0}.$$

For this let $V = (v_1, \ldots, v_p)$ be p linearly independent vectors tangent to $a(y) = a(y^0)$ at the data point $y = y^0$. If we then take directional derivatives

$$\varphi'(\theta) = \frac{d}{dV}\log f(y;\theta) = \frac{d}{ds}\log g(s|a;\theta)$$

we obtain the canonical parameter, or at least an affine equivalent of the obvious canonical parameter. It is then convenient to write

$$\varphi'(\theta) = \ell_{;V}(\theta; y^0) \tag{17}$$

where the subscript denotes directional derivatives taken in the directions given by V.

For the regression model in Section 7 the ancillary has level surfaces, which correspond to the linear spaces $\mathcal{L}(X, \hat{z})$ or $\mathcal{L}(X, y)$, and these have tangent vectors $V = (X, \hat{z})$ at the data point. Thus (16) gives the gradient of the likelihood taken for fixed ancillary.

9 An example with approximate conditioning

Box and Cox (1964) examine the use of transformations of some initial variable y to obtain a new variable $\tilde{y} = h(y, \lambda)$ that conforms more closely to the familiar linear model $\tilde{y} = X\beta + \sigma z$ in regard to linearity of the model, constant error variance, and error normality. They focused on the power transformation $\tilde{y} = y^\lambda$ and thus assumed in effect that the statistical model is defined on the positive real axis. The power transformations are sometimes given in the standardized form $\tilde{y} = (y^\lambda - 1)/\lambda$ but the distinction can be absorbed into the linear model structure provided the design matrix X contains the 1-vector, either explicitly or implicitly. We thus examine the model

$$y = (X\beta + \sigma z)^{1/\lambda}$$

where the power transformation is applied coordinate by coordinate to the n-vector of response values y_i. For simplicity we consider the case of standard normal errors but other error forms are treated easily as indicated in Section 7. And we give formulas for the case where a single independent variable is involved.

Chen, Lockhart and Stephens (2002) give background on the Box and Cox problem and discuss the choice of parameter to estimate, emphasizing the stability of the corresponding estimation procedure; they then give preference to a ratio β/σ of regression parameter to error standard deviation. They also consider tests for normality and develop asymptotic approximations to the distribution of estimators of natural parameters. Yang (2002) investigates confidence intervals for the median response for a particular choice of input variable. Fraser, Wong and Wu (2004) consider the use of recent likelihood theory for the analysis of an appropriate interest parameter of the model.

The likelihood function can be written as a sum $\ell(\theta) = \Sigma_1^n \ell_i(\theta)$ where

$$\ell_i(\theta) = -\log\sigma - \frac{1}{2\sigma^2}(y_i^\lambda - \alpha - \beta x_i)^2 + \log|\lambda| + (\lambda - 1)\log y_i$$

for each y_i. The canonical parameterization is also obtained as a sum of contributions

$$\varphi_i(\theta) = \{-\frac{\lambda y_i^{\lambda-1}}{\sigma^2}(y_i^\lambda - \alpha - \beta x_i) + (\lambda - 1)y_i^{-1}\}(v_{i1}, v_{i2}, v_{i3}, v_{i4})$$

from each y_i where the row vector $V_i = (v_{1i}, v_{2i}, v_{3i}, v_{4i})$ is given as

$$V_i = \frac{y_i}{\hat{\lambda} y_i^{\hat{\lambda}}}(1, x_i, \hat{z}_i, -y_i^{\hat{\lambda}}\log y_i) \tag{18}$$

and $\hat{z}_i = \hat{\sigma}^{-1}(y_i^{\hat{\lambda}} - \hat{\alpha} - \hat{\beta}x_i)$.

Chen, Lockhart and Stephens (2002) consider a large data set with $n = 107$. For illustration suppost we take $x = 32$ litres as a key input value and consider the kilometers per litre or how distance changes with input gasoline at that fuel level:

$$\begin{aligned}\psi(\theta) &= \frac{d}{dx}(\alpha + \beta x)^{1/\lambda}|_{x=32} \\ &= \beta\lambda^{-1}(\alpha + 32\beta)^{(1/\lambda)-1}.\end{aligned}$$

The likelihood function $L^*(\psi)$ and p-value function $p(\psi)$ are then available routinely as we have descibed and give the essential information concerning the kilometers per litre ψ at the input level $x = 32$ litres.

10 Method for approximate conditioning

A familiar idea in elementary probability and statistics involves probability as unit mass that gets mapped or moved when the distribution of some new variable is wanted. For example with the regression model of Section 7 we can write $y = X\beta + \sigma z$ giving the distribution of the response vector y from that for the error vector, or we can use the pivotal form $z = (y - X\beta)\sigma^{-1}$ to consider the reverse mapping. It is then straightforward to think of how parameter change alters the mapping of the error z to the response space. In particular if we calculate $(\partial^2 y/\partial\beta'\partial\sigma)|_y$ for fixed z, we obtain

$$V = \frac{\partial y}{\partial \beta' \sigma} = (X, z)$$

as indicated in Section 8.

This idea of movement of probability was used in Fraser and Reid (2001) to develop approximate ancillaries in a general context. Let $z = z(y, \theta)$ be a full n-dimensional ancillary that typically respects properties such as continuity and independence between observations. We can calculate how parameter change affects a possible response value for fixed pivotal. This is straightforward at the observed data y^0 with corresponding maximum likelihood value $\hat{\theta}^0$ and gives the directions $V = (v_1, \ldots, v_p)$

$$V = -z_y^{-1} z_{;\theta}|_{(y^0, \hat{\theta}^0)} \tag{19}$$

where the subscripts denote partial differentiation; the formula (19) is obtained from the total derivative of the pivot $z(y; \theta)$. Theory shows that in some reasonable generality this gives vectors tangent to a second order ancillary and that this suffices for third order inference.

For the regression example we have

$$z_y = \sigma^{-1} I \qquad z_{;\theta} = \sigma^{-1}(-X, -(y - X\beta)\sigma^{-1})$$

which gives $V = (X, \hat{z})$ at the point $(y^0, \hat{\theta}^0)$ as discussed in Section 7.

For the Box and Cox example in Section 9 it is easier to work coordinate by coordinate. For the ith coordinate we have the pivotal

$$z_i = (y_i^\lambda - \alpha - \beta x_i)\sigma^{-1}$$

which can be solved giving

$$y_i = (\alpha + \beta x_i + \sigma z_i)^{1/\lambda}.$$

The differentiation can then be done directly rather than through the total derivative and we obtain the ith row of V

$$V_i = \frac{y_i}{\hat{\lambda} y_i^{\hat{\lambda}}}(1, x_i, \hat{z}_i, -y_i^{\hat{\lambda}} \log y_i)$$

as recorded at (18).

11 Discussion

For the inference analysis of a statistical model with data, various methods have been proposed for effectively reducing the dimension from that of the original variable to the dimension of the parameter, for then obtaining a pivotal quantity for assessing a parameter of interest, for then obtaining reliable approximations for the distribution of the pivotal quantity, and for then obtaining the observed p-value function for assessing the parameter of interest. The examination of the general asymptotic model and its structure identifies the general route through this process and determines its uniqueness in quite general contexts. Other routes that apply the conditioning and marginalization in a different pattern require specialized model structure but in most cases agree with the general method to high accuracy.

The reduction of the dimension of the variable to that of the parameter is in effect implicit in the extraction of the canonical or exponential reparameterization $\varphi(\theta)$ and its availability with the full loglikelihood $\ell(\theta)$.

The reduction to the dimension of the interest parameter, in the scalar interest case, is obtained in effect by the calculation of the specialized Wald quantity $q(\psi)$ given as (13). The calcualtion of the p-value function $p(\psi)$ is then available from (5) or (6) as if the model was a full exponential model with canonical parameter φ and and cumulant generating function $\ell(\hat{\varphi}) - \ell(\varphi)$ where implicitly it is assumed that the original θ has been reexpressed in terms of φ.

With the general availability of the canonical parameterization φ, using (17) and (19) this indicates the direct availability of p-value functions for scalar interest parameters.

For a vector interest parameters $\psi(\theta)$ of dimension say d a third order likelihood $\ell^*(\psi)$ is availble from (22) using (23). A corresponding φ type reparameterization of ψ has been developed and is in preliminary form: this will enable the direct analysis of the likelihood and φ-type reparameterization exactly in the pattern discusssed above for an initial likelihood and φ parameterization. However for the calculation of the p-value for testing a vector parameter it now seems clear that third order accuracy is unavailable without details of model structure beyond that implicitly available with the likelihood and canonical reparameterization.

12 Appendix

12.1 The surrogate for $\psi(\theta)$

$$\chi(\theta) = \frac{\psi_{\varphi'}(\hat{\theta}_\psi)}{|\psi_{\varphi'}(\hat{\theta}_\psi)|}\varphi(\theta); \tag{20}$$

The row vector multiplying $\varphi(\theta)$ is the unit vector version of the gradient $\psi_{\varphi'}(\hat{\theta}_\psi)$ and is obtained by evaluating

$$\psi_{\varphi'}(\theta) = \frac{\partial\psi(\theta)}{\partial\varphi'} = \frac{\partial\psi(\theta)}{\partial\theta'}\left(\frac{\partial\varphi(\theta)}{\partial\theta'}\right)^{-1} = \psi_{\varphi'}(\theta)\varphi^{-1}_{\theta'}(\theta)$$

at $\hat{\theta}_\psi$; this gives a unit vector perpendicular to $\psi\{\theta(\varphi)\}$ at $\hat{\varphi}_\psi$.

12.2 Information determinants

The information determinants are recalibrated to the φ parameterization

$$\begin{aligned} |\hat{j}_{\varphi\varphi}| &= |\hat{j}_{\theta\theta}||\varphi_\theta(\hat{\theta})|^{-2} \\ |j_{(\lambda\lambda)}(\hat{\theta}_\psi)| &= |j_{\lambda\lambda}(\hat{\theta}_\psi)||\varphi_{\lambda'}(\hat{\theta}_\psi)|^{-2} = |j_{\lambda\lambda}(\hat{\theta}_\psi)||X|^{-2} \end{aligned} \tag{21}$$

where the right hand $p\times(p-1)$ determinant $|X| = |X'X|^{1/2}$ uses $X = \varphi_{\lambda'}(\hat{\theta}_\psi)$ which in the regression context records the volume on the regression surface as a proportion of volume for the regression coefficients.

12.3 Likelihood for a component $\psi(\theta)$

Consider a component parameter $\psi(\theta)$ of dimension d. A third order determinant of likelihood for ψ is obtained from an asymptotic analysis (Fraser, 2003)

$$\ell^*(\psi) = \ell_p(\psi) + \frac{1}{2}\log|j_{(\lambda\lambda)}(\hat{\theta}_\psi)| \tag{22}$$

where $\ell_p(\psi) = \ell(\hat{\theta}_\psi)$ is the profile and $j_{(\lambda\lambda)}(\hat{\theta}_\psi)$ is a special version of the nuisance information calculated with a canonical parameter $\varphi(\theta)$ that has been rescaled to give an identity information at the observed data, $\hat{j}^0 = I$; the needed information determinant can be calculated directly as

$$|j_{(\lambda\lambda)}(\hat{\theta}_\psi)| = |j_{\lambda\lambda}(\hat{\theta}_\psi)||\varphi'_\lambda(\hat{\theta}_\psi)\hat{j}_{\varphi\varphi}\varphi_{\lambda'}(\hat{\theta}_\psi)|^{-1} \tag{23}$$

without the prescribed rescaling.

For the scalar case a corresponding canonical parameter is available; we designate it as $\tilde{\varphi}(\psi)$. We do need the rescaling of the original $\varphi(\theta)$ and assume thus that $\hat{\varphi} = I$. Then

$$\tilde{\varphi}(\psi) = \chi(\hat{\lambda}_\psi, \psi) - \chi(\hat{\lambda}, \hat{\psi})$$

For certain uses as indicated by formulas (12) and (13) the $\tilde{\varphi}$ values do need to be recentered. This allows the analysis of a component parameter ψ using a likelihood $\ell^*(\psi)$ with canonical parameter $\tilde{\varphi}(\psi)$ by direct calculation from $\ell^*(\psi)$, $\tilde{\varphi}(\psi)$ as described in Section 4.

References

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log-likelihood ratio. *Biometrika* **73**, 307-322.

Barndorff-Nielsen, O.E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557-563.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *Jour. Royal Statist. Soc.* **B 26**, 211-252.

Cakmak, J., Fraser, D.A.S., McDunnough, P., Reid, N., and Yuan, X. (1998). Likelihood centered asymptotic model exponential and location model versions. *J. Statist. Planning and Inference* **66**, 211-222.

Chen, G., Lockhart, R.A., and Stephens (2002). Box-Cox transformations in linear models: large sample theory and tests of normality. *Can. Jour. Statist.* **30**, 177-234.

Daniels, H.E. (1954). Saddlepoint approximation in statistics. *Annals Math. Statist.* **25**, 631-650.

Fraser, D.A.S. (1979). Inference and Linear Models, New York: McGraw Hill.

Fraser, D.A.S. (1990). Tail probabilities from observed likelihood. *Biometrika* **77**, 65-76.

Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika***90**, 327-339.

Fraser, D.A.S. & Reid, N. (1995). Ancillaries and third-order significance. *Utilitas Math.* **47**, 33-53.

Fraser, D.A.S. & Reid, N. (2001). Ancillary information for statistical inference. *Empirical Bayes and Likelihood Inference*. Ed. S.E. Ahmed and N. Reid, pp. 185-207. New York: Springer.

Fraser, D.A.S. and Reid, N. (2003). Assessing vector parameters. Technical Report, Dept of Statistics, Univ of Toronto.

Fraser, D.A.S., Reid, N., and Wong, A. (1997). Simple and accurate inference for the mean of the gamma model. *Can. J. Statist.* **25**, 91-99.

Fraser, D.A.S., Reid, N. & Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249-64.

Fraser, D.A.S., Wong, A. and Wu, J. (1999). Regression analysis, normal or nonnormal: accurate p-value from likelihood analysis. *J. Amer. Statist. Assoc.* **94**, 1286-1295.

Fraser, D.A.S., Wong, A. and Wu, J. (2004). Bayes, frequentist and enigmatic examples. *Can. J. Statist.*, submitted.

Lugannani, R. & Rice, S. (1980). Saddlepoint approximation for the distribution function of the sum of independent variables. *Adv. Appl. Prob.* **12**, 475-90.

Yang, Z. (2002). Median estimation through a regression transformation. *Can. J. Statist.* **30**, 235-242.