Regression Analysis, Nonlinear or Nonnormal:
Simple and accurate p-values from Likelihood Analysis

D.A.S. Fraser, A.C.M. Wong, and J. Wu

ABSTRACT

We develop simple approximations for the p-values to use with regression models having linear or nonlinear parameter structure and normal or nonnormal error distribution; computer iteration then gives confidence intervals. Both frequentist and Bayesian versions are given. The approximations are derived from recent developments in likelihood analysis and have third order accuracy. Also for very small and medium sized samples the accuracy can typically be high. The likelihood basis of the procedure seems to provide the grounds for this general accuracy. Examples are discussed and simulations record the distributional accuracy.

KEYWORDS: Asymptotics; Likelihood analysis; Nonlinear; Nonnormal; p-values; Regression
1. INTRODUCTION

Regression analysis is a central technique of statistical methodology, but a large part of the theory is organized around the special case with linear location and normal error. This case of course corresponds to mathematical simplicities and has a long history (e.g. Legendre, 1806; Gauss, 1809; Fisher, 1925).

We examine here the more general case with possibly nonlinear location and possibly nonnormal error:

\[ y = x(\theta) + \epsilon \]  

(1)

where \( y, x, \epsilon \) are \( n \)-vectors, \( x(\theta) \) is a smooth function of an \( r \)-vector \( \theta \), \( \sigma \) is a scaling factor, and \( \epsilon \) has an error distribution \( f(\epsilon) \) on \( R^n \). For inference we restrict attention to the location parameters \( \beta_1, \ldots, \beta_r \).

For this, recent theory focuses on testing a scalar parameter which we take for notational convenience to be \( \beta_r \), the highest order regression coefficient; a test of several coefficients or of several linear constraints on the coefficients is then easily obtained by testing them in sequence holding fixed the preceding ones. The theory can be extended to cover testing \( \sigma \). It can also be extended to cover generalized linear models but some component theory is currently in development. Inference for the parameter \( \beta_r \) will be in the form of a \( p \)-value \( p(\beta_r) \) that records the probability position of the data relative to a value \( \beta_r \) for the parameter. To illustrate this form of inference presentation consider the ordinary linear model \( y = X\beta + \epsilon \) with standard normal error. The \( t \)-value for assessing \( \beta_r \) is

\[ t_r = \frac{\hat{\beta}_r - \beta_r}{s_E c_r^{1/2}} \]  

(2)

where \( \hat{\beta} = CX'y \) is the least squares estimate, \( c_{rr} \) is the \((r,r)\) element of \( C = (X'X)^{-1} \) and \( s_E^2 = s^2/(n-r) \) with \( s^2 = y'y - \hat{\beta}X'y \). Let \( H_f \) be the Student \((f)\) distribution function: then with \( t_r^0 \) as the observed value of \( t_r \) we obtain the \( p \)-value

\[ p(\beta_r) = P(t_r \leq t_r^0) = H_{n-r}(t_r^0), \]  

(3)

which can be viewed as probability left of the data relative to the parameter value \( \beta_r \). The \( p \)-value thus gives the probability position of the data in the range \((0,1)\) of the distribution under the parameter value \( \beta_r \). A test of \( \beta_r \) is obtained by seeing if \( p(\beta_r) \) is close to 0 or close to 1. An estimate of \( \beta_r \) of median type is obtained by solving \( p(\beta_r) = 0.5 \) for say \( \hat{\beta}_r^{0.5} \). A 95\% confidence interval say \((\beta_r^L, \beta_r^U)\) is obtained from \( p(\beta_r^L) = 0.975 \) and \( p(\beta_r^U) = 0.025 \).

The general regression model (1) typically does not have exact measures of departure for testing \( \beta_r \) nor exact distribution theory to give the \( p \)-value \( p(\beta_r) \). The usual approach is to apply first order large sample theory based on likelihood and to use one or other of the signed likelihood ratio \( r \), the Wald departure \( q \), or the score \( z \).

\[ r = \text{sgn} (\hat{\beta}_r - \beta_r)[2(\ell(\hat{\theta}) - \ell(\theta_r))]^{1/2}, \]  

(4)

\[ q = (\hat{\beta}_r - \beta_r)(f'f)^{-1/2}, \quad z = \ell_r(\hat{\theta}_r)(f'f)^{1/2}, \]  

(5)
where \( \ell(\boldsymbol{\theta}) = \log f(y^0; \boldsymbol{\theta}) \) is the observed log likelihood, \( \ell_r(\boldsymbol{\theta}) = (\partial / \partial \beta_r) \ell(\boldsymbol{\theta}) \) is the score for \( \beta_r \), \( \hat{\theta} \) is the maximum likelihood value for \( \boldsymbol{\theta} = (\sigma, \beta_1, \ldots, \beta_r) \), \( \hat{\theta}_r \) is the constrained maximum likelihood value with \( \beta_r \) held fixed at the value \( \hat{\beta}_r \) being tested, \( j_{\theta \theta'} = -(\partial^2 / \partial \theta \partial \theta') \ell(\boldsymbol{\theta}) \bigg|_{\theta = \hat{\theta}} \) is the observed \((r + 1) \times (r + 1)\) information matrix, and \( f^r \) is the \((r + 1, r + 1)\) element of the inverse \( j^{-1} \) of \( j_{\theta \theta'} \). The \( p \)-values are \( \Phi(r) \), \( \Phi(q) \), and \( \Phi(z) \), where \( \Phi(\cdot) \) is the standard normal distribution function; later we will use \( \phi(\cdot) \) for the standard normal density. For implementation reliable maximizing routines are needed for \( \hat{\theta} \) and \( \hat{\theta}_r \) particularly for longer tailed error densities \( f(e) \). In practice the approximate normality of \( r \) is typically much better than that of \( q \) and \( z \); also there are theoretical arguments that \( r \) provides a more sensible measure of departure.

In this paper we develop third order likelihood based methods for obtaining the \( p \)-value \( p(\beta_r) \) for the general model (1). Such third order methods are generally highly accurate and often give excellent results for small sample sizes, attested here by the simulations in Section 5. Third order methods almost always use the signed likelihood ratio \( r \), which is given by (4); this is complemented by a measure of departure \( Q \) that is tailored to individual problems and is here a refinement of the \( q \) or \( z \) in (5). In all cases such \( r \) and \( Q \) are then combined by one or other of the expressions (12), (13) to give the \( p \)-value for testing the scalar parameter of interest. In Section 4 we develop formulas for \( Q \) with the regression model (1): (41) for the frequentist context and (42) for the Bayesian context. These give third order \( p \)-values for the two inference approaches \( \beta \); the nature of the agreement between these is of more general interest and is examined elsewhere. For the frequentist case there are general arguments that the particular third order \( p \)-value is the essentially unique third order measure using all the information.

The likelihood based results for the general regression model are developed in Section 4. Some background on the regression model is recorded in Section 2 and a survey of the relevant likelihood asymptotics is given in Section 3. Section 5 and 6 give examples involving simulations and real data.

2. BACKGROUND: REGRESSION MODEL.

We can check the regression model (1) for linearity by taking a first order Taylor expansion of the location \( x(\beta) \) about a particular point say \( \beta_0 \):  

\[
x(\beta) = x(\beta_0) + X(\beta_0)(\beta - \beta_0) + \cdots 
\]

where \( X(\beta) \) records in \( r \) columns the change in \( x(\beta) \) under \( \beta_1, \ldots, \beta_r \) change, 

\[
X(\beta) = (\partial / \partial \beta' x(\beta) = \left\{ \frac{\partial}{\partial \beta_1} x(\beta), \ldots, \frac{\partial}{\partial \beta_r} x(\beta) \right\} = \{ x_1(\beta), \ldots, x_r(\beta) \},
\]

and \( X(\beta) \) has the role of a design matrix. The linearity of the model can be verified by checking the span of the vectors in \( X(\beta) \), 

\[
\mathcal{L}\{X(\beta)\} = \{X(\beta)b : b \in \mathbb{R}^r\},
\]

which is the tangent plane at \( \beta \). If the tangent plane is free of \( \beta \) then the model is linear; otherwise it is nonlinear.
The statistical analysis of the nonlinear model was examined by Beale (1960) using methods of differential geometry and extended by a succession of authors, Guttman & Meeter (1965), Bates & Watts (1980, 1981, 1988), Hamilton, Watts & Bates (1982), and many others. For the case of normally distributed error, Bates & Watts (1980) proposed two complementing measures of nonlinearity: the maximum relative intrinsic curvature and the maximum relative parameter-effects curvature; these focus respectively on the actual curvature of the surface and on curvature anomalies arising from the choice of parameterization. The measures were extended to the nonnormal case by Amari (1982) and Kass (1984).

By approximating the location surface by a quadratic expression, Hamilton, Watts & Bates (1982) and Hamilton (1986) developed second order inference methods. Two levels of conditioning were proposed in Fraser & Massam (1985), one for the data direction on the tangent plane and one for the signed distance of the data from the location surface; this led to conditional second order inference methods. Abebe et al (1996) examined the nonlinear model with known error scaling, and obtained third order inference within this special formulation.

The statistical analysis of the nonnormal model with linear location was developed by Fraser (1957) and Verhagen (1961) following a pattern proposed by Fisher (1934) for the location-scale model. The linear model after reparameterization can have the form

\[ y = x_0 + X\beta + \sigma e \]  

where \( x_0 \) is an offset say \( x(\beta_0) \) in the original general model notation, and \( \beta \) now designates the departure \( \beta - \beta_0 \) in the original general notation; the specialized notation preceding (2) can then be applied to the adjusted vector \( y - x_0 \) which we now write as just \( y \). The analysis proposed by Fisher is conditional on the standardized residuals,

\[ d = d(y) = (y - X\beta)/s, \]

which are ancillary, having a \( \beta \) and \( \sigma \) free distribution; the conditional distribution of \( b = \beta \) and \( s = s(y) \) is

\[ h^{-1}(d)\sigma^{-n}f[\sigma^{-1}\{X(b - \beta) + sd]\}sn^{-r-1}|X'X|^{1/2} \]

on \( R^r \times R^+ \). If the parameter departures \( \beta - \beta \) are standardized with respect to \( s \) we obtain a \( t \)-type quantity \( T = (\beta - \beta)/s \) with \( T_r = (\beta_r - \beta_r)/s = t_r c_r^{1/2}/(n-r)^{1/2} \) providing the essentially unique measure of departure for \( \beta_r \). The p-value for testing \( \beta_r \) is

\[ p(\beta_r) = \int_{T_r \leq T^0_r} \int_{0}^{\infty} h^{-1}(d)f\{s(XT + d)\}sn^{-r-1}|X'X|^{1/2} \, ds \, dT, \]

where \( T^0_r = (\beta^0_r - \beta_r)/s^0 \) designates the observed value obtained from the data. The evaluation of \( p(\beta_r) \) requires \( r + 1 \) dimensional integration, once with the constraint \( T_r \leq T^0_r \) and once without. An early computer implementation was presented in Fraser (1976) but with larger \( r \) becomes less easily implemented,
thus limiting this direct approach. Asymptotic likelihood methods were applied by DiCiccio, Field & Fraser (1990) giving third order formulas closely related to those obtained from the general case formula in this paper. A combination of one-dimensional integration and Monte Carlo in Fraser, Lee & Reid (1990) also gives third order inference for a location parameter $\theta$. Recent Markov Chain Monte Carlo methods can also be applied to (9) to obtain values for the probability (10). The approximations in Section 4 provide quite generally a simple, straightforward and accurate access to these exact $p$-values. A marginal procedure for the case with Student ($\lambda$) error is proposed in Lange, Little & Taylor (1989): a maximum likelihood estimate is calculated and the standard error is obtained from the expected information matrix; this gives a first order inference procedure. Field & Welsh (1998) examine marginal robustness of confidence intervals by varying a nominal error density used for a conditional calculation and using the approximation in DiCiccio, Field & Fraser (1990).

For the general model in this paper the inference analysis can be viewed as an approximate version of the Fisher pattern above: an approximate ancillary replaces the $d(y)$ and the inference method is conditional on that ancillary.

Some methods for generalized linear models can also apply to cases of the general model (1). For an exponential model with linear canonical parameters, Davison (1988) applied saddlepoint methods; a related procedure was developed by Fraser, Reid & Wong (1991). An inference function approach with conjugate densities was developed by Field & Hampel (1982) and extended by Strawderman, Casella & Wells (1996). A related approach using conditional scores was developed by Waterman & Lindsey (1996). The cases where these models have additive error terms as in (1) are however restricted (Daniels, 1983) to specialized error densities.

3. BACKGROUND: LIKELIHOOD ASYMPTOTICS

Highly accurate likelihood methods have evolved from the saddlepoint method introduced to statistics by Daniels (1954), with follow-up by Barndorff-Nielsen & Cox (1979). Consider a scalar density function $f(y)$ that is of interest but inaccessible and suppose the corresponding cumulant generating function $c(t)$ is available. The saddlepoint point method as introduced is an approximate Fourier inversion that gives an approximation $\hat{f}(y)$ for $f(y)$. As noted initially but not widely followed the method can be expressed entirely in the language of maximum likelihood and we follow accordingly. For use to calculate probability and $p$-values such as $F(y)$ this then requires numerical integration.

Lugannani & Rice (1980) produced a direct approximation $\hat{F}(y)$ for the distribution function $F(y)$. Consider the derived exponential model

$$f(y; \varphi) = \exp\{\varphi y - c(\varphi)\}f(y)$$

(11)

which duplicates $f(y)$ when $\varphi = 0$. The approximation $\hat{F}(y; \varphi)$ to $F(y; \varphi)$ is then an approximate $p$-value at
the data point \( y \) and is given by either of the combining formulas,

\[
\Phi(r) + \phi(r) \left( \frac{1}{r} - \frac{1}{Q} \right),
\]
\[
\Phi \left\{ r - r^{-1} \log \left( \frac{r}{Q} \right) \right\},
\]

where \( r \) is the signed likelihood ratio and \( Q \) here is the maximum likelihood departure \( q \),

\[
r = \text{sgn}(\hat{\varphi} - \varphi)[2\{\ell(\hat{\varphi}; y) - \ell(\varphi; y)\}]^{1/2}
\]
\[
q = (\hat{\varphi} - \varphi)^2_f
\]

corresponding to (4) and (5). The results are third order accurate for moderate derivations and quite accurate even for small samples. For example consider \( f(y; \varphi) = \theta e^{-\theta y} \) on \((0, \infty)\); we take the canonical parameter \( \varphi = \theta \) and canonical variable as \(-y\) with data \( y = 1 \), \( \varphi = 10 \), and we obtain \( q = -9 \), \( r = -3.6599 \), and \( r - r^{-1} \log(r/q) = -3.90579\) which substituted in the standard normal distribution function give \( \Phi(-9) = 0.0^{181} \), \( \Phi(-3.6599) = 0.000126 \), \( \Phi(-3.90579) = 0.000046 \) with exact value 0.000045; the last approximation provides surprising accuracy for a very nonnormal distribution.

Intriguingly however the results are not tied to exponential models nor to moment generating or cumulant generating functions but apply to general statistical models \( f(y; \vartheta) \) with smoothness and asymptotic properties. Consider a statistical model with scalar \( y \) and \( \vartheta \); the third order likelihood based approximation for \( F(y; \vartheta) \) is again given by (12) or (13) using (14) and (15) provided a nominal reparameterization \( \varphi(\vartheta) \) specific to the data point \( y \) is used:

\[
\varphi(\vartheta) = \frac{d}{dy} \ell(\vartheta; y) = \ell_y(\vartheta; y).
\]

The alternative parameterization \( \varphi(\vartheta) \) is the canonical parameter of an exponential model approximating the given model at the data point \( y \) and is obtained as the sample space gradient of likelihood. The approximation was proposed in Fraser (1990) and the third order accuracy established in Fraser & Reid (1993). The operating element is that observed likelihood acts as a cumulant generating function to third order. Note carefully that the approximation uses only the likelihood at the point in question and is calibrated in terms of a reparameterization given as a sample space derivative of likelihood at that point.

Now consider an exponential model with canonical parameter \( \varphi(\vartheta) \) of dimension \( p \), basic parameter \( \vartheta \) of dimension \( p \), and canonical variable \( s(y) \) of dimension \( p \),

\[
f(y; \vartheta) = \exp \{ \varphi(\vartheta)s(y) - c(\vartheta) \} f(y),
\]
a \((p, p)\) exponential model; and suppose interest centers on a scalar parameter \( \psi(\vartheta) \). If \( \psi(\vartheta) \) is a canonical parameter (or ratio of canonical parameters), then a saddlepoint approximation for a conditional density concerning \( \psi(\vartheta) \) is available from Barndoff-Nielsen & Cox (1979) and an approximation for the corresponding
conditional distribution function from Davison (1988) and Fraser, Reid & Wong (1991). If however the
interest parameter \( \psi(\theta) \) is curved and thus departs from being a canonical parameter, the conditional methods
are in general not available and marginal methods are needed to eliminate nuisance parameter effects. Third
order likelihood theory produces a measure of the departure of a data point from a value \( \psi \) for \( \psi(\theta) \) and
it produces the corresponding \( p \)-value. The departure measure does not arise explicitly but is analogous to
the \( t \)-statistic (2). The \( p \)-value is given by the combining formulas (12) or (13) with the standard signed
likelihood ratio \( r \)

\[
r = \text{sgn}(\hat{\psi} - \psi) \left[ 2 \{ \ell(\hat{\theta}; y) - \ell(\hat{\theta}_k; y) \} \right]^{1/2}
\]

(18)

and a special maximum likelihood or score departure \( Q \).

For the frequentist case the \( Q \) is given as

\[
Q_f = \text{sgn}(\hat{\psi} - \psi) |\chi - \chi_\psi| \left\{ \frac{|J_{\varphi \theta}|}{|J_{\lambda \lambda^\prime}(\hat{\theta}_k)|} \right\}^{1/2}
\]

(19)

where \( \chi = \chi(\theta) \) is a scalar linear function of \( \varphi(\theta) \) that is a surrogate for \( \psi(\theta) \) near the tested value \( \psi \)
and \( J_{\lambda \lambda^\prime}(\hat{\theta}_k) \) is the nuisance parameter information based on an equivalence \( \theta' \leftrightarrow \{ \chi(\theta), \psi(\theta) \} \) where the
parentheses are to indicate that it has been recalibrated in terms of the \( \varphi \) parameterization. The surrogate
\( \chi(\theta) \) is constructed by examining \( \psi(\theta) \) as a function of \( \varphi \) near \( \hat{\theta}_k; \) the gradient

\[
\psi_{\varphi}(\theta) = \frac{\partial}{\partial \varphi} \psi(\theta) = \frac{\partial \psi(\theta)}{\partial \theta} \left( \frac{\partial \theta}{\partial \varphi} \right)^{-1}
\]

can be evaluated at \( \hat{\theta}_k \) and used to produce a rotated coordinate of \( \varphi(\theta) \),

\[
\chi(\theta) = \frac{\psi_{\varphi}(\theta_k)}{|\psi_{\varphi}(\theta_k)|} \varphi(\theta),
\]

(20)

that mimics the behaviour of \( \psi(\theta) \) at \( \hat{\theta}_k \). For implementation it is important that the coefficients of the
coordinates of \( \varphi(\theta) \) in (20) are treated as constants based on the data point. This development follows
Fraser, Reid (1995); Barndorff-Nielsen (1986) gives a general formula that can also be specialized to calculate
a \( p \)-value for \( \psi(\theta) \).

For the Bayesian case the \( Q \) is given as

\[
Q_B = \ell(\hat{\theta}_k) \left\{ \frac{|J_{\varphi \theta}|}{|J_{\lambda \lambda^\prime}(\hat{\theta}_k)|} \right\}^{-1/2} \frac{\pi(\theta)}{\pi(\theta_k)}
\]

(21)

where the variable of integration is \( \theta' = (\chi', \psi) \) and the prior is \( \pi(\theta) \). This formula is based on DiCiccio &
Martin (1991) and is expressed in notation from Fraser, Reid & Wu (1999). The Bayesian \( p \)-value is a right
tail posterior probability for \( \psi(\theta) \).

Again we find that the results are widely general and are not tied to exponential models, to cumulant
generating functions or to a fixed dimensional variable such as the sufficient canonical variable \( s = s(y) \).
All that is needed is a second order ancillary statistic to condition on and because only a first derivative is used on the sample space it follows that the only information required concerning the ancillary is given by tangent directions $V = (v_1, \ldots, v_p)$ to the conditioning variable at the data point. These vectors $V$ can be constructed from a fixed dimension coordinate by coordinate pivotal variable $q = q(y; \theta)$ that has a $\theta$ free distribution. The directions $V$ are given by

$$ V = \frac{\partial y}{\partial \theta} \bigg|_\theta = \left\{ \frac{\partial}{\partial y} q(y; \theta) \right\}^{-1} \left\{ \frac{\partial}{\partial \theta} q(y; \theta) \right\} \bigg|_\theta, $$

(22)

where in the first expression the differentiation is for fixed pivotal. The $p$-value for testing $\psi$ in the frequentist case is then given as for the exponential model case but using the following gradient of likelihood at $y$ in the directions $V$,

$$ \psi = \phi(\theta) = \frac{\partial}{\partial \psi} \ell(\theta; y) = \ell_{\cdot V}(\theta; y), $$

(23)

as a stand-in canonical parameter for the $p$-value calculations at the data point $y$.

4. LIKELIHOOD ANALYSIS OF THE GENERAL REGRESSION MODEL

We now use the recent likelihood methods to develop inference methods for the nonlinear nonnormal regression model (1). For this we assume the surface $x(\beta)$ is smooth and that the error density is continuously differentiable to the fourth order. With the interest parameter taken to be $\beta_r$ we write the full parameter as

$$ \theta = (\sigma, \beta) = (\beta', \beta_r) $$

and take a corresponding nuisance parameter to be $\beta' = (\sigma, \beta_1, \ldots, \beta_{r-1})$.

We can allow wide generality for the error distribution but for the present analysis restrict our attention to the case with independent errors $e_i$ with density $f_i(e) = \exp\{-e_i\}$ and assume that the density has been centered so that the slope $s_i(e) = d\ell_i(e)/de$ is zero at the origin: $s_i(0) = \ell'_i(0) = 0$.

The log likelihood function has the form

$$ \ell(\theta; y) = -n \log \sigma + \sum^n_i \ell_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} $$

(24)

which in the standard normal error case becomes

$$ \ell(\theta; y) = -n \log \sigma - \frac{1}{2\sigma^2} \sum^n_i \left\{ y_i - x_i(\beta) \right\}^2. $$

(25)

A primary input for likelihood analysis is the maximum likelihood value $\hat{\theta} = (\hat{\beta}', \hat{\beta}_r)$. This is usually calculated by some iterative procedure applied to the normal equations,

$$ \ell_\sigma(\theta; y) = -\frac{n}{\sigma} - \frac{1}{\sigma^2} \sum^n_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} \{y_i - x_i(\beta)\} = 0 $$

(26)

$$ \ell_{\beta_1}(\theta; y) = -\frac{1}{\sigma} \sum^n_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} X_i(\beta) = 0, $$

(27)
where \( X_i(\beta) = \frac{\partial / \partial \beta'} {x_i(\beta)} \) is the \( i \)th row of the generalized design matrix (7). Now letting

\[
w_i(e) = \frac{-s_i(e)}{e},
\]
we can rewrite the normal equations as

\[
n\sigma^2 - \sum w_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} \left\{ y_i - x_i(\beta) \right\}^2 = 0
\]

(28)

\[
\sum w_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} \left\{ y_i - x_i(\beta) \right\} X_i(\beta) = 0
\]

(29)

in a form suitable for the use of the iteratively reweighted least squares method. The case with \( w_i \equiv 1 \) corresponds to standard normal error. Direct solution of these equations gives the maximum likelihood value \( \hat{\beta} \); the constrained maximum likelihood value \( \bar{\theta}_r = (\lambda_r, \beta_r) \) with \( \beta_r \) fixed is obtained in the same way but with \( \beta_r \) fixed and the \( r \)th equation in (27) or (29) omitted.

The information matrix \( j_{\theta\theta'}(\theta; y) = -\ell_{\theta\theta'}(\theta; y) \) requires the second derivatives

\[
\ell_{\sigma^2} = \frac{n}{\sigma^2} + \frac{2}{\sigma^4} \sum s_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} \left\{ y_i - x_i(\beta) \right\} + \frac{1}{\sigma^4} \sum s_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} \left\{ y_i - x_i(\beta) \right\}^2
\]

(30)

\[
\ell_{\sigma^2 \beta} = \frac{1}{\sigma^2} \sum s_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} X_i(\beta) + \frac{1}{\sigma^3} \sum s_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} \left\{ y_i - x_i(\beta) \right\} X_i(\beta)
\]

(31)

\[
\ell_{\beta \beta'} = \frac{1}{\sigma^2} \sum s_i' \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} X_i(\beta) X_i(\beta) - \frac{1}{\sigma} \sum s_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} (\partial / \partial \beta') X_i(\beta)
\]

(32)

where \( s_i' \) is here the derivative of \( s_i \); from this we calculate \( j_{\theta\theta'} = j_{\theta\theta'}(\theta; y) \) and \( j_{\lambda \lambda'}(\theta_r) = j_{\lambda \lambda'}(\theta_r; y) \) using the full and constrained maximum likelihood values.

An approximate ancillary that conforms to the structure of the model uses a pivotal quantity that coordinate by coordinate indicates how the variable \( y \) measures the parameter \( \theta = (\sigma, \beta') \). With continuity this gives the location scale standardized coordinates in the vector quantity

\[
z = \{ y - x(\beta) \} / \sigma
\]

which has dimension \( n \). The tangent directions for a second order ancillary are then obtained as \( V = \partial y / \partial \beta \) calculated for fixed pivotal and evaluated at the observed \( (y, \beta) \):

\[
V = \frac{\partial x(\beta) + \sigma e}{\partial \beta} \bigg|_{\beta} = \{ e, X(\beta) \},
\]

(33)

which are the standardized residual vector

\[
\tilde{e} = \frac{y - X(\beta)}{\sigma}
\]

(34)

and the tangent plane vector \( X(\beta) \) at the observed data point. These vectors \( V \) are then used to obtain the nominal exponential parametrization \( \varphi(\beta) \)

\[
\varphi(\beta) = \frac{1}{\sigma} \sum s_i \left\{ \frac{y_i - x_i(\beta)}{\sigma} \right\} \{ \tilde{e}_i, X_i(\beta) \}.
\]

(35)
The use of the reparameterization requires the Jacobian \( \varphi(\theta) = (\partial / \partial \boldsymbol{\theta}) \varphi(\theta) \) with elements

\[
\varphi_\alpha = -\frac{1}{\sigma^2} \sum \left\{ s_i \left( \frac{y_i - x_i(\theta)}{\sigma} \right) + \frac{1}{\sigma^2} s_i' \left( \frac{y_i - x_i(\theta)}{\sigma} \right) \right\} \{ \theta_i, X_i(\theta) \}
\]

\[
\varphi_{\beta_i} = -\frac{1}{\sigma^2} \sum s_i' \left\{ \frac{y_i - x_i(\theta)}{\sigma} \right\} X_i(\theta) \{ \theta_i, X_i(\theta) \}.
\]

We can then obtain the recalibrated full information determinant as

\[
|\hat{\varphi}_{\theta'}| = |\hat{\varphi}_{\theta'} \varphi(\theta) - \varphi(\theta)|^{-2}.
\]

The recalibrated nuisance information determinant is

\[
|j_{\lambda \lambda'}(\theta_r)| = |\hat{\varphi}_x(\theta_r) \varphi_{\lambda}(\theta_r)|^{-1}.
\]

The surrogate parameter replacing \( \beta_r \) is

\[
\chi(\theta) = \left\{ \left[ 0, \ldots , 0, 1 \right] \hat{\varphi}_{\theta'}^{-1} \left( \hat{\theta}_r \right) \right\} \varphi(\theta)
\]

and is to be used in the frequentist formula. The score parameter is

\[
\ell_s(\theta) = \ell_s(\theta, \theta) = -\frac{1}{\sigma^2} \sum s_i \left\{ \frac{y_i - x_i(\theta)}{\sigma} \right\} x_i s_i(\theta)
\]

and is to be used in the Bayesian formula; for this the final element of \( X_i \) is designated \( x_{ir} \).

The signed likelihood ratio departure \( r \) is obtained from (24) using the maximum likelihood values \( \hat{\theta} \) and \( \hat{\theta}_r \).

\[
r = \text{sgn} \left( \beta_r - \beta_r \right) \cdot \left\{ 2 \left\{ \Sigma \right\} \left( \hat{\ell}_s(\theta) - \hat{\ell}_s(\theta) \right) - n \log \frac{\sigma}{\sigma_r} \right\}^{1/2},
\]

where \( \hat{\ell}_s(\theta) \) and \( \hat{\ell}_s(\theta) \) are the estimated errors without and with the constraint. The frequentist \( Q_f \) is then

\[
Q_f = \text{sgn} \left( \beta_r - \beta_r \right) \cdot |\chi(\theta) - \chi(\theta_r)| \left\{ \frac{|\hat{j}_{\varphi_{\theta'}}|}{|j_{\lambda \lambda'}(\theta_r)|} \right\}^{1/2}
\]

as obtained from (36), (37), and (38); the Bayesian \( Q_B \) is

\[
Q_B = \ell_s(\theta_r) \left\{ \frac{|\hat{j}_{\varphi_{\theta'}}|}{|j_{\lambda \lambda'}(\theta_r)|} \right\}^{-1/2} \frac{\pi(\theta)}{\pi(\theta_r)}
\]

as obtained from \( j_{\theta \theta'} \) and \( j_{\lambda \lambda'}(\theta_r) \) using (30)-(32) with the maximum likelihood values substituted. With linear location structure, a frequently chosen reference prior is

\[
\pi(\theta) = \sigma^{-1};
\]

in a nonlinear case, the choice would typically relate to the interpretation available for the \( \beta_1, \ldots , \beta_r \) and, of course, for \( \sigma \).
The frequentist p-value $p(\beta_r)$ is then obtained from (12) or (13) with $r$ and $Q_f$ from (40) and (41). The Bayesian right tail posterior probability is obtained from (12) or (13) with $r$ and $Q_B$ from (40) and (42).

5. EXAMPLES: SIMULATIONS TO ASSESS ACCURACY

The likelihood procedure for determining $p$-values involves an implicit measure of departure and an approximation to the corresponding null distribution function. For the case of a normal linear model the measure of departure is quite standard and agrees with the measure obtained from likelihood theory; Example 1 examines the accuracy of the approximation to the null distribution. Examples 2, 3, and 4 then examine the effects of nonnormality and nonlinearity in the model. We are of course investigating a third order likelihood procedure but do our checking at the extreme lower end with sample size as small as possible.

Example 1. The Approximation. Consider a normal linear model with a scalar location parameter. It suffices then to consider the location scale model $y = \mu 1 + \sigma e$ where $e$ is a sample from the standard normal. We test the hypothesis say $\mu = 0$. For a data point $y$ the maximum likelihood value is $\hat{\theta} = (\bar{y}, \hat{\sigma})$ and the constrained value is $\hat{\theta}_0 = (0, \hat{\sigma}_0)$ where $\hat{\sigma}^2 = \Sigma(y_i - \bar{y})^2 / n$ and $\hat{\sigma}_0^2 = \Sigma y_i^2 / n$. The ancillary directions from (37) are $(\hat{e}, 1)$ and agree with the full ancillary defined in the formula preceding (9). The conditional distribution given the ancillary is on the half plane formed by the vectors $\hat{e}, 1$ with a positive coefficient for $\hat{e}$. The elimination of the nuisance parameter effect in both likelihood and exact theory is obtained by projecting probability to the observed nuisance parameter maximum likelihood curve $\Sigma y_i^2 / n = \text{fixed}$ in the preceding half space; this curve is a half circle in the full space. Also in both theories the measure of departure is given in effect by the usual $t$ statistic $(\bar{y} - 0)/(\hat{\sigma}/\sqrt{n})$. We examine the likelihood based $p$-value approximation for the smallest possible sample size $n = 2$, and without loss of generality for simulations take $\sigma = 1$.

With $N = 100,000$ samples of $n = 2$ from the standard normal distribution, we calculate the corresponding likelihood based $p$-values for testing $\mu = 0$ using the formulas (12) and (13) with Section 4 $Q'$s; also for comparison we calculate the ordinary first order likelihood ratio value $\Phi(r)$ and the exact theory $p$-values based on the $t$-statistic. The target for accuracy is of course the uniform $(0, 1)$ distribution. We record the proportion of values in ranges relevant to standard confidence intervals: $(0, 0.005), (0.005, 0.025), (0.025, 0.5), (0.5, 0.975), (0.975, 0.995), (0.995, 1)$; these are recorded in Table 5.1

Table 5.1

<p>| Observed distribution of $p$-values from likelihood ratio (lr), third order (3rd) theory using (12) and (13), and ordinary (t) with $N = 100,000$ for the normal location scale model with $n = 2$. Proportions are recorded; the standard errors for the third order are $0.00022, 0.00044, 0.00158, 0.00158, 0.00044$, and $0.00022$ respectively. | 11 |</p>
<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>lr</td>
<td>0.06029</td>
<td>0.06525</td>
<td>0.37419</td>
<td>0.37558</td>
<td>0.06391</td>
<td>0.06078</td>
</tr>
<tr>
<td>3rd(13)</td>
<td>0.00875</td>
<td>0.02724</td>
<td>0.46507</td>
<td>0.46227</td>
<td>0.02766</td>
<td>0.00901</td>
</tr>
<tr>
<td>3rd(12)</td>
<td>0.00542</td>
<td>0.02277</td>
<td>0.47287</td>
<td>0.46991</td>
<td>0.02338</td>
<td>0.00565</td>
</tr>
<tr>
<td>t</td>
<td>0.00489</td>
<td>0.02010</td>
<td>0.47605</td>
<td>0.47287</td>
<td>0.02105</td>
<td>0.00504</td>
</tr>
<tr>
<td>Nominal</td>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
</tr>
</tbody>
</table>

For this small sample $n = 2$ case the exact null distribution is uniform on the half circle mentioned above. The ability of the asymptotic likelihood theory to give close approximation in this extreme $n = 2$ situation attests well to its reliability in comparison with the likelihood ratio procedure: the Lugannani & Rice combining formula (12) seems to give the closer approximation.

**Example 2. Nonnormality.** We now examine the accuracy of the $p$-values in the context of nonnormal error but retain the linearity of the model. We thus consider the location scale model $y = \mu + \sigma e$ where $e$ is a sample form a nonnormal distribution. The Student family provides longer tails than the normal, and the Student (6) is sometimes viewed as a good approximation for the error pattern in many applications; we take $e$ to be a sample from the Student (6) distribution.

The ancillary directions and exact ancillary halfspace are as described in Example 1. The elimination of the nuisance parameter is also by projecting to the nuisance maximum likelihood curve, which is typically not the half circle as in the normal case. We examine the third order likelihood $p$-values for the smallest $n = 2$ sample size case.

With $N = 100,000$ samples of $n = 2$ from the Student (6) distribution we calculate the corresponding $p$-values for testing $\mu = 0$ using the approximations (12) and (13); also for comparison we calculate the ordinary first order likelihood ratio values $\Phi(r)$. The nominal or target value is of course given by the uniform $(0,1)$ distribution. The results in percentages are recorded in Table 5.2 for the familiar confidence interval ranges. Again the Lugannani & Rice combining formula (12) gives high accuracy.

**Table 5.2**

Observed distribution of $p$-values from likelihood ratio (lr) and third order (3rd) theory using (12) and (13), with $N = 100,000$ for the Student (6) location scale model with $n = 2$. Proportions are recorded; the standard errors for the third order are 0.00022, 0.00044, 0.00158, 0.00158, 0.00044, and 0.00022 respectively.
Example 3. Nonlinearity. We now examine the accuracy of the third order p-values in the context of nonlinearity and do this for normal error. Consider the model \( y = x(\beta) + \sigma e \) where \( e \) is a sample from the standard normal distribution. A simple example suggested by a reader of an earlier report has \( x(\beta) = \beta 1 + h(\beta) x \) which would give a curve in \( \mathcal{L}(1, x) \). By rotating to bring this 2-space to the first two coordinates and using a parabola with \( h(\beta) \) quadratic we obtain

\[
\begin{align*}
y_1 &= \beta + \sigma z_i \\
y_2 &= \beta^2 / 2c + \sigma z_i \\
y_i &= \sigma z_i \quad (i > 2)
\end{align*}
\] (43)

The radius of curvature at \( \beta = 0 \) is \( c \) which in the asymptotic framework would be \( O(n^{1/2}) \).

Consider a data point \( y \) and let \( \hat{\beta} \) be the corresponding location maximum likelihood value. We examine the third order likelihood p-value for the sample size \( n = 3 \) which is the smallest value that retains the general structure of the model. The ancillary direction vectors are then \( (e, k) \) where \( k' = (1, \hat{\beta}/c, 0) \) and \( \hat{e} = (y - x(\hat{\beta}))/\hat{\sigma} \).

With \( N = 100,000 \) samples of \( n = 3 \) from the standard normal we calculate the corresponding p-values for testing \( \beta = 0 \) using formulas (12) and (13); for comparison we calculate the ordinary first order likelihood ratio values \( \Phi(r) \). We also examine two radiuses of curvature, \( c = 5 \), \( c = 3 \). The results are recorded in Table 5.3 for the familiar confidence interval ranges. Again high accuracy for the Lugannani & Rice combining formula (12).

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
</tr>
<tr>
<td>lr</td>
<td>0.05583</td>
<td>0.05919</td>
<td>0.38677</td>
<td>0.38266</td>
<td>0.05831</td>
<td>0.05724</td>
</tr>
<tr>
<td>3rd(13)</td>
<td>0.00085</td>
<td>0.02862</td>
<td>0.46341</td>
<td>0.45800</td>
<td>0.02913</td>
<td>0.01099</td>
</tr>
<tr>
<td>3rd(12)</td>
<td>0.00707</td>
<td>0.02463</td>
<td>0.47018</td>
<td>0.46502</td>
<td>0.02522</td>
<td>0.00788</td>
</tr>
<tr>
<td>Nominal</td>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 5.3

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
</tr>
<tr>
<td>lr</td>
<td>0.05583</td>
<td>0.05919</td>
<td>0.38677</td>
<td>0.38266</td>
<td>0.05831</td>
<td>0.05724</td>
</tr>
<tr>
<td>3rd(13)</td>
<td>0.00085</td>
<td>0.02862</td>
<td>0.46341</td>
<td>0.45800</td>
<td>0.02913</td>
<td>0.01099</td>
</tr>
<tr>
<td>3rd(12)</td>
<td>0.00707</td>
<td>0.02463</td>
<td>0.47018</td>
<td>0.46502</td>
<td>0.02522</td>
<td>0.00788</td>
</tr>
<tr>
<td>Nominal</td>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Observed distribution of p-values from likelihood ratio (lr) and third order (3rd) theory using (12) and (13), with \( N = 100,000 \) for the normal error model (36) with \( n = 3 \) and (a) \( c = 5 \) and (b) \( c = 3 \). Proportions are recorded for the common confidence interval ranges.

(a) Radius of curvature \( c = 5 \)
 Lower  Next  Next  Next  Next  Last  
 0.005  0.02  0.475  0.475  0.02  0.005  

\( \begin{array}{llllll} 
\text{lr} & 0.02853 & 0.04708 & 0.42558 & 0.42367 & 0.04682 & 0.02832 \\
3\text{rd}(13) & 0.00696 & 0.02251 & 0.47167 & 0.46933 & 0.02242 & 0.00711 \\
3\text{rd}(12) & 0.00638 & 0.02153 & 0.47323 & 0.47101 & 0.02119 & 0.00666 \\
\text{Nominal} & 0.005 & 0.02 & 0.475 & 0.475 & 0.02 & 0.005 \\
\end{array} \)

(b) Radius of curvature \( c = 3 \)

 Lower  Next  Next  Next  Next  Last  
 0.005  0.02  0.475  0.475  0.02  0.005  

\( \begin{array}{llllll} 
\text{lr} & 0.02894 & 0.04762 & 0.42431 & 0.42244 & 0.04745 & 0.02924 \\
3\text{rd}(13) & 0.00697 & 0.02274 & 0.47112 & 0.46905 & 0.02267 & 0.00745 \\
3\text{rd}(12) & 0.00628 & 0.02180 & 0.47275 & 0.47081 & 0.02138 & 0.00698 \\
\text{Nominal} & 0.005 & 0.02 & 0.475 & 0.475 & 0.02 & 0.005 \\
\end{array} \)

**Example 4. Nonlinearity; Normal and Nonnormal.** Consider the nonlinear model

\[
y_i = e^{-\beta x_i} + \sigma e_i
\]

with Normal errors and with Student(6) errors \( e_i \). We examine the very small-sample case with \( n = 3 \) and \( x' = (-1, 0, 1) \) and test the hypothesis \( \beta = 1 \).

First we simulate normal data with \( \sigma = 1, \beta = 1 \) and analyze on the basis of a normal model (correct) and on the basis of a student model (incorrect). For \( N = 100,000 \) the results are summarized in Table 5.4.

**Table 5.4**

Observed distribution of \( p \)-values from likelihood ratio (lr), second order formula (2nd) of Hamilton, Watts & Bates (1982), and the third order likelihood (3rd) methods using (12) and (13) based on the model (44) using (a) normal error (correct) and (b) Student (6) error (incorrect). Proportions are recorded; standard errors for the third order are 0.00022, 0.00044, 0.00158, 0.00158, 0.00044, and 0.00022 respectively.

(a) Normal analysis

 Lower  Next  Next  Next  Next  Last  
 0.005  0.02  0.475  0.475  0.02  0.005  

\( \begin{array}{llllll} 
\text{lr} & 0.02953 & 0.04873 & 0.41970 & 0.42631 & 0.04699 & 0.02874 \\
2\text{nd} & 0.00644 & 0.02736 & 0.46425 & 0.46541 & 0.02846 & 0.00805 \\
3\text{rd}(13) & 0.00687 & 0.02307 & 0.46640 & 0.47381 & 0.02251 & 0.00734 \\
3\text{rd}(12) & 0.00679 & 0.02133 & 0.46807 & 0.47545 & 0.02172 & 0.00664 \\
\text{Nominal} & 0.005 & 0.02 & 0.475 & 0.475 & 0.02 & 0.005 \\
\end{array} \)
(b) Student analysis

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>0.02573</td>
<td>0.05143</td>
<td>0.41809</td>
<td>0.42125</td>
<td>0.05189</td>
<td>0.02861</td>
<td></td>
</tr>
<tr>
<td>3rd(13)</td>
<td>0.00866</td>
<td>0.02554</td>
<td>0.46232</td>
<td>0.46799</td>
<td>0.02546</td>
<td>0.01003</td>
</tr>
<tr>
<td>3rd(12)</td>
<td>0.00920</td>
<td>0.02428</td>
<td>0.46291</td>
<td>0.46914</td>
<td>0.02484</td>
<td>0.00963</td>
</tr>
<tr>
<td>Nominal</td>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Next we simulate Student (6) data with $\sigma = 1, \beta = 1$ and analyze on the assumption (a) of a Student (6) model (correct) and (b) of a normal model (incorrect). For $N = 100,000$ the results are summarized in Table 5.5.
Table 5.5

Observed distribution of p-values from likelihood ratio (lr) and (3rd) using (12) and (13) for the model (44) with Student (6) errors: (a) Normal analysis (incorrect) and (b) Student(6) analysis (correct) Percentages are recorded; standard errors for the third order are 0.00022, 0.00044, 0.00158, 0.00158, 0.00044, and 0.00022 respectively.

(a) Normal analysis

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>lr</td>
<td>0.01973</td>
<td>0.0403</td>
<td>0.4427</td>
<td>0.4352</td>
<td>0.0397</td>
<td>0.02201</td>
</tr>
<tr>
<td>3rd(13)</td>
<td>0.00798</td>
<td>0.01796</td>
<td>0.47368</td>
<td>0.47173</td>
<td>0.02069</td>
<td>0.00796</td>
</tr>
<tr>
<td>3rd(12)</td>
<td>0.00995</td>
<td>0.01628</td>
<td>0.47310</td>
<td>0.47277</td>
<td>0.01983</td>
<td>0.00807</td>
</tr>
<tr>
<td>Nominal</td>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
</tr>
</tbody>
</table>

(b) Student analysis

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>lr</td>
<td>0.01975</td>
<td>0.03909</td>
<td>0.44325</td>
<td>0.43959</td>
<td>0.03613</td>
<td>0.02222</td>
</tr>
<tr>
<td>3rd(13)</td>
<td>0.00580</td>
<td>0.01567</td>
<td>0.47777</td>
<td>0.47722</td>
<td>0.01749</td>
<td>0.00665</td>
</tr>
<tr>
<td>3rd(12)</td>
<td>0.00695</td>
<td>0.01422</td>
<td>0.47783</td>
<td>0.47879</td>
<td>0.01655</td>
<td>0.00655</td>
</tr>
<tr>
<td>Nominal</td>
<td>0.005</td>
<td>0.02</td>
<td>0.475</td>
<td>0.475</td>
<td>0.02</td>
<td>0.005</td>
</tr>
</tbody>
</table>

We note that the third order methods and particularly the Lugannani and Rice formula (12) give good accuracy using the correct model for the analysis even for this extremely small sample case.

6. EXAMPLES: DATA FROM THE LITERATURE

We now reexamine some data sets from the nonlinear regression literature.

Example 6.1. Data with \( n = 6 \) was obtained by Marske on biochemical oxygen demand (BOD) as a function of time;

\[
\begin{array}{c|cccccc}
\text{Time t} & 1 & 2 & 3 & 4 & 5 & 7 \\
\text{BOD y} & 8.2 & 10.3 & 19.0 & 16.0 & 15.6 & 19.8 \\
\end{array}
\]

and recorded in Bates & Watts (1988). They analyzed the data using a nonlinear regression model with independent normal errors and location \( E(y) = \beta_1(1 - e^{-\beta_2 t}) \) and reported 95% confidence intervals for \( \beta_1 \) and \( \beta_2 \) using a linear approximation method and using the likelihood ratio method. Table 6.1 records these intervals for \( \beta_1 \) and \( \beta_2 \) and also the intervals obtained by the third order theory discussed in this paper.

Table 6.1

---

16
95% Confidence intervals for $\beta_1$, $\beta_2$ with Marske data based on a linear approximation, the likelihood ratio, and the third order procedure using (12) or (13)

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_1$ Lower</th>
<th>$\beta_1$ Upper</th>
<th>$\beta_2$ Lower</th>
<th>$\beta_2$ Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>12.2</td>
<td>26.1</td>
<td>-0.033</td>
<td>1.095</td>
</tr>
<tr>
<td>Lik. ratio</td>
<td>14.05</td>
<td>37.77</td>
<td>0.132</td>
<td>1.770</td>
</tr>
<tr>
<td>3rd order</td>
<td>13.70</td>
<td>36.36</td>
<td>0.139</td>
<td>1.792</td>
</tr>
</tbody>
</table>

Experience indicates that the likelihood ratio method is better than other first order methods. The third order methods typically remove location and scale bias from the likelihood ratio method itself.

Example 6.2. Bates & Watts (1988) record data ($n = 13$) obtained by Count Rumford in 1798 on the amount of heat generated by friction.

<table>
<thead>
<tr>
<th>Time</th>
<th>4</th>
<th>5</th>
<th>7</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>126</td>
<td>125</td>
<td>123</td>
<td>120</td>
<td>119</td>
<td>118</td>
<td>116</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>24</th>
<th>28</th>
<th>31</th>
<th>34</th>
<th>37.5</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>115</td>
<td>114</td>
<td>113</td>
<td>112</td>
<td>111</td>
<td>110</td>
</tr>
</tbody>
</table>

Bates & Watts (1988) analyzed these data using a nonlinear model with independent normal errors based on Newton’s Law of cooling with mean temperature given as $60 + 70\exp\{-\beta t\}$; they obtained first order likelihood intervals for $\beta$ at levels 80%, 90%, and 95%, using the likelihood ratio method. These are recorded in Table 6.2 together with the corresponding third order intervals based on the third order methods from Section 4.

Table 6.2

Confidence intervals for the decay rate $\beta$ with Rumford data using the likelihood ratio and the third order method

<table>
<thead>
<tr>
<th></th>
<th>80% Lower</th>
<th>80% Upper</th>
<th>90% Lower</th>
<th>90% Upper</th>
<th>95% Lower</th>
<th>95% Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lik. ratio</td>
<td>0.0089</td>
<td>0.0100</td>
<td>0.0087</td>
<td>1.0101</td>
<td>0.0085</td>
<td>0.0103</td>
</tr>
<tr>
<td>3rd order</td>
<td>0.0088</td>
<td>0.0100</td>
<td>0.0087</td>
<td>1.0102</td>
<td>0.0085</td>
<td>0.0104</td>
</tr>
</tbody>
</table>

To the accuracy recorded the third order is quite close to the likelihood ratio which is to be expected with what amounts to high accuracy data. To compare the methods further we omit the first five and last five data points obtaining a set with $n = 3$ data points. From this reduced data set the confidence intervals are recorded in Table 6.3.

Table 6.3

Confidence intervals for $\beta$ with Rumford subset using the likelihood ratio and the third order method
The third order can be viewed as removing location and scale bias from the likelihood ratio intervals; in the present case the third order intervals tend primarily to be wider, thus correcting an overstatement of precision with the likelihood ratio intervals.

**Example 6.3.** Hamilton (1986) examined the following data set recording \( y \) the initial velocity (dpm/min) and \( x \) the concentration (mM) of an enzyme

| \( x \) | 0.02 | 0.02 | 0.06 | 0.06 | 0.11 | 0.11 | 0.22 | 0.22 | 0.56 | 0.56 | 1.10 |
| \( y \) | 67   | 51   | 54   | 86   | 98   | 115  | 124  | 144  | 158  | 160  |

He used a nonlinear regression model with independent normal errors and location

\[
E(y) = \frac{\beta_2x}{\beta_1 + x}
\]

given by the Michaelis-Menten model.

95% Confidence intervals for \( \beta_1 \) were obtained by the likelihood ratio method and by a second order method (Hamilton, 1986) based on intrinsic curvature; exact 95% confidence intervals were also reported based on the normality of \( y_i(\beta_1 + x_i) \). Table 6.4 records these intervals together with the third order 95% confidence interval

**Table 6.4**

<table>
<thead>
<tr>
<th>Method</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lik. ratio</td>
<td>0.0351</td>
<td>0.0953</td>
</tr>
<tr>
<td>2nd order</td>
<td>0.0331</td>
<td>0.0713</td>
</tr>
<tr>
<td>3rd order</td>
<td>0.0315</td>
<td>0.0699</td>
</tr>
<tr>
<td>Exact</td>
<td>0.0314</td>
<td>0.0701</td>
</tr>
</tbody>
</table>

We note a progressive improvement in the intervals from the likelihood ratio through to the exact.

**Example 6.4.** Gallant (1987) examined a data set \( (n = 30) \) organized as a one-way treatment-control design with independent variable \( x_3 \) giving the age of the experimental material, \( x_1 \) giving the one vector, and \( x_2 \) being an indicator variable for treatment group. The first few and last data values are

| \( x_1 \) | 1 | 1 | ... | 1 |
| \( x_2 \) | 1 | 1 | ... | 0 |
| \( x_3 \) | 6.23 | 9.11 | ... | 6.11 |
| \( y \) | 0.98610 | 0.95482 | ... | 0.91840 |
A nonlinear regression model with independent normal errors and location

\[ E(y) = \beta_1 + \beta_2 x_2 + \beta_4 \exp\{\beta_3 x_3\} \]

was used to generate the data and perform the analysis.

Gallant used SAS to obtain 95% confidence intervals for the \( \beta_1 \) parameters. We record these together with the third order intervals from the Section 4 procedure.

| Table 6.5 |
|---|---|---|---|---|---|
| 95% Confidence intervals for Gallant regression parameter |
| \( \beta_1 \) | Lower | Upper | \( \beta_2 \) | Lower | Upper | \( \beta_3 \) | Lower | Upper | \( \beta_4 \) | Lower | Upper |
| Lik. ratio | 0.9953 | 1.0361 | -0.0518 | 0.0001 | -1.4519 | -0.7795 | -0.5576 | -0.4522 |
| 3rd order | 0.9952 | 1.0358 | -0.0511 | 0.0005 | -1.4890 | -0.8330 | -0.5582 | -0.4529 |

The third order method makes moderate adjustments to the likelihood ratio method for the parameter \( \beta_1 \), \( \beta_2 \), \( \beta_4 \), but a substantial adjustment for the important rate parameter \( \beta_3 \).

7. ESTIMATING EQUATIONS AND REGRESSION

Recent third order likelihood methods can also be used with estimating functions. First consider a scalar \( \theta \) and let \( w(y; \hat{\theta}) \) be an estimating function for \( \theta \) with the statistical model \( f(y; \theta) \). The corresponding M-estimator \( \hat{\theta} = \hat{\theta}(y) \) is the solution of \( w(y; \hat{\theta}) = 0 \). We seek the distribution function \( H(\hat{\theta}; \theta) \) of \( \hat{\theta}(y) \) at the observed value \( \hat{\theta} = \hat{\theta}(y^0) \) from data, that is, we seek the observed \( \theta \)-value \( p(\theta) = H(\hat{\theta}; \theta), \) as derived from the M-estimator \( \hat{\theta}(y) \). For this we assume that \( w(y; \theta) \) is monotone increasing in \( \hat{\theta} \) and has asymptotic properties; for example we could have \( w(y; \theta) = \sum_i w_i(y_i; \theta) \) with independent components models for the \( y_i \).

We also assume for \( w \) that the density \( g(w) = g(w; \theta, \theta) \) or the moment generating function \( m(t) = m(t; \hat{\theta}, \theta) \) is available for given \( \hat{\theta} \) and \( \theta \). We then introduce a nominal location model \( g(w - \alpha) \) or nominal exponential model \( g(w; \alpha) = g(w) \exp\{\alpha \theta - \log m(\alpha)\} \); the corresponding distribution function \( G(w) \) for \( \alpha = 0 \) can then be approximated using (12) or (13) with (14), (15) and (16), and underlying parameterization \( \alpha \). At \( w = 0 \), the distribution function value \( G(0) = G(0; \hat{\theta}, \theta) \) will depend on the background \( \hat{\theta} \) and \( \theta \). Then using the familiar equivalence of \( \hat{\theta}(y) \leq \hat{\theta}^0 \) with \( w(y; \hat{\theta}^0) \leq 0 \) coming from the monotonicity, we obtain the third order approximation

\[
p(\theta) = H(\hat{\theta}^0; \theta) = G(0; \hat{\theta}^0, \theta) . \tag{45}\]

The saddlepoint method was applied to estimating functions by Daniels (1983), Field (1982), and Strawderman, Casella & Wells (1996) among others in progressive generality; the last mentioned paper specifically addresses regression problems. In each case the density function was derived and numerical integration used to get \( p \)-values. The present distribution function approach gives an alternative approach and follows the suggestion in Daniels (1983) to use a Lugannani & Rice (1980) type formula. Daniels notes
that “in many practical situations the (moment generating function) cannot be expressed in a neat analytical form”; but the approach suggested here covers also the case with an available density.

Now consider an estimating function \( w(y; \theta) \) of dimension \( p \) for the statistical model \( f(y; \theta) \). Initially suppose that the density \( g(w) = g(w_1; \theta) \) or the moment generating function \( m(t) = m(t_1; \theta) \) for \( w \) is available for given \( \hat{\theta} \) and \( \theta \). Correspondingly this leads to the location model \( g(w - a) \) or the exponential model \( g(w) \exp\{w^\alpha - \log m(\alpha)\} \). By the methods in Section 3 with \( \alpha = 0 \) we can obtain marginal probabilities or conditional probabilities for individual coordinates. These however do not translate into probabilities for \( \hat{\theta}_1, \ldots, \hat{\theta}_r \) unless the estimating equations themselves isolate the coordinates \( \hat{\theta}_1, \ldots, \hat{\theta}_r \). In this case the marginal probabilities for \( w_1, \ldots, w_r \) from the location model for \( w \) would lead to marginal probabilities for the \( \theta_1, \ldots, \theta_r \). These however need not depend respectively on just the corresponding \( \theta_1, \ldots, \theta_r \).

An alternative approach is to apply directly the methods Section 3 to the model \( f(y; \theta) \). This produces the \( p \)-value \( p(\psi) \) for a scalar interest parameter. With this \( p(\psi) \), say as a graph, we can record also the observed maximum likelihood value \( \hat{\psi} \) and the observed estimating equation value \( \hat{\psi} \). This allows comparison of the estimating equation value \( \hat{\psi} \) with the confidence points available from \( p(\psi) \), these latter in some sense giving the full story concerning \( \psi \) from the given model. We do not pursue here this estimating function approach with the regression model.

REFERENCES


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


Fraser, D.A.S. and Reid, N. (1993), Simple asymptotic connections between densities and cumulant generating function leading to accurate approximations for distribution functions, *Statistical Sinica* 3 67–82.


Gauss, K.F. (1809), Theorìa motus corporum coelestium in sectionibus conicis solem ambientium, Hamburg: Perthes and Besser.


