Nonnormal linear regression: An example of significance levels in high dimensions

BY D. A. S. FRASER

Department of Mathematics, York University, Downsview, Ontario, Canada M3J 1P3

H. S. LEE AND N. REID

Department of Statistics, University of Toronto, Toronto, Canada M5S 1A1

SUMMARY

Analysis of nonnormal linear models leads to an initial conditioning on the standardized residuals, giving an unnormed density on $\mathbb{R}^k$, where $k$ is the number of parameters. To obtain an observed level of significance for a single parameter it is then necessary to calculate a marginal probability, thus requiring integration in $k$ dimensions. In this paper a conditional approach to evaluating the observed level of significance is developed, and an importance sampling technique is used to improve the approximation and assess the accuracy of the conditional approximation to the marginal observed level. A further approximation based on the invariant version (Fraser, 1990) of the Lugannani & Rice (1980) formula is also proposed. The approach extends to the evaluation of real pivots and to Bayesian inference for a single parameter component.

Some key words: Bayesian analysis; Conditional inference; Linear model; Nonnormal analysis; Observed level of significance; Regression model.

1. INTRODUCTION

The linear regression model for nonnormal error has the form $y = X\beta + \sigma e$, where $e$ has a density $f_\epsilon(e)$ on $\mathbb{R}^n$, and $X$ is an $n \times p$ design matrix with regression vector $\beta$; for the $i$th coordinate we write $y_i = X_i \beta + \sigma e_i$, where $X_i$ is the $i$th row of $X$. Often the $e_i$'s are taken to be independent and identically distributed with a known density $f(e_i)$. In principle we can allow the density $f(e) = f_\epsilon(e)$ on $\mathbb{R}^n$ to depend on additional parameters, which in particular can allow for correlation among the observations, although in the examples we simply examine inference for several fixed values of $\lambda$.

Standard methods of conditioning for transformation models (Fisher, 1934; Verhagen, 1962; Fraser, 1957; Fraser, 1979, p. 113) use the one-to-one change of variables from $y$ to $\{\tilde{\beta}(y), \tilde{\sigma}(y), d(y)\}$, where $d(y) = (y - X\beta(y))/\tilde{\sigma}(y)$, and $\tilde{\beta}(y)$ and $\tilde{\sigma}(y)$ are estimates of location and scale, for example maximum likelihood or least-squares estimates, that satisfy $\tilde{\beta}(y) = \beta + \sigma \tilde{\beta}(e)$, and $\tilde{\sigma}(y) = \sigma \tilde{\sigma}(e)$. The pair $(\tilde{\beta}, \tilde{\sigma})$ forms a transformation variable (Fraser, 1979, pp. 113, 138). The configuration statistic $d(y) = d(e)$ is a vector of standardized residuals which describes the shape of the sample; it is ancillary with respect to $\beta$ and $\sigma$, and has density

$$h_\lambda (d) = \int_{R^p \times R^n} f_\lambda(Xb + sd) s^{n-p-1}|X'X|^1 \, db \, ds.$$

The conditional distribution of $\tilde{\beta}(y) = b$, $\tilde{\sigma}(y) = s$ given $d$ is

$$f(b, s | d) \, db \, ds = h_\lambda^{-1}(d) f_\lambda([X(b - \beta) + sd] \sigma^{-1}) (s/\sigma)^n s^{-(p+1)} |X'X|^1 \, db \, ds \quad (1.1)$$
on $R^p \times R^+$ (Fraser, 1979, p. 114). A log transformation on $s$ and $\sigma$ to linearize the scaling gives the conditional density of $\tilde{\beta}(y) = b$ and log $\tilde{\sigma}(y) = w$ as

$$f(b, w \mid d) \, db \, dw = h^{-1}_\lambda(d) f_\lambda \left[ e^{\omega \left\{ X(b - \beta) + e^w d \right\}} \right] e^{n(w - \omega)} e^{-\omega w} |X^\prime X|^\frac{1}{2} \, db \, dw,$$  

(1.2)

where $\omega = \log \sigma$. The functionally unique separation of $\beta$ and log $\sigma$ is obtained from the pivots

$$t = (\tilde{\beta}(y) - \beta)/\tilde{\sigma}(y), \quad w = w(y) - \omega;$$

the conditional joint distribution of $t$ and $w$ is

$$f(t, w \mid d) \, dt \, dw = h^{-1}_\lambda(d) f_\lambda \left\{ (Xt + d) e^w \right\} e^{nw} |X^\prime X|^\frac{1}{2} \, dt \, dw.$$  

(1.3)

Inference for any one parameter, say $\beta_p$, or $w$, is based on the marginal density obtained from integrating out the appropriate $p$ components in (1.3). For the case of $\beta_p$ the marginal density is

$$f(t_p) = \int f(t, w) \, dt_1 \ldots dt_{p-1} \, dw;$$  

(1.4)

this pivotal density provides inference for the parameter $\beta_p$ through the pivot $t_p = \{\tilde{\beta}_p(y) - \beta_p\}/\tilde{\sigma}(y)$.

Fraser (1957, 1968) shows that this is essentially the only way to obtain a marginal distribution for $\beta_p$, the component of interest, that is free of the other parameters and conditional on the observed configuration statistic $d$. One way to motivate conditioning on $d$, in addition to its ancillarity, is the fact that $d(y) = d(e)$; that is, $d$ represents a characteristic of the unknown error $e$ that is actually observed.

The functions in (1.1), (1.2), (1.3) are directly available from the error density $f_\lambda$, except for the norming constant $h_\lambda(d)$. When the density depends on a parameter $\lambda$, the coordinates $\tilde{\beta}, \tilde{\sigma}$ are best chosen to be free of $\lambda$, so least-squares estimates are more convenient than constrained maximum likelihood estimates. The distributions (1.3) and (1.4) are essentially independent of this choice of transformation coordinates (Fraser, 1979, p. 112).

In the examples discussed in § 3, $\lambda$ is the degrees of freedom for the Student distribution used for error, and several fixed values are examined. In general, $\lambda$ can be estimated from the data by maximizing $h_\lambda(d)$; this point is discussed further by Fraser (1976; Fraser, 1979, pp. 27, 122, 127).

A simpler version of the regression model, where $y = X\beta + e$ and the error $e$ has a density $f(e)$ on $R^n$, can arise in ordinary regression when the scaling is known, or with exponential and gamma versions of generalized linear models. The ancillary statistic is $d(y) = y - X\tilde{\beta}(y)$, where $\tilde{\beta}(y)$ is a transformation variable satisfying $\tilde{\beta}(y) = \beta + \tilde{\beta}(e)$. The conditional distribution of $\tilde{\beta}(y) = b$ given $d$ is then

$$f(b \mid d) \, db = h^{-1}(d) f\{X(b - \beta) + d\} |X^\prime X|^\frac{1}{2} \, db$$  

(1.5)

on $R^p$, and the normalizing constant is obtained by integration in $p$ dimensions; this is a location family, and the marginal distribution for $b_p - \beta_p$ is obtained by integrating out the remaining $p - 1$ components in (1.5).

In order to obtain an observed significance level for testing the hypothesis $H_0$: $\beta_p = \beta_{p0}$, say, in the general case (1.3), it is necessary to integrate out the other components, and then to integrate the resulting marginal density to obtain a tail area. One method of approximating this $(p + 1)$-dimensional integration is proposed by DiCiccio, Field &
Fraser (1990) based on a change of variable to the likelihood ratio statistics. A related numerical technique for high-dimensional integration is given by Tierney, Kass & Kadane (1989).

In § 2 we consider approximating the marginal density for the component of interest by a conditional density, thus bypassing direct high dimensional integration. The conditional density is constructed to be as independent as possible of the conditioning, and the observed level of significance is then obtained by one-dimensional integration. As a further simplification to avoid the one-dimensional integration, the parameterization invariant tail area approximation of Fraser (1990) can be used, giving results corresponding closely to those of DiCiccio et al. (1990). The extent to which the conditional density is independent of the conditioning is investigated numerically, and leads to a Monte Carlo average that will typically converge to, and thus approximate more closely, the marginal observed level of significance. The results are illustrated on two examples in § 3.

2. Tail probability approximation

We write the component of interest as \( x_r \) and the remaining components as \( x_{0i} \); so, for example, if we want to test \( H_0 : \beta_p = \beta_{p0} \) then \( (x_{0i}, x_r) = (w, t_1, \ldots, t_{p-1}, t_p) \), where \( x_r = t_p = (b_p - \beta_{p0})/s \). If we are interested in inference for \( \sigma \), then \( (x_{0i}, x_r) = (t_1, \ldots, t_p, w) \). If the scale is known, as in (1·5), then \( p = r \) and \( (x_{0i}, x_r) = (b_1 - \beta_1, \ldots, b_p - \beta_p) \). The general problem can then be formulated as evaluating \( \Pr(x_r \geq x_r^{\text{obs}}) \) from the unnormed joint density \( cf(x_{0i}, x_r) \). The method can also be applied to the Bayesian case by taking \( cf(x_{0i}, x_r) \) as the posterior density for a vector parameter, with \( \nu \) denoting the parameter of interest.

We initially propose approximating the marginal density \( f(x_r) \), by a one-dimensional conditional density of the form

\[
\text{cf} \{ \hat{x}_{0i}(x_r), x_r \},
\]

where \( \hat{x}_{0i}(x_r) \) is the value of \( x_{0i} \) that maximizes \( f(x_{0i}, x_r) \) for each \( x_r \). To motivate the form of (2·1), we show how it is related to the conditional density of \( x_r \), given the restricted maximum likelihood estimate \( \hat{\beta}(\beta_p) = \{ \hat{\beta}_1(\beta_p), \ldots, \hat{\beta}_{p-1}(\beta_p) \} \) in the regression model (1·5). This conditional density, given the restricted maximum likelihood estimate, is

\[
\text{cf} \{ \hat{x}_{0i}(x_r) + \hat{\beta}_{0i}(\beta_p) - \beta_{0i}, x_r \}
\]

because of the location form of the joint density (1·5). Evaluation of this expression at \( \beta_{0i} = \hat{\beta}_{0i}(\beta_p) \) gives (2·1).

Conditioning on the restricted maximum likelihood estimate is suggested by Cox & Reid (1987) and Fraser & Reid (1988), in the context of eliminating nuisance parameters; here it provides a starting point for approximating the marginal density of interest.

To improve the approximation of the marginal density by the conditional density, we modify the original joint distribution in a way that does not change the marginal distribution of the component of interest but may improve the approximation of the conditional distribution to the marginal. This amounts essentially to the construction of a pseudo-model for the data that has the same marginal density for the component of interest. Let

\[
\hat{j} = \left[ -\frac{\partial^2}{\partial x \partial x'} I(x) \right]_{(\hat{x}_{0i},\hat{x}_r)}
\]
be the \( r \times r \) negative Hessian of \( l(x) = \log f(x_0, x_r) \) at its overall maximum \((\hat{x}_0, \hat{x}_r)\), and

\[
\hat{f}(x_r) = \left[-\frac{\partial^2}{\partial x_0 \partial x_r} l(x_0, x_r)\right]_{x_0=\hat{x}_0, x_r=\hat{x}_r}
\]  

(2.2)

be the \( r - 1 \times r - 1 \) negative Hessian of \( l(x_0, x_r) \) for fixed \( x_r \) at its restricted maximum. Let \( \hat{f}^{-1}(x_r) \) be the right positive lower triangular square root of \( \hat{f}(x_r) \).

Now define a new variable \( z_{r0} = \hat{f}^{-1}(x_r)\{x_0 - \hat{x}_0\}(x_r)\): the joint density for \((z_{r0}, x_r)\) is

\[
c g(z_{r0}, x_r) = c|\hat{f}(x_r)|^{-1} f\{\hat{x}_0(x_r) + \hat{f}^{-1}(x_r)z_{r0}, x_r\}.
\]  

(2.3)

The conditional density (2.1) obtained from \( g \) is \( c g(0, x_r) \). Hence the observed level of significance, \( pr(x_r | x_r^{\text{obs}}) \) is approximated by

\[
\text{ols } (x_r^{\text{obs}} \mid z_{r0} = 0) = \int_{x_r^{\text{obs}}}^{\infty} g(0, x_r) \, dx_r / \int_{-\infty}^{\infty} g(0, x_r) \, dx_r.
\]  

(2.4)

We refer to this as the ridge observed level of significance.

The density \( g \) is constructed to have a standardized form for each value of \( x_r \). It is maximized at \((0, \ldots, 0, x_r)\), that is, it has a ridge with that coordinate, and the Hessian with respect to \( z_{r0} \) is constant along the ridge. As the location of the maximum and the curvature at the maximum have been standardized, the conditional observed level of significance \( \text{ols } (x_r^{\text{obs}} \mid z_{r0}) \) is independent of the value \( z_{r0} \) to second order in the neighbourhood of the maximum marginal density for \( z_{r0} \).

The exact marginal observed level of significance \( \text{ols } (x_r^{\text{obs}}) \) can be expressed as

\[
\text{ols } (x_r^{\text{obs}}) = E\{\text{ols } (x_r^{\text{obs}} \mid z_{r0})\} = \int \text{ols } (x_r^{\text{obs}} \mid z_{r0}) c h(z_{r0}) \, dz_{r0},
\]  

(2.5)

where \( c h(z_{r0}) = \int c g(z_{r0}, x_r) \, dx_r \) is the marginal density for \( z_{r0} \). To the extent that the conditional observed level of significance accurately approximates the marginal, the values \( \text{ols } (x_r^{\text{obs}} \mid z_{r0}) \) in (2.5) will be constant in \( z_{r0} \). This suggests that we can evaluate the accuracy of the conditional approximation (2.4) by examining the values

\[
\text{ols } (x_r^{\text{obs}} \mid z_{r0}) = \int_{x_r^{\text{obs}}}^{\infty} g(z_{r0}, x_r) \, dx_r / \int_{-\infty}^{\infty} g(z_{r0}, x_r) \, dx_r;
\]  

(2.6)

we describe these as coming from statistically parallel conditional distributions.

Since the marginal density for \( z_{r0} \) is \( r - 1 \) dimensional and the normalizing constant is not available, we sample values \( z_{r0} \) from a constructed marginal density, \( h^*(z_{r0}) \). This gives an importance sampling estimate of (2.5) as

\[
\text{ols } (x_r^{\text{obs}} \mid z_{r0}) \propto \frac{\sum \text{ols } (x_r^{\text{obs}} \mid z_{r0}) h(z_{r0})/ h^*(z_{r0})}{\sum h(z_{r0})/ h^*(z_{r0})},
\]  

(2.7)

where the sums are over \( i = 1, \ldots, N \).

In the examples considered in § 3, it is indeed the case that the individual \( \text{ols } (x_r^{\text{obs}} \mid z_{r0}) \) are relatively constant in \( z_{r0} \), which means that (2.4) is a successful approximation to (2.5) and also that the exact form chosen for \( h^*(z_{r0}) \) may not be too important. For each value of \( z_{r0} \) computation of the \( i \)th term in (2.7) involves one-dimensional numerical integration. We refer to (2.7) as the mean conditional observed level of significance and to the degree that the Monte Carlo sampling is successful, it will provide an improved approximation to the desired marginal observed level of significance.
Nonnormal linear regression

We now record densities $h^*(z_{(0)})$ that seem appropriate to use in the preceding Monte Carlo sampling. As a guide we consider the case of independent normally distributed errors and calculate the conditional distribution of $z_{(0)}$ given $z_r$; since this distribution turns out to be free of $z_r$ it is equal to the marginal distribution. For the linear model with known scale (1.5), the distribution of $z_{(0)}$ with normal errors is that of a sample of $r-1$ from the standard normal. Accordingly we sample $r-1$ independent standard normal observations $(e_1, \ldots, e_{r-1})$, set $z_{(0)} = (e_1, \ldots, e_{r-1})$ and use

$$h^*(z_{(0)}) = c \exp \left( -\frac{1}{2} \sum_{i=1}^{r-1} e_i^2 \right).$$

For the case of unknown scale (1.3), with normal errors, it can be shown that the distribution of $z_{(0)}$ is that of

$$\left\{ (2n)^{k_1} \log \left( \frac{s}{n^{k_1}} \right), n^1 \frac{e_1}{s}, \ldots, n^1 \frac{e_{r-1}}{s}, \right\}, \quad (2.8)$$

where $(e_1, \ldots, e_n)$ is a standard normal sample and $s^2 = e_1^2 + \ldots + e_n^2$. Accordingly for inference in the general model (1.3) we use $z_{(0)}$ equal to the vector (2.8) and

$$h^*(z_{(0)}) = c \exp \left( -\frac{1}{2} \sum_{i=1}^{n} e_i^2 \right) \left( \sum_{i=r}^{n} e_i^2 \right)^{1/2}.$$

A further simplification that avoids the one-dimensional integration of (2.4) can be obtained as follows. If we treat (2.3) as the basic density for a location model, then the parameter invariant version (Fraser, 1990, eqn (3.1), (6.1)) of the Lugannani & Rice (1980) formula gives

$$\text{ols}^* (x_r^{\text{obs}} | 0) \approx 1 - \Phi(z) - \phi(z) (1/z - 1/q), \quad (2.9)$$

where

$$z = \text{sgn} (q) \left[ 2 \{ \log g(0, \tilde{x}_r) - \log g(0, x_r^{\text{obs}}) \} \right]^{1/2},$$

$$q = \tilde{j}^{-1} \left[ -\frac{d}{dx_r} \log g(0, x_r) \right]_{x_r^{\text{obs}}} ,$$

and where $\tilde{j}$ is the negative second derivative of $\log g(0, x_r)$ at its maximum $\tilde{x}_r$. This corresponds to formula (19) of DiCiccio et al. (1990) and provides an approximation to the ridge estimate ols $\left( x_r^{\text{obs}} | 0 \right)$; it avoids the one-dimensional integration and uses values of $g(0, x_r)$ at only the maximum $\tilde{x}_r$ and the observed $x_r^{\text{obs}}$ along the ridge. We refer to it as the two-point ridge estimate.

3. Examples

3.1. Introduction

We now have three approximations for the observed level of significance concerning a hypothesized value for a real parameter: the two-point ridge approximation (2.9), the ridge approximation (2.4), and the mean conditional approximation (2.7); we would expect these to be of increasing accuracy.

In this section we examine two sets of data, a 3-parameter regression model with simulated data and a 6-parameter regression model for 26 observations of house prices. For the first example exact results from direct numerical integration are available, although they have been rounded to two significant digits. For the second example we are able to examine and compare the three approximations in a 6-dimensional integration context.
3.2. Three-parameter regression example

The data for this example are discussed by Fraser (1979, p. 126). For 25 values of an explanatory variable $x$ at unit step size from $-12$ to $+12$, the response values for $y$ are recorded in Table 1. The model used is $y = \alpha + \beta x + \sigma e$, where $e$ is standardized Student on $\lambda$ degrees of freedom.

The 95% confidence interval for $\beta$ based on normal theory, $\lambda = \infty$, is $(0.79, 0.99)$ where original values, now available, have been rounded to two figures. The ordinary value of $t$ for testing $\beta = 1$ is $-2.0920$. Confidence intervals for $\beta$ computed by direct three-dimensional numerical integration were reported by Fraser (1979, p. 130) and are given in line 1 of Table 2, for various choices of $\lambda$; the intervals obtained by the ridge are given in line 2. The data were in fact generated from $y = \alpha + \beta x + \sigma e$ with $\alpha = 20$, $\beta = 1$, $\sigma = 1.1966$, $\lambda = 3$. In accord with this we find that the Student based intervals are more accurate, shorter and better centred. Also shown in Table 2 are the observed levels of significance for testing $\beta = 1$ calculated by the ridge method (2.4) for various $\lambda$ values. The values for the normal case, $\lambda = \infty$, were calculated directly. The smaller values of $\lambda$ tolerate extreme responses better, and give quite different observed levels of significance than the normal theory values.

We also used the methods in §2 to assess the accuracy of the lower confidence bounds $0.92, 0.91, 0.87$ for $\lambda = 1, 3, 6$ degrees of freedom. In Table 3 we record the two-point ridge value (2.9), the ridge value (2.4), and the mean conditional value (2.7) obtained from a Monte Carlo sampling of $N = 30$ parallel distributions. We would expect the Monte Carlo averages to be quite close to the true values and this is indicated by the

<table>
<thead>
<tr>
<th>Table 1. A sample of 25 regression responses with Student, error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$12.8607$</td>
</tr>
<tr>
<td>$23.2321$</td>
</tr>
<tr>
<td>$29.1283$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Confidence intervals for $\beta$ using a Student, analysis with $\lambda = 1, 3, 6, \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 1$</td>
</tr>
<tr>
<td>Integration</td>
</tr>
<tr>
<td>Ridge</td>
</tr>
<tr>
<td>Ridge ols</td>
</tr>
</tbody>
</table>

Integration values are from Fraser (1979, p. 130); ridge values based on (2.4). Also shown, observed levels of significance for testing $\beta = 1$ using ridge method.

<table>
<thead>
<tr>
<th>Table 3. Observed levels of significance for $\beta$ values with $\lambda = 1, 3, 6, \infty$; ordinary $t$ also recorded</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 1$</td>
</tr>
<tr>
<td>$\beta_0$</td>
</tr>
<tr>
<td>$t$</td>
</tr>
<tr>
<td>Two-point ridge</td>
</tr>
<tr>
<td>Ridge</td>
</tr>
<tr>
<td>Mean cond.</td>
</tr>
</tbody>
</table>
standard errors. The ols \((x_r^{obs}|0)\) calculated by one-dimensional integration along the ridge is preferable to the two-point calculation ols* \((x_r^{obs}|0)\) based on the same ridge. The level should be 2.5% if the lower confidence bounds are exact, and the discrepancy of the mean conditional values from this is probably due to rounding errors in the values reported by Fraser (1979, p. 130).

A plot shows that the Student\(_3\) error distribution produced one large response at \(x = -11\) and two low responses at \(x = 6, 9\). These extreme points led to a low normal-theory estimate 0.8952 for the true \(\beta\); the estimate based on the correct Student\(_3\) analysis is 0.9853, much closer to the true value. The confidence intervals in Table 2 show clearly the effect of reducing the \(\lambda\) value for the analysis: with a lower \(\lambda\) value the extreme data points are better tolerated and the lower confidence limit changes towards a value indicated by the main body of the data.

We examine further the ability of the lower \(\lambda\) analyses to tolerate extreme data points by changing the data value 7.9042 to 17.9042 and the data value 32.6893 to 22.6893; these changes reduce the least-squares estimate to 0.711. Table 4 records the 95% confidence intervals for \(\lambda = 1, 3, 6, \infty\), and the observed level of significance for testing the true \(\beta = 1\). Comparison with Table 2 shows clearly how the lower \(\lambda\) analyses accommodate even these modified data points.

Table 4. \textit{Data of Table 1 with value 7.9042 changed to 17.9042 and value 32.6893 changed to 22.6893: 95% confidence intervals; observed levels of significance for testing \(\beta = 1\)}

<table>
<thead>
<tr>
<th>(\lambda)</th>
<th>(\lambda = 1)</th>
<th>(\lambda = 3)</th>
<th>(\lambda = 6)</th>
<th>(\lambda = \infty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge interval</td>
<td>(0.863, 1.021)</td>
<td>(0.731, 1.012)</td>
<td>(0.627, 0.965)</td>
<td>(0.557, 0.864)</td>
</tr>
<tr>
<td>Modal estimate</td>
<td>0.947</td>
<td>0.922</td>
<td>0.824</td>
<td>0.711</td>
</tr>
<tr>
<td>Two-point ridge ols</td>
<td>8.07%</td>
<td>5.46%</td>
<td>1.73%</td>
<td>0.04%</td>
</tr>
<tr>
<td>Ridge ols</td>
<td>7.82%</td>
<td>4.32%</td>
<td>0.75%</td>
<td>0.04%</td>
</tr>
<tr>
<td>Mean cond. ols</td>
<td>7.38 ± 0.27%</td>
<td>4.40 ± 0.06%</td>
<td>0.78 ± 0.025%</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

Ordinary \(t\) value is -3.899.

3.3. \textit{Six-parameter regression analysis}

To illustrate the procedures in higher dimensions we examine a set of housing data (Srivastava & Sen, 1987, p. 33). The variables examined are the selling price \(y\), the number of bedrooms \(x_1\), the floor space \(x_2\), the number of rooms \(x_3\), and the front footage \(x_4\). The model used is

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \sigma e, \]

where \(e\) is taken to be standardized Student with degrees of freedom \(\lambda = 3\), to allow for longer tails and for extreme values. We first calculate the observed level of significance for testing \(\beta_4 = 0\). The two point, the ridge, and the mean conditional values are 12.19%, 13.76% and 12.85%(±0.41%), respectively, based on the Student\(_3\) analysis. For comparison the normal theory value is 7.95%.

The 90%, 95% and 99% confidence intervals for \(\beta_4\) using the ridge approximation and the Student\(_3\) analysis are given in Table 5; the normal theory results are also given.

Location-scale analysis (Fraser, 1979, pp. 37-46) indicates that Student\(_3\) and normal analyses tend to be in close agreement with normal data. However, for extremely
Table 5. Housing data confidence intervals for $\beta_4$ using Student$_3$ analysis and normal analysis

<table>
<thead>
<tr>
<th></th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student$_3$ analysis</td>
<td>$-0.119, 0.578$</td>
<td>$-0.172, 0.655$</td>
<td>$-0.315, 0.835$</td>
</tr>
<tr>
<td>Normal analysis</td>
<td>$-0.053, 0.644$</td>
<td>$-0.126, 0.717$</td>
<td>$-0.278, 0.869$</td>
</tr>
</tbody>
</table>

Fig. 1. Residual plotted against orthogonalized $x_4$ for housing data.

Fig. 2. Observed level of significance plotted against absolute value of $z(0)$ for housing data.

nonnormal data such as that from the Student$_3$ distribution, the normal analyses tend to give results far from those of the correct analysis.

For the present example we note that the Student$_3$ intervals are approximately 0.05 to the left of the normal intervals. This suggests some extreme observations that inflate the least-squares estimate. Figure 1 shows the residuals against $x_4$, orthogonalized with respect to $(x_1, x_2, x_3)$. There is a large negative residual on the left in a group of four data points; this might be expected to be a high leverage point for the estimation of $\beta_4$. The $\lambda = 3$ analysis tolerates this extreme value and adjusts the least-squares estimate downward.

To get a better view of the Monte Carlo sampling of the conditional observed levels, ols ($x_r^{obs} | z(0)$), we plot them against the Euclidean distance $|z(0)|$ of the sampled values $z(0)$ in Fig. 2. The $z(0) = 0$ value is given by the ridge ols ($x_r^{obs} | 0$) which is 0.1376%. For progressively larger values of $|z(0)|$ the observed levels fan out from the preceding value, but are remarkably stable given the complexity of the six-dimensional integration problem.

4. Discussion

We have developed a conditional distribution that approximates a marginal distribution needed for testing a real valued parameter in nonnormal regression models. The sampling
of parallel conditional distributions then provides a validation of the approximate observed level of significance.

The numerical difficulty of integrating out $r - 1$ components for inference on the remaining component is largely bypassed by the ridge approximation. The one-dimensional integration then needed to evaluate the observed level of significance can be bypassed by a Lugannani & Rice type approximation, at the expense of some accuracy in the result. DiCiccio et al. (1990) show that this two-point approximation is accurate to $O(n^{-3/2})$, so the ridge and mean conditional approximations will have at least this order of accuracy.

The technique described here can also be applied to the approximation of marginal posterior densities, and is closely related to the use of the Laplace approximation (Tierney et al., 1989). Accurate approximations to marginal densities based on polynomial approximations to the likelihood function are discussed by DiCiccio et al. (1990), and can be shown to be equal to the two-point ridge approximation. The ridge and Monte Carlo sampling methods then provide further refinements to that approximation.

REFERENCES


[Received July 1989. Revised December 1989]