Title
Properties of Modelling the Error Distribution with An Extra Shape Parameter

Author
Jeremy M. G. Taylor

Publication Date
2011-10-24

Computational Statistics & Data Analysis 13 (1992) 33–46
North-Holland

Properties of modelling the error distribution with an extra shape parameter

Jeremy M.G. Taylor

Department of Biostatistics, School of Public Health, University of California at Los Angeles, Los Angeles, CA 90024, USA

Received March 1990
Revised July 1990

Abstract: A robust extension of normal theory regression is to add an extra parameter to model the kurtosis of the error distribution, for example by using the T-family or the power-exponential family of distributions. The statistical properties of maximum likelihood estimation schemes for both families of models are considered. This article extends the work of Lange et al. (1989) in which the usefulness of the multivariate T-family for modelling data was demonstrated. The cost of adding the extra kurtosis parameter to the standard normal model is considered. This cost is measured by comparing the variance of the quantity of interest when the estimation of the extra parameter is taken into account and the variance when the estimated value of the extra parameter is treated as if it were known. In Lange et al. (1989), it is shown in a general setting that asymptotically there is no cost due to the estimation of the extra parameter, where the quantity of interest is the location or regression parameters. Whether this finding remains valid in small samples for the T and power-exponential families is investigated in this paper using Monte Carlo simulations and a real dataset. The efficiency of parameter estimates and coverage rates are also considered under three different scenarios: when the extra kurtosis parameter is estimated from the data, when the extra parameter is fixed at the true value, and when it is fixed at a wrong value. The expected information matrix is used to estimate the confidence intervals, and the comparisons are based on asymptotic calculations and Monte Carlo simulations. It is found that for the T-family there is very little cost due to estimation of the extra parameter, except for small sample sizes. The inflation in variance due to the estimation of the extra parameter increases as the sample size decreases. In the Monte Carlo simulations of simple regression settings the inflation in variance is found to be at most 14% for the T-family and at most 62% for the power-exponential family. For the T-family, there is a considerable loss of efficiency in fitting a normal model when the true degrees of freedom is small, but only a small loss of efficiency if a model with a low number of degrees of freedom is fit to normal observations. In contrast, the coverage rates of confidence intervals are close to the nominal level if a normal model is fit to the data whatever the true degrees of freedom, but the coverage rates can be too low if a model with a low number of degrees of freedom is fit to normal data. The coverage rates of confidence intervals when the degrees of freedom is estimated from the data are satisfactory, except at small sample size. Similar results are obtained for the power-exponential family. In addition to the larger inflation in variance due to estimation of the extra parameter, the power-exponential family is less satisfactory then the T-family because the extra kurtosis parameter is frequently estimated to be at the boundary of its range in small samples. The findings in this article support the notion that the extra kurtosis parameter should be estimated for large samples and fixed at an appropriate value for small samples.

Keywords: Power-exponential family, T-family, Linear regression, Variance inflation, Robustness, Adaptive estimation.

1. Introduction

A variety of methods exist for the analysis of data containing extreme or outlying observations. One approach is through identification and elimination of outliers followed by standard normal theory methods. Another approach is through accommodation of the extreme observations using classical robust methods such as M-estimation (Huber, 1981). An alternative accommodation approach is through modelling the error distribution by using the power-exponential family (Box and Tiao, 1973) or the T-family (Jeffreys (1939), Fraser (1976)). Some authors have suggested using a T-model with 4 degrees of freedom (Lange et al, 1989), whereas others have suggested 6 degrees of freedom. Lange et al (1989) illustrate the use of the T-family in a wide variety of settings including linear and non-linear regression, repeated measures and pedigree data. This modelling approach is similar in spirit to the adaptive robust approach (Hogg, 1974) in which, loosely speaking, the choice of estimation method is based on the observed residual distribution. The usefulness of this adaptive approach has been demonstrated in simulation studies (Hogg, 1967, Hogg et al, 1972, Yuh and Hogg, 1988).

In this paper statistical properties of the adaptive modelling approach are considered. The efficiency of the estimates and the coverage rates of confidence intervals are considered, but the primary focus is on the effect of the adaptive stage of the analysis on the variance of quantities of interest. For example in a simple regression problem, if the degrees of freedom of the T-family residual distribution is estimated to be five, is it appropriate when considering the standard error or confidence interval of the slopes to assume that the degree of freedom was known and fixed at the value five? Similar issues arise in other families of models. In the power transformation family there have been a number of recent articles discussing whether or not to treat the transformation parameter as if it were fixed after it has been estimated from the data (see Carroll and Ruppert (1981), Bickel and Doksum (1981), Hinkley and Runger (1984), Box and Cox, (1982), Taylor, (1986, 1988)). Another similar situation is the extension of the logistic probability response curve to other sigmoid shapes through the addition of an extra parameter (Taylor, 1989).

This issue is important because statisticians are rightly hesitant to add extra parameters to models to improve the fit, because although it may decrease the bias, it will also tend to increase the variance. However, if it can be shown to increase the variance by only a small amount for the particular model under consideration, then this should reduce the reluctance to add the extra parameters.

The approach to estimating the cost of the adaptive stage of the analysis mimics that used by others in the power transformation family (Bickel and Doksum (1981), Carroll and Ruppert (1981), Taylor, (1986)). The approach is to compare the variance of quantities of interest when the extra parameter is treated as estimated compared to the variance when the extra parameter is treated as fixed.

Section 2 summarizes the asymptotic theory of Lange et al. in the regression setting. Further asymptotic results relating to the efficiency and coverage rates of confidence intervals under misspecification of the kurtosis parameter are also given. Section 3 considers the small sample situation through an example and Monte Carlo simulations.

Throughout this article, the asymptotic and Monte Carlo results are based on maximum likelihood estimates. Three possible algorithms to find the estimates are described by Lange et al. (1989).

2. Asymptotic theory

Consider the linear regression model

$$Y = X\theta + e, \qquad e \sim f,$$

where f is either a T-distribution with density

$$f(e) = \frac{\frac{1}{\sigma}\Gamma(\frac{1}{2}(k+1))}{\sqrt{\pi k}\,\Gamma(\frac{1}{2}k)}\left(1 + \left(\frac{e}{\sigma}\right)^2/k\right)^{-\frac{1}{2}(k+1)}, \qquad 0 < k \leq \infty,$$

characterized by a shape parameter k (degrees of freedom), or a power-exponential distribution.

$$f(e) = \frac{D}{\sigma}\exp\left\{-\frac{1}{2}\left|\frac{e}{\sigma}\right|^{2/1+k}\right\}, \qquad -1 < k \leq 1,$$

characterized by shape parameter k, where $D^{-1} = \Gamma(1 + ((1+k)/2))\ 2^{1+\frac{1}{2}(1+k)}$. Notice that the power-exponential family includes both heavier ($k > 0$) and lighter ($k < 0$) than normal tails ($k = 0$), whereas the T-family only includes heavier ($k < \infty$) than normal tails ($k = \infty$). The T-family and the power-exponential family are special cases of elliptically symmetric families of densities. In Lange et al. (1989) the score vector and the expected information matrix are derived for an arbitrary elliptically symmetric family. The three lemmas below are special cases of these results.

Lemma 1. *For both the T-family and the power-exponential family,* $E[\partial^2 \log L/\partial\theta\partial\sigma] = 0$ *and* $E[\partial^2 \log L/\partial\theta\partial k] = 0$, *where L denotes the likelihood from a sample of n observations.*

Proof. This is a result of symmetry of the error distribution.

The implication of lemma 1 is that functions of the regression parameters θ are asymptotically uncorrelated with the estimated shape parameter, thus at least for large samples we expect there to be little or no cost associated with estimating the extra parameter. In addition the extra shape parameter can be treated as fixed for inference about θ even though it is estimated from the observations.

Lemma 2. *For the T-family model*

$$-E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] = \frac{1}{\sigma^2}[X^T X]\left(\frac{k+1}{k+3}\right). \tag{1}$$

Proof. See Lange et al. (1989).

Lemma 3. *For the power-exponential family*

$$-E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] = \frac{1}{\sigma^2}[X^T X]\frac{2^{-k}}{(1+k)}\frac{\Gamma\left(1+\frac{1-k}{2}\right)}{\Gamma\left(1+\frac{1+k}{2}\right)}. \tag{2}$$

Proof. The result follows from lengthy algebra.

The implications of lemma's 2 and 3 are that asymptotic standard errors are relatively simple to evaluate after the more general model has been fit and the extra parameter estimated.

In addition to considering the properties of the parameter estimates when the model is known to hold, it is also of interest to consider the situation where the model is misspecified. Let $\hat{\theta}^*$ be the maximum likelihood estimate of θ when an analysis is carried out with the extra parameter fixed at a value k^* when in fact k_0 is the true value of k. The asymptotic variance of $\sqrt{n}\,\hat{\theta}^*$ is given by (Godambe and Heyde, 1987)

$$A = E_0\left(\left[\frac{-\partial^2 \log L_*}{\partial \theta^2}\right]^{-1}\right)' E_0\left[\frac{\partial \log L_*}{\partial \theta}\cdot\frac{\partial \log L'_*}{\partial \theta}\right] E_0\left[\frac{-\partial^2 \log L_*}{\partial \theta^2}\right]^{-1}. \tag{3}$$

In this expression, L_* is the likelihood corresponding to k^* and it is evaluated at (θ^*, σ^*), which are the values of (θ, σ) which solve the equations $E_0(\partial \log L_*/\partial\theta) = 0$ and $E_0(\partial \log L_*/\partial\sigma) = 0$. The expectation, denoted by E_0, is with respect to the true distribution of the observations, the parameters of the true distribution are denoted by $(\theta_0, \sigma_0, k_0)$. The values of σ_0 and k_0 may differ from σ^* and k^*.

The expression (3) can be compared with $V = E_0[-\partial^2 \log L_0/\partial\theta^2]^{-1}$. V is the asymptotic variance of $\sqrt{n}\,\hat{\theta}_0$, where $\hat{\theta}_0$ is the estimate when k is fixed at k_0.

Thus, the loss in asymptotic efficiency in using the wrong value of k can be assessed.

The asymptotic coverage rate of confidence intervals can be calculated when the model is fit with k fixed at k^* and the true distribution has $k = k_0$. If the expected information expression ($C = E_*[-\partial^2 \log L_*/\partial\theta^2]^{-1}$) given by equations (1) and (2) is used to construct the confidence intervals, then the asymptotic coverage rate of a nominal 95% confidence interval is given by $P(|Z| < 1.96R)$, where Z is a standard normal random variable and $R^2 = C/A$.

In general, the evaluation of A, V and R require numerical integration, but the special case ($k^* = \infty$) is worth considering for the T-family. When $k^* = \infty$, then

$$-E_0\left[\frac{\partial^2 \log L_*}{\partial\theta^2}\right] = E_0\left[\frac{\partial \log L'_*}{\partial\theta} \cdot \frac{\partial \log L_*}{\partial\theta}\right]$$

$$= \frac{k_0 - 2}{\sigma_0^2 k_0}[X^T X] \Rightarrow A = \left(\sigma_0^2 k_0/(k_0 - 2)\right)[X^T X]^{-1},$$

also $V = (\sigma_0^2(k_0 + 3)/(k_0 + 1))[X^T X]^{-1}$. Thus the asymptotic variance is increased by a factor $k_0(k_0 + 1)/(k_0 - 2)(k_0 + 3)$ if a normal model is fit to the observations. It can also be shown that $R = 1$, so the asymptotic coverage rate attains the correct value of 95%.

The asymptotic results for both the T-family and the power-exponential family for specific choices of k_0 and k_* are shown in the tables together with the small sample Monte Carlo simulation results.

3. Small sample properties

(a) *An example*

The data relating the number of traffic fatalities to the number of drivers in each state (Draper and Smith, 1981, p. 191) was analyzed using the power-exponential model. These data were analyzed using the T-family model in Lange et al. (1989). For these data the explanatory variable is $\log_{10}$ (number of drivers in 1964) and the dependent variable is $\log_{10}$ (number of driving deaths in 1964). A plot of the data indicates that a linear fit is appropriate, with 2 possible outliers (Rhode Island and Connecticut).

Table 1 shows the estimates and standard errors from least squares and the power-exponential model. The maximum likelihood estimate of the shape parameter of $k = 0.56$ suggests that the observations are better explained by a model with heavier than normal tails. A similar result was found using the T-model where the estimated degrees of freedom was 4.6.

The profile likelihood confidence interval for k is $(-0.14, 1.0)$; this wide interval indicates that a large sample size is needed to estimate accurately the shape of the residual distribution. The table also shows the estimates and standard errors of four Bootstrap resampling schemes. In two of the schemes the residuals from the best fitting power-exponential family model with $k = 0.56$ are

Table 1
Parameter estimates and standard errors for Traffic Accident data using Power-exponential model.

	Shape parameter	Intercept	Slope
Least squares	0 (fixed)	−2.937	0.941
		(0.256)	(0.42)
Power-exponential model	0.56	−2.874	0.934
		(0.227)	(0.037)
Bootstrap residuals	0.56 (fixed)	−2.888	0.936
		(0.232)	(0.039)
	estimated	−2.885	0.936
		(0.250)	(0.042)
Bootstrap cases	0.56 (fixed)	−2.863	0.932
		(0.346)	(0.057)
	estimated	−2.844	0.928
		(0.363)	(0.059)

resampled and added to the predicted value from the model and in the other two schemes the pair of observations are resampled. In addition for each scheme for estimating the regression parameters of the model for each Bootstrap sample, k is either fixed at the value 0.56 or estimated from each resampled dataset. The two most striking features of the results from this example are, firstly that Bootstrapping residuals gives estimates of variability comparable with the asymptotic theory based on the expected information matrix, whereas Bootstrapping cases gives much larger variability, and secondly it makes very little difference whether k is estimated or fixed, supporting the asymptotic results in Section 2 that there is no cost associated with estimating the extra parameter.

In a similar analysis of this and other datasets using the T-family presented in Lange et al (1989), qualitatively similar conclusions are reached concerning the Bootstrap schemes.

(b) Monte Carlo simulations

A small Monte Carlo simulation study was performed to ascertain whether the asymptotic results that there is no cost due to estimation of extra parameters remains valid for small samples, and in particular how the inflation in variance depends on the sample size and on the error distribution. In addition, the relative efficiency and coverage rates of confidence intervals were considered in the study.

In the simulation 600 datasets of size $n (n = 20, 40, 80, 160)$ were generated from the model

$$Y = a + bX + e, \qquad e \sim f.$$

The design was $X = \pm i$, $i = 1, 2, \dots, 10$; with an equal number of observations at each of the 20 design points. For the T-family error distribution the values of k and σ were either $(k = 2.5, \sigma = 2)$ or $(k = 5, \sigma = 2)$ or $(k = \infty, \sigma = 2)$. For the power-exponential family error distribution the values of k and σ were either

Table 2
T family Monte Carlo and asymptotic results: Variance and efficiency considerations.

		Variance ratio for estimating k. Var(k estim) ÷ Var(true k) *		Variance ratio: Simulation Var(k estim) ÷ Var(asympt)		Var(k = 2.5, fixed) ÷ Var (true k, fixed)		Var(k = 5, fixed) ÷ Var (true k, fixed)		Var(k = ∞, fixed) ÷ Var (true k, fixed)	
Parameter:		a **	b	a	b	a	b	a	b	a	b
True degrees of freedom (k) = 2.5											
	20	1.07	1.06	0.91	0.91	–	–	1.02	1.00	2.16	2.03
	40	1.03	1.06	0.98	0.98	–	–	1.04	1.07	2.67	2.52
n	80	1.01	1.04	0.99	0.99	–	–	1.02	1.06	2.96	3.22
	160	1.02	1.02	0.98	0.98	–	–	1.07	1.02	2.76	2.59
asymptotic		1.00	1.00	1.00	1.00	–	–	1.05	1.05	3.18	3.18
True degrees of freedom (k) = 5											
	20	1.10	1.08	1.08	1.12	1.01	1.05	–	–	1.28	1.20
	40	1.03	1.05	1.04	1.07	1.05	1.03	–	–	1.21	1.24
n	80	1.03	1.03	0.94	1.09	1.01	1.04	–	–	1.41	1.26
	160	1.01	1.01	1.02	1.10	1.01	1.06	–	–	1.23	1.17
asymptotic		1.00	1.00	1.00	1.00	1.04	1.04	–	–	1.25	1.25
True degrees of freedom (k) = ∞											
	20	1.10	1.06	0.91	0.93	1.34	1.25	1.13	1.09	–	–
	40	1.05	1.04	0.94	0.89	1.25	1.23	1.11	1.08	–	–
n	80	1.02	1.03	0.92	0.91	1.24	1.25	1.10	1.11	–	–
	160	1.02	1.03	1.00	1.00	1.24	1.25	1.11	1.11	–	–
asymptotic		1.00	1.00	1.00	1.00	1.22	1.22	1.09	1.09	–	–

* Var(·) denotes variance of $\hat{a}$ or $\hat{b}$ under the specified assumption for k.
** Linear model, a = intercept, b = regression coefficient.

($k = 0.0$, $\sigma = 2$) or ($k = 0.5$, $\sigma = 2$). The values of a and b were fixed at $a = 0$, $b = 1$. For each simulated data set the parameters a and b were estimated with k fixed at the true value, k fixed at a wrong value or k estimated. In addition 95% confidence intervals were calculated using $\hat{a} \pm 1.96\ \hat{SE}(\hat{a})$ and $\hat{b} \pm 1.96\ \hat{SE}(\hat{b})$; where $\hat{SE}(\hat{a})$ is the estimate of the standard error of $\hat{a}$ obtained from the expected information matrix. That is, equations (1) and (2) were used with k chosen as either the estimated value when it was estimated or as the fixed value assumed in fitting the model when it was fixed. Whether or not this confidence interval contained the true value was noted. The computations were performed using a Newton–Raphson algorithm on an IBM 3090.

Table 2 shows the results for the *T*-family model. The variance ratio for estimating k is the Monte Carlo Variance of the parameter when k is estimated divided by the Monte Carlo Variance of the parameter when k is fixed. Notice that this variance ratio is close to 1 for all situations, however there is a trend for the variance ratio to increase as the sample size decreases. The second column of the table shows the ratio of the Monte Carlo variance of the parameter estimate

Table 3
T family Monte Carlo and asymptotic results; Coverage rates (nominal rate = 95%).

		k estimated		$k = 2.5$, fixed		$k = 5$, fixed		$k = \infty$, fixed	
Parameter:		a *	b	a	b	a	b	a	b
True degrees of freedom $(k) = 2.5$									
	20	87	87	90	91	93	94	94	95
	40	93	93	94	94	95	96	94	94
n	80	94	93	95	93	96	96	97	95
	160	94	96	94	96	95	97	95	96
asymptotic		95	95	95	95	96	96	95	95
True degrees of freedom $(k) = 5$									
	20	90	88	92	90	94	92	93	92
	40	93	92	92	91	95	93	95	93
n	80	96	94	94	92	96	95	95	95
	160	95	94	94	93	95	95	95	95
asymptotic		95	95	93	93	95	95	95	95
True degrees of freedom $(k) = \infty$									
	20	92	91	90	88	92	91	94	94
	40	93	95	91	92	93	94	94	96
n	80	95	96	91	94	94	95	96	96
	160	94	96	90	91	93	95	96	96
asymptotic		95	95	91	91	94	94	95	95

* Linear model, a = intercept, b = regression coefficiet.

to the asymptotic variance calculated using the known values of the parameters. Notice that the ratio is close to one except for $n = 20$. The table also shows the expected result that fitting a normal model to heavy tailed data is very inefficient, but that fitting a T-model with a low number of degrees of freedom to normal data gives a much smaller loss in efficiency. The results for the asymptotic efficiency loss are in reasonable, but not perfect, agreement with the results for small samples.

The coverage rates of confidence intervals are given in Table 3. The coverage rates of the confidence intervals when k is fixed at the correct value are reasonably close to 0.95, except possibly when $n = 20$. When k is estimated from the data the confidence intervals are somewhat optimistic, although they might be considered satisfactory for all values of n except for $n = 20$. When k is fixed at too small a value, the coverage rates are slightly too low even in large samples. In contrast, if a normal model is fit, the coverage rates are adequate for all sample sizes.

Tables 4 and 5 show the results for the power-exponential family. The conclusions are qualitatively similar to those from the T-family. The main differences are that the variance ratios are bigger, although probably acceptably close to 1 except for $n = 20$, and the coverage rates are worse when k is estimated and would not be considered satisfactory for $n = 20$, or 40.

Table 4
Power exponential family Monte Carlo results; Variance and efficiency considerations.

	Variance ratio for estimating k Var(k estim) ÷ Var(true k) *		Variance ratio: Simulation Var(k estim) ÷ Var(asympt)		Var(False k, fixed) ÷ Var(true k, fixed)	
Parameter:	a **	b	a	b	a	b
True shape parameter (k) = 0.5						
20	1.53	1.62	1.74	1.98	1.11	1.09
40	1.26	1.22	1.39	1.40	1.02	1.11
n 80	1.13	1.10	1.25	1.13	1.23	1.17
160	1.05	1.06	1.07	1.19	1.18	1.15
asymptotic	1.00	1.00	1.00	1.00	1.22	1.22
True shape parameter (k) = 0.0						
20	1.54	1.37	1.61	1.50	1.24	1.12
40	1.20	1.29	1.30	1.24	1.15	1.17
n 80	1.13	1.13	1.25	1.24	1.13	1.15
160	1.04	1.05	0.97	1.16	1.18	1.17
asymptotic	1.00	1.00	1.00	1.00	1.16	1.16

* Var(·) denotes variance of $\hat{a}$ or $\hat{b}$ under the specified assumption for k.
** Linear model, a = intercept, b = regression coefficient.

An additional troublesome feature of maximum likelihood estimates for the power-exponential was that for small samples frequently the estimate of k was at the boundary of the parameter space. Neither of the associated densities, uniform ($k = -1$) or double-exponential ($k = 1$) has an appealing shape for real data.

Table 5
Power exponential family Monte Carlo results; Coverage rates (nominal rate = 95%).

	k estimated		$k = 0.5$, fixed		$k = 0.0$, fixed	
Parameter:	a *	b	a	b	a	b
True shape parameter (k) = 0.5						
20	68	67	90	90	91	91
40	87	86	92	93	92	94
n 80	91	93	93	95	93	95
160	93	94	94	95	94	95
asymptotic	95	95	95	95	95	95
True shape parameter (k) = 0.0						
20	60	62	86	86	92	90
40	81	79	88	87	91	93
n 80	89	90	89	90	94	94
160	96	95	92	91	96	95
asymptotic	95	95	91	91	95	95

* Linear model, a = intercept, b = regression coefficient.

Table 6
Monte Carlo simulation and asymptotic results – T family (4 regression parameters).

		β_0 **	β_1	β_2	β_3	β_0	β_1	β_2	β_3
True degrees of freedom (k) = 5									
		Variance * ratio for estimating k; (k estimated/k = 5 fixed)				Variance * ratio. (k estimated/asymptotic)			
	20	1.11	1.12	1.08	1.09	1.21	1.24	1.24	1.27
Sample	40	1.09	1.05	1.07	1.10	1.16	1.13	1.16	1.07
size	80	1.03	1.01	1.04	1.05	1.10	1.10	1.11	1.05
	160	1.02	1.02	1.01	1.01	1.02	1.18	1.04	1.13
asymptotic		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		Relative efficiency * $\frac{\text{Var } k=\infty}{\text{Var } k=5}$				Coverage rate (k = 5, fixed)			
	20	1.18	1.23	1.21	1.17	88	90	88	87
Sample	40	1.19	1.12	1.20	1.29	92	93	93	95
size	80	1.20	1.21	1.21	1.16	94	94	93	95
	160	1.21	1.29	1.23	1.24	95	94	95	96
asymptotic		1.25	1.25	1.25	1.25	95	95	95	95
		Coverage rate (k estimated)				Coverage rate (k = ∞, fixed)			
	20	79	79	78	78	90	91	89	88
Sample	40	90	90	91	91	93	94	94	93
size	80	92	92	93	94	94	93	95	95
	160	95	94	95	94	96	93	97	94
asymptotic		95	95	95	95	95	95	95	95

* Variance of the estimated regression parameter under specified assumptions for k.
** Linear model, β_0 = intercept, β_1, β_2, β_3 = regression coefficient.

To investigate whether the number of parameters in the model influenced the results a further simulation study was performed. Six hundred datasets of size $n(n = 20, 40, 80, 160)$ were generated from the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e,$$

where the error term e had a T distribution with $k = 5$ or ∞, and $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 1, 1, 1)$. For $n = 20$, $X_{1i} = i - 11$ if $i \leq 10$ and $X_{1i} = i - 10$ if $i > 10$; $X_{2i} = -5$ if $i \leq 5$ or $i > 15$ and $X_{2i} = 5$ otherwise; and $X_{3i} = -5$ if $i < 5$ or $10 < i \leq 15$ and $X_{3i} = 5$ otherwise. For the other sample sizes the design was replicated. The results are given in Tables 6 and 7. Again the variance ratio for estimating k is close to 1 especially for large sample sizes. The coverage rates when k is estimated are lower than the values in Table 3 and would probably not be considered satisfactory for $n = 20$ or $n = 40$. As in Tables 2 and 3 the accuracy of the asymptotic calculations does not seem to be affected by whether the true $k = 5$ or $k = \infty$.

Discussion

Based on asymptotic theory and some Monte Carlo simulations, it is shown in this article that the adaptive stage of fitting an error distribution can as a first approximation be ignored in making statistical inferences concerning the regres-

Table 7
Monte Carlo simulation and asymptotic results – T family (4 regression parameters).

		β_0 **	β_1	β_2	β_3	β_0	β_1	β_2	β_3
True degrees of freedom $(k) = \infty$									
		Variance * ratio for estimating k; (k estimated/$k = \infty$ fixed)				Variance * ratio. (k estimated/astymptotic)			
	20	1.14	1.14	1.13	1.12	1.12	1.19	1.20	1.15
Sample	40	1.02	1.04	1.05	1.02	1.03	1.03	1.05	1.07
size	80	1.02	1.01	1.03	1.02	1.04	1.05	1.11	1.12
	160	1.01	1.01	1.01	1.02	1.03	1.09	1.12	1.12
asymptotic		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		Relative Efficiency * $\frac{\text{Var } k = 5}{\text{Var } k = \infty}$				Coverage Rate ($k = \infty$, fixed)			
	20	1.11	1.10	1.11	1.07	91	91	88	89
Sample	40	1.08	1.06	1.08	1.09	92	93	94	92
size	80	1.09	1.08	1.08	1.07	94	94	94	94
	160	1.10	1.09	1.09	1.10	94	95	95	94
asymptotic		1.09	1.09	1.09	1.09	95	95	95	95
		Coverage rate (k estimated)				Coverage rate ($k = 5$, fixed)			
	20	83	81	81	83	88	85	86	88
Sample	40	90	91	93	91	90	90	92	90
size	80	93	93	93	94	91	92	92	93
	160	94	95	94	94	93	93	92	92
asymptotic		95	95	95	95	94	94	94	94

* Variance of the estimated regression parameter under specified assumptions for k.
** Linear model, β_0 = intercept, β_1, β_2, β_3 = regression coefficients.

sion parameters provided the sample size is not small, and particularly if the T rather than the power-exponential distribution is used This is a useful result which together with their good robustness properties suggest that the adaptive robust techniques should have more of a role in practical statistics than they currently enjoy. However, caution should be exercised in extending this result to situations very different than those considered in the Monte Carlo study or to other more complex situations. For example the variance ratio for finite samples may not be so close to one for non-linear, or multivariate models using the T-distribution or if the design of the explanatory variables are very irregular.

Comparison of the results in Table 6 and 7 and Tables 2 and 3 suggests that for correct statistical inference in small samples the addition of extra regression parameters is of as much concern as is the inclusion of an extra parameter to model the kurtosis. From the tables it can be seen that the coverage rates when k is fixed and known are lower when 4 regression parameters are estimated compared to when 2 regression parameters are estimated. Thus the inadequacy of the coverage rates when $n = 20$ in Tables 6 and 7 is a result of both the estimation of the extra kurtosis parameter (k) and the estimation of the extra regression parameters.

A similar type of simulation study was performed by Hogg et al. (1972). They considered an adaptive scheme for estimating a location parameter from a single sample. The adaptive scheme considered is a decision rule followed by either the mean, the median or the mid range. The decision rule is a modified version of the maximum likelihood rule for selecting between the normal, the double-exponential and the uniform distribution. They found that if the true distribution was normal, the ratio of the Monte Carlo variance of the adaptive scheme compared to that of the mean is 1.12. If the true distribution was double-exponential the variance ratio of the adaptive scheme to the median is 1.21. If the true distribution was uniform the variance ratio of the adaptive scheme to the mid-range is 2.06. The values for the normal and the double-exponential are similar to those in Table 4 and encouragingly small. The variance ratio for the uniform is significantly larger than 1, however for practical situations it is maybe not that relevant because such short tailed distributions as the uniform are usually not considered likely (or of serious consequence) for real data.

In later work, Hogg et al. (1988) and Yuh and Hogg (1988) compared a variety of adaptive and non-adaptive robust estimates. One of the adaptive schemes, which had arguably the best properties, was based on the T-distribution. In this scheme, depending on the value of a selector statistic which measures peakedness of the distribution either a T_{11} or a T_3 model was used. In addition, these authors use a one-step approximation to the maximum likelihood estimates of the regression parameters.

A significant difference between the adaptive scheme of Hogg et al. (1978, 1988), and Yuh and Hogg (1988) and the model based scheme used in this paper, is that these authors allow only a finite number of possible estimates of the location parameter, whereas in this paper there are essentially an infinite number of possible estimators depending on the estimates of k.

One interesting finding from this study is that the inflation in variance due to the estimation of the extra shape parameter in small samples is less for the T-family than for the power-exponential family. The probably reason for this is that the T-family covers a narrower range of distributions not including any with lighter than normal tails.

The coverage properties of the confidence intervals were encouraging especially for the T-family, except for small samples. One improvement for small samples would be to use the interval $\hat{a} \pm c.\ \hat{SE}(\hat{a})$ where c is a critical point of the t distribution with N-P degrees of freedom and P is the number of regression parameters, rather than the interval $\hat{a} \pm 1.96\ \hat{SE}(\hat{a})$. This value of c would be the natural choice in small samples for the normal model. The coverage rates of such confidence intervals were evaluated for the simulations presented in Tables 6 and 7. For $n = 20$ it was found that the coverage rate increased by approximately 3 percentage points, and for $n = 40$ the coverage rate increased by about 1 percentage point. So although the coverage properties of confidence intervals are improved by this adjustment, they are still inadequate for $n = 20$.

Another possible improvement is to base the estimate of the standard error on the observed information matrix rather than the expected information matrix. Su

(1988) has shown that when using the multivariate T distribution with missing data and known degrees of freedom the expected information tends to give slightly optimistic confidence intervals, whereas the confidence intervals based on the observed information tend to have better coverage rates. The arguments in Efron and Hinkley (1978) would also support this view. Alternatively, an estimate of the standard error could be based on empirical approximations to the expression given in equation (3).

Boyer and Kolson (1983) and Prescott (1978) have suggested that for larger samples a continuous adapting scheme is a useful robust estimator, but for smaller samples adapting too closely may not be worth the effort; although it could be argued that is really isn't any more effort to estimate an extra parameter than employ a discrete decision rule. The results in Tables 2 to 7 of this paper, for both the efficiency considerations and the coverage rates, would support this suggestion. For $n \geq 40$, the procedure of fitting a T-model and estimating the degrees of freedom has good statistical properties, whereas for $n = 20$, fitting a T model with 5 degrees of freedom is a reasonable choice.

Acknowledgement

This work was partially supported by NIH grants CA16042 and CA45216. The author thanks Rod Little for useful comments.

References

P.J. Bickel and K.A. Doksum, An analysis of transformation revisited, *J. Am. Statist. Assoc.*, **76** (1981) 296–311.

J.E. Boyer and J.O. Kolson, Variance for adaptive trimmed means *Biometrika*, **70** (1) (1983) 97–102.

G.E. Box and D.R. Cox, An analysis of transformation, *J.R. Statist. Soc. Ser.* B. **26** (1964) 211–232.

G.E. Box and D.R. Cox, An analysis of transformation revisited, rebutted, *J. Am. Statist. Assoc.*, **77** (1982) 209–210.

G.E. Box, and G.C. Tiao, *Bayesian Inference in Statistical Analysis* (Addison-Wesley, Reading, Mass., 1973).

R.J. Carroll and D. Ruppert, Prediction and the power transformation family, *Biometrika*, **68** (1981) 609–616.

N.R. Draper and H. Smith, *Applied Regression Analysis* (2nd edition) (Wiley, New York, 1981).

B. Efron and D.V. Hinkley, Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, *Biometrika*, **65** (1978) 457–487.

D.A.S. Fraser, Necessary and adaptive inference, *J. Am. Statist. Assoc.*, **71** (1976) 99–113.

V.P. Godambe and C.C. Heyde, Quasi-likelihood and Optimal estimation, *International Statistical Review*, **55** (1987) 231–244.

D.V. Hinkley and G. Runger, The analysis of transformed data, *J. Am. Statist. Assoc.*, **79** (1984) 302–309.

R.V. Hogg, Some observations on robust estimation, *J. Am. Statist. Assoc.*, **62** (1967) 1179–1186.

R.V. Hogg, Adaptive Robust Procedures: A partial review and some suggestions for future applications and theory, *J. Am. Statist. Assoc.*, **69** (1974) 909–926 (with discussion).

R.V. Hogg, G.K. Bril, S.M. Han and L. Yuh, An argument for Adaptive Robust estimation, in: J.N. Srivastava (Ed.), *Probability and Statistics, essays in honor of F.A. Graybill* (North-Holland, Amsterdam, 1988) 135–148.

R.V. Hogg, V.A. Uthoff, R.H. Randles and A.S. Davenport, On the selection of the underlying distribution and adaptive estimation, *J. Am. Statist. Assoc.*, **67** (1972) 596–600.

P.J. Huber, *Robust Statistics* (Wiley, New York, 1981).

H. Jeffreys, *Theory of Probability* (Clarendon Press, Oxford, 1939).

K.L. Lange, R.J.A. Little and J.M.G. Taylor, Robust statistical modeling using the T distribution, *J. Am. Statist. Assoc.*, **84** (1989) 881–896.

P. Prescott, Selection of trimming proportions for robust adaptive trimmed means, *J. Am. Statist. Assoc.*, **73** (1978) 133–140

H-L. Su, Estimation of standard errors in some multivariate models when some observations are missing, Ph.D. Dissertation, Division of Biostatistics, University of California at Los Angeles (1988).

J.M.G. Taylor, The retransformed mean after a fitted power transformation, *J. Am. Statist. Assoc.*, **81** (1986) 114–18.

J.M.G. Taylor, The cost of generalizing logistic regression, *J. Am. Statist. Assoc.*, **83** (1988) 1078–1083.

L. Yuh, and R.V. Hogg, On adaptive M-regression, *Biometrics*, **44** (1988) 433–445.