Robust Statistical Modeling Using the *t* Distribution

KENNETH L. LANGE, RODERICK J. A. LITTLE, and JEREMY M. G. TAYLOR*

The *t* distribution provides a useful extension of the normal for statistical modeling of data sets involving errors with longer-than-normal tails. An analytical strategy based on maximum likelihood for a general model with multivariate *t* errors is suggested and applied to a variety of problems, including linear and nonlinear regression, robust estimation of the mean and covariance matrix with missing data, unbalanced multivariate repeated-measures data, multivariate modeling of pedigree data, and multivariate nonlinear regression. The degrees of freedom parameter of the *t* distribution provides a convenient dimension for achieving robust statistical inference, with moderate increases in computational complexity for many models. Estimation of precision from asymptotic theory and the bootstrap is discussed, and graphical methods for checking the appropriateness of the *t* distribution are presented.

KEY WORDS: Bootstrap; Elliptical distributions; EM algorithm; Maximum likelihood; Nonlinear regression; Outliers; Pedigree analysis; Regression; Repeated-measures data.

1. INTRODUCTION

Statistical inference based on the normal distribution (univariate or multivariate) is known to be vulnerable to outliers. Despite this fact and the considerable interest in robust procedures in the mathematical statistical literature, most applied statistical analysis continues to be based on the normal model. Even in linear regression, where robustness concerns have penetrated statistical software widely available to practitioners, procedures are mainly directed at detecting outliers. For example, see the regression diagnostic procedures in BMDP (Dixon 1983), SAS (1982), or SPSS (1983). After editing outliers, subsequent analysis is often still restricted to least squares based on the normal linear model. A serious problem with this approach is that resulting inferences fail to reflect uncertainty in the exclusion process; in particular, standard errors tend to be too small.

Reasons for the slow adoption of robust estimation procedures by practitioners may include the bewildering choice of alternative procedures, and a lack of published applications to real complicated data. Rather than comparing many alternative methods in a relatively simple data setting, in this article we apply a single method to robust inference on a variety of real data sets. Our approach is to replace the normal distribution by the *t* distribution in statistical models. Specifically, suppose that sample data y_i $(1 \le i \le n)$ are recorded for n units. Typically, one assumes that the y_i are independent normal random vectors. If $N_k(\mu, \Sigma)$ denotes the k-variate normal distribution with mean μ and covariance matrix Σ, then

$$y_i \underset{\text{ind}}{\sim} N_{v_i}\{\mu_i(\theta), \Sigma_i(\varphi)\}, \tag{1}$$

where v_i is the number of components of y_i, which may vary from unit to unit in some applications, μ_i is a $(v_i \times 1)$ vector mean function of known form indexed by a set of unknown parameters θ, and Σ_i is a $(v_i \times v_i)$ covariance matrix of known form indexed by a set of unknown parameters φ. The functions μ_i and Σ_i may involve known fixed covariates x_i, recorded for each unit i. We propose to replace (1) with the model

$$y_i \underset{\text{ind}}{\sim} t_{v_i}\{\mu_i(\theta), \Psi_i(\varphi), \nu\}, \tag{2}$$

where $t_k\{\mu, \Psi, \nu\}$ denotes the k-variate *t* distribution (Cornish 1954; Dunnett and Sobel 1954) with location vector μ, scale matrix Ψ, ν df, and density

$$p(y \mid \mu, \Psi, \nu) = \frac{|\Psi|^{-1/2}\Gamma\{(\nu + k)/2\}}{\{\Gamma(1/2)\}^k\Gamma(\nu/2)\nu^{k/2}} \times \left(1 + \frac{(y - \mu)^T\Psi^{-1}(y - \mu)}{\nu}\right)^{-(\nu+k)/2}.$$

Inferences about θ and φ in the multivariate *t* setting can proceed by likelihood methods analogous to those for the normal model (1).

The following known facts about the multivariate *t* are instructive and used later. Suppose that $y \mid u \sim N_k(\mu, \Psi/u)$ for scalar $u \sim \chi^2_\nu/\nu$, where ν is positive and may be a noninteger. We then have the following properties.

Property 1. $y \sim t_k(\mu, \Psi, \nu)$.

Property 2. $E(y) = \mu$ $(\nu > 1)$ and $\text{cov}(y) \equiv \Sigma = \nu\Psi/(\nu - 2)$ $(\nu > 2)$.

Property 3. $u \mid y \sim \chi^2_{\nu+k}/\{\nu + \delta^2\}$, where $\delta^2 = (y - \mu)^T\Psi^{-1}(y - \mu)$.

Property 4. $\delta^2/k \sim F_{k,\nu}$ (Box and Tiao 1973, eq. 2.7.21).

Note that the multivariate *t* distribution approaches the normal distribution with covariance matrix Ψ as $\nu \to \infty$. When $\nu < \infty$, maximum likelihood (ML) estimation of θ and certain functions of φ are robust in the sense that outlying cases with large Mahalanobis distances $\delta_i^2 = (y_i - \mu_i)^T\Psi_i^{-1}(y_i - \mu_i)$ are downweighted. In particular, ML estimates of θ (with q components, say) for the normal model (1) satisfy the likelihood equation $\partial l/\partial\theta = \sum_{i=1}^n A_i\Sigma_i^{-1}(y_i - \mu_i) = 0$, where l denotes the log-likelihood and

* Kenneth L. Lange and Roderick J. A. Little are Professors, Department of Biomathematics, School of Medicine, and Jeremy M. G. Taylor is Associate Professor, Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90024. This research was supported by National Institutes of Health Grant CA 16042 and National Institute of Mental Health Grant USPHS MH 37188. The authors thank H. Schelbert and C. Nienaber for supplying the positron emission tomography data in Examples 3 and 4, and Wil Dixon, Mark Schluchter, an associate editor, and three referees for many useful suggestions. Author listing is alphabetical.

A_i is the $(q \times v_i)$ matrix of partial derivatives of μ_i with respect to θ. ML estimates of θ under the t model (2) satisfy $\sum_{i=1}^{n} w_i A_i \Psi_i^{-1}(y_i - \mu_i) = 0$, where

$$w_i = (\nu + v_i)/(\nu + \delta_i^2) \tag{3}$$

is the weight assigned to case i; w_i clearly decreases with increasing δ_i^2. (See Prop. 1 in App. B.)

Note that the degree of downweighting of outliers in (3) increases with decreasing ν. If ν is fixed a priori at some reasonable value, it is a robustness tuning parameter. ML estimation of θ based on (2) is a form of M estimation (Huber 1981), yielding robust estimates of location with a redescending influence function. With sufficient data, ν may be estimated from the data by ML, yielding an *adaptive* robust procedure in the sense used by Hogg (1974); see Yuh and Hogg (1988). This approach follows Box and Tiao (1973) and Box (1980) in embedding the normal model in a larger model with a parameter (here ν) that affords protection against nonnormality. A different approach modifies normal-theory estimators and/or their standard errors to allow for nonnormal errors. In particular, Zellner's (1976) linear regression model leads to least squares estimates of the regression coefficients, but with inflated standard errors (see Sec. 2.1). Tyler (1983), Browne (1984), and Shapiro and Browne (1987) studied modifications of normal covariance-structure tests under elliptically distributed errors.

Hampel, Ronchetti, Rousseeuw, and Stahel (1986) noted that the use of the t distribution for describing a single sample dates back at least to Jeffreys (1939), who fitted it to series of astronomical data. Fraser (1976, 1979) discussed one-sample and general linear models with t errors, using structural inference to estimate the parameters. A related approach to the one-sample problem appears in the article by Relles and Rogers (1977), who obtained good results from a Bayesian analysis using the t model with a uniform prior on ν. West (1984) extended this approach to regression. Maronna (1976) discussed ML estimation of the mean and covariance matrix under t errors. Rubin (1983) and Sutradhar and Ali (1986) considered ML for multivariate regression with multivariate t errors, and Little (1988a) extends this work to handle incomplete data. Pendergast and Broffitt (1985) mentioned the multivariate t in connection with M estimation for growth-curve models, and Masreliez and Martin (1977) applied the t distribution to Kalman filtering in time series. We include novel applications of the t approach to nonlinear regression, unbalanced repeated-measures data, and pedigree analysis.

By Property 1 the t model (2) can be derived by mixing a multivariate normal deviate y_i with a scaling variable u_i:

$$y_i \mid u_i \sim N_{v_i}(\mu_i(\theta), \Psi_i(\varphi)/u_i), \qquad u_i \sim \chi_\nu^2/\nu.$$

Other models with errors having longer-than-normal tails are obtained by choosing other distributions for u. In particular, for univariate y Rogers and Tukey (1962) argued that u should have finite support, and they proposed modeling u with cdf $u^{\nu+1}$ $(0 \le u \le 1)$. When $\nu = 0$, u is uniform, and y has the so-called slash distribution. This model may have better properties than the t for data sets with gross outliers, but in some respects it is less convenient than the t computationally. Specifically, the likelihood involves the computation of incomplete gamma functions for each observation, and the weighting function (3) and expected information are more complicated than for the t model. The distribution of the Mahalanobis distance is no longer F as in Property 4, which complicates the computation of residual plots (discussed in Sec. 4).

Another variant of the mixture approach models u as a binary variable, taking value 1 with probability π and $\lambda < 1$ with probability $1 - \pi$. The marginal distribution of y is then a mixture of two normals. Choosing π close to 1 (say .95) and λ small (say .1) yields a suitable model when the data are contaminated with a small fraction of outliers. This model is easy to work with computationally and can be effective in the presence of extreme outliers. But the multivariate t compared favorably with this model in simulations in Little (1988a), and the contaminated normal model requires two robustness parameters (π and λ) to be specified or estimated, rather than just one for the t model. Berkane and Bentler (1988) presented moment estimators of π and λ for the contaminated multivariate normal model with constant v_i and unstructured mean and covariance matrix.

The mixing approach produces families of distributions with longer-than-normal tails. The exponential power family (Box and Tiao 1973) is an example of a class of distributions with tails that are both longer and shorter than the normal. Our limited experience with ML for this model (Taylor 1989) indicates that it has more computational problems than the t model because of a tendency of the robustness parameter to approach one of its boundary values. In addition, the uniform and double-exponential distributions, which occur when the extra parameter attains its boundary values, are not very appealing for modeling real data. Since the model is not obtainable by mixing the normal with a scale variable, EM algorithms that treat the scaling variable as missing are not available. The penalty for estimating the robustness parameter in terms of increased variance for estimates of location appears greater than the penalty for the t model (Taylor 1989). The fact that the model includes distributions with shorter tails than the normal is an advantage, but in applications departures from normality in this direction seem to be less frequent, with less serious consequences.

We do not deny the value of these or other approaches to robust inference (Hampel et al. 1986; Huber 1981). Nevertheless, we think that inference based on a parametric model such as (2) combines conceptual simplicity with generality, since it can be applied in a wide range of settings. The model supplies a single parameter ν for robustness, like Tukey's (1949) single degree of freedom for nonadditivity or Box and Cox's (1964) single parameter for power transformations. Given sufficient data, ν can be estimated from the data by likelihood methods, and the improvement over the normal model can be tested using standard methods such as the likelihood ratio test. For small samples another option is to set ν a priori at some

sensible value; we have found that the value $\nu = 4$ has worked well in many of our applications.

Asymptotic estimates of precision are readily available from the information matrix, and they can incorporate uncertainty in estimating ν if ν is treated as a parameter to be estimated. In Appendix B we show that the expected information matrix based on Model (2) (and other models with elliptically symmetric error distributions) is block diagonal between the mean parameters θ and the scale and kurtosis parameters φ and ν. Hence ML estimates of θ and (φ, ν) are asymptotically uncorrelated, and asymptotic standard errors of estimates of θ are unaffected by estimating the scale matrix or the degrees of freedom.

In small samples, alternatives to ML such as profile likelihood plots for particular parameters, the Bayesian approach of Relles and Rogers (1977), or the related structural approach of Fraser (1976) may yield better tests and interval estimates than methods based on asymptotic theory. These approaches are more computationally demanding, however.

Of course, t modeling is not a panacea for all robustness problems. In particular, data with shorter-than-normal tails or asymmetric error distributions, varying degrees of long-tailedness among the variables, or extreme outliers are not well modeled by (2). An advantage of the t modeling approach, however, is that a clear statement of assumptions is incorporated in the model specification, and a critical assessment of them can yield modifications of the model that deal with some of its limitations (e.g., by allowing different degrees-of-freedom parameters for different variables).

We now present some examples of (2). Section 3 considers methods for computing standard errors, and Section 4 develops graphical diagnostic checks of the model. Section 5 states conclusions. Alternative computational approaches to ML estimation are outlined in Appendix A, and Appendix B derives the score and expected information matrix for (2), including some results for the more general elliptically symmetric family of distributions.

2. SPECIAL CASES OF THE GENERAL MODEL

2.1 Univariate Regression

Let $v_i = 1$, and introduce a $(k \times 1)$ vector of covariates x_i for each unit i. An important case of (1) is the normal regression model $y_i \sim_{\text{ind}} N_1(\mu(\theta; x_i), \sigma^2/c_i)$, where μ_i depends on i only through x_i and the residual variance is inversely proportional to a known constant c_i. The analogous t model

$$y_i \underset{\text{ind}}{\sim} t_1(\mu(\theta; x_i), \psi^2/c_i, \nu) \tag{4}$$

can be used for robust regression. Model (4) is different from that of Zellner (1976), who placed a multivariate t distribution on the vector of errors $\{y_1 - \mu(\theta; x_1), \ldots, (y_n - \mu(\theta; x_n)\}$. Whereas (4) yields robust estimates of θ, Zellner's model yields the standard least squares estimates of θ, but with estimated standard errors that are inflated by the factor $[\nu/(\nu - 2)]^{1/2}$. Our first example of (4) is for linear regression, where $\mu(\theta; x_i) = x_i^T\theta$.

Example 1: Stack-Loss Data. The stack-loss data set presented by Brownlee (1965) has been subjected to robust methods by numerous authors, including Andrews (1974) and Ruppert and Carroll (1980). Table 1 shows slopes of the regression of Y = stack loss on X_1 = air flow, X_2 = temperature, and X_3 = acid, calculated for the linear regression model with t errors and ν ranging from ∞ (normal) to .5. A row is included for the ML estimate of ν, namely $\hat{\nu} = 1.1$. Four other sets of estimates from Ruppert and Carroll (1980) are presented, two from trimmed least squares ($\hat{\theta}_{KB}$ and $\hat{\theta}_{PE}$) and M estimates proposed by Huber and Andrews.

Maximized log-likelihoods from the t models are given in the second column of the table; they describe the profile log-likelihood as a function of ν. The difference in maximized log-likelihood between the best-fitting t and the normal model is 2.72. Doubling this number yields a likelihood ratio (LR) chi-squared statistic 5.44 on 1 df, an apparently significant improvement in fit, although asymptotic theory cannot be trusted because the sample size is small (21 cases). Estimates for the $\nu = 1.1$ model are similar to those of Andrews (1974), which Ruppert and Carroll (1980) favored based on closeness of fit to the bulk of the observations. Estimation of standard errors of the parameter estimates is a tricky issue, given the small sample size (see Sec. 3).

The t fits were obtained by iteratively reweighted least squares (as discussed in App. A, Sec. A.2). Cases with particularly small final weights are outliers. The weights from the Cauchy model ranged from .017 to 1.985. The four observations with the smallest weights (.017, .023, .047, .052) are cases 21, 4, 3, and 1, respectively, which are also identified as outliers in the least squares analysis of these data by Daniel and Wood (1971). Least squares analysis with the four points removed yields similar estimates to the Cauchy fit to all of the data, as can be seen in Table 1.

Example 2: Radioimmunoassay Data. Nonlinear least squares fitting requires an iterative algorithm, so the additional computational effort in incorporating t errors is less pronounced than in linear models. Tiede and Pagano (1979) provided an application of robust nonlinear regression to radioimmunoassay. Their methods were illustrated on TSH standards data with two measurements for each dose, plotted in Figure 1. The continuous line shows the least squares fit of the mean model $\mu(\theta; x_i) = \theta_0 + \theta_1/(1 + \theta_2 x_i^{\theta_3})$ and is distorted by the clear outlier at $x = 20$. The dotted line shows the ML fit of this mean function, assuming constant variance and t errors with 4 df. Final weights [Eq. (3)] from this model are .01 for the outlier and range from .59 to 1.25 for the other 13 points. (The constant-variance assumption seems dubious for counted data, but Tiede and Pagano stated that it is supported by empirical evidence.) The t_4 fit is similar to that obtained by Tiede and Pagano using the robust sine M estimator (Andrews, Bickel, Hampel, Huber, Rogers, and Tukey 1972).

ML estimation of ν for these data is not very satisfactory,

Table 1. Regression of Stack-Loss Data: Estimates From 13 Methods

Method	Log-likelihood	Intercept (θ_0)	Air flow (θ_1)	Temperature (θ_2)	Acid (θ_3)
Normal	−33.0	−39.92	.72	1.30	−.15
t, $\nu = 8$	−32.7	−40.71	.81	.97	−.13
t, $\nu = 4$	−32.1	−40.07	.86	.75	−.12
t, $\nu = 3$	−31.8	−39.13	.85	.66	−.10
t, $\nu = 2$	−31.0	−38.12	.85	.56	−.09
t, $\hat{\nu} = 1.1$	−30.3	−38.50	.85	.49	−.07
t, $\nu = 1$	−30.3	−38.62	.85	.49	−.07
t, $\nu = .5$	−31.2	−40.82	.84	.54	−.04
Normal minus four outliers		−37.65	.80	.58	−.07
$\hat{\theta}_{KB}$		−42.83	.93	.63	−.10
$\hat{\theta}_{PE}$		−40.37	.72	.96	−.07
Huber		−41.00	.83	.91	−.13
Andrews		−37.20	.82	.52	−.07

reflecting the fact that the t model is not well suited to data with extreme outliers. The ML estimate is $\hat{\nu} = .29$, a very low value. Although the fit (the dashed line in Fig. 1) is similar to that for $\nu = 4$, the final weights are concentrated on a small number of data points: Four have weights close to 3.4, one has a weight of .4, and the others have weights of .1 or less. This overconcentrated distribution of weights reflects the unappealing shape of the t distribution when ν is small, which has a spike at 0 and is close to 0 elsewhere. The low value of ν results from attempting to accommodate the extreme outlier: When the outlier is removed, $\hat{\nu}$ increases to 1.2. In general, we agree with Fraser (1979, p. 45) that ML estimation of ν is not advisable when $\hat{\nu}$ goes much below 1.

Example 3: Nonlinear Calibration of Blood-Flow Data. Estimation of ν is more successful in our second nonlinear regression example, which concerns calibration of two measures of blood flow in the canine myocardium, measured in milliliters per minute per 100 grams (ml/min/100g). The first measurement (x) is regional myocardial blood flow (RMBF), from a standard invasive measurement procedure using radioactively labeled microspheres; the second (y) is extraction times blood flow (EF), based on noninvasive N-13 ammonia images from positron emission tomography (PET) (Schelbert et al. 1981). Figure 2 presents data on 251 determinations of x and two versions of y, one (y_1) from integrating the results of PET scans taken up to 60 seconds, and the other (y_2) from the results of PET scans taken up to 510 seconds. The solid lines in Figure 2 show nonlinear least squares regression fits based on the univariate models UNj:

$$y_{ij} \underset{\text{ind}}{\sim} N(\mu(x_i; \theta_j), \varphi_j^2), \qquad \mu_j(x_i; \theta_j) = x_i\{1 - \theta_{1j}e^{-\theta_{2j}/x_i}\}, \quad i = 1, \ldots, 251,$$

for $j = 1, 2$. The form of μ_j is based on theory of Renkin (1959) and Crone (1963). A model that fits the same curve for y_1 and y_2 (i.e., $\theta_{11} = \theta_{12}$ and $\theta_{21} = \theta_{22}$) does not fit the data. Note that in Figure 2 the residuals of y_2 are more scattered than the residuals of y_1, and display some evidence of outliers.

Estimates from these univariate normal models (UN1, UN2) are given in the first two rows of Table 2, with the maximized log-likelihood and standard errors based on a numerical approximation of the observed information matrix (App. A). The third and fourth rows show results from the corresponding univariate t models $y_{ij} \sim_{\text{ind}} t(\mu_j(x_i; \theta_j), \varphi_j^2, \nu)$ ($i = 1, \ldots, 251$), with ν estimated from the data (UT1, UT2). For UT1, $\hat{\nu} = 10.3$, and the increase in log-likelihood over the normal model is marginal (LR $\chi_1^2 = 5.2$). For UT2, $\hat{\nu} = 3.4$, and the increase in log-likelihood is highly significant (LR $\chi_1^2 = 59.9$), confirming the presence of nonnormal errors. The parameter estimates of location from the normal and t models are similar, but note the reduced standard errors for $\hat{\theta}_{12}$ and $\hat{\theta}_{22}$ that result from switching from a normal to a t model.

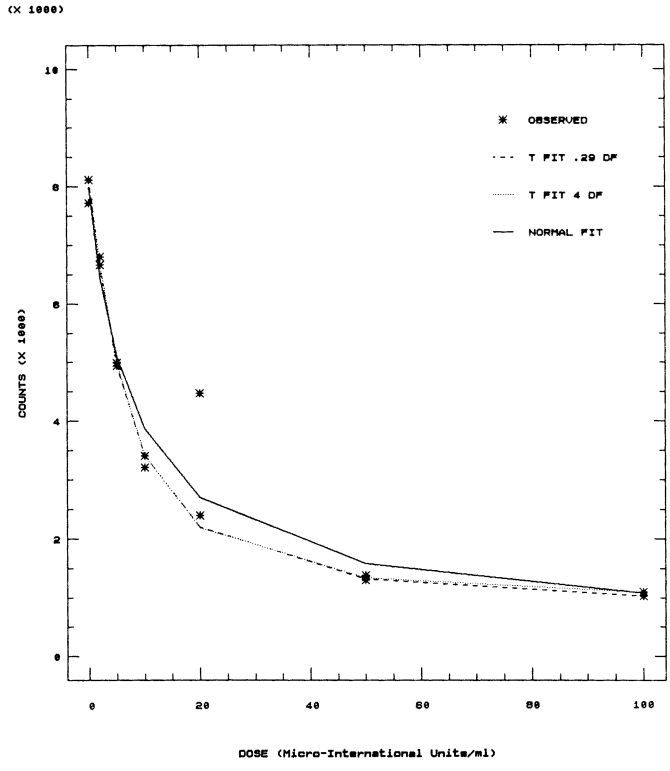

Figure 1. Example 2 Radioimmunoassay Data: Observed Values From Three Models.

2.2 Robust Multivariate Regression

In Section 1 we noted earlier applications of the multivariate t model to multivariate linear regression with bal-

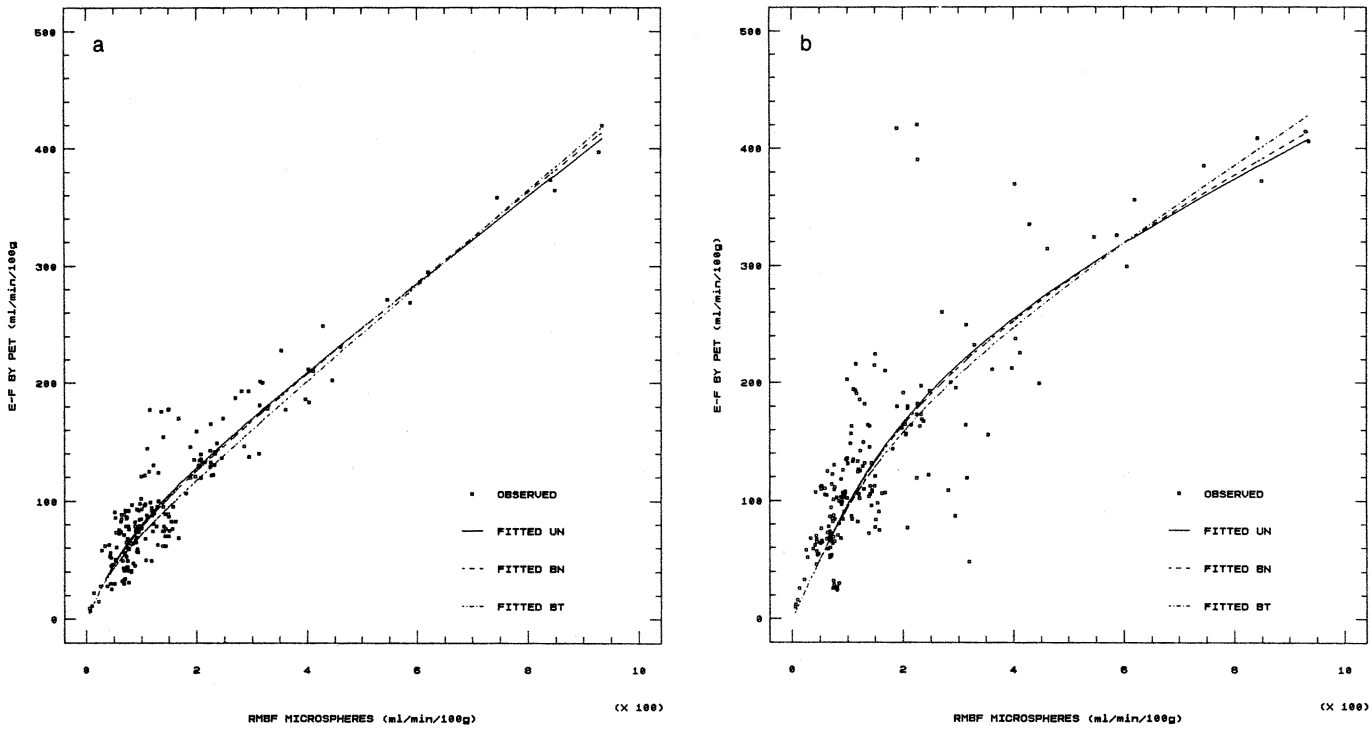

Figure 2. Examples 2 and 3 Blood-Flow Data, Observed and Fitted Data: (a) y_1—Blood Flow From PET Scans up to 60 Seconds; (b) y_2—Blood Flow From PET Scans up to 510 Seconds.

anced data. Our first multivariate example extends the nonlinear regression analysis of Example 3 to model correlation between the responses.

Example 4: Multivariate Nonlinear Regression of Blood-Flow Data (Ex. 3 continued). The analysis of Example 3 failed to consider that measurements of y_1 and y_2 were based on the same PET study, and hence are correlated. This correlation can be modeled using the bivariate normal model (BN)

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \underset{\text{ind}}{\sim} N_2 \left\{ \begin{pmatrix} \mu_1(x_i;\, \theta_1) \\ \mu_2(x_i;\, \theta_2) \end{pmatrix}, \begin{pmatrix} \varphi_1^2 & \rho\varphi_1\varphi_2 \\ \rho\varphi_1\varphi_2 & \varphi_2^2 \end{pmatrix} \right\},$$

and the model obtained by replacing the bivariate normal distribution by a bivariate t (BT). Fitted values from BN and BT are displayed as the broken lines in Figure 2, and parameter estimates and maximized log-likelihood values are shown in Table 2. Note that the correlation is clearly significantly greater than 0. In addition, the t model fits much better than the normal model (LR $\chi_1^2 = 162.9$), with smaller standard errors of the location parameters and increased estimated correlation. A scatterplot of the residuals from the BT model is given in Figure 3, and clearly shows the outliers that are downweighted by the t model.

2.3 Robust Analysis of Unbalanced Multivariate Data

Many multivariate statistical analyses involve reduction of the data to a sample mean and covariance matrix. These statistics are sufficient under the multivariate normal model $y_i \sim_{\text{ind}} N_v\{\mu, \Sigma\}$, which is obtained from (1) by setting $v_i = v$, θ equal to the set of unconstrained means $\{\mu_j;\ 1 \le j \le v\}$, and φ equal to the set of unconstrained variances and covariances $\{\sigma_{jk};\ 1 \le j \le k \le v\}$. Robust estimation of the mean and covariance matrix can be achieved by ML

Table 2. Parameter Estimates and Standard Errors, Blood-Flow Data, Examples 3 and 4

Model	Maximum log-likelihood	Parameter estimates							
		θ_{11}	θ_{21}	θ_{12}	θ_{22}	φ_1^2	φ_2^2	ρ	ν
UN1	−908.69	.636	113.9			513			
		(.016)	(9.1)			(46)			
UN2	−1,093.12			.782	306.0		2,231		
				(.054)	(37.6)		(199)		
UT1	−906.09	.629	106.2			412			10.3
		(.014)	(8.8)			(56)			(5.4)
UT2	−1,063.15			.746	274.7		943		3.4
				(.035)	(28.0)		(151)		(.8)
BN	−1,966.59	.622	102.9	.758	287.8	516	2,233	.500	
		(.015)	(8.5)	(.050)	(34.1)	(46)	(200)	(.048)	
BT	−1.885.13	.598	74.2	.701	241.4	320	896	.739	3.2
		(.011)	(6.1)	(.029)	(21.6)	(40)	(113)	(.036)	(.5)

NOTE: Standard errors are in parentheses.

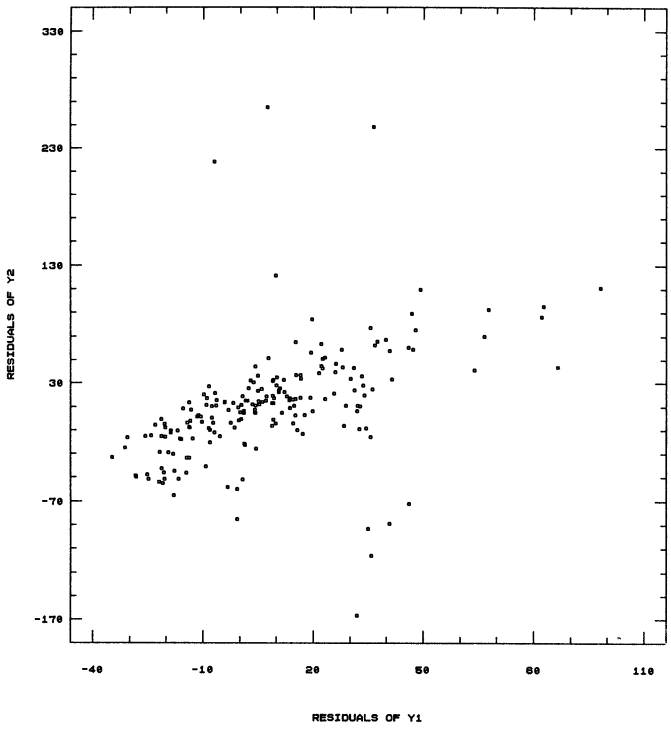

Figure 3. Example 3 Blood-Flow Data: Scatterplot of Residuals of y_1 and y_2 From the BT Model.

estimation under the multivariate t model $y_i \sim_{\text{ind}} t_v\{\mu, \Psi, v\}$ (Maronna 1976; Rubin 1983). Little (1988a) extends this approach to situations where the data matrix $\{y_{ij}\}$ is incomplete, with any pattern of missing values. ML estimates for the multivariate t and contaminated normal models are calculated by the EM algorithm (see App. A) and involve minor modifications of the EM algorithm for an incomplete multivariate normal sample. Since ML estimation for the normal model requires iterative fitting procedures, the mixture-modeling approach achieves robust estimation with minor increments in cost.

An important extension is to model the mean and covariance matrix further. Jennrich and Schluchter (1986) described ML estimation for the flexible normal model

$$y_i \underset{\text{ind}}{\sim} N_v\{X_i\theta, \Sigma(\varphi)\}, \tag{5}$$

where X_i is a known $(v \times q)$ design matrix for case i, θ is a $(q \times 1)$ vector of unknown parameters, Σ is a known function of covariance-structure parameters φ, and some components of y_i may be missing. Growth-curve models and analysis of variance models with more than one error term are special cases of (5). Robust modeling of the incomplete data can be achieved by replacing (5) with

$$y_i \underset{\text{ind}}{\sim} t_v\{X_i\theta, \Psi(\varphi), v\}, \tag{6}$$

as suggested in Little (1988b). The next example is an application of this idea.

Example 5: Repeated Lung-Function Measures With Missing Values. LaVange and Helms (1983, table 1) reported data from a longitudinal study of lung function conducted on 72 children aged 3 to 12 years at the Frank Porter Graham Child Development Center. The variables consist of race (black or white), gender, and log $\text{vmax}_{75\%}$ for single-year ages from 3 to 12, where $\text{vmax}_{75\%}$ is the maximum expiratory flow rate after 75% of the forced vital capacity has been exhaled. Of the 10 possible measurements of $\text{vmax}_{75\%}$ for each child, the number actually recorded range from 1 to 8, with an average of 4.3 per child; thus the amount of missing data was substantial. Some combinations of early and late ages were never observed together, so the covariance matrix was not estimable without placing restrictions on the parameters.

The results in Table 3 show whether there are differences in the growth curves of $\text{vmax}_{75\%}$ over time between males and females. Let y_{ij} denote the value of $\text{vmax}_{75\%}$ for individual i at age $j + 2$, for $1 \le j \le 10$. The first row of the table shows the maximized log-likelihood 164.43 for the following version of (6) (Model 1T): $y_i \sim_{\text{ind}} t_{10}(\mu_i(\theta), \Psi(\varphi), v)$, where $\mu_i(\theta)$ has jth component

$$\begin{aligned}\mu_{ij} &= \theta_0 + \theta_1\text{age}_j + \theta_2\text{age}_j^2 \quad \text{if sex}_i = \text{male}\\ &= \theta_3 + \theta_4\text{age}_j + \theta_5\text{age}_j^2 \quad \text{if sex}_i = \text{female},\end{aligned}$$

modeling separate quadratic curves relating lung function to age among males and females. The quadratic terms model nonlinearity, in the absence of any theory-based functional form for the curves. The scale matrix $\Psi = \{\psi_{jk}\}$ is modeled as $\psi_{jk} = \varphi_1\{\varphi_2 + (1 - \varphi_2)\varphi_3^{|k-j|}\}$ $(1 \le j, k \le 10)$, where φ_1 is the total dispersion, φ_2 is a heritability parameter, and φ_3 is an environmental decay constant. See Hopper and Mathews (1982) and Lange (1986) for moti-

Table 3. Models of Lung-Function Data, Example 5: Summary Fits for 12 Models

	Multivariate t models								
	Model description			Model fit		Multivariate normal models			
	Constraints		No. of				No. of	Model fit	
Model	Means	Cov	parameters	Log-likelihood	χ^2 (df)	Model	parameters	Log-likelihood	χ^2 (df)
1T	—	—	10	187.1	0 (0)	1N	9	164.4	0 (0)
2T	$\theta_2 = \theta_5 = 0$	—	8	186.2	2.0 (2)	2N	7	164.1	.7 (2)
3T	$\theta_2 = \theta_5 = 0, \theta_1 = \theta_4$	—	7	184.8	4.7 (3)	3N	6	161.9	5.2 (3)
4T	$\theta_2 = \theta_5 = 0, \theta_0 = \theta_3, \theta_1 = \theta_4$	—	6	184.2	5.9 (4)	4N	5	161.4	6.2 (4)
5T	—	$\varphi_3 = 0$	9	183.1	8.1* (1)	5N	8	163.5	2.0 (1)
6T	—	$\varphi_2 = 0$	9	181.5	9.2* (1)	6N	8	156.4	16.5* (1)

NOTE: Parameter constraints for the normal models are as for the t model in the same row.
* Significantly worse fit than Model 1N (normal models) or 1T (t models) at the 1% level (asymptotic LR chi-squared test).

Table 4. Models of Lung-Function Data, Example 5: Parameter Estimates and Standard Errors for Models 4N and 4T

Model	θ_0	θ_1	φ_1	φ_2	φ_3	ν
4N	−.3650	.0637	.1671	.3618	.1747	∞
	(.075)	(.0102)	(.0173)	(.076)	(.092)	(—)
4T	−.2858	.0608	.1087	.4061	.3040	4.4
	(.069)	(.0090)	(.0172)	(.092)	(.102)	(1.2)

NOTE: Standard errors are in parentheses.

vation of this form for the scale matrix. Overall, the model has 10 parameters: 6 for the mean function, 3 for the covariance matrix, and 1 for the ν df, which is estimated.

Rows 2T to 6T in Table 3 show results for parsimonious versions of the first model, obtained by placing the constraints on the parameters indicated in the table. The models are ranked by maximized log-likelihood. Models 5T and 6T fit the full mean structure but simplify the covariance structure. In particular, 6T sets the heritability parameter φ_2 to 0, yielding the autoregressive covariance structure fitted by LaVange and Helms (1983). The LR chi-squared statistic ($\chi^2 = 9.2$ on 1 df) is significant, suggesting that this covariance structure does not fit the data. Model 5T sets the decay parameter φ_3 to 0, yielding the compound symmetry structure. This model also does not fit well (LR $\chi^2_1 = 8.1$).

Models 2T to 4T retain φ_2 and φ_3 but progressively simplify the mean structure, dropping the quadratic terms (Model 2T), the slope times sex interaction, yielding a model with common slope and separate intercepts for males and females (Model 3T), and finally dropping the sex effect altogether (Model 4T). None of these simplifications significantly worsens the fit as measured by LR tests, so the simplest model (4T) seems a reasonable summary of the data. That is, the lung-function curves appear linear, with no differences between males and females.

The right side of Table 3 shows the fit of the same set of models with normal rather than t errors. The normal models fit much worse than their t counterparts (LR $\chi^2 >$ 40 on 1 df). Comparisons within the normal models again lead to Model 4 as the best-fitting model, taking into account parsimony. One interesting difference between the normal and t analyses emerges: Unlike the t analysis, setting $\varphi_3 = 0$ (Model 5N) does not significantly worsen the fit of the normal model (1N). It appears that for the normal model outliers are obscuring the (expected) decline in the covariances as the time between measurements increases. Parameter estimates from the best-fitting normal and t models (4N and 4T) are shown in Table 4, with asymptotic standard errors based on a numerical approximation of the observed information matrix (App. A). Note that the best-fitting t has between 4 and 5 df and increases the size and statistical significance of the φ_3 parameter, as expected from the model comparisons. The slopes of the regression lines are similar for the normal and t fits, but the intercept for the t fit is noticeably smaller.

2.4 Inherently Nonrectangular Multivariate Data

The models discussed in Section 2.3 concerned an underlying rectangular data matrix, although particular entries in the matrix might be missing. In other situations the data are inherently nonrectangular, and it makes little sense to refer to an underlying rectangular structure. One such situation is pedigree analysis, where the units of observation are extended families or pedigrees and pedigree sizes and complexity vary widely (Lange and Boehnke 1983). An example follows.

Example 6: Gc Measured Genotype Data. The Gc locus is known to determine qualitative variation in the human group component (Gc) protein, a transport protein for vitamin D. A question of interest is whether the genotypes at the Gc locus also determine quantitative differences in plasma Gc concentrations. The data of Daiger, Miller, and Chakraborty (1984) addressed this question. The Gc concentration and Gc genotype (1/1, 1/2, or 2/2), age, and gender are measured on 133 individuals, consisting of 31 monozygous twin pairs, 13 dizygous twin pairs, and 45 unrelated controls. The mean structure is specified by five parameters, three mean Gc concentrations for each genotype, a regression coefficient on the covariate age, and a main effect for sex. The covariance structure is specified by a total dispersion φ^2, assumed the same for all concentration measurements, and correlations φ_M and φ_D between concentrations of monozygous and dizygous twins, respectively.

Boerwinkle, Chakraborty, and Sing (1986) fitted normal models to these data. Table 5 shows fits to four models, fitted using both t and normal error structures. Model 1 is the model just described; the mean structure is denoted by $[G, A, S]$ to signify additive effects of genotype, age, and sex on concentration. The second model adds the constraint $\varphi_M = 2\varphi_D$ for the correlations, with insignificant deterioration in fit. In genetic terms, this suggests that a simple model with only additive genetic variance and ran-

Table 5. Models of Gc Locus Data, Example 6: Summary Fits for Three Models

	Multivariate t models					Multivariate normal models			
	Model description			Model fit				Model fit	
Model	Mean	Cov	No. of parameters	Log-likelihood	χ^2 (df)	Model	No. of parameters	Log-likelihood	χ^2 (df)
1T	G, A, S	$\varphi_M \neq 2\varphi_D$	9	−213.0	0 (0)	1N	8	−217.6	0 (0)
2T	G, A, S	$\varphi_M = 2\varphi_D$	8	−213.3	.6 (1)	2N	7	−217.7	.2 (1)
3T	A, S	$\varphi_M \neq 2\varphi_D$	7	−229.2	32.4* (2)	3N	6	−230.3	25.3* (2)
4T	G	$\varphi_M = 2\varphi_D$	6	−214.4	2.8 (3)	4N	5	−218.5	1.7 (3)

* Significantly worse fit than Model 1N (Normal models) or 1T (t models) at the 1% level (asymptotic LR chi-squared test).

Table 6. Models of Gc Locus Data, Example 6: Parameter Estimates and Standard Errors for Models 4N and 4T

Model	1/1 Mean	1/2 Mean	2/2 Mean	φ^2	φ_M	φ_D	ν
4N	31.74	29.47	26.00	12.76	.805	.403	∞
	(.47)	(.62)	(1.01)	(1.76)	(.056)	(.028)	(—)
4T	31.80	28.96	26.26	7.78	.845	.423	4.1
	(.41)	(.69)	(.85)	(1.95)	(.057)	(.028)	(1.8)

NOTE: Standard errors are in parentheses.

dom environmental variance may adequately fit the data. The third model sets the effects of genotype (G) to 0 and is decisively rejected under both the t and normal analyses (LR $\chi^2_2 = 32.4$ and 25.3, respectively). The final model includes genotype, constrains the correlations, and sets the age and sex effects equal to 0. This model fits well relative to the first model and is preferred on grounds of parsimony. For the models that include the effects of genotype (1, 2, 4), the t fits are significantly better than the normal fits (LR $\chi^2_1 = 9.3, 8.8$, and 8.2, respectively). Table 6 shows parameter estimates and large-sample standard errors for the normal and t fits to Model 4. The t fit yields an estimate of 4.1 df. Apart from the scale estimates that are not directly comparable, estimates of the other parameters are similar for the normal and t models, although the estimates of φ_M and φ_D are slightly larger for the t fit.

3. ESTIMATING PRECISION OF PARAMETER ESTIMATES

There is a variety of methods for estimating standard errors after fitting a t-family model. The observed or expected information matrix could be used, or a bootstrap resampling scheme could be employed. If ν is estimated from the data, the analyst can treat the ν df as fixed or (more appropriately) allow for its estimation. This second issue is less important than it might first appear, since the block-diagonal structure of the expected information matrix, derived for a general elliptically symmetric family of distribution in Appendix B, implies that the ML estimates of θ and ν are asymptotically uncorrelated.

In our implementation the observed information matrix is obtained by numerical differentiation of the score vector and matrix inversion of the resulting Hessian. The expected information matrix can be written explicitly (App. B). In particular, for the linear regression model $y_i \sim_{\text{ind}} t_1(x_i\theta, \psi^2, \nu)$, inverting the expected information matrix yields $\text{cov}(\hat{\theta}) = (\nu + 3)/(\nu + 1)(X^TX)^{-1}\psi^2$, whether or not ν is estimated, where X is the design matrix.

For the bootstrap schemes one can resample whole cases or resample residuals. For unbalanced multivariate problems such as Examples 5 and 6, resampling residuals is not possible, although it is possible to resample within patterns of the missing data as an alternative to equiprobable whole-case resampling (Su 1988), thus preserving the missingness structure of the data. Given large samples, all of these methods should give similar standard errors. Nevertheless, in small samples they can yield quite different answers, as the following two regression examples illustrate.

Example 7 (Ex. 1 continued): Precision of Estimated Regression Coefficients for Stack-Loss Data. Table 7 shows standard errors of the slopes and intercepts in Example 1 from a variety of methods. For the bootstrap standard errors 1,000 samples were generated, and ν was fixed at the ML estimate in estimating the parameters from the bootstrap samples.

For the normal models, eliminating the outliers more than halves the standard errors. The reduced standard errors are appropriate if outliers are known not to belong in the population, but are probably too optimistic, given uncertainty in the outlier-detection process. The asymptotic standard errors for the estimates from the t model seem small compared with the normal models, and are about 40% smaller than those obtained by bootstrapping the residuals. An interesting finding is that bootstrapping whole cases gives much larger standard errors than the other schemes. We expect bootstrapping cases to give larger standard errors than bootstrapping residuals because it introduces variability in the design, on top of the usual error variability (Wu 1986). The size of the discrepancy between the two sets of standard errors is surprising, however; it reflects in part the fact that the bootstrap distributions from bootstrapping cases had a high kurtosis. Su (1988), in an extensive simulation comparison of t and normal fits to incomplete trivariate data, found a similar pattern of results, with the information-based standard errors slightly optimistic and the bootstrap standard errors

Table 7. Standard Errors of Three Sets of Estimated Slopes From Table 1

	Method of estimation											
	Normal				t, $\nu = 1.1$				Normal minus four outliers			
Standard error method	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3	θ_0	θ_1	θ_2	θ_3
Asymptotic (observed information)	10.7	.121	.331	.141	3.8	.051	.134	.055	4.4	.059	.145	.054
Asymptotic (expected information)	10.7	.121	.331	.141	4.7	.054	.147	.063	4.4	.059	.145	.054
Bootstrap (cases)	8.2	.171	.477	.116	11.0	.208	.506	.156	4.4	.090	.166	.065
Bootstrap (residuals)	12.4	.141	.384	.163	7.5	.083	.223	.096	4.2	.059	.146	.055

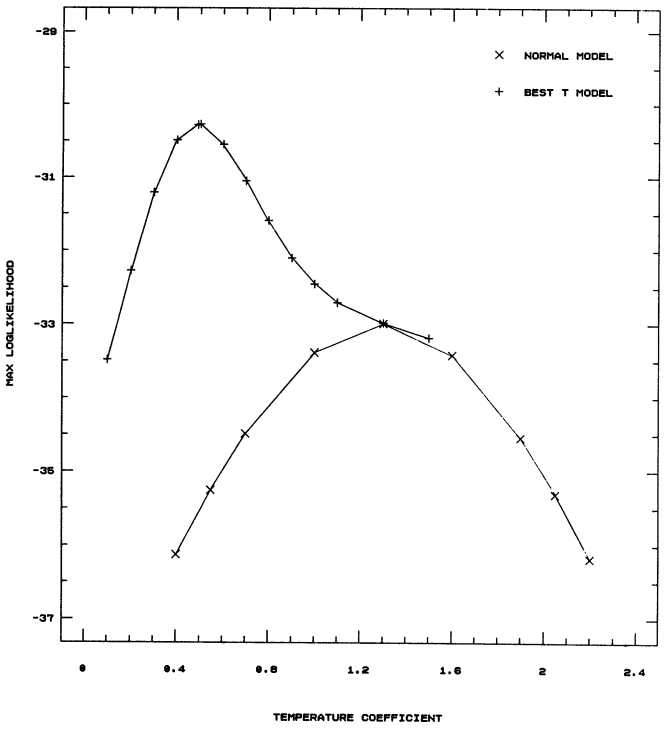

Figure 4. Example 1 Stack-Loss Data: Profile Likelihood Plots of θ_2—Normal and t Models.

slightly conservative. On his simulated data sets the differences between the methods were much smaller, and dissipated with increasing sample size.

Asymptotic confidence intervals may be computed by multiplying the standard error by the appropriate normal percentile. Nevertheless, some adjustment for the small sample size is advisable. An obvious strategy is to replace the normal percentile with the t, with $n - 4 = 17$, yielding the correct inference for the normal error model. A better approach is to plot the profile likelihood and calculate the interval corresponding to a fixed drop in the maximized log-likelihood based on the χ^2 approximation. Figure 4 shows the profile likelihood of θ_2 for the normal and best-t models; θ_2 was selected because of the appreciable effect of robust estimation on this parameter. For the normal model the plot is very symmetric and yields a 95% profile likelihood confidence interval very close to the standard t interval $(1.30 \pm (2.1)(.331)) = (.60, 2.0)$. For the t model the plot is asymmetric and the 95% profile likelihood confidence interval is (.21, .93), compared with t intervals $(.49 \pm (2.1)(.134)) = (.21, .77)$ based on the observed information and $(.49 \pm (2.1)(.223)) = (.02, .96)$ based on the residuals bootstrap. If the profile likelihood interval is regarded as approximating the truth (admittedly a dubious assumption), then the t interval based on the observed information is too narrow, and the interval based on the residuals bootstrap is a bit too wide.

Example 8: Traffic-Accident Data. Our final illustration of uncertainty calculations is a simple linear regression example with one explanatory variable, taken from Draper and Smith (1981, p. 191). The dependent variable is $\log_{10}$(driving deaths in 1964) and the independent variable is $\log_{10}$(number of drivers in 1964), for each of the $n = 50$ states. A plot of the data indicates that a straight line is appropriate with two possible outliers (Rhode Island and Connecticut). The least squares estimates of the intercept (θ_0) and the slope (θ_1) are -2.94 and .941, respectively. For the t model $\hat{\theta}_0 = -2.82$, $\hat{\theta}_1 = .925$, and $\hat{\nu} = 4.6$; a profile likelihood plot indicates a 95% confidence interval for ν of $(1.6, \infty)$.

Standard errors of estimates of θ_0 and θ_1 are shown in Table 8. The bootstrap and simulation results are based on 600 replications. The last two rows of the table are results from a simulation study in which 600 data sets are generated according to the model $y_i \sim_{\text{ind}} t_1(\hat{\theta}_0 + \hat{\theta}_1 x_i, \hat{\psi}^2, \hat{\nu})$, where parameter values are the ML estimates from the t-model fit. Each simulated data set is then fit by the t model with ν either estimated or fixed at 4.6. The values in the table indicate rough similarity between the results based on the information matrices, bootstrapping residuals, and the simulations. As in the previous example, bootstrapping cases results in appreciably larger standard errors than bootstrapping residuals. In both examples, the standard errors based on the normal after deleting the outliers appear too small. There is some indication that the standard errors based on the expected information matrix may be small. Standard errors are relatively unaffected by whether ν is estimated, supporting the result that asymptotically there is no price to pay for estimating ν and then ignoring that it is estimated. In both simulations the coverage of nominal 95% confidence intervals for θ_0 and θ_1 based on the expected information matrix was assessed. In all four cases the coverage was between 94% and 96%.

The t-model standard errors for this example show less variation between methods than those in Example 7. This might be expected, because we have more cases (50 compared to 21), fewer parameters (4 compared to 6), and $\hat{\nu}$ is larger (4.6 compared to 1.1), all of which suggest that the asymptotic theory is likely to be more accurate.

4. GOODNESS OF FIT AND OUTLIER ANALYSIS

Any statistical analysis should include a critical analysis of model assumptions. In this section we consider diag-

Table 8. Standard Errors of Intercept and Slope for Traffic-Accident Data

Method	Intercept (θ_0)	Slope (θ_1)	Method	Intercept (θ_0)	Slope (θ_1)
Normal	.256	.0420	Bootstrap cases (ν = 4.6, fixed)	.311	.0505
Normal minus two outliers	.206	.0338	Bootstrap residuals (ν varying)	.226	.0377
t ($\hat{\nu}$ = 4.6) (observed information)	.266	.0429	Bootstrap residuals (ν = 4.6, fixed)	.223	.0373
t ($\hat{\nu}$ = 4.6) (expected information)	.211	.0347	Simulation standard deviation (ν estimated)	.238	.0399
Bootstrap cases (ν varying)	.317	.0511	Simulation standard deviation (ν = 4.6, fixed)	.234	.0392

nostics for checking the fit of the t models. We first consider statistics defined for each case i, and then briefly mention statistics defined for the values within each case.

For the normal model (1), a natural measure of closeness of the ith observation to the center of the distribution is the Mahalanobis-like distance $\delta_i^2(\theta, \varphi) = \{y_i - \mu_i(\theta)\}^T \Sigma_i^{-1}(\varphi)\{y_i - \mu_i(\theta)\}$, which under (1) has a chi-squared distribution with v_i df. Substituting ML estimates of θ and φ yields $d_i^2 \equiv \delta_i^2(\hat{\theta}, \hat{\varphi})$, which has asymptotically the same chi-squared distribution as δ_i^2. A check on normality is achieved by transforming each d_i^2 to an asymptotically standard normal deviate and then plotting the ordered values against expected normal order statistics; for the special case of univariate least squares regression, this is the familiar half-normal plot. Deviations from the 45-degree line suggest lack of normality; in particular, larger-

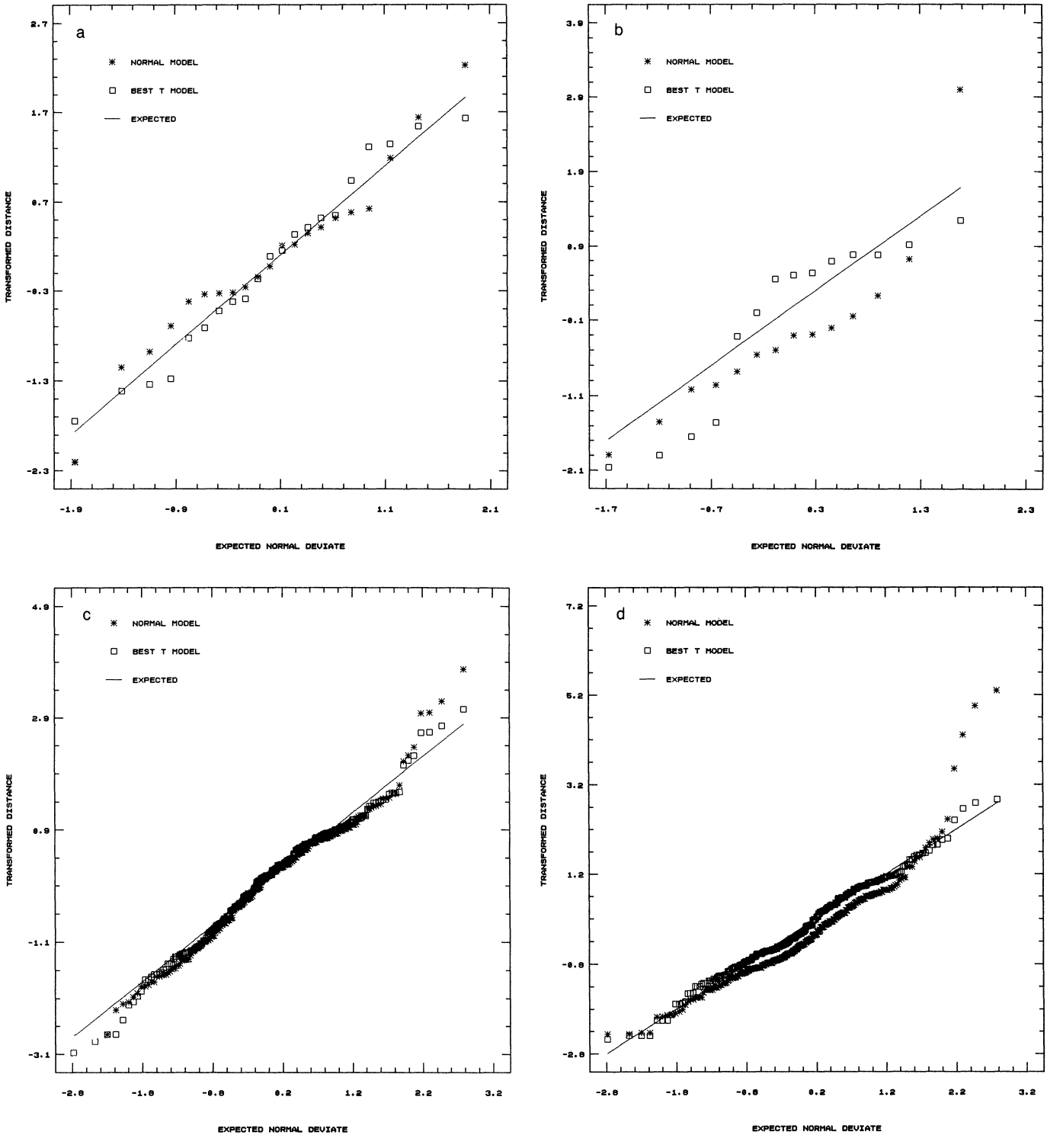

Figure 5. Plots of Transformed Distances for the Normal and Best-t Models: (a) Example 1 Stack-Loss Data; (b) Example 2 Radioimmunoassay Data; (c) Example 3 Blood-Flow Data, y_1; (d) Example 3 Blood-Flow Data, y_2; (e) Example 4 Blood-Flow Data, y_1 and y_2; (f) Example 5 Lung-Function Data; (g) Example 6 Genotype Data.

than-expected values of d_i^2 identify outlying cases (Gnanadesikan 1977; Hopper and Mathews 1982; Little 1988a; Little and Smith 1987). We carried out the transformation to normality by numerical integration. Alternatively, approximations good enough for plotting can be obtained by transforming to approximate normality using a cube-root or fourth-root transformation of d_i^2 (Hawkins and Wixley 1986).

If this plot reveals outliers, one might fit the multivariate t model (2). For this model define the distances

$$\delta_i^2(\theta, \varphi) = \{y_i - \mu_i(\theta)\}^T \Psi_i^{-1}(\varphi)\{y_i - \mu_i(\theta)\} \quad (7)$$

and $d_i^2 \equiv \delta_i^2(\hat{\theta}, \hat{\varphi})$. By Property 4, under this model δ_i^2/v_i is F-distributed with v_i and v df. In addition, d_i^2/v_i has the same F distribution asymptotically. The latter F statistics can be transformed to standard normal deviates and plot-

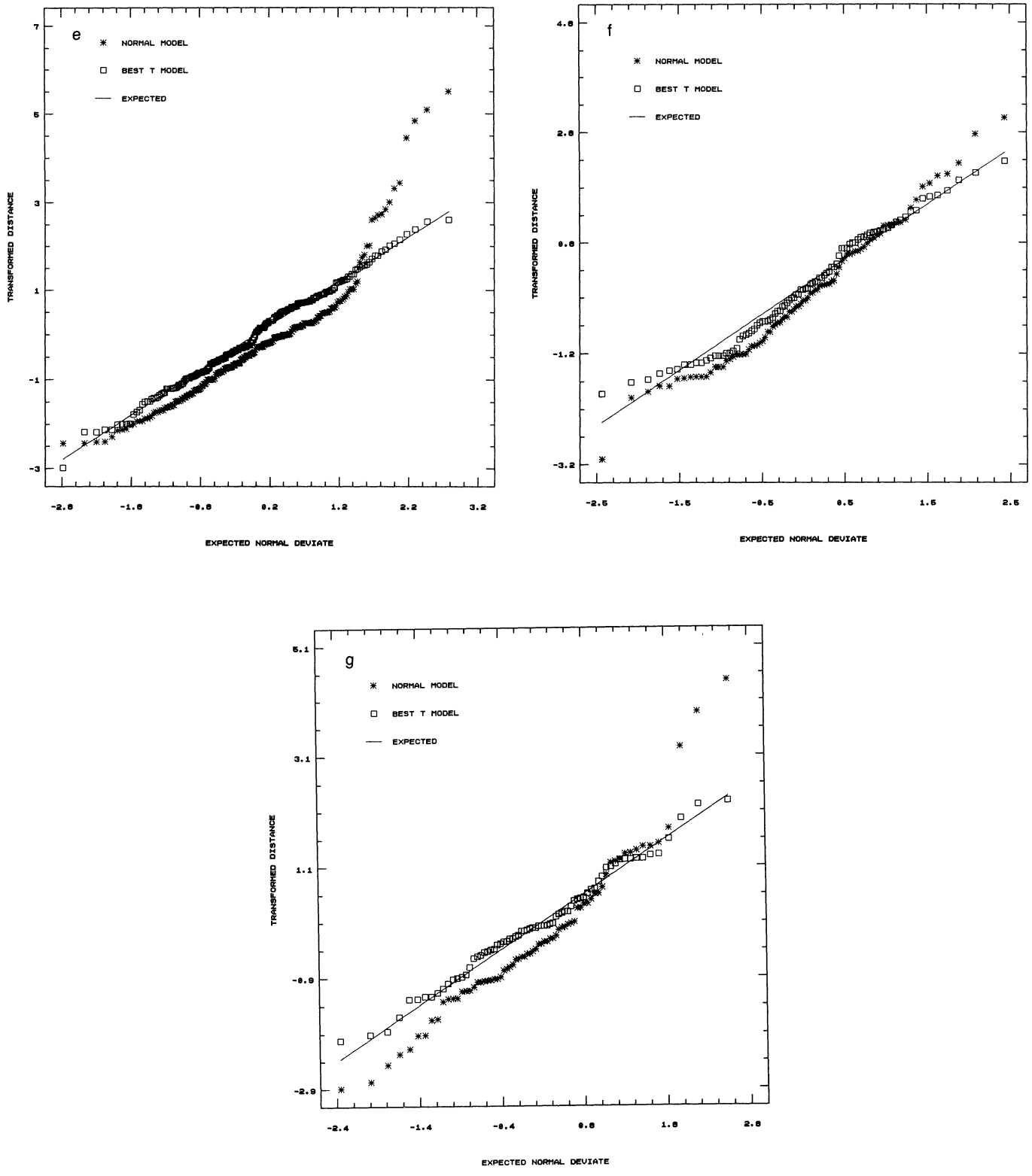

Figure 5 (continued).

ted against their expectations to assess the fit of the t model. Approximate transformations from F to normality are discussed in Little (in press).

Figure 5 shows these *transformed distance plots* for Examples 1–6 in Section 2. Plots for the best-fitting normal and t models are superimposed. Note that the larger transformed distances from the normal models drift above the 45-degree line, indicating longer-than-normal tails, particularly for Examples 3, 4, and 5, where the normal model fits poorly. The plots from the t models all follow the 45-degree line more closely, confirming the better fit.

In Section 2 we noted that the best t fit for Example 2 had very low degrees of freedom (.29) and was unsatisfying in that it placed nearly all of the weight on just four points. This unsatisfactory property is reflected in the transformed distance plot for this model, shown in Figure 5b: the points with the four smallest transformed distances for the best t-model line are markedly below the line, suggesting a closer fit than expected. The plot with $\nu = 1$ (not shown) looks better for this data set. Thus in this example visual inspection of the transformed distance plot suggests a model that does not maximize the likelihood, tempering our enthusiasm for a choice of degrees of freedom based exclusively on the likelihood. In defense of such a strategy, we note that it appears satisfactory in the other examples discussed here. In addition, in problems with small sample sizes such as Example 2 one might fix the degrees of freedom at some predetermined value (such as 4), rather than attempting to estimate degrees of freedom simultaneously with the other parameters. Integration over a prior for ν is another possibility (Relles and Rogers 1977).

The plots shown here provide a means for identifying outlying cases from the normal model and assessing the success of the t models in dampening their influence. For identifying outlying values within cases, Little and Smith (1987) propose a stepwise procedure, where values are identified that yield successively the greatest reduction in d_i^2 when removed.

5. CONCLUSION

This article illustrates the ability of models based on the t distribution to handle outliers in a wide range of settings. The incorporation of this approach in existing statistical software appears feasible and should enhance the ability to address robustness concerns fairly routinely when conducting multivariate analyses of quantitative outcomes. We conclude with some remarks on the limitations of the approach, and mention some areas that seem to us to require further study.

1. Our basic model (2) is parametric. A referee noted that extensions to nonparametric and semiparametric models, where the mean structure is an unspecified or partially specified smooth function of covariates, seem possible in principle. For example, the robust locally weighted regression algorithms of Cleveland (1979) could be modified to model t errors by changing the robustness weights (δ_k in Cleveland's article) to the form (3) (with $v_i = 1$) appropriate for univariate t errors. Properties of such procedures remain to be considered.

2. As noted in Section 1, models other than (2) also yield robust estimates, such as the contaminated normal, the generalization of the slash distribution of Rogers and Tukey (1962), or the exponential power family (Box and Tiao 1973); we chose the t mainly on grounds of familiarity and computational simplicity. Studies comparing these alternative models might be useful, particularly in multivariate settings where previous work appears limited.

3. Like most robust methods, the t family still works within a symmetric class of distributions. Sometimes a transformation to symmetry is available. If not, the appropriate analysis of location when the distribution is skewed is an issue that needs careful attention.

4. In multivariate applications, the same parameter ν models nonnormality in all of the variables. A useful generalization allows ν to vary across the variables to model differing degrees of nonnormality. For example, for the data in Examples 3 and 4 a model with a different value of ν for y_1 and y_2 might improve the fit slightly. Little (1988a) discusses such extensions for the multivariate t model with unconstrained location vector and scale matrix.

5. Although ML estimation provides the machinery for computing large-sample standard errors that take into account the presence of outliers, exactly when the data set is large enough for asymptotic theory to apply, and what to do when asymptotic theory does not apply, require further attention.

6. When should ν be fixed a priori at some sensible value, and when should it be estimated from the data? General principles of parsimony suggest that ν should be fixed for small data sets and estimated for large ones. Our regression examples and theory suggest surprisingly little added variance from estimating ν rather than treating it as known. But Example 2 suggests that estimated values of ν below 1 should be regarded with suspicion. And inferences about ν itself might be improved by transforming to $1/\nu$ or $\log(\nu)$.

7. Although we found few numerical problems in the data sets we studied, widely distributed software should recognize and deal with the possibility of multiple maxima of the likelihood, particularly with sparse data sets and small values of ν.

APPENDIX A: LIKELIHOOD FITTING

A.1 Maximum Likelihood Algorithms

The log-likelihood for Model (2) is, ignoring constants, $l(\theta, \varphi, \nu) = \sum_{i=1}^{n} l_i(\theta, \varphi, \nu)$, where

$$l_i(\theta, \varphi, \nu) = \frac{1}{2}\ln|\Psi_i(\varphi)| - \frac{1}{2}(\nu + v_i)\ln\left(1 + \frac{\delta_i^2(\theta, \varphi)}{\nu}\right) - \frac{1}{2}v_i\ln(\nu) + \ln\left[\Gamma\left(\frac{\nu + v_i}{2}\right)\right] - \ln\Gamma\left(\frac{\nu}{2}\right);$$

Γ denotes the gamma function and $\delta_i^2(\theta, \varphi)$ is given by (7).

We have experience in three iterative methods for maximizing the log-likelihood: a quasi-Newton (QN) algorithm, a scoring algorithm with variable step length, and an EM algorithm; the QN algorithm is implemented in a FORTRAN program named Fisher (Lange, Boehnke, and Weeks 1987). Given estimates $\gamma^{(t)} = (\theta^{(t)}, \varphi^{(t)}, \nu^{(t)})$ at iteration t, the QN and scoring algorithms compute $\gamma^{(t+1)} = \gamma^{(t)} + \lambda_t Q^{-1}(\gamma^{(t)})S(\gamma^{(t)})$ where (a) $S(\gamma)$ is the

score vector of $l(\gamma)$, given in Proposition 1 in Appendix B; (b) for the scoring algorithm $Q(\gamma)$ is the expected information, given in Proposition 4 in Appendix B; (c) for the QN algorithm $Q(\gamma)$ is an approximation to the observed information matrix, based on the successive rank-two modifications of Broyden, Fletcher, Goldfarb, and Shanno, as explained by Powell (1978). The term $Q^{-1}(\gamma^{(t)})S(\gamma^{(t)})$ furnishes the direction of the current increment, and $\lambda_t > 0$ determines its length. Both algorithms are ascent algorithms in the sense that taking λ_t sufficiently small forces an increase in $l(\gamma)$. The QN algorithm handles parameter bounds and linear constraints using the method sketched by Jennrich and Sampson (1978).

The EM algorithm (Dempster, Laird, and Rubin 1977; Little and Rubin 1987) augments the data $Y = (y_1, \ldots, y_n)$ by additional hypothetical data $Z = (z_1, z_2, \ldots, z_n)$ such that ML estimates of γ, given $Y^* = (Y, Z)$, are easy to compute. Given estimates $\gamma^{(t)}$ at iteration t, the $(t + 1)$st iteration of EM consists of an expectation (E) step and a maximization (M) step. The E step computes the expected value of the complete-data log-likelihood $l(\gamma \mid Y, Z)$ with respect to the conditional distribution of Z, given Y and $\gamma^{(t)}$. The M step maximizes the resulting function with respect to γ, yielding new estimates $\gamma^{(t+1)}$. Under mild conditions each iteration of EM increases the log-likelihood $l(\gamma \mid Y)$.

EM is particularly useful when the M step is noniterative, since then the algorithm does not require inversion of an information matrix at each iteration, which is burdensome when the number of parameters is large. When the M step is iterative, EM may still be useful if the maximization of the complete-data likelihood is available using existing software. Sections A.2 and A.3 discuss two special cases of (2) where EM is useful.

Starting values for θ and φ were obtained by fitting a normal model by least squares, with covariances in the multivariate models set to 0. In univariate models the starting value for ν can be determined from the kurtosis of residuals from the normal fit; alternatively, a grid search over values of ν can be used. Multiple maxima of the likelihood seem possible, particularly when ν is small; however, we did not find any for our problems.

A.2 EM for a General Univariate Regression Model With t Errors

Let y_i denote a scalar outcome. By Property 1, the model

$$y_i \mid u_i \underset{\text{ind}}{\sim} N\{\mu(\theta; x_i), \psi^2/(c_i u_i)\}, \qquad u_i \underset{\text{ind}}{\sim} \chi^2_\nu/\nu \tag{A.1}$$

yields the t regression model (4) for y_i, where $\mu(\ )$ is a regression function with covariates x_i and a $(p \times 1)$ vector of regression coefficients θ, c_i is a known constant, and ψ^2 is an unknown scale parameter. ML estimation for Model (3) can be achieved by applying the EM algorithm with missing data $\{u_i: i = 1, \ldots, n\}$.

If ν is assumed known, EM is iteratively reweighted least squares. The E step computes the weight

$$w_i^{(t)} = \mathrm{E}(u_i \mid y_i, \theta^{(t)}, \psi^{(t)}) = \frac{\nu + 1}{\nu + \delta_i^2(\theta^{(t)}, \psi^{(t)})}, \tag{A.2}$$

where $\delta_i^2(\theta, \psi) = c_i\{y_i - \mu(\theta, x_i)\}^2/\psi^2$, by Property 3. [Compare (3) with $v_i = 1$.] The M step finds $\theta^{(t+1)}$ to minimize the weighted sum of squares $\Sigma_i c_i w_i^{(t)}\{y_i - \mu(\theta, x_i)\}^2$, and $\psi^{(t+1)} = \Sigma_i c_i w_i^{(t)}\{y_i - \mu(\theta^{(t+1)}, x_i)\}^2/n$. Since the M step is weighted least squares, it is noniterative if the regression is linear (Dempster et al. 1977) and otherwise can be carried out by any nonlinear regression program that handles weights for the cases.

If ν is treated as a parameter, it can be estimated by repeating the aforementioned algorithm over a grid of values of ν. Alternatively, EM can be applied to estimate ν simultaneously with θ and ψ. Given current estimates $\gamma^{(t)} = (\theta^{(t)}, \psi^{(t)}, \nu^{(t)})$, the E step computes $w_i^{(t)}$ using (A.2) with $\nu = \nu^{(t)}$, and in addition

$$\begin{aligned} q_i^{(t)} &= \mathrm{E}(\ln u_i \mid y_i, \gamma^{(t)}) \\ &= \mathrm{DG}\left(\frac{\nu^{(t)}}{2} + \frac{1}{2}\right) - \ln(\nu^{(t)} + \delta_i^2(\theta^{(t)}, \psi^{(t)})/2), \end{aligned}$$

where $\mathrm{DG}(x) = d/dx \ln \Gamma(x)$, the digamma function [from Property 3 in Sec. 1]. The M step computes $\theta^{(t+1)}$ and $\psi^{(t+1)}$ by weighted least squares with weights $w_i^{(t)}$ and finds $\nu^{(t+1)}$ that maximizes

$$\begin{aligned} &l_2(\nu) \\ &= \frac{n\nu}{2} \ln\left(\frac{\nu}{2}\right) - n \ln\left\{\Gamma\left(\frac{\nu}{2}\right)\right\} + \left(\frac{\nu}{2} - 1\right) \sum_{i=1}^n q_i^{(t)} - \frac{\nu}{2} \sum_{i=1}^n w_i^{(t)}, \end{aligned}$$

using a one-dimensional search such as Newton's method.

A.3 EM for Multivariate t Models

Suppose that y_i is now $(k \times 1)$, and

$$y_i \mid u_i \underset{\text{ind}}{\sim} N_k(\mu_i(\theta), \Psi(\varphi)/u_i), \qquad u_i \underset{\text{ind}}{\sim} \chi^2_\nu/\nu, \tag{A.3}$$

yielding the multivariate t model

$$y_i \underset{\text{ind}}{\sim} t_k(\mu_i(\theta), \Psi(\varphi), \nu), \tag{A.4}$$

a special case of (2) where y_i has the same dimension and the same scale matrix for all i. Given data $\{y_i; i = 1, \ldots, n\}$ and Model (A.4), ML estimates of the parameters can be found by EM, treating $\{u_i\}$ as missing data. Furthermore, EM can be extended to handle missing values in the $\{y_i\}$ by treating both $\{u_i\}$ *and* the missing components of $\{y_i\}$ as missing data. Details for the case where ν is known ($\mu_i = \theta^T x_i$) and Ψ unconstrained (i.e., multivariate regression with t errors and missing y data) are given in Little (1988a). The algorithm can be extended to handle simultaneous estimation of ν by a simple generalization of the univariate case in Section A.2.

For other choices of mean and covariance structures, the E step remains essentially the same, and the M step becomes equivalent to ML for the complete-data model (A.3). Again, for EM to be useful the M step should be noniterative, or available using existing software. For example, suppose that the mean and covariance matrix have the linear structure $\mu_i(\theta) \equiv \mu = X\theta$ and $\Psi(\varphi) = \Sigma_{g=1}^r \varphi_g G_g$, where X is a known $(k \times q)$ matrix, θ is $(q \times 1)$, $G_1, \ldots, G_r$ are known linearly independent $(k \times k)$ matrices, and $\varphi_1, \ldots, \varphi_r$ are covariance-structure parameters. Szatrowski (1980a) gave necessary and sufficient conditions on X and $G_1, \ldots, G_r$ for the existence of explicit ML estimates of θ and φ. Szatrowski (1980b) applied these conditions to provide explicit estimates for balanced data in mixed-model analyses of variance.

APPENDIX B: DERIVATION OF THE SCORE AND EXPECTED INFORMATION

It is not difficult to carry through most of the computations for an arbitrary elliptically symmetric family of densities (Chmielewski 1981) of the form $p(y \mid \mu, \Psi, \nu) = |\Psi|^{-1/2} g((y - \mu)^T \Psi^{-1} (y - \mu), \nu)$, where y and $\mu = \mu(\theta)$ are $(k \times 1)$, $\Psi = \Psi(\varphi)$ is $(k \times k)$, and ν is a scalar parameter modeling kurtosis. For the k-variate t,

$$g(s, \nu) = \frac{\Gamma((\nu + k)/2)}{\Gamma(1/2)^k \Gamma(\nu/2) \nu^{k/2}} \left(1 + \frac{s}{\nu}\right)^{-(\nu + k)/2}.$$

Another example is afforded by the generalized power-exponential family with $g(s, \nu) = c(\nu) e^{-s^\nu/2}$ (Box and Tiao 1973).

Proposition 1. Let $l(\theta, \varphi) = \ln p(y \mid \mu(\theta), \Psi(\varphi), \nu)$ be the contribution of a typical observation to the log-likelihood. The

contribution to the score has components

$$\frac{\partial l}{\partial \theta_i} = -2\frac{g_1}{g}\frac{\partial \mu^T}{\partial \theta_i}\Psi^{-1}(y-\mu),$$

$$\frac{\partial l}{\partial \varphi_i} = -\frac{1}{2}\operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\right) - \frac{g_1}{g}(y-\mu)^T\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\Psi^{-1}(y-\mu),$$

and

$$\frac{\partial l}{\partial v} = \frac{g_2}{g},$$

where g_1 and g_2 denote partial derivatives with respect to the first and second entries of g, respectively.

Proof. These expressions are straightforward to compute using the rules

$$\frac{\partial}{\partial \theta_i}[(y-\mu)^T\Psi^{-1}(y-\mu)] = -2\frac{\partial \mu^T}{\partial \theta_i}\Psi^{-1}(y-\mu),$$

$$\frac{\partial}{\partial \varphi_i}\ln|\Psi| = \operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\right),$$

and

$$\frac{\partial \Psi^{-1}}{\partial \varphi_i} = -\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\Psi^{-1}.$$

To exploit symmetry we change variables to $z = \Psi^{-1/2}(y-\mu)$. Clearly, given $\|z\| = r$, z is uniformly distributed on the sphere $\|z\| = (\sum_{i=1}^k z_i^2)^{1/2} = r$. In addition,

$$\frac{\partial l}{\partial \theta_i} = -\frac{2g_1}{g}\frac{\partial \mu^T}{\partial \theta_i}\Psi^{-1/2}z$$

and

$$\frac{\partial l}{\partial \varphi_i} = -\frac{1}{2}\operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\right) - \frac{g_1}{g}z^T\Psi^{-1/2}\frac{\partial \Psi}{\partial \varphi_i}\Psi^{-1/2}z.$$

Now, let J denote the contribution of the current observation to the expected information.

Proposition 2. J is block diagonal with the mean components θ in one block and the scale components φ and v in another block.

Proof. $J_{\theta_i\varphi_j} = E((\partial l/\partial\theta_i)(\partial l/\partial\varphi_j)) = 0$, since for $\|z\|$ fixed $\partial l/\partial\theta_i$ is an odd function of z and $\partial l/\partial\varphi_j$ is an even function of z. Similarly, $J_{\theta_i v} = 0$.

Lemma 1 [after Graybill (1983, p. 366) and Huber (1981, pp. 231, 232)]. For any $k \times k$ matrices A and B,

$$E\left(\frac{z^T}{\|z\|}A\frac{z}{\|z\|}\,\middle|\,\|z\|\right) = \frac{1}{k}\operatorname{tr}(A) \tag{B.1}$$

and

$$E\left(\frac{z^T}{\|z\|}A\frac{z}{\|z\|}\frac{z^T}{\|z\|}B\frac{z}{\|z\|}\,\middle|\,\|z\|\right) = \frac{1}{k(k+2)}[2\operatorname{tr}(AB) + \operatorname{tr}(A)\operatorname{tr}(B)]. \tag{B.2}$$

Proof. Consider (B.2). It suffices to replace z by $w \sim N_k(0, I)$. Then, by direct computation, $E(w^TAww^TBw) = 2\operatorname{tr}(AB) + \operatorname{tr}(A)\operatorname{tr}(B)$. When $A = B = I$, $E(\|w\|^4) = 2k + k^2$. Since

$$E(w^TAww^TBw) = E(\|w\|^4)E\left(\frac{w^T}{\|w\|}A\frac{w}{\|w\|}\frac{w^T}{\|w\|}B\frac{w}{\|w\|}\,\middle|\,\|w\| = 1\right),$$

(B.2) follows immediately upon division by $E(\|w\|^4)$.

Proposition 3.

$$J_{\theta_i\theta_j} = 4E\left(\|z\|^2\left(\frac{g_1}{g}\right)^2\right)\frac{1}{k}\frac{\partial \mu^T}{\partial \theta_i}\Psi^{-1}\frac{\partial \mu}{\partial \theta_j}, \tag{B.3}$$

$$J_{\varphi_i\varphi_j} = \frac{C}{4} + \frac{C}{k}E\left[\|z\|^2\left(\frac{g_1}{g}\right)\right] + \frac{1}{k(k+2)}\left[C + 2\operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_j}\right)\right] \times E\left[\|z\|^4\left(\frac{g_1}{g}\right)^2\right], \tag{B.4}$$

where $C = \operatorname{tr}(\Psi^{-1}\,\partial\Psi/\partial\varphi_i)\operatorname{tr}(\Psi^{-1}\,\partial\Psi/\partial\varphi_j)$,

$$J_{\varphi_i v} = E\left(\|z\|^2\frac{\partial}{\partial v}\left(\frac{g_1}{g}\right)\right)\frac{1}{k}\operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\right), \tag{B.5}$$

and

$$J_{vv} = E\left(-\frac{\partial}{\partial v}\left(\frac{g_2}{g}\right)\right). \tag{B.6}$$

Proof. We prove (B.3) as a typical example. Using (B.1),

$$\begin{aligned} J_{\theta_i\theta_j} &\equiv E\left(\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right) \\ &= E\left[\left(-2\frac{g_1}{g}\right)^2 z^T\Psi^{-1/2}\frac{\partial \mu}{\partial \theta_i}\frac{\partial \mu^T}{\partial \theta_j}\Psi^{-1/2}z\right] \\ &= 4E\left[\|z\|^2\left(\frac{g_1}{g}\right)^2 E\left(\frac{z^T}{\|z\|}\Psi^{-1/2}\frac{\partial \mu}{\partial \theta_i}\frac{\partial \mu^T}{\partial \theta_j}\Psi^{-1/2}\frac{z}{\|z\|}\,\middle|\,\|z\|\right)\right] \\ &= 4E\left[\|z\|^2\left(\frac{g_1}{g}\right)^2\right]\frac{1}{k}\operatorname{tr}\left(\Psi^{-1/2}\frac{\partial \mu}{\partial \theta_i}\frac{\partial \mu^T}{\partial \theta_j}\Psi^{-1/2}\right) \\ &= 4E\left(\|z\|^2\left(\frac{g_1}{g}\right)^2\right)\frac{1}{k}\frac{\partial \mu^T}{\partial \theta_j}\Psi^{-1}\frac{\partial \mu}{\partial \theta_i}. \end{aligned}$$

In effect, Proposition 3 reduces the computation of J to the evaluation of one-dimensional integrals, since for any well-behaved function $f(r)$, $E(f(\|z\|)) = \int_0^\infty f(r)g(r^2, v)r^{k-1}c_k\,dr$, where c_k is the surface area of the unit sphere in $\mathbf{R}^k$. We now specialize to the t distribution.

Proposition 4. For the t distribution,

$$J_{\theta_i\theta_j} = \frac{v+k}{v+k+2}\frac{\partial \mu^T}{\partial \theta_i}\Psi^{-1}\frac{\partial \mu}{\partial \theta_j},$$

$$J_{\varphi_i\varphi_j} = \frac{v+k}{v+k+2}\frac{1}{2}\operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_j}\right) - \frac{1}{2(v+k+2)}\operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\right)\operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_j}\right),$$

$$J_{\varphi_i v} = -\frac{1}{(v+k+2)(v+k)}\operatorname{tr}\left(\Psi^{-1}\frac{\partial \Psi}{\partial \varphi_i}\right),$$

and

$$J_{vv} = -\frac{1}{2}\left[\frac{1}{2}TG\left(\frac{v+k}{2}\right) - \frac{1}{2}TG\left(\frac{v}{2}\right) + \frac{k}{v(v+k)} - \frac{1}{v+k} + \frac{v+2}{v(v+k+2)}\right],$$

where $TG(x) = d^2/d^2x \ln\Gamma(x)$ is the trigamma function.

Proof. For the t distribution,

$$\frac{g_1}{g} = -\frac{1}{2}\frac{v+k}{v+\|z\|^2},$$

$$\frac{\partial}{\partial v}\left(\frac{g_1}{g}\right) = -\frac{1}{2}\frac{1}{v+\|z\|^2} + \frac{1}{2}\frac{v+k}{(v+\|z\|^2)^2},$$

and

$$-\frac{\partial}{\partial v}\left(\frac{g_2}{g}\right) = -\frac{1}{2}\left[\frac{1}{2}TG\left(\frac{v+k}{2}\right) - \frac{1}{2}TG\left(\frac{v}{2}\right) + \frac{1}{v}\frac{\|z\|^2}{v+\|z\|^2} - \frac{1}{v+\|z\|^2} + \frac{v+k}{(v+\|z\|^2)^2}\right].$$

Integration yields

$$E\left(\left[1+\frac{\|z\|^2}{v}\right]^{-m}\right) = \frac{(v/2+m-1)\cdots(v/2)}{((v+k)/2+m-1)\cdots((v+k)/2)},$$

and in particular

$$E\left(\frac{\|z\|^2}{v}\left(1+\frac{\|z\|^2}{v}\right)^{-1}\right) = E\left(1-\left(1+\frac{\|z\|^2}{v}\right)^{-1}\right) = \frac{k}{v+k},$$

$$E\left(\frac{\|z\|^2}{v}\left[1+\frac{\|z\|^2}{v}\right]^{-2}\right) = \frac{vk}{(v+k+2)(v+k)},$$

and

$$E\left(\left[\frac{\|z\|^2}{v}\right]^2\left[1+\frac{\|z\|^2}{v}\right]^{-2}\right) = \frac{k(k+2)}{(v+k+2)(v+k)}.$$

The proposition follows by applying these facts to the expressions in Proposition 3.

Notes. Summing expressions over observations gives the expected information matrix. As $v \to \infty$, one recovers the expected information matrix for the corresponding normal distribution.

[*Received December 1987. Revised May 1989.*]

REFERENCES

Andrews, D. F. (1974), "A Robust Method for Multiple Linear Regression," *Technometrics,* 16, 523–531.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), *Robust Estimates of Location: Survey and Advances,* Princeton, NJ: Princeton University Press.

Berkane, M., and Bentler, P. M. (1988), "Estimation of Contamination Parameters and Identification of Outliers in Multivariate Data," *Sociological Methods and Research,* 17, 55–64.

Boerwinkle, E., Chakraborty, R., and Sing, C. F. (1986), "The Use of Measured Genotype Information in the Analysis of Quantitative Phenotypes in Man, I: Models and Analytical Methods," *Annals of Human Genetics,* 50, 181–194.

Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modeling and Robustness," *Journal of the Royal Statistical Society,* Ser. A, 143, 383–430.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society,* Ser. B, 26, 211–252.

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

Browne, M. W. (1984), "Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures," *British Journal of Mathematical and Statistical Psychology,* 37, 62–83.

Brownlee, K. A. (1965), *Statistical Theory and Methodology in Science and Engineering* (2nd ed.), New York: John Wiley.

Chmielewski, M. A. (1981), "Elliptically Symmetric Distributions: A Review and Bibliography," *International Statistical Review,* 49, 67–74.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association,* 74, 829–836.

Cornish, E. A. (1954), "The Multivariate t-Distribution Associated With a Set of Normal Standard Deviates," *Australian Journal of Physics,* 7, 531–542.

Crone, C. (1963), "Permeability of Capillaries in Various Organs as Determined by the Indicator Diffusion Method," *Acta Physiologica Scandinavica,* 58, 292.

Daiger, S. P., Miller, M., and Chakraborty, R. (1984), "Heritability of Quantitative Variation at the Group-Specific Component (Gc) Locus," *American Journal of Human Genetics,* 36, 663–676.

Daniel, C., and Wood, F. S. (1971), *Fitting Equations to Data,* New York: John Wiley.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 39, 1–38.

Dixon, W. J. (ed.) (1983), *BMDP Statistical Software,* Berkeley: University of California Press.

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

Dunnett, C. W., and Sobel, M. (1954), "A Bivariate Generalization of Student's t-Distribution With Tables for Certain Special Cases," *Biometrika,* 41, 153–169.

Fraser, D. A. S. (1976), "Necessary Analysis and Adaptive Inference" (with discussion), *Journal of the American Statistical Association,* 71, 99–113.

——— (1979), *Inference and Linear Models,* New York: McGraw-Hill.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations,* New York: John Wiley.

Graybill, F. A. (1983), *Matrices With Applications to Statistics* (2nd ed.), Monterey, CA: Wadsworth.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions,* New York: John Wiley.

Hawkins, D. M., and Wixley, R. A. J. (1986), "A Note on the Transformation of Chi-Squared Variables to Normality," *The American Statistician,* 40, 296–298.

Hogg, R. V. (1974), "Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory" (with discussion), *Journal of the American Statistical Association,* 69, 909–926.

Hopper, J. L., and Mathews, J. D. (1982), "Extensions of the Multivariate Normal Models for Pedigree Analysis," *Annals of Human Genetics,* 46, 373–383.

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

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

Jennrich, R. I., and Sampson, P. F. (1978), "Some Problems Faced in Making a Variance Component Algorithm Into a General Mixed Model Program," in *Proceedings of Computer Science and Statistics, Symposium on the Interface* (Vol. 11), Raleigh: North Carolina State University, pp. 56–63.

Jennrich, R. I., and Schluchter, M. D. (1986), "Unbalanced Repeated-Measures Models With Structured Covariance Matrices," *Biometrics,* 42, 805–820.

Lange, K. (1986), "Cohabitation, Convergence, and Environmental Covariances," *American Journal of Medical Genetics,* 24, 483–491.

Lange, K., and Boehnke, M. (1983), "Extensions of Pedigree Analysis, IV: Covariance Components Models for Multivariate Traits," *American Journal of Medical Genetics,* 14, 513–524.

Lange, K., Boehnke, M., and Weeks, D. (1987), "Programs for Pedigree Analysis," report, University of California, Los Angeles, Dept. of Biomathematics.

LaVange, L. M., and Helms, R. W. (1983), "The Analysis of Incomplete Longitudinal Data With Modeled Covariance Structures," Mimeo 1449, University of North Carolina, Inst. of Statistics.

Little, R. J. A. (1988a), "Robust Estimation of the Mean and Covariance Matrix From Data With Missing Values," *Applied Statistics,* 37, 23–39.

——— (1988b), "Analysis of Data With Missing Values: Discussion," *Statistics in Medicine,* 7, 347–355.

——— (in press), "Editing and Imputation of Multivariate Data: Issues and New Approaches," in *Theory and Pragmatics of Data Quality Control,* eds. G. Liepens and V. R. R. Uppuluru, New York: Marcel Dekker.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data,* New York: John Wiley.

Little, R. J. A., and Smith, P. J. (1987), "Editing and Imputation for Quantitative Survey Data," *Journal of the American Statistical Association,* 82, 58–68.

Maronna, R. A. (1976), "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 4, 51–67.

Masreliez, C. J., and Martin, R. D. (1977), "Robust Bayesian Estimation for the Linear Model and Robustifying the Kalman Filter," *IEEE Transactions on Automatic Control,* 22, 361–371.

Pendergast, J. F., and Broffitt, J. D. (1985), "Robust Estimation in Growth Curve Models," *Communications in Statistics—Theory and Methods,* 14, 1919–1939.

Powell, M. J. D. (1978), "A Fast Algorithm for Nonlinearly Constrained Optimization Calculations," in *Proceedings of the 1977 Dundee Conference on Numerical Analysis,* ed. G. A. Watson, Berlin: Springer-Verlag, pp. 144–157.

Relles, D. A., and Rogers, W. H. (1977), "Statisticians Are Fairly Robust Estimators of Location," *Journal of the American Statistical Association,* 72, 107–111.

Renkin, E. M. (1959), "Transport of Potassium-42 From Blood Tissue in Isolated Mammalian Skeletal Muscles," *American Journal of Physiology,* 197, 1205.

Rogers, W. H., and Tukey, J. W. (1962), "Understanding Some Long-Tailed Distributions," *Statistica Neerlandica,* 26, 211–226.

Rubin, D. B. (1983), "Iteratively Reweighted Least Squares," in *Encyclopedia of the Statistical Sciences* (Vol. 4), New York: John Wiley, pp. 272–275.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association,* 75, 828–838.

SAS Institute Inc. (1982), *SAS User's Guide: Statistics,* Cary, NC: Author.

Schelbert, H. R., Phelps, M. E., Huang, S.-E., MacDonald, N. S., Hansen, H., Selin, C., and Kuhl, D. E. (1981), "N-13 Ammonia as an Indicator of Myocardial Blood Flow," *Circulation,* 63, 1259–1272.

Shapiro, A., and Browne, M. W. (1987), "Covariance Structures Under Elliptical Distributions," *Journal of the American Statistical Association,* 82, 1092–1097.

SPSS (1983), *SPSSX User's Guide,* New York: McGraw-Hill.

Su, H.-L. (1988), "Estimation of Standard Errors in Some Multivariate Models When Some Observations Are Missing," unpublished Ph.D. dissertation, University of California, Los Angeles, Div. of Biostatistics.

Sutradhar, B. C., and Ali, M. M. (1986), "Estimation of Parameters of a Regression Model With a Multivariate t Error Variable," *Communications in Statistics—Theory and Methods,* 15, 429–450.

Szatrowski, T. H. (1980a), "Necessary and Sufficient Conditions for Explicit Solutions in the Multivariate Normal Estimation Problem for Patterned Means and Covariances," *The Annals of Statistics,* 8, 802–810.

——— (1980b), "Explicit Maximum Likelihood Estimates From Balanced Data in the Mixed Model of the Analysis of Variance," *The Annals of Statistics,* 8, 811–819.

Taylor, J. M. G. (1989), "The Inflation Variance Due to Modeling the Error Distribution With an Extra Shape Parameter," technical report, University of California, Los Angeles, School of Public Health.

Tiede, J. J., and Pagano, M. (1979), "The Application of Robust Calibration to Radioimmunoassay," *Biometrics,* 35, 567–574.

Tukey, J. W. (1949), "One Degree of Freedom for Non-Additivity," *Biometrics,* 5, 232–242.

Tyler, D. E. (1983), "Robustness and Efficiency Properties of Scatter Matrices," *Biometrika,* 70, 411–420.

West, M. (1984), "Outlier Models and Prior Distributions in Bayesian Linear Regression," *Journal of the Royal Statistical Society,* Ser. B, 46, 431–439.

Wu, C. F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *The Annals of Statistics,* 14, 1261–1295.

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

Zellner, A. (1976), "Bayesian and Non-Bayesian Analysis of the Regression Model With Multivariate Student-t Error Terms," *Journal of the American Statistical Association,* 71, 400–405.