Outlier Robust Nonlinear Mixed Model Estimation

[1]James D. Williams, [2]Jeffrey B. Birch and [3]Abdel-Salam G. Abdel-Salam

[1]*Business Analytics, Dow AgroSciences, 9330 Zionsville Rd. Indianapolis, IN 46268, USA.*
[2]*Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0439, USA.*
[3] *Department of Mathematics, Statistics and Physics, College of Arts and Sciences, Qatar University, Qatar.*
[3] *Department of Statistics, Faculty of Economics and Political Science, Cairo University, Egypt.*

Abstract

In standard analyses of data well-modeled by a nonlinear mixed model (NLMM), an aberrant observation, either within a cluster, or an entire cluster itself, can greatly distort parameter estimates and subsequent standard errors. Consequently, inferences about the parameters are misleading. This paper proposes an outlier robust method based on linearization to estimate fixed effects parameters and variance components in the NLMM. An example is given using the 4-parameter logistic model and bioassay data, comparing the robust parameter estimates to the nonrobust estimates given by SAS$^{\circledR}$.

Keywords: Dose-response; Linearization; Robust Estimation, m-estimation

1 Introduction

Nonlinear mixed-effects models are very useful to analyze repeated measures datasets and are used in variety of applications. Dealing with the longitudinal data is an essential issue in agricultural, environmental, medical and biological applications. Mixed-effects models in general are efficient and flexible

to model the within-subject correlation often present in this type of applications. Several methodologies have already been introduced in the frame work of linear and nonlinear mixed models to analyze such data ([1] and [2]).

Most researchers find that in the presence of the outliers, the Gaussian Quasi-Maximum Likelihood estimators are very inaccurate and that a robust procedure is needed. Mancini et.al. [3] and Muler and Yohia [4] proposed a robust M-estimator that assigns a much lower weight to the outliers than the Gaussian Maximum Likelihood estimators does. Pinheiro et al. [5] and Staudenmayer et al. [6] introduced robust estimation techniques in which both random effects and errors have multivariate Student-t distributions, while Rosa et al. [7] discussed Markov chain Monte Carlo (MCMC) implementations considering a Bayesian formulation. Nevertheless, few alternatives have been studied for outlier accommodation in the context of nonlinear mixed-effects models.

To date, Yeap and Davidian [8] is one of the very few references that address this case. They proposed a two stage approach for robust estimation in nonlinear mixed effects when outliers are present within and between individuals. Meza et. al. [9] presented an extension of a Gaussian nonlinear mixed-effects model considering a class of heavy-tailed multivariate distributions for both random effects and residual errors. We introduce a one step approach by utilizing a robust version of the linearized Gaussian likelihood for the nonlinear mixed model, extending the linear mixed model case discussed in Gill [10]. The details are provided in section 2. One advantage of our proposed method is that it can be easily computed using standard statistical and mathematical functions found in computing software such as SAS/IML® or R.

Nonlinear mixed model (NLMM) is given by

$$Y_{ij} = f(\mathbf{x}'_{ij}, \boldsymbol{\beta}, \mathbf{u}_i) + \epsilon_{ij} \tag{1}$$

where $i = 1, ..., s$ and $j = 1, ..., n_i$. In this model, Y_{ij} is the jth response (or measurement) of the ith subject or cluster, s is the number of clusters, n_i is the number of responses in cluster i, $\mathbf{x}'_{ij}$ is the regressor vector for the jth response of subject i, $\boldsymbol{\beta}$ is the $b \times 1$ vector of fixed effects parameters, and $\mathbf{u}_i$ is the vector of random effects for subject i, $\mathbf{u}_i = (u_{i1}, u_{i2}, ..., u_{iq})$. Here, q_i is the number of random effects in cluster i (usually the same for all i), and it is assumed that $\mathbf{u}_i \sim N_{q_i}(\mathbf{0}, \mathbf{G}_i)$. The variance-covariance matrix $\mathbf{G}_i$ is a matrix whose elements are the variance and covariance components

of the random effects, which we denote in vector notation as $\boldsymbol{\theta}_G$. A more complete notation would be $\mathbf{G}_i(\boldsymbol{\theta}_G)$, but for parsimony we simply write $\mathbf{G}_i$. It is usually assumed that $\mathbf{G}_i$ is the same for all i. If the observations within a cluster are stacked, we have the following model

$$\mathbf{Y}_i = \mathbf{f}(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{u}_i) + \boldsymbol{\epsilon}_i \tag{2}$$

where

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}'_{i1} \\ \mathbf{x}'_{i2} \\ \vdots \\ \mathbf{x}'_{in_i} \end{bmatrix}.$$

and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{in_i})'$. Assume that $\boldsymbol{\epsilon}_i \sim N_{n_i}(\mathbf{0}, \mathbf{R}_i)$ are independent for $i = 1, \ldots, s$, where $\mathbf{R}_i$ is the matrix whose elements are the variances and covariances of the error vector conditional on the random effects. Denote the variances and covariances in vector notation as $\boldsymbol{\theta}_R$. If we stack the data for each cluster, then we have the model

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}, \boldsymbol{\beta}, \mathbf{u}) + \boldsymbol{\epsilon} \tag{3}$$

where

$$\mathbf{f}(\mathbf{X}, \boldsymbol{\beta}, \mathbf{u}) = \begin{bmatrix} \mathbf{f}(\mathbf{X}_1, \boldsymbol{\beta}, \mathbf{u}_1) \\ \mathbf{f}(\mathbf{X}_2, \boldsymbol{\beta}, \mathbf{u}_2) \\ \vdots \\ \mathbf{f}(\mathbf{X}_s, \boldsymbol{\beta}, \mathbf{u}_s) \end{bmatrix}.$$

Let $\mathbf{R} = blockdiag(\mathbf{R}_i)$ and $\mathbf{G} = blockdiag(\mathbf{G}_i)$. Then $\boldsymbol{\epsilon} \sim N_N(\mathbf{0}, \mathbf{R})$ and $\mathbf{u} \sim N_q(\mathbf{0}, \mathbf{G})$, where $N = \sum n_i$ and $q = \sum q_i$.

The nonlinear mixed-effects model permits two types of correlation for the measurements within a cluster. The first source of correlation within each cluster comes from the random effects, while the second source of correlation comes from the within-cluster variance-covariance matrix of errors R.

Different parametric methods can be utilized to estimate the parameters in (3), but in this work we concentrate only on the maximum likelihood (ML) and the restricted maximum likelihood ($REML$) methods. For an introduction to some of the advantages and disadvantages of ML and $REML$ compared to the other commonly used estimation methods, see Harville [11].

The ML and $REML$ estimators of parameters of (3) are dependent on the marginal density of Y_i, but since there is no closed form solution to (2),

even if the random errors and random effects are assumed to be Gaussian distributions, numerical methods are used to approximate the integral [12]. The two most common methods are the linearization and the integral approximation methods, although these are not the only possible methods. Another method is the two stage approach, also known as the individual estimates method [1].

The aim of this paper is to propose a new estimation technique which is robust to outliers. This paper is organized as follows. In Section 2 we discuss robust estimation methods for the nonlinear mixed models. In Section 3 we present inference about the parameters in the nonlinear mixed effects models. A real application is given as example for the proposed technique in Section 4. Section 5, includes simulation results for the proposed technique comparing to the classical methods. Finally, some conclusions are given and possible future work is discussed in Section 6.

2 Robust Estimation of the NLMM

2.1 Linearize the NLMM

There are three ways in which to estimate parameters in the NLMM. One method, described by Davidian and Giltinan [13] is known as two-stage hierarchical modeling, where within cluster variance component estimates are obtained, whereupon between cluster estimates of the fixed effects are obtained. This method relies heavily on the assumption that there are sufficient within cluster observations to adequately estimate within cluster variability. A second method to estimating NLMM parameters is through integral approximation. See Schabenberger and Pierce [14] for details. This method is the most exact method, but can be computationally unfeasible. An alternative method is linearization.

Applying the first-order multivariate Taylor series approximation to $\mathbf{f}$ gives

$$\mathbf{f}(\mathbf{X}, \boldsymbol{\beta}, \mathbf{u}) \approx \mathbf{f}(\mathbf{X}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}) + \mathbf{D}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{Z}(\mathbf{u} - \hat{\mathbf{u}}) \qquad (4)$$

where

$$\mathbf{D} = \frac{\partial \mathbf{f}(\mathbf{X}, \boldsymbol{\beta}, \mathbf{u})}{\partial \boldsymbol{\beta}'}$$

and

$$\mathbf{Z} = \frac{\partial \mathbf{f}(\mathbf{X}, \boldsymbol{\beta}, \mathbf{u})}{\partial \mathbf{u}'}$$

evaluated at $(\boldsymbol{\beta}, \mathbf{u}) = (\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}})$. Substituting (4) into (1) gives

$$\mathbf{Y} \approx \mathbf{f}(\mathbf{X}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}) + \mathbf{D}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{Z}(\mathbf{u} - \hat{\mathbf{u}}) + \boldsymbol{\epsilon}$$

$$\mathbf{Y} - \mathbf{f}(\mathbf{X}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}) + \mathbf{D}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}} \approx \mathbf{D}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

$$(5)$$

Or

$$\tilde{\mathbf{Y}} \approx \mathbf{D}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \tag{6}$$

where $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{f}(\mathbf{X}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}) + \mathbf{D}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}$, $\tilde{\mathbf{Y}} \overset{A}{\sim} N_N(\mathbf{D}\boldsymbol{\beta}, \mathbf{V})$, and $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. $\tilde{\mathbf{Y}}$ is known as the *pseudo response* and equation (5) is known as the *pseudo model* because the actual observations $\mathbf{Y}$ are replaced with the pseudo responses $\tilde{\mathbf{Y}}$ through the iterative estimation process. Since $\mathbf{V}$ is the variance-covariance matrix of the pseudo responses, it depends on both $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_R$. Denote $\boldsymbol{\theta} = (\boldsymbol{\theta}'_G, \boldsymbol{\theta}'_R)'$ as the $p \times 1$ vector of variance and covariance parameters to be estimated. To accentuate the dependency of $\mathbf{V}$ on $\boldsymbol{\theta}$, we will denote $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$.

2.2 Robust Maximum Likelihood Estimation

2.2.1 Robust Likelihood

The log-likelihood function of the pseudo data, $\ell(\boldsymbol{\beta}, \boldsymbol{\theta} | \tilde{\mathbf{y}})$, is given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta} | \tilde{\mathbf{y}}) = -\frac{1}{2} N \log(2\pi) - \frac{1}{2} \log |\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2} (\tilde{\mathbf{y}} - \mathbf{D}\boldsymbol{\beta})' \mathbf{V}(\boldsymbol{\theta})^{-1} (\tilde{\mathbf{y}} - \mathbf{D}\boldsymbol{\beta}). \tag{7}$$

Since $\mathbf{V}(\boldsymbol{\theta})$ is a symmetric positive definite matrix, $\mathbf{V}(\boldsymbol{\theta})^{-1}$ is symmetric positive definite, and we may write $\mathbf{V}(\boldsymbol{\theta})^{-1} = \mathbf{V}(\boldsymbol{\theta})^{-1/2} \mathbf{V}(\boldsymbol{\theta})^{-1/2}$ for the symmetric square root matrix $\mathbf{V}(\boldsymbol{\theta})^{-1/2}$. Further, define $\mathbf{r} = \mathbf{V}(\boldsymbol{\theta})^{-\frac{1}{2}} (\tilde{\mathbf{y}} - \mathbf{D}\boldsymbol{\beta})$ as the *standardized pseudo-residual vector*. Using $\mathbf{r}$, rewrite the likelihood function (7) as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta} | \tilde{\mathbf{y}}) = -\frac{1}{2} N \log(2\pi) - \frac{1}{2} \log |\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2} \mathbf{r}' \mathbf{r}. \tag{8}$$

Note that $\mathbf{r}' \mathbf{r} = \sum r_i^2$, where r_i is the ith element of $\mathbf{r}$.

Following the m-estimation theory expressed in Huber and Ronchetti [15], the likelihood ℓ is robustified by replacing $\sum r_i^2$ with $\sum \rho(r_i)$, where $\rho(r_i)$ is a function of r_i that increases much slower than r_i^2 (see [16]) . For example, a
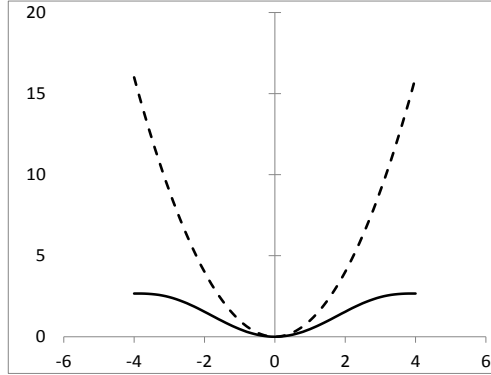
Figure 1: The Tukey Bisquare ρ Function with $C_t = 4$ (solid) vs. squared errors (dashed)

ρ function suggested by Andrews, et al. [17], called the *Bisquare* ρ function, is

$$\rho_{C_t}(r_i) = \begin{cases} -\frac{(C_t)^2}{6}\left((1 - (\frac{r_i}{C_t})^2)^3 - 1\right) & |r_i| \le C_t \\ \frac{(C_t)^2}{6} & \text{otherwise} \end{cases},$$

where C_t is known as the *tuning constant*. For an example of this function, see Figure 1.

The overall effect of replacing $\sum r_i^2$ with $\sum \rho(r_i)$ in the likelihood function is to limit or "down-weight" the influence that any one data point can exert on the parameter estimates.

The first derivative of $\rho_{C_t}(r_i)$ with respect to r_i, denoted $\psi_{C_t}(r_i) = \frac{\partial \rho_{C_t}}{\partial r_i}$, is given by

$$\psi_{C_t}(r_i) = \begin{cases} r_i(1 - (\frac{r_i}{C_t})^2)^2 & |r_i| \le C_t \\ 0 & \text{otherwise} \end{cases}.$$

An alternative ρ function is the one suggested by Huber [18], called the *Huber* ρ function, defined by

$$\rho_H(r_i) = \begin{cases} \frac{1}{2}r_i^2 & |r_i| \le C_h \\ C_h|r_i| - \frac{1}{2}(C_h)^2 & \text{otherwise} \end{cases},$$

where C_h is the Huber tuning constant. The first derivative of $\rho_H(r_i)$ with

respect to r_i, denoted $\psi_H(r_i) = \frac{\partial \rho_H}{\partial r_i}$, is given by

$$\psi_H(r_i) = \left\{ \begin{array}{cc} r_i & |r_i| \leq C_h \\ sign(r_i)\, C_h & \text{otherwise} \end{array} \right. .$$

Subsequent derivations are made assuming the Huber ρ function.

Replacing $\mathbf{r}'\mathbf{r}$ with $\sum_{i=1}^{N} \rho(r_i)$ in the likelihood function (8) gives the robustified version of the likelihood, m, given by

$$m(\boldsymbol{\beta}, \boldsymbol{\theta}|\tilde{\mathbf{y}}) = -\frac{k_1}{2} N \log(2\pi) - \frac{k_1}{2} \log |\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2} \sum_{i=1}^{N} \rho(r_i) \qquad (9)$$

where $k_1 = E(\psi(r_i) r_i)$ and the expectation is taken over the approximate distribution of r_i, $N(0,1)$. The constant k_1 is sometimes called the *consistency correction factor*, and is needed in order to make the estimating equations have zero expectation. Note that as the tuning constant approaches infinity, the robustified likelihood approaches the original likelihood.

2.2.2 Solution for β

To find the solution for $\boldsymbol{\beta}$, we take first derivatives of m with respect to $\boldsymbol{\beta}$ and set the result equal to $\mathbf{0}$, resulting in the estimating equations given by

$$\frac{\partial m}{\partial \boldsymbol{\beta}} = \mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1/2} \boldsymbol{\psi}(\mathbf{r}) \equiv \mathbf{0}$$

where $\boldsymbol{\psi}(\mathbf{r}) = (\psi(r_1), \psi(r_2), \ldots, \psi(r_N))$, and $\psi(r_i) = \frac{\partial \rho(r_i)}{\partial r_i}$.

Next, we take second derivatives in order to employ the multivariate Newton-Raphson optimization method with Fisher Scoring. The second derivative of m with respect to $\boldsymbol{\beta}$ is given by

$$
\begin{aligned}
\mathbf{H}_{\beta\beta'} = \frac{\partial^2 m}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= \frac{\partial}{\partial \boldsymbol{\beta}'} \left[\mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1/2} \boldsymbol{\psi}(\mathbf{r}) \right] \\
&= -\mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1/2} \boldsymbol{\Psi}(\mathbf{r}) \mathbf{V}(\boldsymbol{\theta})^{-1/2} \mathbf{D} + \frac{\partial \mathbf{D}'}{\partial \boldsymbol{\beta}'} \mathbf{V}(\boldsymbol{\theta})^{-1/2} \boldsymbol{\psi}(\mathbf{r}) \quad (10)
\end{aligned}
$$

where $\boldsymbol{\Psi}(\mathbf{r}) = \langle \frac{\partial \psi(r_i)}{\partial r_i} \rangle = \langle \frac{\partial^2 \rho(r_i)}{\partial r_i^2} \rangle$. Now, the approximate Fisher Information matrix is the expected Hessian matrix, given by

$$
\begin{aligned}
\mathbf{I}_{\beta\beta'} = E(-\mathbf{H}_{\beta\beta'}) &= \mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1/2} E(\boldsymbol{\Psi}(\mathbf{r})) \mathbf{V}(\boldsymbol{\theta})^{-1/2} \mathbf{D} \qquad (11) \\
&= k_1 \mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{D} \qquad (12)
\end{aligned}
$$

since $E\left(\boldsymbol{\psi}(\mathbf{r})\right) \approx 0$.

The solution for $\boldsymbol{\beta}$ is found iteratively as

$$\boldsymbol{\beta}^{(h+1)} = \boldsymbol{\beta}^{(h)} + \left[\mathbf{I}_{\beta\beta'}^{(h)}\right]^{-1}\frac{\partial m^{(h)}}{\partial\boldsymbol{\beta}}. \tag{13}$$

2.2.3 Solution for $\boldsymbol{\theta}$

We calculate first derivatives of m, $m(\boldsymbol{\beta}, \boldsymbol{\theta}|\tilde{\mathbf{y}})$, with respect to θ_i to obtain the estimating equations for θ_i, given by

$$
\begin{aligned}
\frac{\partial m}{\partial\theta_i} &= -\frac{k_1}{2}tr\left[\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\right] - \boldsymbol{\psi}(\mathbf{r})'\frac{\partial\mathbf{V}(\boldsymbol{\theta})^{-1/2}}{\partial\theta_i}(\tilde{\mathbf{y}} - \mathbf{D}\boldsymbol{\beta}) \\
&= -\frac{k_1}{2}tr\left[\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\right] + \frac{1}{2}\boldsymbol{\psi}(\mathbf{r})'\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{r} = \mathbf{0}.
\end{aligned}
$$

(see appendix of Richardson and Welsh [19])

Element-wise second derivatives of m with respect to θ_j are given by,

$$
\begin{aligned}
\frac{\partial^2 m}{\partial\theta_j\partial\theta_i} &= \frac{\partial}{\partial\theta_j}\left[\frac{\partial m}{\partial\theta_i}\right] \\
&= \frac{\partial}{\partial\theta_j}\left[-\frac{k_1}{2}tr\left(\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\right) + \frac{1}{2}\boldsymbol{\psi}(\mathbf{r})'\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{r}\right] \\
&= \frac{k_1}{2}tr\left[\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\right] \\
&\quad - \frac{1}{4}\boldsymbol{\psi}'\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{r} \\
&\quad - \frac{1}{2}\mathbf{r}'\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{V}(\boldsymbol{\theta})^{-1}\boldsymbol{\psi} \\
&\quad - \frac{1}{4}\mathbf{r}'\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}\mathbf{V}(\boldsymbol{\theta})^{-1}\boldsymbol{\Psi}(\mathbf{r})\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{r}.
\end{aligned}
$$

Note that $\dfrac{\partial^2\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_i} = \mathbf{0}$ assuming the Huber ρ as noted above.

Now we collect first derivatives into a vector,

$$
\frac{\partial m}{\partial\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial m}{\partial\theta_1} \\ \frac{\partial m}{\partial\theta_2} \\ \vdots \\ \frac{\partial m}{\partial\theta_p} \end{bmatrix}
$$

8

and collect second derivatives into a matrix, $\mathbf{H}_{\theta\theta'}$, as

$$\mathbf{H}_{\theta\theta'} = \frac{\partial^2 m}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2 m}{\partial\theta_1^2} & \frac{\partial^2 m}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 m}{\partial\theta_1\partial\theta_p} \\ \frac{\partial^2 m}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 m}{\partial\theta_2^2} & \cdots & \frac{\partial^2 m}{\partial\theta_2\partial\theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 m}{\partial\theta_p\partial\theta_1} & \frac{\partial^2 m}{\partial\theta_p\partial\theta_2} & \cdots & \frac{\partial^2 m}{\partial\theta_p^2} \end{bmatrix}.$$

The approximate expected Hessian matrix for $\boldsymbol{\theta}$, denoted $\mathbf{I}_{\theta\theta'}$, has ij^{th} element, $\mathbf{I}_{\theta\theta'} = \{I_{ij}\}$, is given by

$$I_{ij} = E\left\{-\frac{\partial^2 m}{\partial\theta_i\partial\theta_j}\right\} = \frac{k_1}{2}tr\left[\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_j}\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\right]. \tag{14}$$

(See Appendix A for details.) We can find the iterative Fisher Scoring estimates of $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta}^{(h+1)} = \boldsymbol{\theta}^{(h)} + \left[\mathbf{I}_{\theta\theta'}^{(h)}\right]^{-1}\frac{\partial m^{(h)}}{\partial\boldsymbol{\theta}}. \tag{15}$$

2.2.4 Estimation of Both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$

It is important to note that the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be done separately, since $E\left[\dfrac{\partial^2 m}{\partial\theta_i\partial\boldsymbol{\beta}}\right] = \mathbf{0}$. (See Appendix A for details.) Therefore, the estimation algorithm is less complicated. At the $h+1$ iteration of the Newton-Raphson algorithm with Fisher Scoring,

$$\boldsymbol{\beta}^{(h+1)} = \boldsymbol{\beta}^{(h)} + \left[\mathbf{H}_{\beta\beta'}^{(h)}\right]^{-1}\frac{\partial m^{(h)}}{\partial\boldsymbol{\beta}}.$$

After $\boldsymbol{\beta}^{(h+1)}$ is obtained $\boldsymbol{\theta}^{(h+1)}$ is calculated using the updated parameter vector $\boldsymbol{\beta}^{(h+1)}$ and the vector $\boldsymbol{\theta}^{(h)}$ from the previous iteration. So,

$$\boldsymbol{\theta}^{(h+1)} = \boldsymbol{\theta}^{(h)} + \left[\mathbf{H}_{\theta\theta'}^{(h)}\right]^{-1}\frac{\partial m^{(h)}(\boldsymbol{\theta}^{(h+1)},\boldsymbol{\theta}^{(h)})}{\partial\boldsymbol{\theta}}.$$

3 Inference for $\boldsymbol{\beta}$

3.1 Wald Test

Now that we have obtained robust estimates we can test them using a Wald-type test. The general form of the Wald test for testing the hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ verses the alternative hypothesis $H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}$ is

$$W = [\mathbf{C}\boldsymbol{\beta} - \mathbf{d}]' \left[\mathbf{C} \, Var(\hat{\boldsymbol{\beta}})^{-1}\mathbf{C}'\right]^{-1} [\mathbf{C}\boldsymbol{\beta} - \mathbf{d}].$$

The asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is normal with mean of $\boldsymbol{\beta}$ and covariance matrix

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \frac{E\left[\psi^2(r_i)\right]}{\left[E\left[\psi'(r_i)\right]\right]^2}\mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{D}, \tag{16}$$

so W has an asymptotic χ^2 distribution with degrees of freedom equal to $rank(\mathbf{C})$. Similarly, dividing W by $rank(\mathbf{C})$ gives

$$F = \frac{W}{rank(\mathbf{C})}$$

which has an asymptotic F-distribution with $rank(\mathbf{C})$ numerator degrees of freedom and denominator degrees of freedom equal to the number of observations minus the number of parameters being estimated.

3.2 T-tests on Individual Fixed-Effects Parameters

To test a hypothesis of the form $H_0 : \beta_j = d$ verses the alternative hypothesis $H_1 : \beta_j \neq d$, the Wald-type test above takes the form

$$F = (\hat{\beta}_j - d)^2 [\hat{Var}\,(\hat{\beta}_j)]^{-1},$$

where $\hat{Var}\,(\hat{\beta}_j)$ is the jth diagonal element of $Var(\hat{\boldsymbol{\beta}})$. The F statistic has an asymptotic F distribution with 1 numerator degree of freedom and denominator degrees of freedom equal to the number of observations minus the number of parameters being estimated. The associated t statistic is given by

$$t = sign(\hat{\boldsymbol{\beta}})\sqrt{F},$$

which has an asymptotic t distribution with degrees of freedom equal to the number of observations minus the number of parameters being estimated.
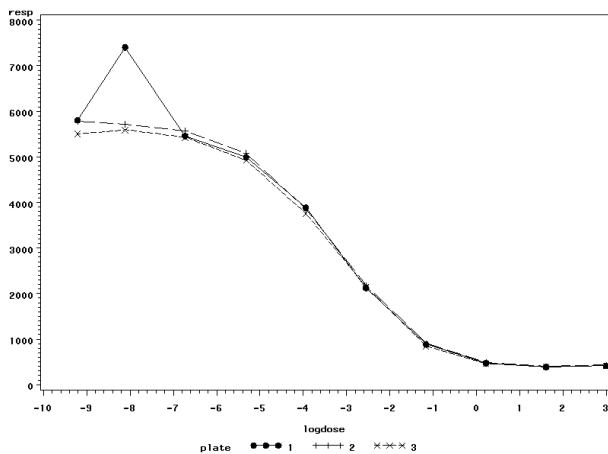
Figure 2: The Dose-Response (BDR) data

4 Example

We illustrate the proposed methods using a real bioassay dose-response data set (BDR) from a biopharmaceutical company, not disclosed here for confidentiality reasons. The data is given in Table 1 in Appendix B. The data consist of ten doses of a drug administered to ten subjects in each of three independent plates (clusters) chosen at random. A plot of the response against the logarithm of the dose for each of the three plates is given in Figure 2.

It is obvious that there is an outlier in the first plate, second dose. A robust estimation procedure is needed.

4.1 Standard Analysis of the BDR Data

The 4-Parameter Logistic model with one random effect is a common model used to represent this data of this type. This model has the form

$$y_{ij} = A + \frac{D - A}{1 + \left(\frac{x_{ij}}{C}\right)^B} + \epsilon_{ij}, \tag{17}$$

where y_{ij} is the jth measured response of the subject exposed to dose x_{ij}, A is the upper asymptote parameter, D is the lower asymptote parameter, C is the ED_{50} parameter (the dose required to elicit 50% response), and B is the rate parameter. A plot of this function is given in Figure 3.
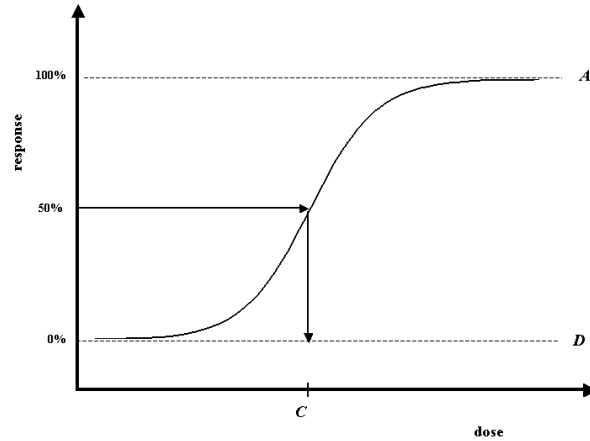
11

Figure 3: The 4-Parameter Logistic Function

In order to account for the plate-to-plate variability, a random effect for plate is added to the model. For the BDR data, the most appropriate random effect to add to the model is the a random effect added to the upper asymptote, A, as

$$y_{ij} = (A + a_i) + \frac{D - (A + a_i)}{1 + \left(\frac{x_{ij}}{C}\right)^B} + \epsilon_{ij}. \tag{18}$$

See [20] for an example of using this model. Since the A, B, C, and D parameters are fixed effects and the a parameter is a random effect, model (18) is a nonlinear mixed model. The variance model contains one variance component and independent errors with containt variance σ^2. Using this model and the BDR data the following parameter estimates were obtained, as displayed in Table 1.

Note that the estimate of the A parameter is much higher than expected, because of the undue influence of the outlier in the upper asymptote region. Figure 4 gives a plot of the estimated population average (PA) curve given by SAS® .

The PA curve systematically overestimates the mean response in the upper asymptote. Such bias can greatly distort subsequent inferences about the quality of the drug and its physical properties.

Table 1: SAS PROC NLMIXED Output for the BDR Data

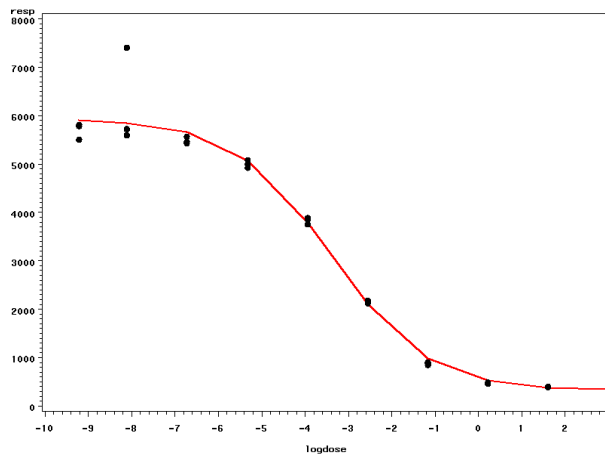| Parameter	Estimate	Standard Error
A	5933.61	144.78
B	-0.8953	0.07751
C	0.03295	0.003551
D	321.10	95.7084
σ_a^2	29796	31299
σ^2	60470	12377

Figure 4: PROC NLMIXED Population Average Curve Fit to Dose-Response Data

4.2   Robust Analysis of the BDR Data

Using the Robust 4-Parameter Logistic Model with the Huber ρ function, robust parameter estimates were obtained. Both robust and SAS® estimates are shown in Table 2 for comparison.

Notice that the fixed effects parameter estimates and standard errors change dramatically when the robust estimation is applied. A plot of the PA curves for both the robust procedure and for SAS® is given in Figure 5.

The robust parameter estimates do not systematic over-estimate the PA near the upper asymptote. It is clear that the outlier has not influenced the estimates of the robust model as greatly as the standard estimates.

13

Table 2: Comparison of SAS PROC NLMIXED and Robust NLMM Output for the BDR Data

	Robust		PROC NLMIXED	
Parameter	Estimate	Standard Error	Estimate	Standard Error
A	5696.72	34.459	5933.61	144.78
B	-0.9612	0.0285	-0.8953	0.07751
C	0.0368	0.0012	0.03295	0.003551
D	342.43	30.6746	321.10	95.7084
σ_a^2	634.789	1840.14	29,796	31,299
σ^2	6456.56	1776.838	60,470	12,377

5 Simulation Results

A Monte Carlo simulation study was conducted to examine the asymptotic distribution of the robust fixed effects parameters and their associated standard errors when an outlier is present. A simulation with 1000 data sets was generated using the mean-shift outlier model applied to model (18) with three independent clusters. The outlier was present at the second dose of one of the plates (clusters) for each data set, with an average of 1200 units greater than the PA curve at that dose. The doses used are the same doses as in the BDR data set. Population parameters are as given in Table 3.

Table 3: Population Parameters for Simulation Study

A	B	C	D	σ_a^2	σ^2
5500	-0.95	0.04	340	600	6500

The t statistic based on the robust estimation to test the null hypothesis $H_0 : A = 5500$ verses the alternative hypothesis $H_a : A \neq 5500$, is given by

$$T_{test} = \frac{\hat{A}_{robust} - 5500}{ese(\hat{A}_{robust})}$$

where $ese(\hat{A}_{robust})$ is the estimated standard error of $\hat{A}_{robust}$. This T statistic has an asymptotic t distribution with 24 degrees of freedom (30 observations
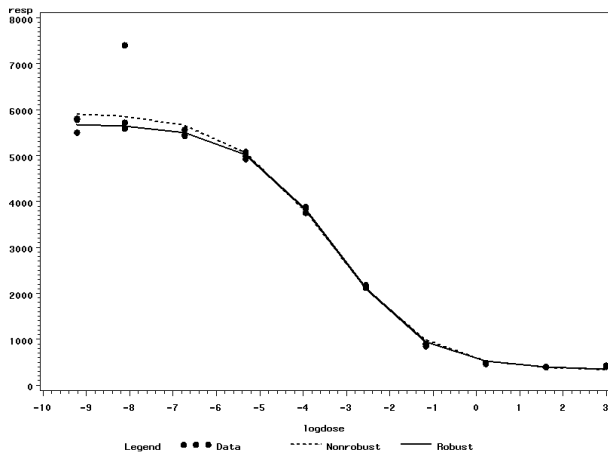
Figure 5: Comparison of PROC NLMIXED Fit and Robust Fit to Dose-Response Data

minus 6 parameters estimated). Similarly, the T statistics for the B, C, and D parameters were formulated. The 1000 T statistics for each of the 4 fixed effects parameters were plotted against the $t(24)$ distribution in a Q-Q plot, given in Figures 6 and 7, in Appendix C. Note that the T statistics for the four fixed effects parameters follow a noncentral t distribution. The reason for this is that the imposed outlier shifts the population mean slightly, but the robust estimation of the fixed effects is influenced little. By comparison, the same 1000 datasets were analyzed using SAS®, and the 1000 T statistics were plotted against the $t(24)$ distribution in a Q-Q plot, in Figures 8 and 9, in Appendix C. Note that the non-robust estimates deviate substantially from the t distribution. The outlying observation distorted the statistical inferences for the non-robust estimation procedure.

6   Discussion and Conclusion

Nonlinear mixed models are very useful statistical tools in analyzing nonlinear data with random effects. However, when outliers are present in the data, current non-robust estimation methods may produce erroneous or misleading results. A robust procedure is needed.

   A robust nonlinear mixed model is proposed by first using a Taylor series to approximate the nonlinear mixed model. The log likelihood is altered by

15

incorporating an appropriate loss function, ρ, of the residuals. Robust estimates of the parameters for the means model and for the variance components can then be obtain via an iterative technique based on Fisher scoring. Asymptotic theory leads to using Wald Inference to testing hypotheses of interest.

The robust procedure is illustrated to be successful on a case study involving actual dose-response data. A short simulation experiment supports the procedure and shows the superiority of the robust nonlinear mixed models over the classical nonlinear mixed models.

Appendix A: Derivative Calculations

$$\frac{\partial \mathbf{r}}{\partial \theta_i} = \frac{\partial \mathbf{V}(\boldsymbol{\theta})^{-1/2}}{\partial \theta_i}(\tilde{\mathbf{y}} - \mathbf{D}\boldsymbol{\beta})$$

$$= -\frac{1}{2}\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{V}(\boldsymbol{\theta})^{-1/2}}{\partial \theta_i}\mathbf{V}(\boldsymbol{\theta})^{-1/2}(\tilde{\mathbf{y}} - \mathbf{D}\boldsymbol{\beta}) = -\frac{1}{2}\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{V}(\boldsymbol{\theta})^{-1/2}}{\partial \theta_i}\mathbf{r}$$

$$\frac{\partial \mathbf{V}(\boldsymbol{\theta})^{-1}}{\partial \theta_i} = -\mathbf{V}(\boldsymbol{\theta})\frac{\partial \mathbf{V}(\boldsymbol{\theta})}{\partial \theta_i}\mathbf{V}(\boldsymbol{\theta})^{-1}$$

$$\frac{\partial \boldsymbol{\psi}(\mathbf{r})}{\partial \theta_i} = \frac{\partial \boldsymbol{\psi}(\mathbf{r})}{\partial \mathbf{r}'}\frac{\partial \mathbf{r}}{\partial \theta_i} = \boldsymbol{\Psi}\frac{\partial \mathbf{r}}{\partial \theta_i} = -\frac{1}{2}\boldsymbol{\Psi}\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{V}}{\partial \theta_i}\mathbf{r}, \text{ where } \boldsymbol{\Psi} = \langle\frac{\partial \psi(r_i)}{\partial r_i}\rangle$$

$$\frac{\partial^2 m}{\partial \theta_i \partial \boldsymbol{\beta}} = \frac{\partial}{\partial \theta_i}\left[\mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1/2}\boldsymbol{\psi}(\mathbf{r})\right]$$

$$= -\frac{1}{2}\mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{V}}{\partial \theta_i}\mathbf{V}(\boldsymbol{\theta})^{-1/2}\boldsymbol{\psi} + \mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1/2}\boldsymbol{\Psi}(\mathbf{r})\left[-\frac{1}{2}\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{V}}{\partial \theta_i}\mathbf{r}\right]$$

$$E\left[\frac{\partial^2 m}{\partial \theta_i \partial \boldsymbol{\beta}}\right] = \mathbf{0} - \frac{1}{2}\mathbf{D}'\mathbf{V}(\boldsymbol{\theta})^{-1/2}E\left[\langle\frac{\partial \psi(r_i)}{\partial r_i}\rangle\mathbf{V}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{V}}{\partial \theta_i}\mathbf{r}\right] = \mathbf{0}, \text{ because}$$

$\psi(r_i)$ is an odd function (see class notes from Birch [21]) and we are assuming that the underlying distribution of the errors is symmetric about zero. Hence, $E\left[\frac{\partial \psi(r_i)}{\partial r_i}r_i\right] = 0$.

Appendix B: BDR Data

Appendix C: Simulation Plots

Dose	Plate 1	Plate 2	Plate 3
20	412.834	429.970	416.690
5	392.792	399.475	397.989
1.25	473.593	485.651	466.704
0.3125	886.381	904.203	850.612
0.0781	2127.067	2180.242	2148.495
0.0195	3887.744	3856.684	3756.269
0.0049	4993.997	5082.415	4925.234
0.0012	5459.608	5568.591	5430.150
0.0003	7407.658	5723.422	5593.813
0.0001	5808.212	5786.643	5503.879

References

[1] Demidenko, E. Mixed Models : Theory and Applications. John Wiley and Sons, New York, 2004.

[2] Hastie, T., Tibshirani, R., and Friedman, J.. The Elements of Statistical Learning: DataMining, Inference, and Prediction. Springer, New York, Inc, 2009.

[3] Mancini, L., Ronchetti, E. and F. Trojani. Optimal conditionally unbiased bounded-influence inference in dynamic location and scale models. *Journal of the American Statistical Association* 2005; **105**: 628-641.

[4] Muler, N. and V.J. Yohai. Robust estimates for GARCH models. *Journal of Statistical Planning and Inference* 2008; **138**: 2918-2940.

[5] Pinheiro, J., Liu, C., Wu, Y. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *J. Comput. Graph. Stat.* 2001; **10**: 249-276.

[6] Staudenmayer, J., Lake, E.E., Wand, M.P. Robustness for general design mixed models using the t-distribution. *Stat. Model.* 2009; **9**: 235-255.

[7] Rosa, G.J.M., Gianola, D., Padovani, C.R. Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. *J. Appl. Stat.* 2004; **31**: 855-873.
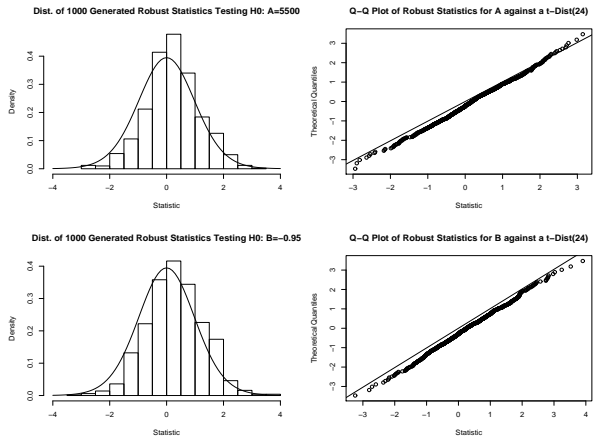
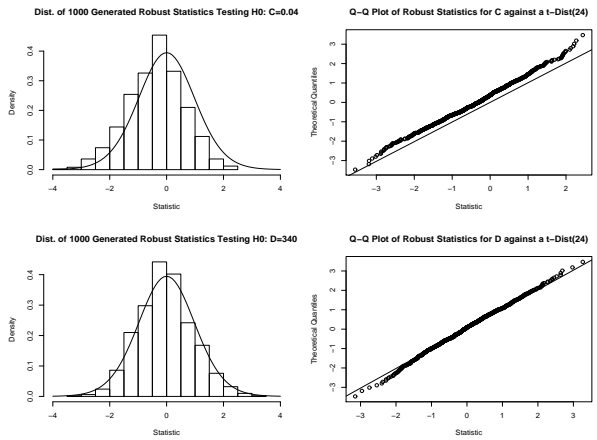Figure 6: Q-Q Plot of Robust t Statistics Verses $t(24)$ for the A and B Parameters

Figure 7: Q-Q Plot of Robust t Statistics Verses $t(24)$ for the C and D Parameters
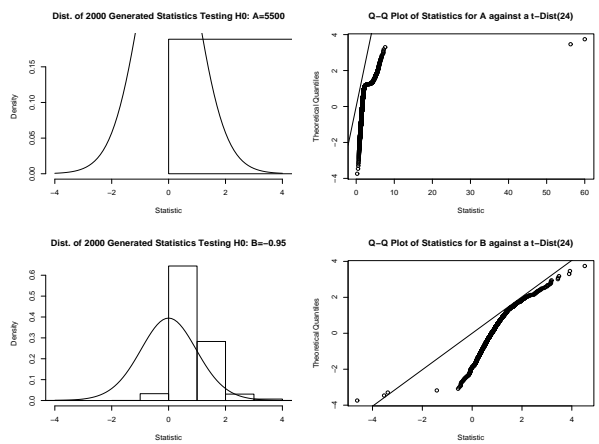
18

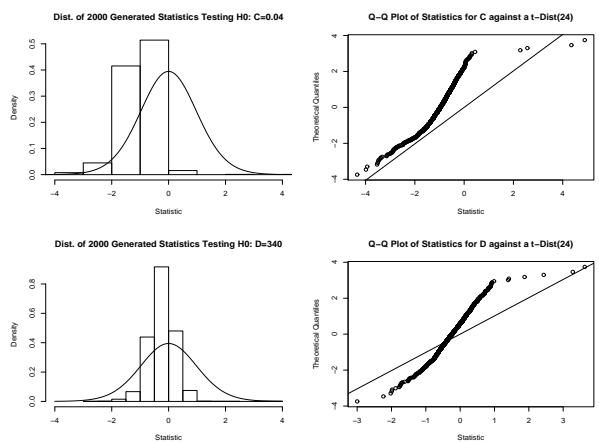Figure 8: Q-Q Plot of SAS t Statistics Verses $t(24)$ for the A and B Parameters

Figure 9: Q-Q Plot of SAS t Statistics Verses $t(24)$ for the C and D Parameters

19

[8] Yeap, B.Y., Davidian, M. Robust two-stage estimation in hierarchical nonlinear models. *Biometrics* 2001; **57**: 266-272.

[9] Meza, A., Osorio, F., and De la Cruz, R. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing* 2012; **22**: 121139.

[10] Gill, P.S. A Robust Mixed Linear Model Analysis For Longitudinal Data. *Statistics in Medicine* 2000;**19**: 975-987.

[11] Harville, D. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* 1977; **72**: 320-338.

[12] Pinheiro, J. C. and Bates, D. M. Mixed Effects Models in S and S-Plus. Springer, New York, Inc., 2000.

[13] Davidian, M. and Giltinan, D.M. Nonlinear Models for Repeated Measures Data. Chapman and Hall, New York, 1995.

[14] Schabenberger, O. and Pierce, F.J. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, New York, 2002.

[15] Huber, P.J. and Ronchetti, E. M. Robust Statistics (2nd edn). Wiley and Sons, Hoboken, New Jersey, 2009.

[16] Huggins, R.M. A Robust Approach to the Analysis of Repeated Measures. *Biometrics* 1993; **49**: 715-720.

[17] Andrew, D.R. P.J. Bickel, F.R. Hampel, P.J. Huber, W. H. Rogers, J. W. Tukey. *Robust Estimates of Location: Survey and Advances.* New Jersey: Princeton University Press, 1972.

[18] Huber, P.J. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics* 1964; **35**: 73-101.

[19] Richardson, A.M., and Welsh, A.H. Robust Restricted Maximum Likelihood in Mixed Linear Models. *Biometrics* 1995; **51**: 1429-1439.

[20] Williams, J.D., Birch, J.B., and Walfish, S. A Statistical Method to Account for Plate-to-Plate Variability in Multiple-Plate Bioassays. *BioPharm*, 2003; **16**(4): 44-54.

[21] Birch, J.B. Exploratory and Robust Data Analysis class, 2002. notes.