Outlier Robust Nonlinear Mixed Model Estimation

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

1Business Analytics, Dow AgroSciences, 9330 Zionsville Rd. Indianapolis, IN 46268, USA.
2Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0439, USA.
3Department of Mathematics, Statistics and Physics, College of Arts and Sciences, Qatar University, Qatar.
3Department 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®.

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 Stau- denmayer 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

\[ Y_{ij} = f(x'_{ij}, \beta, u_i) + \epsilon_{ij} \]  

where \( i = 1, ..., s \) and \( j = 1, ..., n_i \). In this model, \( Y_{ij} \) is the \( j^{th} \) response (or measurement) of the \( i^{th} \) subject or cluster, \( s \) is the number of clusters, \( n_i \) is the number of responses in cluster \( i \), \( x'_{ij} \) is the regressor vector for the \( j^{th} \) response of subject \( i \), \( \beta \) is the \( b \times 1 \) vector of fixed effects parameters, and \( u_i \) is the vector of random effects for subject \( i \), \( u_i = (u_{i1}, u_{i2}, ..., u_{iq}) \). Here, \( q_i \) is the number of random effects in cluster \( i \) (usually the same for all \( i \)), and it is assumed that \( u_i \sim N_{q_i}(0, G_i) \). The variance-covariance matrix \( G_i \) is a matrix whose elements are the variance and covariance components
of the random effects, which we denote in vector notation as $\theta_G$. A more complete notation would be $G_i(\theta_G)$, but for parsimony we simply write $G_i$. It is usually assumed that $G_i$ is the same for all $i$. If the observations within a cluster are stacked, we have the following model

$$Y_i = f(X_i, \beta, u_i) + \epsilon_i$$

where

$$X_i = \begin{bmatrix} x_{i1}' \\ x_{i2}' \\ \vdots \\ x_{in_i}' \end{bmatrix}.$$ 

and $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{in_i})'$. Assume that $\epsilon_i \sim N_{n_i}(0, R_i)$ are independent for $i = 1, \ldots, s$, where $R_i$ 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 $\theta_R$. If we stack the data for each cluster, then we have the model

$$Y = f(X, \beta, u) + \epsilon$$

where

$$f(X, \beta, u) = \begin{bmatrix} f(X_1, \beta, u_1) \\ f(X_2, \beta, u_2) \\ \vdots \\ f(X_s, \beta, u_s) \end{bmatrix}.$$ 

Let $R = \text{blockdiag}(R_i)$ and $G = \text{blockdiag}(G_i)$. Then $\epsilon \sim N_N(0, R)$ and $u \sim N_q(0, G)$, 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 approx-
imation 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 hi-
erarchical modeling, where within cluster variance component estimates are
obtained, whereupon between cluster estimates of the fixed effects are ob-
tained. 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 ap-
proximation. See Schabenberger and Pierce \[14\] for details. This method is
the most exact method, but can be computationally unfeasible. An alterna-
tive method is linearization.

Applying the first-order multivariate Taylor series approximation to \( f \) gives

\[
f(X, \beta, u) \approx f(X, \hat{\beta}, \hat{u}) + D(\beta - \hat{\beta}) + Z(u - \hat{u}) \tag{4}
\]

where

\[
D = \frac{\partial f(X, \beta, u)}{\partial \beta'}
\]

and

\[
Z = \frac{\partial f(X, \beta, u)}{\partial u'}
\]
evaluated at \((\beta, u) = (\hat{\beta}, \hat{u})\). Substituting (4) into (1) gives
\[
Y \approx f(X, \hat{\beta}, \hat{u}) + D(\beta - \hat{\beta}) + Z(u - \hat{u}) + \epsilon
\]
\[
Y - f(X, \hat{\beta}, \hat{u}) + D\hat{\beta} + Zu \approx D\beta + Zu + \epsilon
\]
(5)

Or
\[
\tilde{Y} \approx D\beta + Zu + \epsilon
\]
(6)

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

### 2.2 Robust Maximum Likelihood Estimation

#### 2.2.1 Robust Likelihood

The log-likelihood function of the pseudo data, \(\ell(\beta, \theta|\tilde{y})\), is given by
\[
\ell(\beta, \theta|\tilde{y}) = -\frac{1}{2}N \log(2\pi) - \frac{1}{2} \log |V(\theta)| - \frac{1}{2}(\tilde{y} - D_\beta)^{'}V(\theta)^{-1}(\tilde{y} - D\beta). \tag{7}
\]

Since \(V(\theta)\) is a symmetric positive definite matrix, \(V(\theta)^{-1}\) is symmetric positive definite, and we may write \(V(\theta)^{-1} = V(\theta)^{-1/2}V(\theta)^{-1/2}\) for the symmetric square root matrix \(V(\theta)^{-1/2}\). Further, define \(r = V(\theta)^{-\frac{1}{2}}(\tilde{y} - D\beta)\) as the standardized pseudo-residual vector. Using \(r\), rewrite the likelihood function (7) as
\[
\ell(\beta, \theta|\tilde{y}) = -\frac{1}{2}N \log(2\pi) - \frac{1}{2} \log |V(\theta)| - \frac{1}{2}r^{'}r. \tag{8}
\]

Note that \(r^{'}r = \sum r_{i}^{2}\), where \(r_{i}\) is the \(i^{th}\) element of \(r\).

Following the m-estimation theory expressed in Huber and Ronchetti [15], the likelihood \(\ell\) 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
The Tukey Bisquare $\rho$ Function with $C_t = 4$ (solid) vs. squared errors (dashed)

$\rho$ function suggested by Andrews, et al. [17], called the Bisquare $\rho$ function, is

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

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

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

An alternative $\rho$ function is the one suggested by Huber [18], called the Huber $\rho$ function, defined by

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

where $C_h$ is the Huber tuning constant. The first derivative of $\rho_H(r_i)$ with
respect to \( r_i \), denoted \( \psi_H(r_i) = \frac{\partial \rho}{\partial r_i} \), is given by

\[
\psi_H(r_i) = \begin{cases} 
    r_i & |r_i| \leq C_h \\
    \text{sign}(r_i) C_h & \text{otherwise}
\end{cases}
\]

Subsequent derivations are made assuming the Huber \( \rho \) function.

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

\[
m(\beta, \theta| \tilde{y}) = -\frac{k_1}{2} N \log(2\pi) - \frac{k_1}{2} \log |V(\theta)| - \frac{1}{2} \sum_{i=1}^{N} \rho(r_i)
\]

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.

\[2.2.2 \text{ Solution for } \beta\]

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

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

where \( \psi(r) = (\psi(r_1), \psi(r_2), \ldots, \psi(r_N)) \), 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 \( \beta \) is given by

\[
H_{\beta \beta'} = \frac{\partial^2 m}{\partial \beta \partial \beta'} = \frac{\partial}{\partial \beta'} [D'V(\theta)^{-1/2}\psi(r)]
\]

\[
= -D'V(\theta)^{-1/2}\Psi(r)V(\theta)^{-1/2}D + \frac{\partial D'}{\partial \beta'} V(\theta)^{-1/2}\psi(r)
\]

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

\[
I_{\beta \beta'} = E(-H_{\beta \beta'}) = D'V(\theta)^{-1/2}E(\Psi(r))V(\theta)^{-1/2}D
\]

\[
= k_1 D'V(\theta)^{-1}D
\]
since $E(\psi(r)) \approx 0$.

The solution for $\beta$ is found iteratively as

$$\beta^{(h+1)} = \beta^{(h)} + \left[I_{\beta\beta}^{(h)}\right]^{-1} \frac{\partial m^{(h)}}{\partial \beta}. \quad (13)$$

### 2.2.3 Solution for $\theta$

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

$$\frac{\partial m}{\partial \theta_i} = -\frac{k_1}{2} tr \left[V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} \right] - \psi(r) \frac{\partial V(\theta)^{-1/2}}{\partial \theta_i}(\tilde{y} - D\beta)$$

$$= -\frac{k_1}{2} tr \left[V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} \right] + \frac{1}{2} \psi(r)'V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} r = 0. \quad (see \ appendix \ of \ Richardson \ and \ Welsh \ [19])$$

Element-wise second derivatives of $m$ with respect to $\theta_j$ are given by,

$$\frac{\partial^2 m}{\partial \theta_j \partial \theta_i} = \frac{\partial}{\partial \theta_j} \left[ \frac{\partial m}{\partial \theta_i} \right]$$

$$= \frac{\partial}{\partial \theta_j} \left[ -\frac{k_1}{2} \left( \frac{V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i}}{\partial \theta_i} \right) + \frac{1}{2} \psi(r)'V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} r \right]$$

$$= \frac{k_1}{2} tr \left[V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_j} V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} \right]$$

$$- \frac{1}{4} \psi'V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_j} V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} r$$

$$- \frac{1}{2} r \frac{\partial V(\theta)}{\partial \theta_j} V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} V(\theta)^{-1} \psi$$

$$- \frac{1}{4} r \frac{\partial V(\theta)}{\partial \theta_j} V(\theta)^{-1} \Psi(r)V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} r. \quad (14)$$

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

Now we collect first derivatives into a vector,

$$\frac{\partial m}{\partial \theta} = \begin{bmatrix} \frac{\partial m}{\partial \theta_1} \\ \frac{\partial m}{\partial \theta_2} \\ \vdots \\ \frac{\partial m}{\partial \theta_p} \end{bmatrix}$$
and collect second derivatives into a matrix, $H_{\theta\theta'}$, as

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

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

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

(14)

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

$$\theta^{(h+1)} = \theta^{(h)} + \left[ I_{\theta\theta'}^{(h)} \right]^{-1} \frac{\partial m^{(h)}}{\partial \theta}.$$ 

(15)

2.2.4 Estimation of Both $\beta$ and $\theta$

It is important to note that the estimation of $\beta$ and $\theta$ can be done separately, since $E \left[ \frac{\partial^2 m}{\partial \theta_i \partial \beta} \right] = 0$. (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,

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

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

$$\theta^{(h+1)} = \theta^{(h)} + \left[ H_{\theta\theta'}^{(h)} \right]^{-1} \frac{\partial m^{(h)}}{\partial \theta} (\theta^{(h+1)}, \theta^{(h)}).$$
3 Inference for $\beta$

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 $H_0 : C\beta = d$ versus the alternative hypothesis $H_1 : C\beta \neq d$ is

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

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

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

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

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

which has an asymptotic $F$-distribution with $\text{rank}(C)$ 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 $H_0 : \beta_j = d$ versus the alternative hypothesis $H_1 : \beta_j \neq d$, the Wald-type test above takes the form

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

where $Var(\hat{\beta}_j)$ is the $j^{th}$ diagonal element of $Var(\hat{\beta})$. 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

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

which has an asymptotic $t$ distribution with degrees of freedom equal to the number of observations minus the number of parameters being estimated.
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 \( j^{th} \) 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.
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 random effect added to the upper asymptote, $A$, as

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

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 $\sigma^2$. 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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5933.61</td>
<td>144.78</td>
</tr>
<tr>
<td>B</td>
<td>-0.8953</td>
<td>0.07751</td>
</tr>
<tr>
<td>C</td>
<td>0.03295</td>
<td>0.003551</td>
</tr>
<tr>
<td>D</td>
<td>321.10</td>
<td>95.7084</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>29796</td>
<td>31299</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>60470</td>
<td>12377</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Robust Estimate</th>
<th>Robust Standard Error</th>
<th>PROC NLMIXED Estimate</th>
<th>PROC NLMIXED Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5696.72</td>
<td>34.459</td>
<td>5933.61</td>
<td>144.78</td>
</tr>
<tr>
<td>B</td>
<td>-0.9612</td>
<td>0.0285</td>
<td>-0.8953</td>
<td>0.07751</td>
</tr>
<tr>
<td>C</td>
<td>0.0368</td>
<td>0.0012</td>
<td>0.03295</td>
<td>0.003551</td>
</tr>
<tr>
<td>D</td>
<td>342.43</td>
<td>30.6746</td>
<td>321.10</td>
<td>95.7084</td>
</tr>
<tr>
<td>$\sigma_a^2$</td>
<td>634.789</td>
<td>1840.14</td>
<td>29,796</td>
<td>31,299</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>6456.56</td>
<td>1776.838</td>
<td>60,470</td>
<td>12,377</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>$\sigma_a^2$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5500</td>
<td>-0.95</td>
<td>0.04</td>
<td>340</td>
<td>600</td>
<td>6500</td>
</tr>
</tbody>
</table>

The $t$ statistic based on the robust estimation to test the null hypothesis $H_0 : A = 5500$ versus 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
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
incorporating an appropriate loss function, $\rho$, of the residuals. Robust estimates of the parameters for the means model and for the variance components can then be obtained 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

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

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

\[
\frac{\partial V(\theta)^{-1}}{\partial \theta_i} = -V(\theta) \frac{\partial V(\theta)}{\partial \theta_i} V(\theta)^{-1}
\]

\[
\frac{\partial \psi(r_i)}{\partial \theta_i} = \frac{\partial \psi(r)}{\partial r} \frac{\partial r}{\partial \theta_i} = \Psi \frac{\partial r}{\partial \theta_i} = -\frac{1}{2} \Psi V(\theta)^{-1} \nabla^2 V(\theta) \frac{\partial V(\theta)^{-1}}{\partial \theta_i} r, \text{ where } \Psi = \langle \frac{\partial \psi(r_i)}{\partial r} \rangle
\]

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

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

\[
E \left[ \frac{\partial^2 m}{\partial \theta_i \partial \beta} \right] = 0 - \frac{1}{2} D'V(\theta)^{-1/2} E \left[ \frac{\partial \psi(r_i)}{\partial r} V(\theta)^{-1} \frac{\partial V(\theta)}{\partial \theta_i} r \right] = 0, \text{ because } \psi(r_i) \text{ is an odd function (see class notes from Birch [21]) and we are assuming that the underlying distribution of the errors is symmetric about zero. Hence, } E \left[ \frac{\partial \psi(r_i)}{\partial r} r_i \right] = 0.
\]

Appendix B: BDR Data

Appendix C: Simulation Plots
<table>
<thead>
<tr>
<th>Dose</th>
<th>Plate 1</th>
<th>Plate 2</th>
<th>Plate 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>412.834</td>
<td>429.970</td>
<td>416.690</td>
</tr>
<tr>
<td>5</td>
<td>392.792</td>
<td>399.475</td>
<td>397.989</td>
</tr>
<tr>
<td>1.25</td>
<td>473.593</td>
<td>485.651</td>
<td>466.704</td>
</tr>
<tr>
<td>0.3125</td>
<td>886.381</td>
<td>904.203</td>
<td>850.612</td>
</tr>
<tr>
<td>0.0781</td>
<td>2127.067</td>
<td>2180.242</td>
<td>2148.495</td>
</tr>
<tr>
<td>0.0195</td>
<td>3887.744</td>
<td>3856.684</td>
<td>3756.269</td>
</tr>
<tr>
<td>0.0049</td>
<td>4993.997</td>
<td>5082.415</td>
<td>4925.234</td>
</tr>
<tr>
<td>0.0012</td>
<td>5459.608</td>
<td>5568.591</td>
<td>5430.150</td>
</tr>
<tr>
<td>0.0003</td>
<td>7407.658</td>
<td>5723.422</td>
<td>5593.813</td>
</tr>
<tr>
<td>0.0001</td>
<td>5808.212</td>
<td>5786.643</td>
<td>5503.879</td>
</tr>
</tbody>
</table>

References


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
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


