Chapter 6

Nonlinear Mixed Models

In our experience with profile monitoring, we have found it more likely that the profile is best described by a nonlinear function than by a linear function. Nonetheless, the majority of existing profile monitoring research deals with linear profile. Williams, Woodall, and Birch (2003) give a broad treatment of nonlinear profile monitoring where separate nonlinear (NL) regression models are fit to each profile. Williams et al. (2006a) gave an application of NL profile monitoring to dose-response data. Thus an important extension of the work shown in Chapters 4 and 5 is the use of a nonlinear mixed (NLM) model to account for profile data that is nonlinear in nature and has the two levels of correlation discussed in Section 4.1. We discuss here the NL and NLM model and show in Chapter 7 the results of simulation studies that evaluated the effectiveness of a NLM model approach when compared to the NL approach. We hypothesized that at a minimum, the superiority of the NLM model over the NL model will occur for the same situations that the LMM is superior to the LS approach, and found that the NLM model is clearly superior for all the scenarios that we investigated.

6.1 NL Model Formulation

Just as we did for the LMM, we assume that we have m profiles of data, each of which has n_i measurements where i refers to the i^{th} profile. We can then fit a separate NL model to each profile. Let y_{ij} refer to the j^{th} measurement for the i^{th} profile. The model for each of the separate nonlinear regressions is given by

$$y_{ij} = f(x_{ij}, \boldsymbol{\theta_i}) + \epsilon_{ij} \text{ for } i = 1, 2, \ldots, m, j = 1, 2, \ldots, n_i, \tag{6.1}$$

where $f(.)$ is some nonlinear function, x_{ij} is a regressor variable for the measurement, $\boldsymbol{\theta_i}$ is a $px1$ vector of parameters for each profile, and ϵ_{ij} is the error associated with the j^{th} measurement of the i^{th} profile. The errors are often assumed to be independent and normally distributed, that is $\epsilon_{ij} \sim N(0, \sigma_i^2)$, implying that the measurements within a profile are uncorrelated. We will assume throughout that the profiles have the same variability in the error term, thus $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \ldots, m$.

If the responses for the i^{th} profile are stacked to form a vector, $\mathbf{y}_i$, then we have the alternative form of the model from (6.1) given by

$$\mathbf{y}_i = f(\mathbf{x}_i, \boldsymbol{\theta_i}) + \boldsymbol{\epsilon}_i \text{ for } i = 1, 2, \ldots, m, \tag{6.2}$$

where $\mathbf{x}_i$ is a vector of the values of the predictor variable and $\boldsymbol{\epsilon}_i$ is a vector of errors that has a multivariate normal distribution, $\boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \mathbf{R}_i)$ with $\mathbf{R}_i$ being a n by n positive definite variance-covariance matrix. If the errors are correlated, $\mathbf{R}_i$ is often assumed to be a simple form such as compound symmetry (CS) or autoregressive (AR) in order to reduce the number of covariance parameters that need to be estimated, just as we did in the LMM. If the measurements within a profile are uncorrelated then $\boldsymbol{\epsilon}_i \sim MN(0, \sigma_i^2\mathbf{I})$.

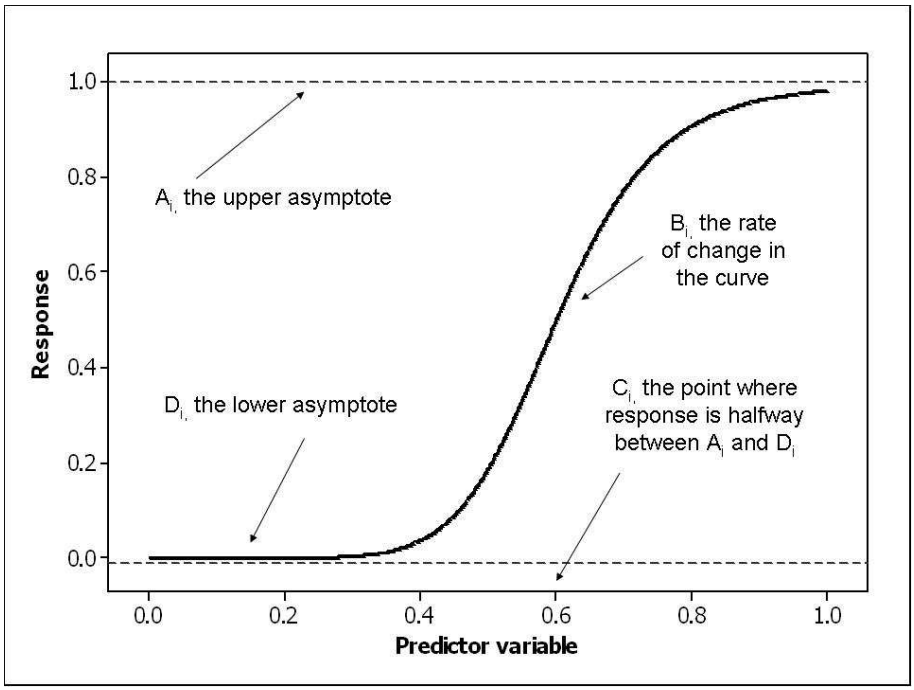
There are a wide variety of nonlinear functions that have been used for applications. A

number of books covering nonlinear models (Gallant, 1987; Ratkowsky, 1990; Schabenberger and Pierce, 2002; Seber and Wild, 2003) covered many of the different types of functions that have been utilized. For example, consider the 4-parameter logistic model, which has been used frequently for dose-response studies. This model is given by

$$y_{ij} = A_i + \frac{D_i - A_i}{1 + \left(\frac{x_{ij}}{C_i}\right)^{B_i}} + \epsilon_{ij} \text{ for } i = 1, 2, \ldots, m, j = 1, 2, \ldots, n_i, \tag{6.3}$$

where A_i is the upper asymptote, D_i is the lower asymptote, C_i is the point where the curve reaches halfway between A_i and D_i, and B_i is a parameter representing the rate of increase or decrease from D_i to A_i. The larger the value of B_i, the steeper will be the slope of the curve. See Figure 6.1 for an example of the 4-parameter logistic curve where $A_i = 1$, $B_i = 8$, $C_i = .6$, and $D_i = 0$.

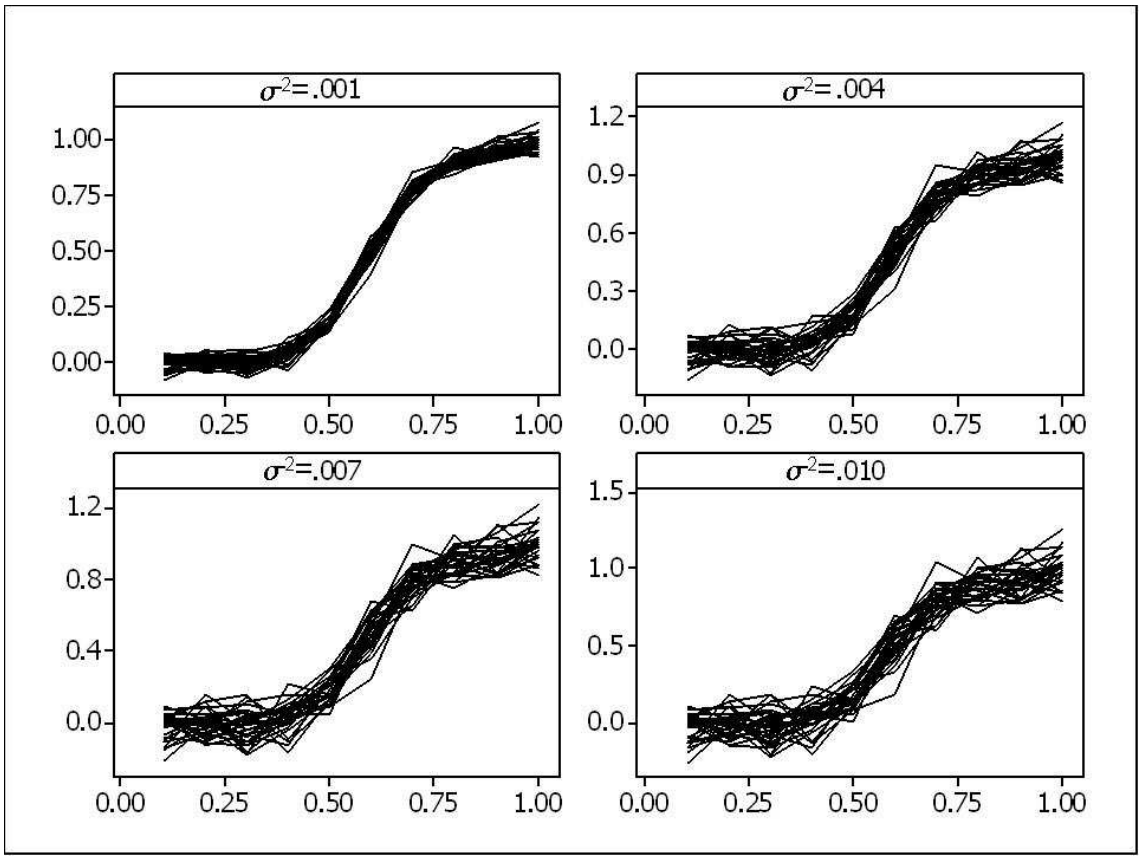
Figure 6.1: Example of a 4-parameter logistic curve where $A_i = 1$, $B_i = 8$, $C_i = .6$, and $D_i = 0$.



When the parameters in (6.2) are all fixed, then the only variability in the responses

will be due to the $\boldsymbol{\epsilon}_i$ term. To illustrate the impact of the variability of the errors on profiles, consider Figure 6.2, which shows randomly generated 4-parameter logistic curves with uncorrelated errors with different values of σ^2 for all generated profiles. Here $m = 30$, $n = 10$, and the values of A_i, B_i, C_i, and D_i are the same as those in Figure 6.1. We see that the larger the variability of the errors the greater the differences between the profiles.

Figure 6.2: Illustration of the difference in the generated profiles due to differences in the variability of the errors. For these profiles, $m = 30$, $n = 10$, and the values of A_i, B_i, C_i, and D_i are the same as those in Figure 6.1.



6.2 NL Model Estimation

For the NL model in (6.1) and (6.2), the maximum likelihood estimator (MLE)) of $\boldsymbol{\theta_i}$, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}$, is that estimator that minimizes the residual sum of squares given by

$$\begin{aligned} SS(\boldsymbol{\theta_i}) &= \sum_{j=1}^{n_i} [y_{ij} - f(x_{ij}, \boldsymbol{\theta_i})]^2 \text{ for } i = 1, 2, \ldots, m \\ &= [\mathbf{y}_i - f(\mathbf{x}_i, \boldsymbol{\theta_i})]' [\mathbf{y}_i - f(\mathbf{x}_i, \boldsymbol{\theta_i})] \text{ for } i = 1, 2, \ldots, m. \end{aligned} \tag{6.4}$$

Because of the nonlinearity introduced in (6.4) by $f(.)$ there is not a closed form expression for the estimator that minimizes $SS(\boldsymbol{\theta_i})$. There are two major iterative algorithms used for obtaining the parameter estimates for this NL model.

The first, the Gauss-Newton (GN) algorithm, replaces $f(\mathbf{x}_i, \boldsymbol{\theta_i})$ in (6.4) with a Taylor series approximation about $\boldsymbol{\theta_i}$ so that the minimization can then proceed in an iterative fashion. We denote $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,0}$ as the vector of initial starting values for the iterative algorithm, and denoting the matrix of derivatives of the NL function evaluated at the initial starting values as

$$\widehat{\mathbf{F}}_i = \frac{\partial\, f(\mathbf{x}_i, \boldsymbol{\theta_i})}{\partial \boldsymbol{\theta_i}}\Big|_{\boldsymbol{\theta_i} = \widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,0}}. \tag{6.5}$$

With the GN algorithm one computes

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,1} = \widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,0} + \left(\widehat{\mathbf{F}}_i{}' \widehat{\mathbf{F}}_i^{\,-1}\right) \widehat{\mathbf{F}}_i{}' \left[\mathbf{y}_i - f(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,0})\right]. \tag{6.6}$$

After each iteration $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,0}$ is replaced by $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,1}$ until $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,0} \approx \widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,1}$. Thus the algorithm stops when the change in estimated parameters from one iterate to the next is sufficiently small. In this case the algorithm has converged and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} = \widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\,1}$. Adjustments to this basic algorithm are often implemented in practice to avoid numerical issues and ensure convergence (Schabenberger and Pierce, 2002, Chapter 5; Seber and Wild, 2003, Chapter 14).

The second, the Newton-Raphson (NR) algorithm, replaces the entire function in (6.4) by a Taylor series approximation and similar to the GN method, iteratively solves for the parameter estimates. For simple NL models, there will often not be much difference in the estimates obtained by either of the two algorithms. Thus the default method in $SAS^{®}$, the GN algorithm, will be our choice for what follows.

With the normality assumption in the NL model, we can state the distributional properties of the vector of obtained parameter estimators, $\widehat{\boldsymbol{\theta}_i}$. In contrast to the standard linear model, normality of the parameter estimator will only hold asymptotically, that is,

$$\widehat{\boldsymbol{\theta}_i} \overset{A}{\sim} MN\left[\boldsymbol{\theta}_i, \sigma_i^2(\mathbf{F}_i'\mathbf{F}_i)^{-1}\right] \text{ for } i = 1, 2, \ldots, m. \tag{6.7}$$

The result in (6.7) holds whether or not the errors are uncorrelated (Seber and Wild, 2003, Section 12.2).

6.3 NLM Model Formulation

In the NLM model we extend the NL model in (6.2) to allow for random effects. In vector form it is given by

$$\mathbf{y}_i = f(\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{b}_i) + \boldsymbol{\epsilon}_i \text{ for } i = 1, 2, \ldots, m, \tag{6.8}$$

where $\boldsymbol{\theta}$ is a vector of fixed effects common to all profiles, $\mathbf{b}_i$ is the r by 1 vector of random effects with $\mathbf{b}_i \sim MN(\mathbf{0}, \mathbf{D})$. The matrix $\mathbf{D}$ is assumed to be a diagonal matrix with the diagonal elements referred to as the variance components. A good introduction and review of the literature on the NLM model can be found in Davidian and Giltinan (2003) or books by Pinheiro and Bates (2000), Schabenberger and Pierce (2002), and Demidenko (2004).

Demidenko (2004, Section 6.1) proposed a restriction of (6.8) by forcing the random effects

to enter the model linearly so that we can rewrite (6.8) as

$$\mathbf{y}_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \text{ for } i = 1, 2, \ldots, m. \tag{6.9}$$

where $\mathbf{Z}_i$ is a n_i by r matrix of values corresponding to the random effects. While this may simplify the computational effort required to obtain estimates, we will not pursue it here because it is less flexible.

Just as the LMM, the model in (6.8) allows for two levels of correlation for the measurements within a profile. The first results from the random effects which cause all the measurements within a profile to be correlated to each other. The second results from the within-profile variance-covariance matrix of the errors, $\mathbf{R}_i$. A NLM model that uses neither of the two levels of correlation is simply the NL model in (6.2) with uncorrelated errors because $\mathbf{Z}_i = \mathbf{0}$ and $\boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \sigma_i^2\mathbf{I})$. See Davidian and Giltinan (2003, pp. 395-400) for more discussion of the within profile correlation and its interpretation in the NLM model.

In contrast with the LMM where often all the fixed effects will have a corresponding random effect without resulting computational difficulties, the NLM model poses a more formidable computational challenge when including multiple random effects. Thus it is recommended that the model builder begin with a smaller number of random effects and add additional effects only if needed at the cost of increased computing effort. Demidenko (2004, Section 8.17) goes even further by stating that only a single random effect should be added at a time and that a good initial model would be one with no random effects. Subject matter expertise and knowledge of the data collection process can also help determine the appropriateness of adding random effects to a NL model. In Chapter 7 we will show how use of the NL model can be used to determine which random effects should be added.

6.4 NLM Model Estimation

As noted by Schabenberger and Pierce (2002), given the random effects in the NLM model, one can write the marginal density of $\mathbf{y}_i$ as

$$g(\mathbf{y}_i) = \int g(\mathbf{y}_i|\mathbf{b}_i)\, g(\mathbf{b}_i) d\mathbf{b}_i. \tag{6.10}$$

where $g(.)$ is the assumed probability density function. The evaluation of this integral is required for inference but the distribution $g(\mathbf{y}_i)$ is not known even when the errors and random effects have multivariate normal distributions. As a result, numerical methods are needed. The most common methods are the linearization and the integral approximation methods.

Earlier work on estimation of NLM models focused on the linearization approach which is an approximation of the nonlinear function in (6.8) by some linear function. This replacement results in a multivariate normal density function for which an estimator can be obtained. On the other hand, the integral approximation approach has increased in popularity in recent years due to advances in computing power and existence of Monte Carlo (MC) based methods for directly evaluating the integral in (6.10). The nlmixed procedure of $SAS^{®}$ uses the integral approximation approach with adaptive MC procedures that are more computationally efficient than standard MC procedures. Thus our analysis of the NLM model will be based on the integral approximation method.

6.5 Diagnostics and Robust Estimation in NL and NLM Models

Davidian and Giltinan (1995, p. 328) noted that diagnostic methods for NL and NLM models are underdeveloped and nearly non-existent, but would be very useful. While there are

computational difficulties in dealing with the nonlinearity of the models, they recommended more research in the area of diagnostics and noted that new research will increase the utilization of NL and NLM models. We review in this section some of the methods that have appeared, but believe the area to still be underdeveloped.

Pinheiro and Bates (2000) proposed to check assumptions of the NLM model by using the classical regression diagnostic plots such as the normal probability plot of errors and a scatter plot of residuals versus the fitted values. They do not give justification for why such plots would work other than that they have been used for classical regression problems. Demidenko (2004, Section 9.5) discussed some methods to determine the influence of unusual data points on NL regression models. Lee and Xu (2004) considered diagnostic methods for the NLM model based on the case deletion and local influence approach. However, they had to use MC methods in order to compute the diagnostics and this would seem to limit their applicability.

An alternative to diagnostic methods are robust estimation methods that minimize the impact of outliers and other departures from model assumptions. This area of research also appears to be underdeveloped but we give some pertinent references here.

Hartford and Davidian (2000) performed a simulation study to study the robustness of the estimators of fixed effects to the misspecification of the distribution of the random effects. They found that if the underlying distribution of random effects is unimodal and not drastically different from a normal distribution that the estimation of the fixed effects is not severely impacted. However, their study compared two different linearization methods and did not consider an integral based method so it is unclear if their results would have changed under a more complete comparison. In addition, their focus was the impact of misspecification of the random effects distribution on the fixed effects. Thus it remains to

be seen how the misspecification of the distribution of errors can impact the estimation of the fixed effects.

Yeap and Davidian (2001) proposed a two-stage robust procedure for estimation for non-linear models with random effects. In the first stage, estimation of the profile specific effects and its covariance parameters is performed by solving a simultaneous set of weighted functions of the residuals. In the second stage, the robust estimates of profile specific effects are used to obtain an overall effect and the covariance parameters of the random effects. Here the weights in the robust estimation procedure are based on a robust version of the Mahalanobis distance. Yeap and Davidian (2001) recommended using the weights of the first stage to determine outlying observations within a profile and the weights of the second stage to determine outlying profiles but these weights appear to be more of an ad hoc diagnostic tool. They acknowledged that the weights will only detect a small number of outliers that are not severely different from the main portion of the data.

Our approach, which combines the separate NL regressions with a NLM model, differs from these previous approaches appearing in the literature. It will be detailed in Section 7.6. Prior to explaining the approach, we compare approaches based on the NL and NLM models in Chapter 7.