Chapter 5

Profile Monitoring for Linear Mixed Models

Here we propose a method of fitting the profiles for data where the within-profile measurements are correlated with each other, thus relaxing the assumption of independent errors. We do so by fitting a linear mixed model (LMM), which allows us to account for the correlation within profiles. The LMM also allows us to consider the profiles as a random sample of profiles from a common population distribution, which may be a more realistic assumption than assuming that the profiles are completely independent of each other.

The estimated parameter vector for the i^{th} profile is given by

$$\widehat{\boldsymbol{\beta}}_{i,MIX} = \widehat{\boldsymbol{\beta}}_{MIX} + \widehat{\mathbf{b}}_i \text{ for } i = 1, 2, \ldots, m, \tag{5.1}$$

which can be found using (4.7) and (4.8). Once the LMM has been fit to the profiles, we have reduced the data to a time-ordered series of vectors, $\widehat{\boldsymbol{\beta}}_{i,MIX} = \widehat{\boldsymbol{\beta}}_{MIX} + \widehat{\mathbf{b}}_i$ for $i = 1, 2, \ldots, m$ where $\widehat{\boldsymbol{\beta}}_{MIX}$ is the fixed portion that is the same for all profiles and $\widehat{\mathbf{b}}_i$ is the deviation from the fixed portion. The control chart application consists of replacing the $\mathbf{a}_i$ vector in (2.4) with the $\widehat{\boldsymbol{\beta}}_{i,MIX}$ vector in (5.1). Thus rather than using the actual data in the T^2 statistic we use the estimates of the model coefficients obtained from the data. We are reducing the

problem of detecting changes in the data profiles to detecting changes in the parameters that summarize the profiles. This is a more efficient approach because we are monitoring a smaller number of parameters rather than a larger number of data observations. Of course, our approach is based on the assumption that the fitted model adequately describes the profile data.

5.1 T^2 Methods for LMM

Using (5.1) the T^2 statistics based on the sample mean vector and classical moment estimator of the variance-covariance matrix are given by

$$T^2_{1,i,MIX} = (\widehat{\boldsymbol{\beta}}_{i,MIX} - \overline{\boldsymbol{\beta}}_{MIX})' \mathbf{S}^{-1}_{1,MIX} (\widehat{\boldsymbol{\beta}}_{i,MIX} - \overline{\boldsymbol{\beta}}_{MIX}) \text{ for } i = 1, 2, \ldots, m, \tag{5.2}$$

where

$$\mathbf{S}_{1,MIX} = \frac{\sum_{i=1}^{m} (\widehat{\boldsymbol{\beta}}_{i,MIX} - \overline{\boldsymbol{\beta}}_{MIX})(\widehat{\boldsymbol{\beta}}_{i,MIX} - \overline{\boldsymbol{\beta}}_{MIX})'}{m-1}, \tag{5.3}$$

and where

$$\overline{\boldsymbol{\beta}}_{MIX} = \frac{\sum_{i=1}^{m} \widehat{\boldsymbol{\beta}}_{i,MIX}}{m}. \tag{5.4}$$

$T^2_{1,i,MIX}$ will be proportional to a beta distribution if the matrix $\mathbf{V}_i$ is known, and asymptotically proportional to a beta distribution if a consistent estimate of $\mathbf{V}_i$ is obtained. As mentioned in Section 2.2 this statistic will not be effective in detecting out-of-control data, nonetheless we use it here as a reference statistic.

Similarly we have

$$T^2_{2,i,MIX} = (\widehat{\boldsymbol{\beta}}_{i,MIX} - \overline{\boldsymbol{\beta}}_{MIX})' \mathbf{S}^{-1}_{2,MIX} (\widehat{\boldsymbol{\beta}}_{i,MIX} - \overline{\boldsymbol{\beta}}_{MIX}) \text{ for } i = 1, 2, \ldots, m, \tag{5.5}$$

where

$$\mathbf{S}_{2,MIX} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (\widehat{\boldsymbol{\beta}}_{i+1,MIX} - \widehat{\boldsymbol{\beta}}_{i,MIX})(\widehat{\boldsymbol{\beta}}_{i+1,MIX} - \widehat{\boldsymbol{\beta}}_{i,MIX})'. \tag{5.6}$$

$T^2_{2,i,MIX}$ will be proportional or asymptotically proportional to a χ^2 distribution when the number of profiles is at least twice as large as p (Williams et al., 2006b).

In Appendix B we show that $\sum_{i=1}^{m} \widehat{\mathbf{b}}_i = \mathbf{0}$ as long as the $\mathbf{Z}_i$ matrix is contained within the $\mathbf{X}_i$ matrix. This is true even if the $\mathbf{Z}_i$ and $\mathbf{X}_i$ matrices are different from profile to profile. In addition we show in Appendix B, that when the eblups sum to zero, that $\overline{\boldsymbol{\beta}}_{MIX} = \widehat{\boldsymbol{\beta}}_{MIX}$.

As a result, we show in Appendix C that the T^2 statistics shown in (5.2) and (5.5) can be expressed as a function of the eblups. We can rewrite the expressions in (5.2) and (5.5) as

$$T^2_{1,i,MIX} = (\widehat{\mathbf{b}}_i)' \left[\frac{\sum_{i=1}^{m} (\widehat{\mathbf{b}}_i)(\widehat{\mathbf{b}}_i)'}{m-1} \right]^{-1} (\widehat{\mathbf{b}}_i) \text{ for } i = 1, 2, \ldots, m, \tag{5.7}$$

and

$$T^2_{2,i,MIX} = (\widehat{\mathbf{b}}_i)' \left[\frac{\sum_{i=1}^{m-1} (\widehat{\mathbf{b}}_{i+1} - \widehat{\mathbf{b}}_i)(\widehat{\mathbf{b}}_{i+1} - \widehat{\mathbf{b}}_i)'}{2(m-1)} \right]^{-1} (\widehat{\mathbf{b}}_i) \text{ for } i = 1, 2, \ldots, m. \tag{5.8}$$

These expressions simplify the computation of the T^2 statistics when fitting a LMM to the data.

For the situation where $\mathbf{Z}_i$ is not contained within $\mathbf{X}_i$, we can still obtain a simplification of the T^2 statistics even though the eblups will not necessarily sum to zero. This proof is shown in Appendix C and the simplification of (5.2) is given by

$$T^2_{1,i,MIX} = (\widehat{\mathbf{b}}_i - \overline{\mathbf{b}})' \left[\frac{\sum_{i=1}^{m} (\widehat{\mathbf{b}}_i - \overline{\mathbf{b}})(\widehat{\mathbf{b}}_i - \overline{\mathbf{b}})'}{m-1} \right]^{-1} (\widehat{\mathbf{b}}_i - \overline{\mathbf{b}}), \tag{5.9}$$

and the simplification of (5.5) is given by

$$T^2_{2,i,MIX} = (\widehat{\mathbf{b}}_i - \overline{\mathbf{b}})' \left[\frac{\sum_{i=1}^{m-1} (\widehat{\mathbf{b}}_{i+1} - \widehat{\mathbf{b}}_i)(\widehat{\mathbf{b}}_{i+1} - \widehat{\mathbf{b}}_i)'}{2(m-1)} \right]^{-1} (\widehat{\mathbf{b}}_i - \overline{\mathbf{b}}), \tag{5.10}$$

.

5.2 Impact of Ignoring LMM Structure

A naive approach to compare m linear profiles is to ignore the correlation structure and the random effects and obtain the estimates of the model parameters via the LS estimator of the classical linear model (LM) in (4.1) even though the data follow the model (4.3). We refer to this naive approach as the LS approach to distinguish it from the LMM approach. In the LS approach the fixed parameters are estimated separately for each profile and are obtained via the expression in (4.2).

When taking the LS approach, we have

$$T^2_{1,i,LS} = (\widehat{\boldsymbol{\beta}}_{i,LS} - \overline{\boldsymbol{\beta}}_{LS})' \mathbf{S}^{-1}_{1,LS} (\widehat{\boldsymbol{\beta}}_{i,LS} - \overline{\boldsymbol{\beta}}_{LS}) \text{ for } i = 1, 2, \ldots, m, \tag{5.11}$$

where

$$\mathbf{S}_{1,LS} = \frac{\sum_{i=1}^{m} (\widehat{\boldsymbol{\beta}}_{i,LS} - \overline{\boldsymbol{\beta}}_{LS})(\widehat{\boldsymbol{\beta}}_{i,LS} - \overline{\boldsymbol{\beta}}_{LS})'}{m-1}, \tag{5.12}$$

and where

$$\overline{\boldsymbol{\beta}}_{LS} = \frac{\sum_{i=1}^{m} \widehat{\boldsymbol{\beta}}_{i,LS}}{m}, \tag{5.13}$$

and we have

$$T^2_{2,i,LS} = (\widehat{\boldsymbol{\beta}}_{i,LS} - \overline{\boldsymbol{\beta}}_{LS})' \mathbf{S}^{-1}_{1,LS} (\widehat{\boldsymbol{\beta}}_{i,LS} - \overline{\boldsymbol{\beta}}_{LS}) \text{ for } i = 1, 2, \ldots, m, \tag{5.14}$$

where

$$\mathbf{S}_{2,LS} = \frac{1}{2(m-1)} \sum_{i=1}^{m} (\widehat{\boldsymbol{\beta}}_{i+1,LS} - \widehat{\boldsymbol{\beta}}_{i,LS})(\widehat{\boldsymbol{\beta}}_{i+1,LS} - \widehat{\boldsymbol{\beta}}_{i,LS})'. \tag{5.15}$$

Nonetheless, even if the data have random effects, the least squares estimator in (4.2) is an unbiased estimator when $\mathbf{X}_i = \mathbf{Z}_i$ because $E(\widehat{\boldsymbol{\beta}}_{i,LS}) = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{X}_i\boldsymbol{\beta} = \boldsymbol{\beta}$. This means that all m coefficient estimates are each an estimate of $\boldsymbol{\beta}$ and therefore that $\frac{\sum_{i=1}^{m} \widehat{\boldsymbol{\beta}}_{i,LS}}{m} = \overline{\boldsymbol{\beta}}_{LS}$ is also an unbiased estimator of $\boldsymbol{\beta}$.

Consequently, even when the data have correlated errors and random effects we can obtain an unbiased estimate of the fixed coefficients via least squares. As noted by Demidenko (2004, Section 3.9), the standard errors of the estimators will be inflated when ignoring the LMM structure in the data. This will have a negative impact on hypothesis tests and confidence intervals for the fixed coefficients. To investigate the impact of ignoring the LMM structure on profile monitoring we performed a simulation study described in the next section.

While both the LMM and LS approaches allow some common-cause variability between the profiles, the LMM approach has several advantages over the LS approach, some of which were noted by Verbeke and Molenberghs (2000). First, the LMM approach can be easily used for balanced and unbalanced data and is better than the LS approach when the number of observations per profile is small. The LMM approach combines information from the profiles to achieve the model fit with fewer parameters than fitting separate regression models for each profile. Second, the LMM approach is capable of handling profiles with missing data, even for situations where the number of observations for a particular profile is less than the number of parameters that would be needed to fit a regression model to that individual profile.

5.3 Simulation Study Setup

We now explain the general procedure for the simulation studies used to compare the LS and LMM approaches. Multivariate data that follow a linear mixed model structure were generated where the random errors follow some specific structure. This is accomplished by generating univariate normal data and using the Cholesky decomposition to transform the generated univariate data to multivariate data. The data are fit with a LMM using

proc mixed of $SAS^{®}$ with the correct model specification. We included both correlated and uncorrelated errors in our comparisons.

The control limit is established using the appropriate percentiles of the beta or χ^2_p distributions so that the probability of signal for the in-control data is .05, the nominal value. The actual probability of signal is estimated by the proportion of datasets where there was a signal. That is, a signal occurs when at least one of the m T^2 statistics exceeds the control limit.

Here we consider the case of simple linear regression with a random slope and intercept so we have $p = 2$ and $\mathbf{X}_i = \mathbf{Z}_i$. For the studies performed, a total of $10,000$ datasets (Monte Carlo repetitions) were generated for each combination of interest unless otherwise noted. We can assume without loss of generality that $\boldsymbol{\beta} = [0, 1]$. This assumption results from the regression equivariance property of the estimators in both the classical LM and the LMM, as discussed in Appendix D.

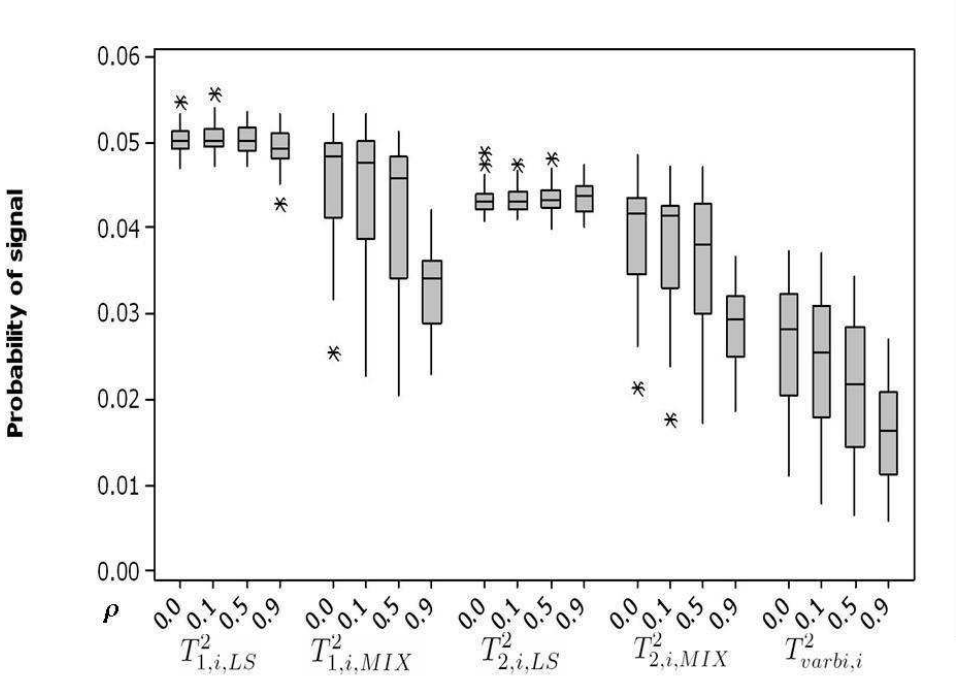
In some of our scenarios, there were no problems with non-convergence when obtaining the LMM results so no starting values were needed. In other scenarios where non-convergence was more likely to occur, we used as starting values the known parameters used to generate the data as was discussed in Section 4.7.

5.4 Balanced, Equally Spaced Data

Our initial study is for the balanced, equally spaced data scenario. Here a commonly used error structure is an AR(1) model where ρ represents the amount of autocorrelation between error terms for successive equally spaced observations. We show here the results for different values of m, n and ρ where $\sigma_0^2 = .1$, $\sigma_1^2 = .1$, and $\sigma^2 = .1$. The matrix $\mathbf{D} = diag(\sigma_0^2, \sigma_1^2)$

represents the variability in the intercept and slope respectively from profile to profile. The statistics $T^2_{1,i}$ and $T^2_{2,i}$ are calculated from the LMM approach (the "right way" in that both the random effects and the correct correlation structure are accounted for) and the LS approach (the "naive way" where both the random effects and the correlation structure are ignored). We also included in our study the version of the T^2 statistic shown in (4.15) as proposed by Waternaux, Laird, and Ware (1989) with $Var(\widehat{\mathbf{b}}_i)$ instead of $Var(\widehat{\mathbf{b}}_i - \mathbf{b}_i)$. We found that use of $Var(\widehat{\mathbf{b}}_i - \mathbf{b}_i)$ resulted in extremely large probabilities of signal for in-control data. Thus we use the expression found in (4.15) rather than the expression in (4.14) with $\mathbf{D}$ and $\mathbf{V}_i$ replaced with their estimated values. Table 5.1 shows the proportion of the $10,000$ datasets that had a signal on the control charts for the various T^2 statistics.

Figure 5.1: Boxplots of the probability of signal of balanced, equally spaced data for various combinations of m, n, σ_0^2, σ_1^2, and σ^2 for the different versions of the T^2 statistic.



We see from Table 5.1 that for the in-control situation, it appears that the more compli-

Table 5.1: Proportion of datasets with a signal for in-control data for the balanced, equally spaced data scenario.

m	n	ρ	$T^2_{1,i,LS}$	$T^2_{1,i,MIX}$	$T^2_{2,i,LS}$	$T^2_{2,i,MIX}$	$T^2_{varbi,i}$
30	5	0	0.0507	0.0481	0.0464	0.0436	0.0193
30	5	0.1	0.0515	0.0465	0.0468	0.0427	0.0159
30	5	0.5	0.0506	0.0445	0.0438	0.0377	0.0134
30	5	0.9	0.0515	0.0334	0.0450	0.0274	0.0100
30	10	0	0.0524	0.0514	0.0413	0.0407	0.0215
30	10	0.1	0.0526	0.0506	0.0419	0.0426	0.0193
30	10	0.5	0.0496	0.0464	0.0456	0.0410	0.0159
30	10	0.9	0.0495	0.0355	0.0468	0.0317	0.0103
60	5	0	0.0497	0.0491	0.0441	0.0434	0.0298
60	5	0.1	0.0503	0.0495	0.0432	0.0425	0.0278
60	5	0.5	0.0524	0.0471	0.0456	0.0418	0.0263
60	5	0.9	0.0512	0.0342	0.0460	0.0307	0.0177
60	10	0	0.0497	0.0496	0.0450	0.0449	0.0303
60	10	0.1	0.0502	0.0503	0.0458	0.0456	0.0310
60	10	0.5	0.0504	0.0490	0.0451	0.0433	0.0281
60	10	0.9	0.0494	0.0407	0.0411	0.0331	0.0204
90	5	0	0.0492	0.0490	0.0426	0.0425	0.0349
90	5	0.1	0.0501	0.0499	0.0417	0.0403	0.0318
90	5	0.5	0.0518	0.0479	0.0444	0.0429	0.0299
90	5	0.9	0.0489	0.0340	0.0459	0.0310	0.0210
90	10	0	0.0531	0.0531	0.0445	0.0445	0.0374
90	10	0.1	0.0532	0.0527	0.0445	0.0440	0.0372
90	10	0.5	0.0528	0.0510	0.0445	0.0455	0.0345
90	10	0.9	0.0487	0.0415	0.0439	0.0368	0.0261

cated mixed model analysis makes little difference in terms of a probability of a signal when the data are balanced and equally spaced. This is true for different values of m, n, and ρ not shown here. We note that the use of $T^2_{2,i,LS}$ and $T^2_{2,i,MIX}$ statistics gives slightly smaller probability of signal than the nominal .05 level and that the statistics based on the LMM approach have slightly smaller probabilities than those based on the LS approach. While the probability of a signal is lower for the $T^2_{varbi,i}$ statistics, we expect the out-of-control performance of this statistic to be similar to those of $T^2_{1,i,LS}$ and $T^2_{1,i,MIX}$ with little ability to detect step changes or outliers.

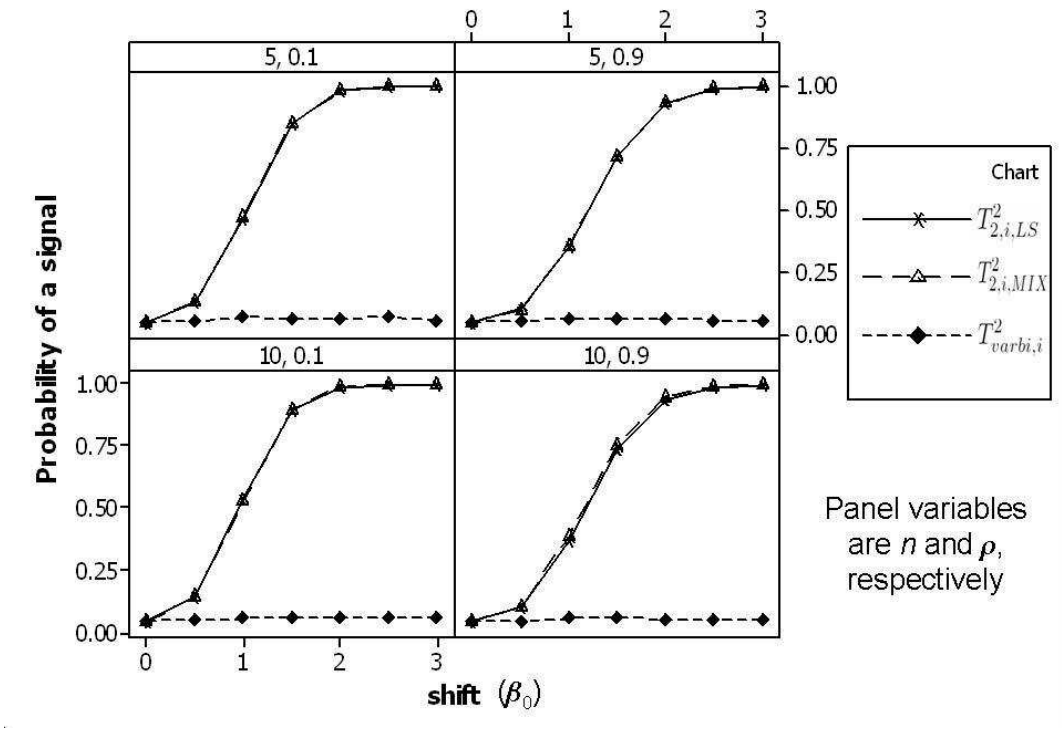
We performed similar studies with differences in the σ_0^2, σ_1^2, and σ^2 values used to generate the data in order to investigate their impact. We generated data for all possible combinations of 2 levels for each of $n = (5, 10)$, $\sigma_0^2 = (.1, 1)$, $\sigma_1^2 = (.1, 1)$, and $\sigma^2 = (.1, 1)$, 3 levels for m, $(30, 60, 90)$, and 4 levels for ρ, $(0, .1, .5, .9)$. Thus we have expanded the 24 combinations shown in Table 5.1 to a total of $4 * 3 * 2 * 2 * 2 * 2 = 192$ combinations. Rather than present the full table of results, we summarize the results by showing in Figure 5.1 the boxplots of the estimated false alarm probabilities for the four T^2 statistics. Each boxplot contains false alarm rates for all the combinations of m, n, σ_0^2, σ_1^2, and σ^2 and represents 48 combinations of settings, each of which is based on $10,000$ generated datasets.

We note that for the T^2 statistics based on the LS approach, there is little variability in the probability of signal for different values of m, n, σ_0^2, σ_1^2, and σ^2. There is more variability in the probability of signal for the T^2 statistics based on the LMM approach and for the $T^2_{varbi,i}$ statistics. There will be relatively little difference in the estimated probability of signal for the LS and LMM approaches, however, when m, n, σ_0^2, σ_1^2, and σ^2 are changed.

We next consider the probability of signal for data that comes from an out-of-control process. These power studies were performed by introducing a step change in the mean vector, $\boldsymbol{\beta}$. Because the probability of signal is not always .05 for the in-control data, the power studies were based on a simulated control limit to ensure that the probability of a signal for in-control data will be the same for all the charts and close to the nominal .05 level.

For the m profiles of data, the first l of them were generated from the in-control distribution with $\boldsymbol{\beta} = [0, 1]$ and the last $m - l$ were generated from the same distribution and same settings of the design factors, except that $\boldsymbol{\beta} = [\beta_0, 1]$ for $\boldsymbol{\beta} > 0$. Thus we have introduced a step change in the intercept causing the last $m - l$ profiles to be shifted away from the first

Figure 5.2: Probability of signal of out-of-control, balanced, equally spaced data for different values of n and ρ, for the $T^2_{2,i,LS}$, $T^2_{2,i,MIX}$, and $T^2_{varbi,i}$ charts where $m = 30$, $\sigma^2 = .1$, $\sigma_0^2 = .1$ and $\sigma_1^2 = .1$.



l profiles.

Figure 5.2 shows some of the results of the power studies. Here $m = 30$, $\sigma^2 = .1$, $\sigma_0^2 = .1$ and $\sigma_1^2 = .1$ where the step change occurred after the fifth profile. We do not show the results for the charts based on the $T^2_{1,i,LS}$ and $T^2_{1,i,MIX}$ values because it is known that these statistics will not perform well in detecting step changes (Sullivan and Woodall, 1996; Vargas, 2003).

The curves corresponding to the use of $T^2_{2,i,MIX}$ and $T^2_{2,i,LS}$ practically coincide, indicating that the two methods will perform similarly in detecting the step change in the intercept. This is true regardless of the amount of correlation in the errors or the number of observations per profile. On the other hand $T^2_{varbi,i}$ performs poorly, with little ability to signal the shift.

This is because the expression for $Var(\widehat{\mathbf{b}}_i)$ in (4.13) is only correct if all the data are in-control and come from the same distribution. When a step change is present the estimated values of $Var(\widehat{\mathbf{b}}_i)$ are inflated reducing the ability to detect that change.

To further explain the similarity of the results obtained from the LS and LMM approaches, Demidenko (2004, Section 2.3) showed that in the balanced data scenario with uncorrelated errors that the estimator of fixed effects in the LMM reduces to an estimator based on ordinary least squares. Recall from (4.7) that the estimator of the fixed effects depends on the variance-covariance matrix, $\mathbf{V}_i$. Because $\mathbf{X}_i = \mathbf{Z}_i = \mathbf{X}_\star$ for $i = 1, 2, \ldots, m$, the result of Demidenko (2004, Section 2.3) shows that (4.7) can be written as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{MIX} &= \left(\sum_{i=1}^{m} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{m} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{y}_i\right) \\
&= \left(\mathbf{X}_\star' \mathbf{V}_i^{-1} \mathbf{X}_\star\right)^{-1} \mathbf{X}_\star' \mathbf{V}_i^{-1} \frac{\sum_{i=1}^{m} \mathbf{y}_i}{m} \\
&= \left(\mathbf{X}_\star' \mathbf{X}_\star\right)^{-1} \mathbf{X}_\star' \overline{\mathbf{y}}
\end{aligned}$$

It should be noted, that the previous result does not imply that $\widehat{\boldsymbol{\beta}}_{i,MIX} = \widehat{\boldsymbol{\beta}}_{MIX} + \widehat{\mathbf{b}}_i$ and $\widehat{\boldsymbol{\beta}}_{i,LS}$ are equivalent. It does imply that $\frac{\sum_{i=1}^{m} \widehat{\boldsymbol{\beta}}_{i,MIX}}{m} = \widehat{\boldsymbol{\beta}}_{MIX} + \frac{\sum_{i=1}^{m} \widehat{\mathbf{b}}_i}{m} = \frac{\sum_{i=1}^{m} \widehat{\boldsymbol{\beta}}_{i,LS}}{m}$. Thus, in the balanced data scenario, the average of the estimates obtained by separate simple linear regression will be equal to the obtained estimate of the fixed effects.

Similar simulation results, not shown here, were obtained when the step change occurred at some other point, or when the shift occurred in the slope. We found the conclusions stated here for the out-of-control data will hold for changes in the slope or intercept, no matter the value of l.

5.5 Balanced, Unequally Spaced Data

We now consider the situation where the $\mathbf{X}_i$ and $\mathbf{Z}_i$ matrices are the same for each profile but the observations are not necessarily equally spaced from each other. The observations are taken at certain fixed points depending on the nature of the process generating the profile data. For example, some profiles may have more observations near the extremes than in the middle and vice versa. To investigate the impact of observations not taken at random points, but set at certain values, we performed a similar simulation study as that of the previous section.

Table 5.2: Proportion of datasets with a signal for in-control data for the balanced, unequally spaced data scenario.

ρ	σ_1^2	σ_0^2	σ^2	$T^2_{1,i,LS}$	$T^2_{1,i,MIX}$	$T^2_{2,i,LS}$	$T^2_{2,i,MIX}$	$T^2_{varbi,i}$
0.1	0.1	0.1	0.1	0.0490	0.0463	0.0440	0.0402	0.0202
0.1	0.1	0.1	1.0	0.0485	0.0331	0.0436	0.0285	0.0154
0.1	0.1	1.0	0.1	0.0514	0.0493	0.0476	0.0438	0.0201
0.1	0.1	1.0	1.0	0.0488	0.0399	0.0420	0.0360	0.0200
0.1	1.0	0.1	0.1	0.0529	0.0480	0.0442	0.0409	0.0197
0.1	1.0	0.1	1.0	0.0505	0.0346	0.0443	0.0320	0.0141
0.1	1.0	1.0	0.1	0.0519	0.0507	0.0438	0.0433	0.0207
0.1	1.0	1.0	1.0	0.0490	0.0463	0.0440	0.0402	0.0202
0.5	0.1	0.1	0.1	0.0492	0.0468	0.0431	0.0393	0.0223
0.5	0.1	0.1	1.0	0.0506	0.0391	0.0409	0.0331	0.0256
0.5	0.1	1.0	0.1	0.0521	0.0484	0.0456	0.0430	0.0206
0.5	0.1	1.0	1.0	0.0479	0.0419	0.0405	0.0335	0.0238
0.5	1.0	0.1	0.1	0.0508	0.0502	0.0423	0.0393	0.0194
0.5	1.0	0.1	1.0	0.0521	0.0429	0.0434	0.0389	0.0244
0.5	1.0	1.0	0.1	0.0507	0.0500	0.0442	0.0424	0.0211
0.5	1.0	1.0	1.0	0.0492	0.0468	0.0431	0.0393	0.0223
0.9	0.1	0.1	0.1	0.0494	0.0467	0.0449	0.0436	0.0217
0.9	0.1	0.1	1.0	0.0490	0.0482	0.0391	0.0369	0.0212
0.9	0.1	1.0	0.1	0.0505	0.0434	0.0460	0.0397	0.0199
0.9	0.1	1.0	1.0	0.0474	0.0474	0.0416	0.0379	0.0215
0.9	1.0	0.1	0.1	0.0523	0.0522	0.0439	0.0421	0.0210
0.9	1.0	0.1	1.0	0.0500	0.0482	0.0425	0.0412	0.0205
0.9	1.0	1.0	0.1	0.0522	0.0468	0.0469	0.0411	0.0195
0.9	1.0	1.0	1.0	0.0494	0.0466	0.0449	0.0435	0.0217

For correlated errors in unequally spaced data, the AR(1) model is no longer reasonable

because the correlation between error terms would be assumed to be equal between successive observations, no matter their distance from each other. A more appropriate error structure for unequally spaced data is an exponential model (Schabenberger and Pierce, 2002) that takes into account the distance between measurements. The exponential model is called the power model in $SAS^{®}$. In the case where the data are equally spaced, the exponential model reduces to the AR(1) model (Schabenberger and Pierce, 2002). We set the parameters to be $m = 30$ and $n = 5$ and vary σ_0^2, σ_1^2, σ^2, and ρ. Table 5.2 shows the probability of signal for in-control data when

$$\mathbf{X}_i = \mathbf{Z}_i = \begin{bmatrix} 1 & 0 \\ 1 & .45 \\ 1 & .5 \\ 1 & .55 \\ 1 & 1 \end{bmatrix}.$$

This corresponds to 3 points near the center of the profile and a single point at each of the extremes. We performed this type of study for different types of unequally spaced profiles with less extreme spacing and found essentially the same conclusions as those to be shown here.

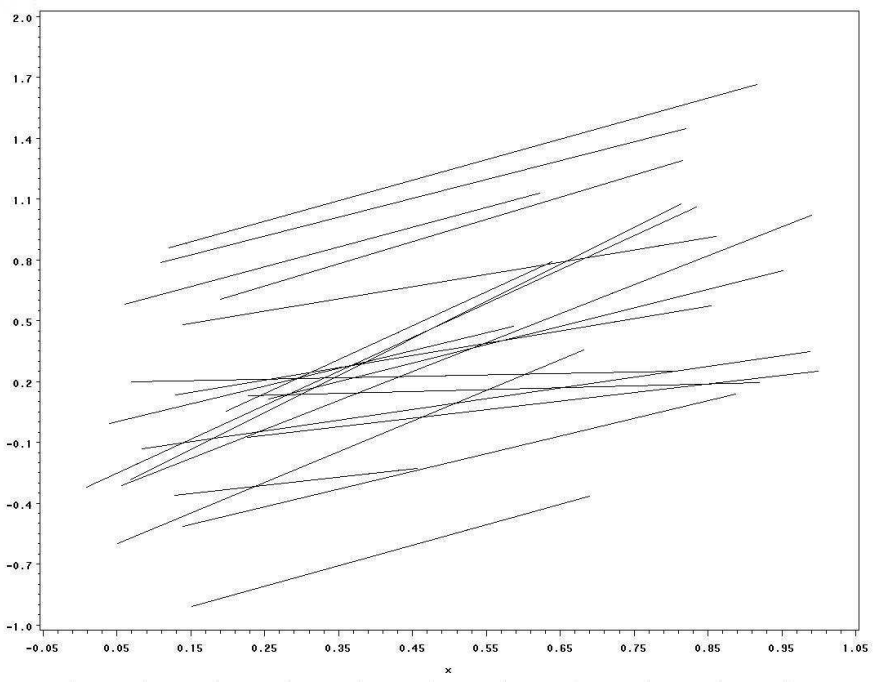
We see that similar to the balanced, equally spaced data scenario, the profiles with unequally spaced data will not result in a large difference between the LMM and LS approaches. This is true no matter the correlation in the errors or the variability of the random effects and errors. Power studies for this situation were performed with very similar results to those shown in Figure 5.2. Thus they are not shown here.

From the simulation study results we conclude that when the data are balanced (equally or unequally spaced) as will often be the case for control charts applications, there appears to be no advantage in modeling correlation and/or including random effects.

5.6 Unbalanced Data

Because we found little difference using LS and LMM for balanced data, we considered unbalanced data. This consists of cases where $\mathbf{X}_i$ and $\mathbf{Z}_i$, although equal to each other for the same profile, are different from profile to profile. For simplicity, we kept $n_i = n$ for all the profiles. Similar to the balanced data study, we considered different combinations of m, n, and ρ where $\sigma^2 = .1$, $\sigma_0^2 = .1$ and $\sigma_1^2 = .1$. The setup is the same as for balanced data but now, the locations along the profile where data are collected were randomly generated within a fixed interval. Once the locations were generated, they were held fixed for each of the Monte Carlo repetitions. Figure 5.3 shows an example of the data generated. It shows the separate simple linear regression fits to the randomly generated data for 20 profiles, each of which has 5 measurements.

Figure 5.3: Separate simple linear regression fit to an example of unbalanced data where $m = 20$, $n = 5$, and $\rho = .1$.



The $T^2_{1,i}$ and $T^2_{2,i}$ values are calculated from LMM and LS approaches and the control limit is obtained using the appropriate percentiles from the beta or χ^2 distributions. Table 5.3 shows the proportion of the $10,000$ datasets that had a signal on the control charts for the various T^2 statistics.

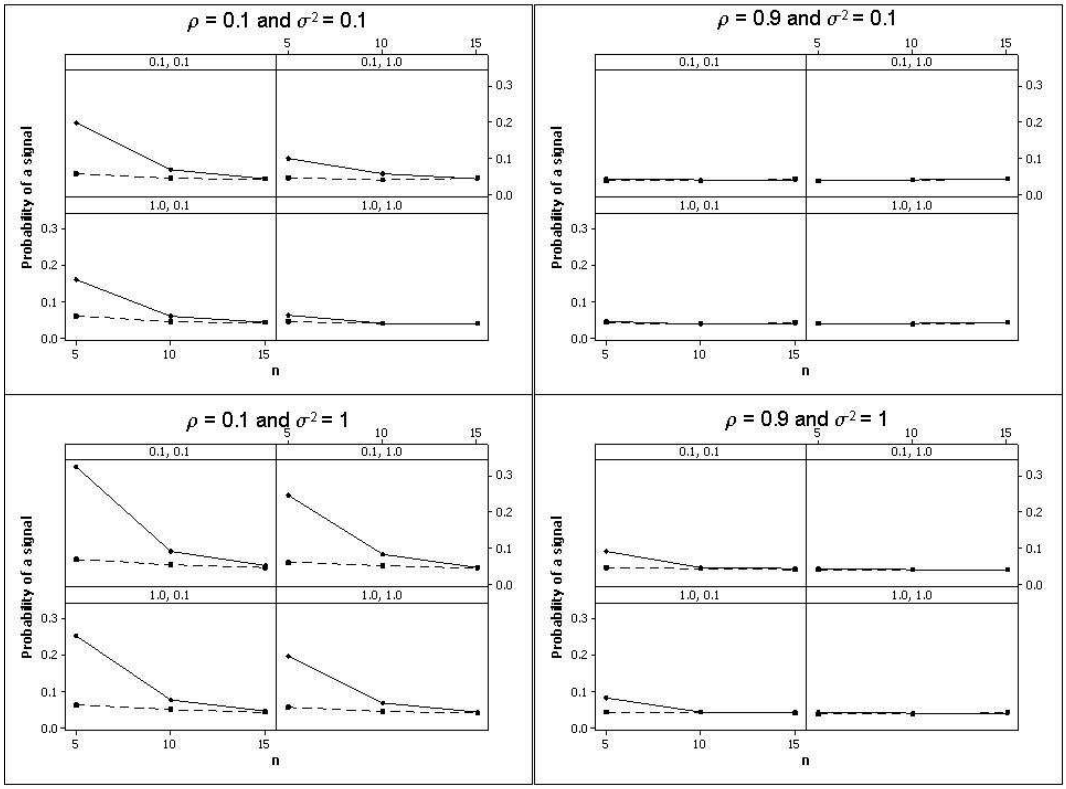
Table 5.3: Proportion of datasets with a signal for in-control data for unbalanced data scenario.

m	n	ρ	$T^2_{1,i,LS}$	$T^2_{1,i,MIX}$	$T^2_{2,i,LS}$	$T^2_{2,i,MIX}$	$T^2_{varbi,i}$
30	5	0	0.3746	0.0907	0.2635	0.0678	0.0199
30	10	0	0.0672	0.0604	0.0517	0.0467	0.0214
30	15	0	0.0543	0.0551	0.0436	0.0426	0.0203
30	5	0.1	0.1452	0.0712	0.0931	0.0563	0.0166
30	10	0.1	0.0608	0.0556	0.0441	0.0414	0.0142
30	15	0.1	0.0525	0.0542	0.0412	0.0393	0.0163
30	5	0.5	0.0895	0.0590	0.0587	0.0492	0.0202
30	10	0.5	0.0535	0.0511	0.0395	0.0412	0.0176
30	15	0.5	0.0508	0.0517	0.0401	0.0405	0.0206
30	5	0.9	0.0559	0.0505	0.0431	0.0426	0.0208
30	10	0.9	0.0509	0.0498	0.0391	0.0392	0.0190
30	15	0.9	0.0507	0.0511	0.0415	0.0412	0.0211
60	5	0	0.3452	0.0880	0.2772	0.0702	0.0288
60	10	0	0.0912	0.0602	0.0739	0.0507	0.0302
60	15	0	0.0653	0.0596	0.0522	0.0456	0.0292
60	5	0.1	0.2541	0.0764	0.1977	0.0585	0.0236
60	10	0.1	0.0852	0.0595	0.0701	0.0471	0.0254
60	15	0.1	0.0591	0.0545	0.0447	0.0433	0.0286
60	5	0.5	0.1282	0.0641	0.0927	0.0462	0.0262
60	10	0.5	0.0576	0.0504	0.0487	0.0437	0.0298
60	15	0.5	0.0549	0.0559	0.0436	0.0439	0.0307
60	5	0.9	0.0551	0.0514	0.0433	0.0399	0.0303
60	10	0.9	0.0485	0.0488	0.0414	0.0398	0.0306
60	15	0.9	0.0507	0.0536	0.0420	0.0450	0.0294
90	5	0	0.3937	0.1070	0.3524	0.0850	0.0337
90	10	0	0.1119	0.0727	0.0963	0.0598	0.0360
90	15	0	0.0631	0.0594	0.0545	0.0532	0.0359
90	5	0.1	0.2796	0.0799	0.2428	0.0665	0.0295
90	10	0.1	0.0943	0.0642	0.0798	0.0544	0.0350
90	15	0.1	0.0562	0.0516	0.0464	0.0431	0.0270
90	5	0.5	0.1282	0.0647	0.1070	0.0536	0.0316
90	10	0.5	0.0687	0.0583	0.0553	0.0495	0.0378
90	15	0.5	0.0503	0.0468	0.0434	0.0396	0.0305
90	5	0.9	0.0568	0.0525	0.0508	0.0483	0.0366
90	10	0.9	0.0545	0.0550	0.0468	0.0464	0.0363
90	15	0.9	0.0471	0.0466	0.0402	0.0389	0.0322

Here we see that in some situations that using the LS approach will result in a much larger probability of signal than the nominal value. This increased probability occurs when there is a smaller number of observations per profile ($n = 5$) and increases as the number of

profiles gets larger. It is also higher when the correlation in the errors is smaller. In contrast, using the LMM approach maintains the probability of a signal closer to its nominal .05 level. When the number of observations per profile is larger, there will be little difference between the LS and LMM approaches.

Figure 5.4: Probability of signal of in-control unbalanced data for ρ, σ^2, σ_0^2 and σ_1^2 for the $T_{2,i}^2$ statistic where $m = 60$. The solid line represents the probability of signal for the LS approach and the dashed line for the LMM approach. The smaller panel variables are σ_0^2 and σ_1^2 respectively.



In contrast with the balanced data scenarios of the previous section, the difference in the LS and LMM approaches depends on the values of ρ, σ^2, σ_0^2 and σ_1^2. To see this, consider Figure 5.4 which shows the probability of signal for various combinations of ρ, σ^2, σ_0^2 and σ_1^2 for the $T_{2,i}^2$ statistic obtained via the LS (solid line) and LMM (the dashed line). Here $m = 60$, the horizontal axis is n, and there are four larger panels that show various combinations of

ρ and σ^2. Within the larger panels are smaller panels which show the combinations of σ_0^2 and σ_1^2 respectively.

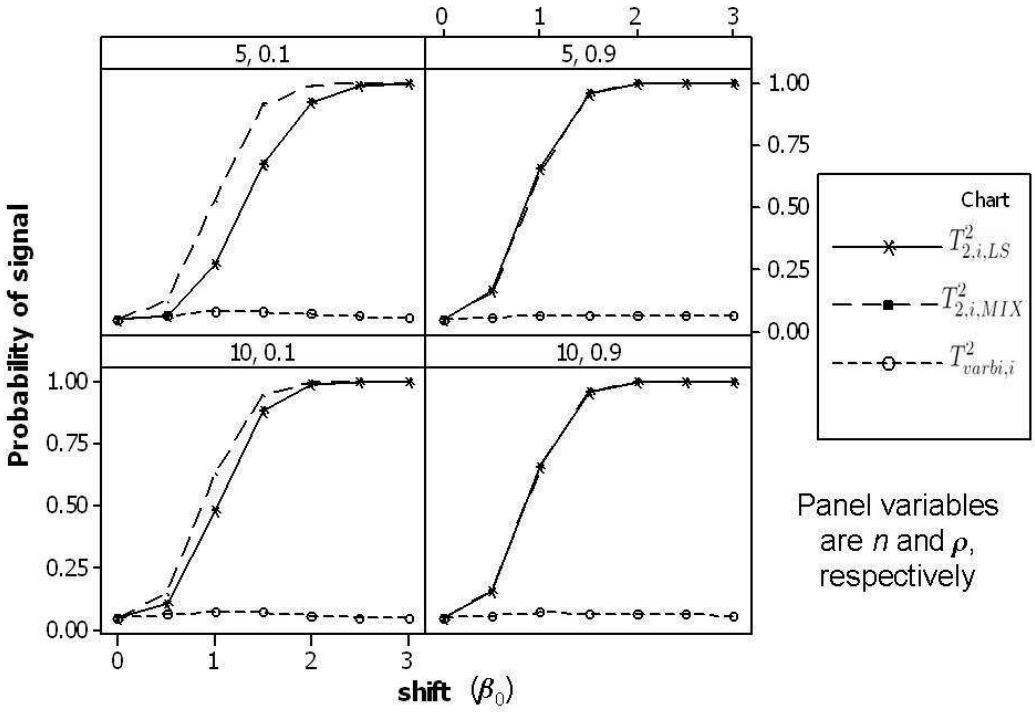
Here we see that the difference between the LS and LMM approaches is greater when the correlation is low, the variability in the errors is high, and $n = 5$. When $n = 10$ there will be slight differences in the LS and LMM approaches. At $n = 15$ there will be no difference regardless of the level of correlation or the variability in the random effects and errors. Thus the LMM approach preserves the appropriate Type I error probability of a signal even for an increased amount of error. Similar conclusions hold for other values of m and for methods based on $T^2_{1,i,MIX}$ and $T^2_{2,i,MIX}$ values.

In results not shown here, we have repeated multiple times the simulation study for the unbalanced data scenario with different sets of randomly generated $\mathbf{X}_i$ matrices. We found that the conclusions obtained from Table 5.3 and Figure 5.4 hold with a different set of randomly generated locations.

In order to do the power studies for this scenario, we simulated the control limit to ensure that the charts will have the same probability of signal for in-control data. Figure 5.5 shows some of the results of the power studies for unbalanced data where $m = 30$, $\sigma^2 = .1$, $\sigma_0^2 = .1$ and $\sigma_1^2 = .1$ and where the step change occurred after the fifth profile. Again we do not show the results for the charts based on $T^2_{1,i,LS}$ and $T^2_{1,i,MIX}$ values.

We see that just as for in-control data the LMM approach will be superior for smaller amounts of correlation than for a larger level of correlation of the errors. The larger the number of observations per profile, the smaller the difference will be between the LMM and LS approaches. The $T^2_{varbi,i}$ chart performs poorly, just as it did for the balanced data scenario.

Figure 5.5: Probability of signal of unbalanced data for n and ρ, for the $T^2_{2,i,LS}$, $T^2_{2,i,MIX}$, and $T^2_{varbi,i}$ charts where the step change in the intercept, β_0, occurred after the 5^{th} profile and where $m = 30$, $\sigma^2 = .1$, $\sigma_0^2 = .1$ and $\sigma_1^2 = .1$.



At first glance this results may seem counterintuitive. The two approaches perform similarly when the correlation in the errors is higher. Intuition suggests that since the LMM is taking into account the correlation of the errors, it would have higher probabilities of signal for correlated out-of-control data than the LS approach, which is the reverse of what our results show.

To explain this phenomenon, recall from Section 4.2 that an increased amount of correlation in the errors makes the profiles smoother and more similar to each other. As a result, both the LS and LMM approaches give similar results when the correlation is high.

Thus when the data are unbalanced, a mixed model approach will be beneficial, par-

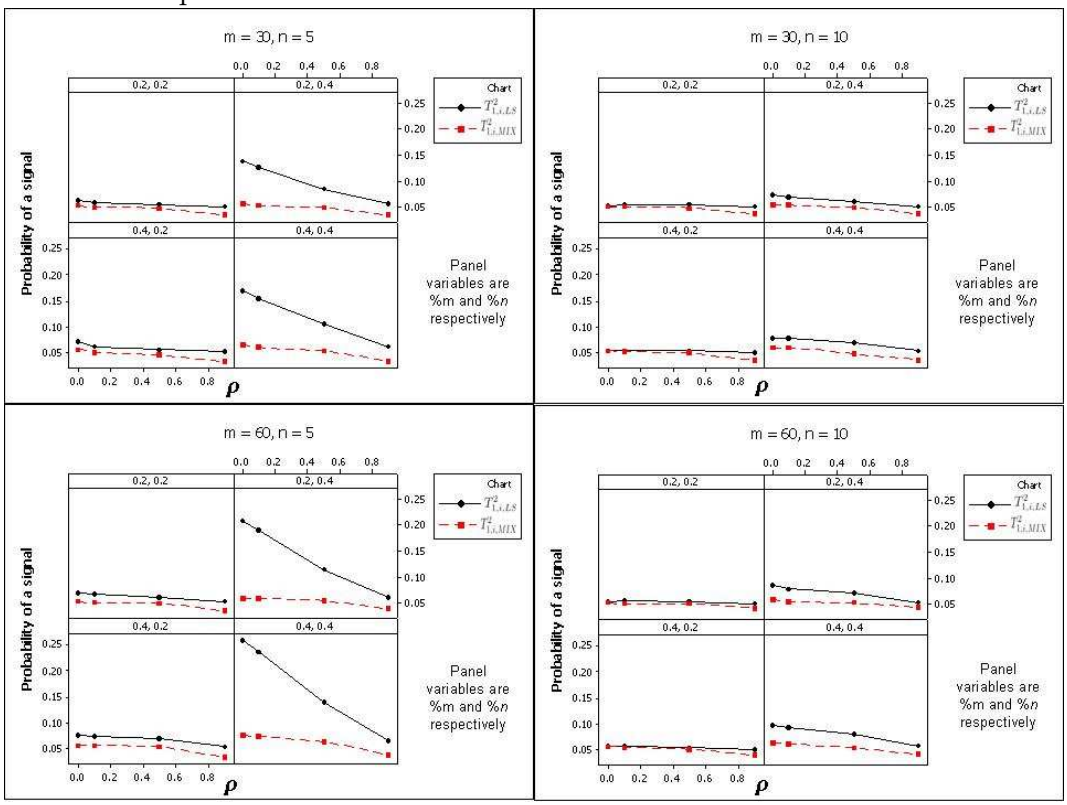
ticularly when the number of observations per profile and the amount of correlation are small.

5.7 Missing Data

The LMM approach has an advantage over the LS approach when there are missing data. Because the LMM approach pools information together from the profiles, it uses information from the profiles with full data to fit the profiles that have missing data points. The LMM approach can even be used to fit a curve to profiles that do not have enough data points to fit its own separate model. For example, profiles with only a single point could not be used when fitting separate simple linear regressions to each profile. Such profiles could be used in the LMM approach. While we do not advocate making a decision about whether or not a profile is outlying based on a single point, we do want to investigate the impact of missing data on the LMM and LS approaches.

We did a simulation study to evaluate the impact of missing data where the data within a profile is assumed to be missing at random (MAR). While this may be a simplistic assumption, it will serve here to illustrate the differences in the LMM and LS approaches. If the missing data are due to some underlying phenomena, for example, due to dropout in a longitudinal study, then the MAR assumption will not be met. More information on the different types of missing data can be found in Verbeke and Molenberghs (2000, Chapters 14-16) and in Little and Rubin (1987). Likelihood based inference is still valid when the data are MAR so no changes are needed in the SAS coding (Vonesh and Chinchilli, 1997, p. 264). Studying the impact of missing data in profile monitoring requires consideration of both the proportion of profiles that have missing observations, referred to here as %m,

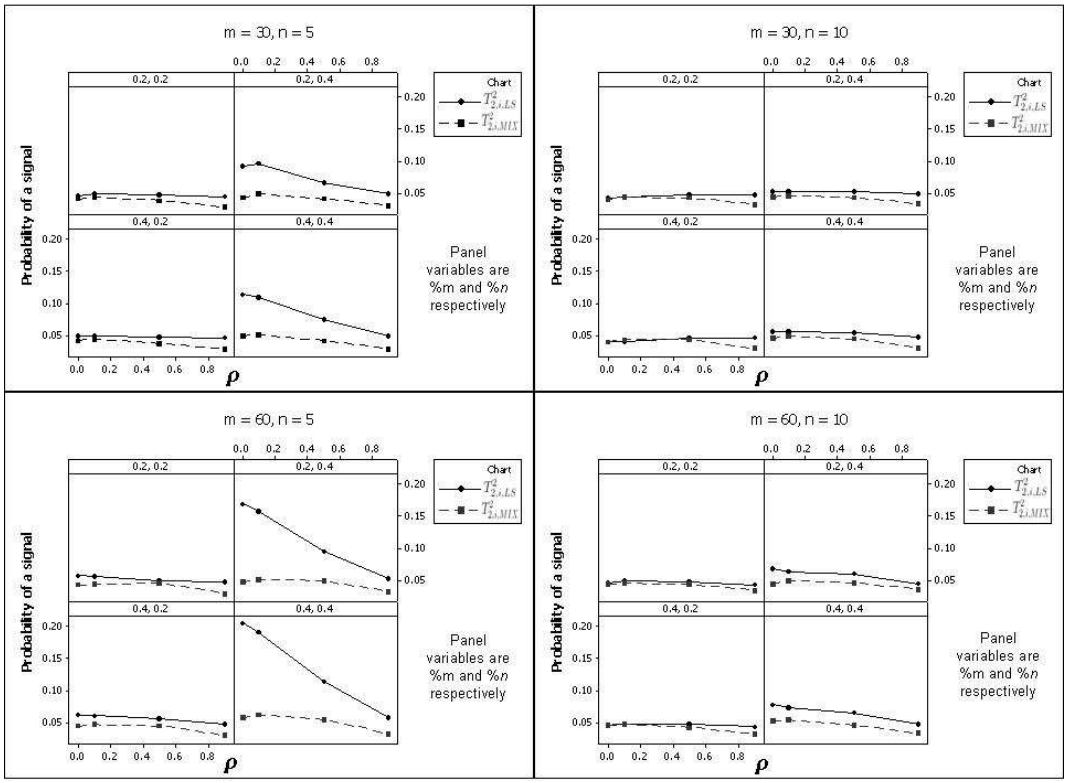
Figure 5.6: Probability of signal for $T^2_{1,i,LS}$ and $T^2_{1,i,MIX}$ when data are missing at random from the in-control profiles.



and the proportion of observations missing within the profiles, referred to here as %n. For example, if $m = 30$, $n = 5$, $\%m = .2$, and $\%n = .4$, there will be 6 profiles that have missing observations and for those 6 profiles, there will be two observations missing for each profile.

We considered the balanced equally spaced data scenario where the data were first generated, then a subset of profiles was selected at random, and then the missing observations were selected at random locations for that subset of profiles. The missing observations occur at different points for the profiles. We show here the results for $\sigma_0^2 = .1$, $\sigma_1^2 = .1$, and $\sigma^2 = .1$ with the control limit obtained from the corresponding beta and chi-square distributions. We considered several values of ρ for the correlation in the errors and also included the case where $\rho = 0$. We examined two levels of %m and %n, .2 and .4.

Figure 5.7: Probability of signal for $T^2_{2,i,LS}$ and $T^2_{2,i,MIX}$ when data are missing at random from the in-control profiles.

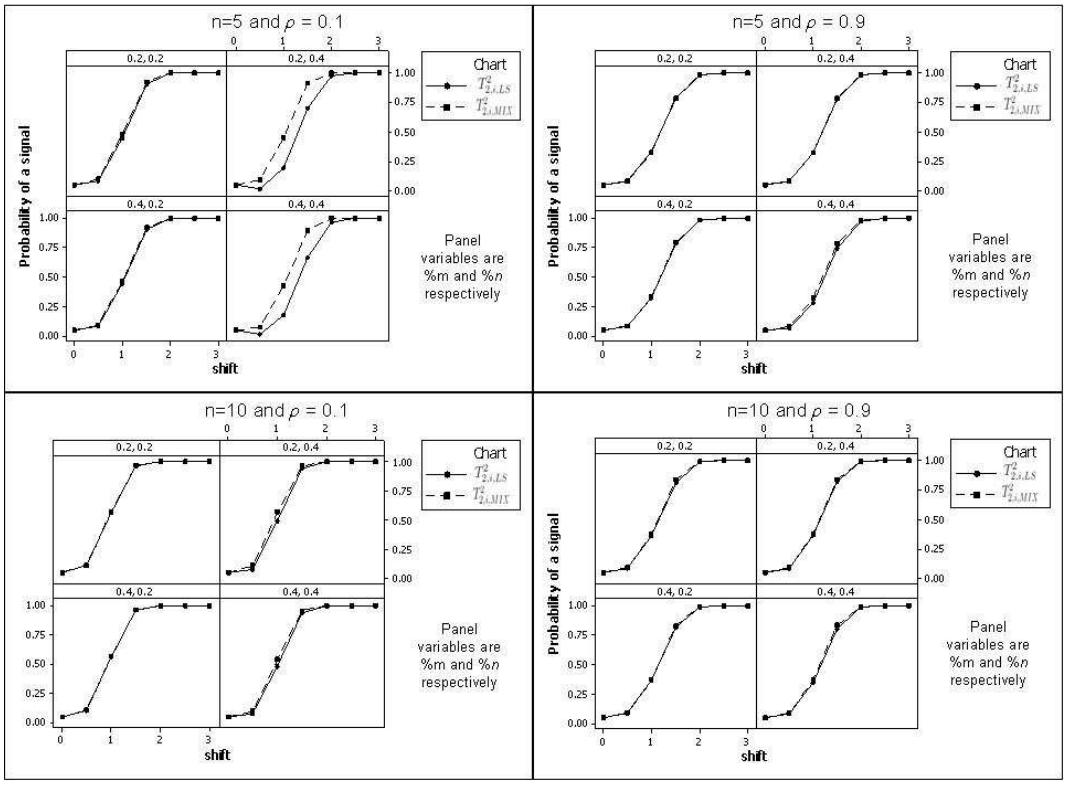


The probability of signal is shown in Figures 5.6 and 5.7 for data that comes from an in-control process. We see that use of the T^2 values give a higher probability of signal than the nominal .05 level when $n = 5$ and %n is larger. The larger the number of observations per profile then the less drastic will be the impact of missing data. An increase in %m only has a minor impact on the probability of signal. As the number of profiles increases the probability of a signal (frequency of a false alarm) increases.

As we did for other data scenarios, we considered the performance for out-of-control data. As before, we introduce a step change in the intercept and compare the probability of a signal for the methods based on $T^2_{2,i,LS}$ and $T^2_{2,i,MIX}$ values because of their ability to detect step changes. Figure 5.8 shows the probability of a signal where $m = 60$, $\sigma_0^2 = .1$, $\sigma_1^2 = .1$, and

$\sigma^2 = .1$ for different combinations of ρ, n, $\%m$ and $\%n$.

Figure 5.8: Probability of signal for n and ρ where out-of-control data are missing at random, for the $T^2_{2,i,LS}$ and $T^2_{2,i,MIX}$ charts where $m = 60$, $\sigma^2_0 = .1$, $\sigma^2_1 = .1$, and $\sigma^2 = .1$.

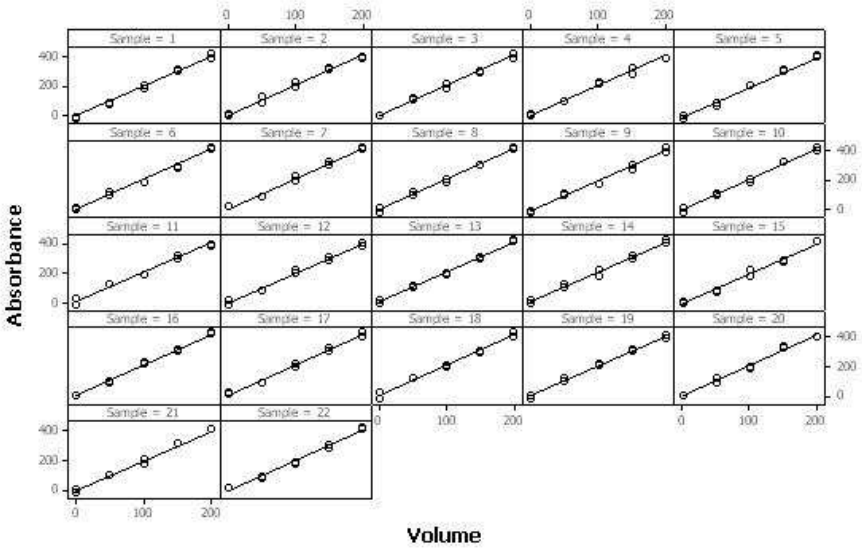


Just as we observed for the in-control data, the biggest difference between the LMM and LS approaches for MAR data occurs when the correlation in the errors is low, the number of observations is small, and when the proportion of missing data is larger. Earlier we noted that for unbalanced data that the difference between the LMM and LS approaches is greater when there is an increasing amount of variability in the errors. This is also true here. While not shown in Figure 5.8, when σ^2 increases, the LMM approach shows greater superiority over the LS approach.

5.8 Example

To illustrate the control chart procedures discussed here we use the calibration dataset first analyzed in Mestek, Pavlik, and Suchánek (1994) and later analyzed in Mahmoud and Woodall (2004). The data consist of 22 calibration curves each of which relates an absorbance measure of a chemical solution to the volume at which the solution was prepared. The purpose is to determine if the calibration curves are stable over time. There are 5 volumes and 2 replicate measurements for each volume so each profile has a total of 10 measurements. The raw data profiles for the calibration data along with their simple linear regression fits are shown in Figure 5.9. The calibration curves are very similar to each other and have more variability in the intercepts than in the slopes. These data are balanced, equally spaced, and have no missing observations.

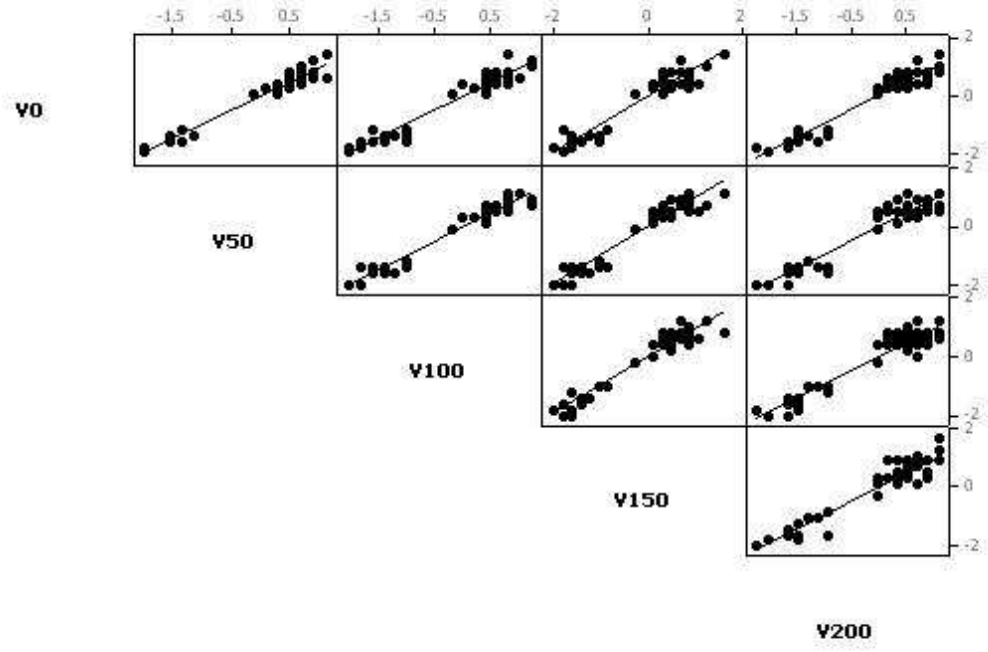
Figure 5.9: The raw data and the fitted linear regression lines for the 22 profiles from the calibration dataset. Each profile had 2 measurements at each of 5 locations.



We first investigated to see if the measurements within a profile are correlated across

the different volumes. To determine the appropriate correlation structure we employed the graphical methods discussed in Dawson, Gennings, and Carter (1997). After centering and scaling the data by volume we obtained the draftman's display shown in Figure 5.10. The draftman's display is a plot of the residuals at volume j versus those at volume $j+1$, for $j = 0 : 200(50)$.

Figure 5.10: Draftman's display of the calibration data showing a compound symmetry correlation structure.



Based on the examples shown in Dawson, Gennings, and Carter (1997), we conclude that the calibration dataset has a compound symmetry (CS) structure. This is evident from the positive linear trend in the individual scatterplots on the draftman's display of Figure 5.10. The strength of the trend is consistent for the plots closest to the diagonal and for the plots in the upper right hand corner. If the strength of the trend were to weaken or decrease for the plots further away from the diagonal, then we would have concluded that the calibration

dataset had errors that followed an AR structure.

Figure 5.11: T^2 charts for the profiles of the calibration dataset that have been fit by the LS approach.

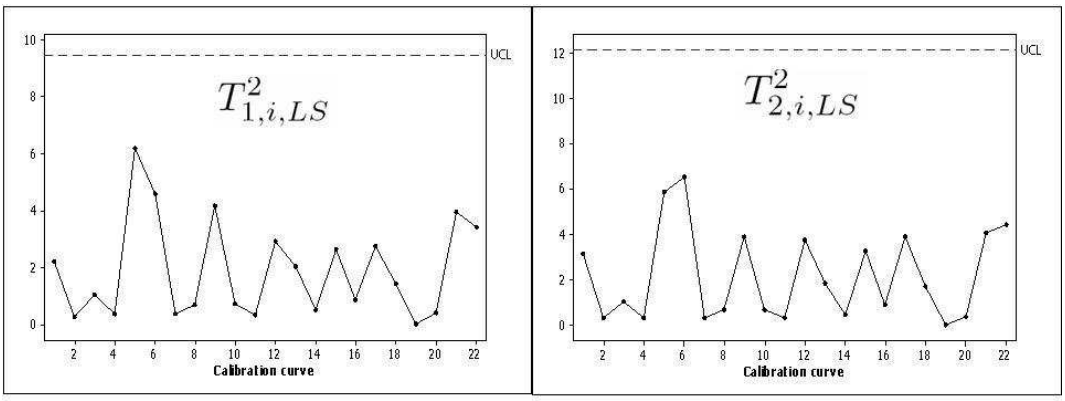
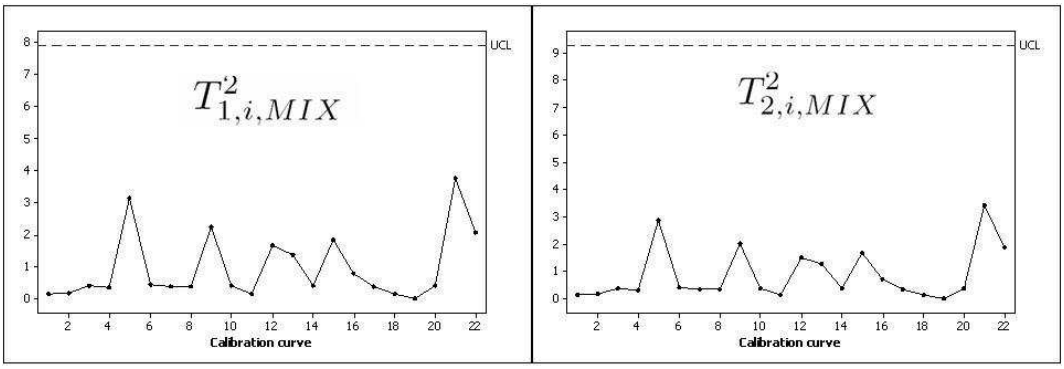


Figure 5.12: T^2 charts for the profiles of the calibration dataset that have been fit by the LMM approach.



We then fit the profiles with the LS and LMM approaches. Figure 5.11 shows the T^2 charts for the LS approach. Figure 5.12 shows the T^2 charts for the LMM approach. When fitting the LMM, we found that the only necessary random effect was that for the intercept. A likelihood ratio test of a random effect on the slope leads us to conclude that a random effect is not needed for the slope because the slopes are so similar to each other. In addition, once the random effect for intercept is included, the estimated errors are no longer correlated

with each other, thus we can safely analyze the data with the assumption of independent errors.

We see that for this dataset, the LS and LMM approaches give very similar results. There are no signals, suggesting that the 22 calibration curves come from an in-control process. All of them can be used to set the control limit for Phase II. We did not expect to have any drastic differences in the LS and LMM approaches because of the results of our simulation studies showed that there is little difference for balanced, equally spaced data as we have here for the calibration data.

5.9 Conclusions

To summarize the results of this chapter, we have found that in all the Phase I scenarios investigated the LMM approach has either equivalent or superior performance when compared to the LS approach. When the data are balanced, there is little difference between the two approaches but we have found using simulation that the advantage of the LMM over the LS approach is greatest when the data are unbalanced or when there are missing observations. For unbalanced or missing data, the LMM is better, especially for smaller levels of correlation and for a smaller number of observations per profile. Thus our results concur with the conclusions of Verbeke and Molengberghs (2000), although they were not considering profile monitoring.