Common Industrial Applications of Mixed Models

Bob Obenchain, Eli Lilly

ABSTRACT

The two types of linear models with random coefficients that I use most frequently in pharmaceutical applications are (i) models for within and between batch variation and (ii) random slopes regression models. Although very simple conceptually, these basic models illustrate some fundamental distinctions and computational difficulties that have fueled controversies over fixed versus random coefficient models for at least the last 25 years.

1. HISTORICAL PERSPECTIVE. The following are a number of quotations from Frank Yates' June 1968 presidential address to the Royal Statistical Society [Yates(1968), Section 2, Irrelevance of Much Present-Day Statistical Literature]:

"Only when radically new situations arise are extensions of theory necessary."

"A publishes a paper which is irrelevant, or based upon unrealistic premises. This stimulates B and C to further thoughts on the same lines. And in a few years a body of literature is built up."

"In case you think I exaggerate, let me give you an example, that of 'fixed' and 'random' effects in experimental design. This dichotomy has now a considerable vogue in America, and is spreading to this country."

"Differences in formulae, arising from differences in definition, soon intruded, and before long it was represented that tests of significance which could be correctly applied would differ for the two models."

Twenty five years ago, it may not have been clear that Yates was "off-base" in his assessment of potential roles for mixed models. But statistical literature on random effects has certainly continued to expand over the last quarter century. Some important papers in this field are:

Rao(1971a, 1971b)	...MINQUE and MIVQUE estimates of variance components;
Patterson and Thompson(1971)	...REML estimation;
Henderson(1975, 1984, 1990)	...mixed model equations and BLUP theory;
Searle(1971, 1988)	...linear models and variance components;
Harville(1977, 1988, 1990)	...maximum likelihood theory and algorithms; prediction;
Robinson(1990)	...BLUP review article; and
McLean, Sanders and Stroup(1991)	...broad and narrow inference spaces.

From today's perspective, we have every right to think that mixed models represent a truly important and practically useful extension to linear models theory. In fact, we can be proud of this distinctly **Made-In-America** flavor of modeling!

What are the particular strengths and weaknesses of random coefficient methods relative to traditional fixed effect methods?

> Random coefficient estimates have the weakness that they have to be found via numerical search algorithms. Completely general, closed form expressions for, say, maximum likelihood variance component estimates apparently don't (and cannot) exist. Therefore, random coefficient analyses are frequently much more computationally intensive than traditional approaches; they can end up consuming more computer time and resources than, say, least-squares analyses. Furthermore, random coefficient algorithms can be sensitive to start-up values and/or fail to converge!
>
> The strengths of the new, random coefficient methods are that they can not only lead to point estimates and statistical inferences which differ from those given by classical ANOVA but also that they can be more realistic in certain situations. Differences between fixed and random approaches are, perhaps, most likely to surface when one's data are **unbalanced**. In fact, some users view random coefficient methods as a sort of panacea for "missing" response problems, claiming that this approach allows them to analyze "messy" datasets without making any extra, unnecessary assumptions. Others view random coefficient methods as establishing a **common language** for describing and inter-relating classical and Bayesian approaches to linear models, Kalman filtering, Kreiging, shrinkage estimation, etc., etc.

The emergence of statistical software supporting mixed model estimation has played a key role in getting industrial statisticians interested in mixed models. The first widely used commercial product allowing random coefficients for continuous variables (as well as for categorical effects) was BMDP® program 5V, Jennrich and Schluchter(1986). Today, these sorts of mixed models are also treated by version 1.1 of the GLMM©, system for General Linear Mixed Models on IBM-compatible personal computers, Blouin and Saxton(1990), as well as by SAS/STAT® procedure MIXED.

For industrial statisticians wishing a good introduction to mixed model theory, including a wide range of illustrations of potential applications, I heartily recommend the review article by Wolfinger, Tobias, and Sall(1991.) Here, I will only have space to discuss two extremely simple special cases: (i) models for within and between batch variation and (ii) random slopes regression models.

2. WITHIN-BATCH and BETWEEN-BATCH VARIATION. Consider the situation where one or more measurements of final potency or yield are made on each batch of product produced in a certain step of a stable manufacturing process, i.e. a process that is "in control" statistically. Furthermore, suppose that the next manufacturing step is a mixing or dilution step that requires an estimate of the true potency or yield from that previous step.

This situation arises, for example, in filling and finishing operations in the pharmaceutical industry. Our analytical testing is not only technically complex and time consuming but is usually also destructive of the test sample. To avoid any variation due to within-batch product heterogeneity, representative samples from different parts of each batch are usually composited; assay samples are then taken from this composite.

Even when assaying from a composite, estimated measurement variation can still be much larger than estimated batch-to-batch product variation. Testing, using independent replicates, usually continues until a decision can be reached about whether a batch is within certain in-house limits. As a result, different numbers of replicates may be performed on different batches. Electronic laboratory information systems may store only the average assay result for each batch; results from individual replicates may be recorded only in laboratory notebooks.

2.1 Mixed Model. The model most appropriate for this sort of situation seems to be of the form:

$$A_{ij} = \pi_i + \alpha_{ij} ,$$

where A_{ij} is the result of the j-th replicate assay on the i-th batch; π_i, is the unknown true average potency of the i-th batch; α_{ij} is the assay error in the j-th replicate on the i-th batch; the π_i, are random effects with a fixed, unknown mean, $E(\pi_i) = \pi$, and a constant, unknown variance, σ_B^2 ; and the α_{ij} are random effects with mean 0 and constant, unknown variance, σ_W^2 . [Equivalently, the π fixed effect could be added to the right-hand-side of the above model equation; the random π_i would then have mean zero.] Now suppose that the within-batch variance, σ_W^2 / R, of the average assay over R independent replicates is larger than the between-batch variance, σ_B^2 . In this situation, a long-range average assay value, $\overline{A}_{..}$, can actually be a more precise predictor of the random π_i, than is the average assay on the i-th batch, $\overline{A}_{i.}$. In other words, individual batch predictions are best "shrunken" all of the way to their long-range average.

A somewhat more conservative strategy, called "standard release" potency estimation, is used at Eli Lilly when process capability studies reveal that σ_W^2 / R is estimated to exceed σ_B^2 . Specifically, whenever $\overline{A}_{i.}$ is within, say, $\pm 2 \cdot \sigma_W / \sqrt{R}$ of the long-range average assay, $\overline{A}_{..}$, the batch is assigned a standard potency of $\overline{A}_{..}$ and the next manufacturing step can use a standard setup. But, if $\overline{A}_{i.}$ is not within $\pm 2 \cdot \sigma_W$ of $\overline{A}_{..}$, then that batch is assigned its measured potency of $\overline{A}_{i.}$ and the next manufacturing step must use a customized setup.

2.2 Numerical Example. Getting reliable estimates of the within-batch and between-batch variance components, σ_W^2 and σ_B^2, , is clearly the key step in developing a potency prediction strategy. Suppose, then, that the only data available to estimate σ_W^2 and σ_B^2, are the following.

Batch 1:	3 Assay Replicates	5.9, 4.2, and 7.1
Batch 2:	Average of 3 Replicates	5.6
Batch 3:	4 Assay Replicates	4.8, 4.2, 5.1, and 5.7
Batch 4:	Average of 2 Replicates	6.8
Batch 5:	3 Assay Replicates	3.8, 4.2, and 5.7
Batch 6:	2 Assay Replicates	6.4 and 5.5

Suppose we place our data in a SAS dataset named ONEWAY with 14 observations on 3 variables: BATCH, ASSAY, and WGT. BATCH is a class variable with values 1 through 6 and WGT=1 for all observations except WGT=3 for ASSAY=5.6 and WGT=2 for ASSAY=6.8.

SAS/STAT procedure GLM and JMP® software use a "method of moments" approach to estimate variance components associated with random effects. Mean squares are equated to their expected values to estimate between and within batch variance components. Unfortunately, this approach does nothing to prevent the between-batch variance estimate from going negative. In fact, this estimate is highly likely to be negative whenever σ_B^2 is relatively small.

Because BATCH is a class variable, we could use SAS/STAT procedure VARCOMP. Unfortunately, these results will be biased because VARCOMP does not allow us to use weights for observations.

```
proc glm data=oneway;
  class batch;
  model assay = batch;
  random batch;
  weight wgt;
run;
```

```
proc varcomp method=mivque0 data=oneway;
  class batch;
  model assay = batch;
run;
```

Thus SAS/STAT procedure MIXED provides the most appropriate (restricted maximum likelihood and mivque0) analyses of the most realistic model for the above data. In situations where the "method-of-moments" estimate for the between-batch variance component is negative, the reml and mivque0 estimates are frequently zero rather than being strictly positive. But this still represents at least a small step in the "right" direction!

```
proc mixed method=reml data=oneway;
  class batch;
  model assay = ;
  random batch;
  weight wgt;
run;
```

Here are numerical results for our simple example:

Variance Component Estimates:	GLM or JMP:	VARCOMP: mivque0 (biased)	VARCOMP: reml (biased)	MIXED: mivque0	MIXED: reml
Between Batch	0.207	0.0694	0.0825	0.140	0.183
Within Batch	0.979	0.989	0.979	1.051	1.010

Our conclusions for this small numerical example are that one would probably have to average more than 4 true replicate assays from this process to get better per batch estimates than the long range average assay.

3. REGRESSIONS with RANDOM INTERCEPTS and SLOPES. By far the most important application of regression models in the pharmaceutical industry is to stability prediction (also called shelf-life estimation or expiration dating.) In fact, the US Food and Drug Administration (FDA) has published detailed guidelines (1987) on how to use least-squares regression lines and confidence bands in this context. Interest in using random coefficient models in stability studies is relatively recent; see Murphy and Weisman(1990) and Chow and Shao(1991). In order to display statistical models for the time variation in assay measurements from several different batches of a pharmaceutical product, we will need the following notation:

B = Number of distinct batches under study. All of these batches are assumed to include the same, single dosage form; to be in identical packaging; and to have been maintained under identical temperature/humidity storage conditions since packaging.

i = Batch number index: $1 \le i \le B$.

T_i = Number of distinct time points at which samples of product from batch i were assayed.

t = Time point index: $1 \le t \le T_i$.

x_{it} = Shelf-age of product from batch i at time-point t. (x_{i1} is usually 0.)

n_{it} = Number of replicate assays performed on batch i at time-point t.

j = Assay index: $1 \le j \le n_{it}$.

y_{itj} = Result from assay j on batch i at time-point t. Assay results are frequently expressed in terms of the amount of active ingredient detected (milligrams or micrograms) or as a percentage of label-claim; linear fits to these units imply zero-order degradation kinetics. But, when y represents the logarithm of the amount or percentage, linear fits correspond to first-order degradation kinetics.

The simple regression model is then...

$$y_{itj} = \mu_i + x_{it} \cdot \beta_i + \varepsilon_{itj}$$

where

μ_i = unknown initial potency of batch i (i.e. at zero shelf-age.)

β_i = unknown (linear) degradation rate in potency for batch i.

ε_{itj} = unknown error term for assay j at time-point t within batch i.

3.1 Classical Fixed-Effect Estimates and Confidence Bands. A common practice in the pharmaceutical industry is to form mean values over all assay replicates taken at each time-point for a batch, $\bar{y}_{it.} = \sum y_{itj} / n_{it}$, and to fit regression lines using only these mean values. In reality, the numbers of assay replicates, n_{it}, usually vary between time-points within a batch (more replicates being taken at, say, time zero and at two years of shelf-

age than at intermediate quarterly assays.) On the other hand, it is also common practice to ignore the resulting heteroscedasticity of assay averages.

Note also that we are not really estimating B separate, unrelated simple-linear-regressions. We are pooling assay means from B batches into one regression with 2×B regression coefficients [an initial potency(intercept) and degradation rate(slope) for each of the B batches] and a **common** error variance. In this pooled model, the 1987 FDA guideline "Test for Batch Similarity" is a F-test of the general linear hypothesis: $\beta_1 = \beta_2 = \ldots = \beta_B$ at the 25% significance level suggested by Bancroft(1964). [The numerator degrees-of-freedom of this F-test are $B-1$ and its denominator degrees-of-freedom are $T_1+T_2+\ldots+T_B-2\cdot B$.]

FDA guidelines also essentially define the allowable shelf-life of a pharmaceutical product to be the length of time within which a one-sided or two-sided 95% confidence band for a fitted linear regression line deviate from the product's stated label-claim on potency by no more than 10% of that label-claim.

Suppose that the batch-similarity-hypothesis proves to be acceptable not only for degradation slopes but also for initial potencies. Under these circumstances, FDA guidelines allow us to use ordinary, un-weighted least-squares methodology to fit a single line to (mean) assay values. A two-sided confidence band would then be of the following form at time x_0 :

$$\hat{\mu}+\hat{\beta}\cdot x_0 \pm \tau_{N-2}^{\alpha/2}\cdot\hat{\sigma}_\varepsilon\cdot\sqrt{\frac{1}{N}+\frac{(x_0-\bar{x}_{\cdot\cdot})^2}{\sum_i\sum_t (x_{it}-\bar{x}_{\cdot\cdot})^2}}$$

where $N = T_1+\ldots+T_B$ is the total number of assays used to estimate the regression, $\hat{\sigma}_\varepsilon$ is the square root of the pooled residual-mean-square for error, and $\tau_{N-2}^{\alpha/2}$ is the tabulated upper $100(1-\alpha/2)\%$ point of Student's t-distribution with degrees-of-freedom$= N-2$.

3.2 Nested Error Structures. Unfortunately, the multiple assays run at each time point in stability studies are rarely true replicates; see Fuller and Battese(1973), Carstensen and Nelson(1976), and Norwood(1986). Instead of being independent, assay errors typically consist of a sum of two components:

$$\varepsilon_{itj} = \nu_{it} + \xi_{itj} \quad ,$$

where

ν_{it} = unknown assay "setup" effect at time-point t within batch i. This "deviation" from linearity is due to one or more of the following potential causes: cleaning of equipment, wear/aging of equipment, changes in solvent/reagent batches, recalibration, changes in weather or environment, changes in operators/technicians, etc. These setup effects are assumed to be random with mean zero and unknown, constant variance σ_ν^2 .

ξ_{itj} = unknown error term for assay j within setup number t for batch i. These error terms are assumed to be random with mean zero and unknown, constant variance σ_ξ^2.

In the pharmaceutical industry, this error structure is commonly called "nested errors." In SAS/STAT procedure MIXED, this error structure is described in the REPEATED statement as TYPE=CS (compound symmetry) and SUBJECT=BATCH.

3.3 Random Effects Models. The great advantage of the random coefficient approach to stability modeling is that any batch-to-batch variation in intercepts and slopes is an integral part of the model. Suppose that all correlations between random effects (other than that between the intercept and slope within an individual batch) are assumed to be zero. Because the random effect prediction equation is of the form

$$\hat{y}_{it.} = \hat{\mu}_i + x_{it} \cdot \hat{\beta}_i + \hat{v}_{it} + \hat{\xi}_{it.}$$

with nested errors and correlation ρ between $\hat{\mu}_i$ and $\hat{\beta}_i$, the variance of prediction will be

$$\begin{aligned} V(\hat{y}_{it.}|x_{it.}) &= \begin{bmatrix} 1 & x_{it.} \end{bmatrix} \begin{bmatrix} \sigma_\mu^2 & \rho\sigma_\mu\sigma_\beta \\ \rho\sigma_\mu\sigma_\beta & \sigma_\beta^2 \end{bmatrix} \begin{bmatrix} 1 \\ x_{it.} \end{bmatrix} + \sigma_v^2 + \sigma_\xi^2 / n_{it} \ , \\ &= \sigma_\mu^2 + 2 \cdot \rho \cdot \sigma_\mu \cdot \sigma_\beta + \sigma_\beta^2 + \sigma_v^2 + \sigma_\xi^2 / n_{it} \ , \\ &= \sigma_\mu^2 \cdot (1-\rho^2) + \sigma_\beta^2 \cdot (x_{it} + \rho \cdot \sigma_\mu / \sigma_\beta)^2 + \sigma_v^2 + \sigma_\xi^2 / n_{it} \ . \end{aligned}$$

Note that the variance of $\hat{y}_{it.}$ changes with x_{it} and is a minimum at $x_{it} = -\rho \cdot \sigma_\mu / \sigma_\beta$, which is a positive shelf-age in the common situation where ρ is negative. When intercepts and slopes are uncorrelated, $\rho = 0$, the random coefficient confidence band will always be most narrow at time zero, $x_{i1} = 0$.

In SAS/STAT procedure MIXED, correlated intercepts and slopes are described in the RANDOM statement as TYPE=UN (unstructured) and SUBJECT=BATCH. Uncorrelated intercepts and slopes are the default type but can also be specified as TYPE=SIM (simple).

3.4 Numerical Examples. Ruberg and Stegeman(1991) present a pair of numerical examples which demonstrate that current FDA guidelines can potentially reward stability protocols that generate sparse data and/or relatively imprecise assay results. Each example consists of only average assay results on B=6 batches. In the first dataset of N=37 averages, assay results are relatively "consistent" in the senses that [1] assay averages at different shelf-ages lie fairly close to the regression line for their batch and [2] the six different regression lines lie fairly close to each other. In the second dataset of N=35 averages, assay results are relatively "weak" in the senses that [1] individual

averages are much more variable than in the first dataset and [2] the six different regression lines have widely differing slopes.

Estimated Degradation Slopes (%/year)	Consistent Data: Fixed Effects	Consistent Data: Random Effects	Weak Data: Fixed Effects	Weak Data: Random Effects
Batch 1	-1.515	-1.516	-0.109	-0.201
Batch 2	-1.449	-1.450	-0.449	-0.397
Batch 3	-1.781	-1.742	-0.778	-0.606
Batch 4	-1.393	-1.481	+0.194	-0.331
Batch 5	-1.999	-1.729	-2.218	-0.587
Batch 6	-1.701	-1.706	-1.045	-0.539
Variance Components:				
Intercepts		0.0485		1.6924
Slopes		0.0301		0.1018
Measurement Error	0.0492	0.0128	0.7947	0.7713
Test for Equality of Batch Slopes:				
Observed Significance Level	0.186	0.086	0.359	0.806
Standard Error of Prediction:				
Age Zero	0.083 %	0.106 %	0.364 %	0.575 %
1 Year	0.058 %	0.123 %	0.260 %	0.572 %
2 Years	0.085 %	0.189 %	0.398 %	0.671 %
4 Years	0.191 %	0.358 %	0.898 %	1.033 %

In the above table, fixed-effect estimates were computed with SAS/STAT procedure GLM while mixed-model estimates come from SAS/STAT procedure MIXED. When asked to estimate the correlation between intercepts and slopes, MIXED iterations failed to converge; the above random-effect estimates result from assuming $\rho = 0$.

Note that random-effect estimates for the "consistent" dataset are rather close to their fixed-effect counterparts. But the random-effect slopes for the "weak" dataset are "shrunken" towards their estimated mean of -0.443; actually, the slope of Batch 2 "over-shoots" this mean, changing from -0.449 to -0.397. Furthermore, by switching from fixed-effect analyses to random-effects inferences, the "consistent" batch slopes become statistically **more** different from each other while the "weak" batch slopes become statistically **less** different from each other.

Ruberg and Stegeman(1991) point out the FDA guidelines for testing batch similarity at the 25% significance level allow pooling in the "weak" dataset but not in the "consistent" dataset. Here we see that this same conclusion would result from testing for equality of random-coefficients. But there is **no need for any such test** in the random-coefficient

approach because, again, batch-to-batch variation in intercepts and slopes is an integral part of the random-coefficient model.

Note also that random-coefficient confidence bands are consistently wider than fixed-effect bands. This will be especially true, of course, near the average shelf-age because the fixed-effect prediction standard error is a minimum at his point (approximately 1 year for both datasets.). Curiously, procedure MIXED gives a slightly smaller random-coefficient prediction standard error estimate (0.572%) at shelf-age 1 year than at time zero (0.575%) for the "weak" dataset; this should not happen when the intercept-slope correlation is zero.

4. SUMMARY. Because much analytical testing in the pharmaceutical industry is destructive and, thus, highly variable, the traditional expected mean squares approach to estimation of between-batch variance components in one-way ANOVA frequently gives negative estimates. Modern likelihood approaches not only avoid this embarrassment but also easily handle cases where data are unbalanced or missing.

When the random slopes and intercepts of simple linear regression models are assumed to be statistically independent, the resulting confidence bands will be most narrow at $x = 0$. On the other hand, modern software systems will at least attempt to estimate this intercept-slope correlation from the data. And a negative correlation estimate would place the most narrow point on the resulting confidence band at a positive x. This is much more like the fixed coefficient band that always has its most narrow point at the average x .

Using the random-coefficients approach (rather that classical, fixed effect formulas) to generate regression confidence bands for stability studies appears to be a step-in-the-right-direction for a number of good reasons. This approach makes "preliminary testing of batch similarity" moot and, thus, would greatly simplify not only administration of stability protocols but also interpretation of results. And this random batches approach would genuinely discourage protocols that produce sparse data and/or relatively imprecise assay results.

REFERENCES

Bancroft, T. A. (1964). "Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance," **Biometrics**, 20, 427-442.

Blouin, D. C. and Saxton, A. M. (1990). **General Linear Mixed Models: GLMM User's Manual**, Version (1.1) Baton Rouge: Louisiana State University. (38 pages.)

Carstensen, J. T., and Nelson, E. (1976). "Terminology regarding labeled and contained amounts in dosage forms," **Journal of Pharmaceutical Sciences**, 65, 311-312.

Chow, S-C. and Shao, J. (1991). "Estimating drug shelf-life with random batches," **Biometrics**, 47, 1071-1979.

Food and Drug Administration (1987). **Guideline for Submitting Documentation for Stability Studies of Human Drugs and Biologics**, Rockville, Maryland: Center for Drugs and Biologics, United States Department of Health and Human Services.

Fuller, W. A. and Battese, G. E. (1973). "Transformations for estimation of linear models with nested error structure," **Journal of the American Statistical Association** 68, 626-632.

Harville, D. A. (1977). "Maximum likelihood approaches to variance component estimation and to related problems," **Journal of the American Statistical Association** 72, 320-338.

Harville, D. A. (1986). "Using least squares software to compute combined intra-interblock estimates of treatment contrasts," **The American Statistician** 40, 153-157.

Henderson, C. R. (1975). "Estimation of variance and covariance components," **Biometrics** 9, 226-252.

Jennrich, R. I. and Schluchter, M. D. (1986). "Unbalanced repeated-measures models with structured covariance matrices," **Biometrics**, 42, 805-820.

McLean, R.A., Sanders, W.L., and Stroup, W.W. (1991). "A unified approach to mixed linear models." **The American Statistician** 45, 54-64.

Murphy, J. R. and Weisman, D. (1990). "Using random slopes for estimating shelf life," **Proceedings of the Biopharmaceutical Section,** 196-203. (Paper presented at the 1990 Annual Meetings, Anaheim, California.) Alexandria, VA 22314: American Statistical Association.

Norwood, T. E. (1986). "Statistical analysis of pharmaceutical stability data," **Drug Development and Industrial Pharmacy** 12, 553-560.

Robinson, G. K. (1991). "That BLUP is a good thing: the estimation of random effects," (with discussion.) **Statistical Science** 6, 15-51.

Ruberg, S. J. and Stegeman, J. W. (1991). "Pooling data for stability studies: testing the equality of batch degradation slopes," **Biometrics** 47, 1059-1069.

Searle, S. R. (1971). **Linear Models**. New York: John Wiley and Sons.

Searle, S. R. (1988). "Mixed models and unbalanced data: wherefrom, whereat, and whereto?" **Communications in Statistics - Theory and Methods** 17, 935-968.

Wolfinger, R., Tobias, R. and Sall, J. (1991). "Mixed models: a future direction," **SUGI 16 Proceedings** (SAS Users Group International, 16th Annual Conference, New Orleans, Louisiana) 1380-1388.

Yates, F. (1968). "Theory and practice in statistics," **Journal of the Royal Statistical Society** A, 131, 463-477.

Bob Obenchain, Lilly Corporate Center, Drop 2233, Indianapolis, IN 46285; Internet: OBENCHAIN_ROBERT_L@LILLY.COM; voice: (317) 276-3150; fax:(317) 276-3220.