

REPEATED MEASURES USING PROC MIXED INSTEAD OF PROC GLM

James H. Roger and Michael Kenward
Live Data and Reading University, U.K.

Abstract

The new procedure Mixed in Release 6.07 of the SAS® System fits mixed general linear models. These are linear models which include both fixed effects and random effects. This paper reviews the use of mixed models for repeated-measures data, where an observation is taken repeatedly, through time or space, on the same subject.

Several standard tools for analyzing repeated-measures data are available in the SAS procedure GLM. These can be implemented quite simply in the SAS procedure MIXED. However, the random-effects models available in the MIXED procedure extend the type of situations which can be handled; the inclusion of subjects who are only observed at a subset of the periods (missing data), the inclusion of covariates in the model which vary across the repeated observations on a single subject, and observations which are measured at continuous rather than discrete time points.

An example application of random-coefficient regression models is given. The paper also highlights some of the outstanding problems with using the MIXED procedure in practice — specifically, problems of convergence of the algorithm, and problems associated with testing the fixed-effects parameters. These problems also apply to other applications of the MIXED procedure, such as the case of imbalance in cross-over trials.

Overview

Repeated measures occur whenever the same observation is taken sequentially on the same subject, usually through time. In many cases the covariate information, including applied treatments, is measured on the subjects as a unit, rather than at each period. Some typical applications are;

- Weights of animals at monthly intervals
- Monitoring of blood pressure following drug application
- Lead levels in an air pollution study
- Systematic designs in field trials.

The main complication is that we expect observations on the same subject will be correlated. This may be due to a simple subject effect. On the other hand, there may be stronger correlation between observations which are

adjacent in the time sequence, compared to the correlation between one at the beginning and one at the end of the sequence.

There are several different methods for the analysis of repeated measures. They are based upon different assumptions about the processes which induce this correlation between the repeated observations. The MIXED procedure extends the class of models beyond those which can be fitted using the GLM procedure. The GLM procedure can handle situations where the analysis can be split effectively into separate between-subject and within-subject analyses. On the other hand, the MIXED procedure allows the random variation to be modelled at both within and between subject levels concurrently.

The notation which is used in this paper is the same as that used in the SAS Technical Report P-229, *SAS/Stat® Software: Changes and Enhancements*. The fixed effects model which is fitted by the GLM procedure can be expressed in matrix form as;

$$Y = X\beta + \epsilon$$

where

β is the vector of unknown fixed-effect parameters with known model matrix X ,

and

ϵ is the vector of residual effects — the difference between the modelled value and the observed value for each observation. The ϵ_i are assumed to be independently distributed with zero mean and the same unknown variance σ^2 .

The mixed model is similarly expressed in matrix form as;

$$Y = X\beta + Z\nu + \epsilon$$

where

β is the vector of unknown fixed-effect parameters with known model matrix X , as before

ν is an unknown vector of random effects

Z is the known model matrix associated with ν and ϵ is vector of residual effects, declared in the same way as before, but with different assumptions about its distribution.

The distribution of vectors ν and ϵ are assumed to be independent with zero means and respectively variance-

covariance matrices G and R . As a result the expected value of the response variable Y is $X\beta$ and the variance is $(ZGZ' + R)$. The matrices G and R contain unknown variance and covariance parameters which are usually estimated from the data.

In the MIXED procedure the statements CLASS, MODEL, ESTIMATE, CONTRAST and LSMEANS have basically the same role as the equivalent statements in GLM. However, the RANDOM and REPEATED statements have both a different syntax and a different purpose to statements with the same name in the GLM procedure. The individual parts of the mixed model are specified to the MIXED procedure in the following way;

- The values for the matrix X and the structure of the vector β are specified by the MODEL statement.
- The values for the matrix Z and the structure of the vector ν are specified using the parameters of the RANDOM statement.
- The structure of matrix G is specified using the TYPE= option on the RANDOM statement.
- The structure of the matrix R is specified using the TYPE= option on the REPEATED statement.

The reader should not conclude that the REPEATED statement is used whenever the data are of the repeated-measures type. The RANDOM and REPEATED statements are used to control separate parts of the mixed model equations. There will be instances when a repeated-measures model is better expressed using the RANDOM statement, while there are also random-effect models which do not fall into the repeated-measures class but where the REPEATED statement is the simplest tool for expressing them to the MIXED procedure.

An example

This example of repeated-measures data concerns the measurement of lung function in children (Ashton, 1984). The study was carried out over a ten year period at the MRC Pneumoconiosis Unit in Wales, United Kingdom.

Twenty two twin-pairs of children aged 7 to 17 years in 1970-72 were observed. Several measurements were made on each child. The following measurements are discussed in this paper.

- Forced Expiratory Volume within one second (FEV_1)
- Age measured in years
- Height measured in metres
- Height standardised sitting height (Sitting height / Height)
- Obesity ($\text{Weight} / \text{Height}^2$) measured in Kgm^{-2} .

Observations were repeated at three year intervals. The second observation was taken in the period 1973-74, the third in the period 1977-78 and the final observation was taken in the period 1980-81. As a result of the length of this study, it was not possible to follow up all the children in each period. Eleven twin-pairs (22 children) were observed in all four periods. Four twin-pairs (8 children) were observed three times. Three twin-pairs (6 children) were observed twice, while four twin-pairs (8 children) were observed only once.

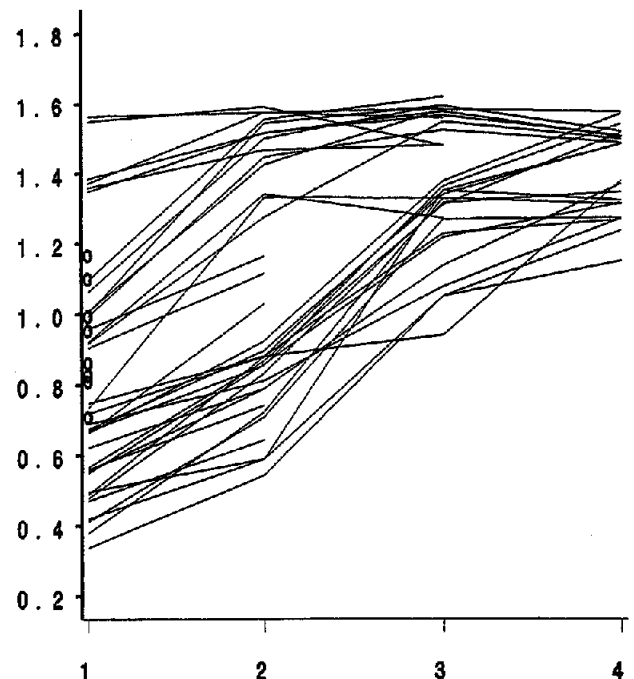


Figure 1. $\text{Log}(FEV_1)$ against observation period number.

Previous studies of FEV_1 indicate that it is sensible to use a log transformation. The response variable in this example is $\text{Log}(FEV_1)$. In Figure 1 the values of $\text{Log}(FEV_1)$ are shown at each observation period. The values for each child are linked by straight lines. The eight children with only one observation are shown as circles. Note the twin pairs in this graph.

A SAS data set containing the data, ready for use with the MIXED procedure, is generated using the following program;

```
DATA CD.D ;
INFILE 'c:\user\MRCFEV1.dat' ;
INPUT Twin Sub Time Y Age Ht Sht Obe ;
LABEL Twin = 'Twin-pair No' ;
```

```

LABEL Sub = 'Subject No' ;
LABEL Time = 'Period number' ;
LABEL Y = 'Log of FEV1' ;
LABEL Age = 'Age (years)' ;
LABEL Ht = 'Height (metres)' ;
LABEL Sht = 'Sitting height (Height standardised)' ;
LABEL Obe = 'Obesity' ;
RUN ;

```

There are ideally four records in the data set for each child — one for each period. Those children who were observed less than four times have missing records.

Facilities in the SAS procedure GLM

The REPEATED statement in the GLM procedure allows four linked approaches to analyzing repeated-measures data. The response variable must be measured at a fixed set of time points, such as the four periods in this example. Covariates which vary from period to period, such as Height and Obesity, cannot be accommodated. The only possible covariate in this study, which is constant within subject, is the twin-pair reference number Twin.

The following discussion of the facilities for repeated measures in the GLM procedure assumes that the data is held with a single record for each subject. The Log(FEV₁) values are held in the four variables LFEV_1 to LFEV_4. Data is only recorded for the 22 children observed in all four periods, because the GLM procedure cannot handle subjects who are not observed at all periods. The analysis centres on the covariate Twin which has 11 levels, one for each of the eleven twin-pairs with complete data.

The REPEATED statement in the GLM procedure requires a name for the classification which runs across the four time periods. To be consistent with the later MIXED programs we declare the name as Time and select polynomial contrasts, using the following REPEATED statement.

```
REPEATED Time 4 POLYNOMIAL / SUMMARY ;
```

The four main approaches are as follows.

1. A separate univariate analysis within each observation period. In our example, it is an analysis of LFEV_1 followed by LFEV_2 etc. up to LFEV_4. Here we are looking at the main effect of the covariate Twin. Does Log(FEV₁) in period 1 vary less from child to child within a twin-pair compared to between

unrelated children? There is a highly significant difference.

2. An analysis of the data, as if they came from a split-unit experiment, where the model includes an effect for each subject. The assumption is that the correlations between the response variables at any two time periods are the same. This is often called the *sphericity condition* and is tested by the GLM procedure. It also implements the Greenhouse-Geiser and Huynh-Feldt adjustments to the F tests to accommodate any divergence from the sphericity assumption. Simulation studies have shown that these are effective in most practical situations.

This approach tests the equivalent of the main effect of Time — "Does Log(FEV₁) change from period to period?". Also it tests the interaction of Time with Twin, which sees whether the pattern across twin-pairs varies from period to period — "Does a twin-pair which responds with high FEV₁ in period 1 also respond high in later periods?".

The main effect of Twin in this split-unit analysis is equivalent to a main-unit treatment and is tested using the contrast;

$$(LFEV_1 + LFEV_2 + LFEV_3 + LFEV_4) / 4$$

This test does not rely on any assumption about sphericity. It is an "average" of the four tests on the individual variables in the first approach.

3. A multivariate analysis of variance where each time period is regarded as a separate variable.

This also tests the main effect of Time and the interaction between Time and the covariates. However it does not make any assumptions about the correlation between the responses in the four periods.

4. An analysis of specific contrasts across the periods — for instance, the difference between each value and the mean of the values for subsequent periods (HELMERT option). The procedure GLM offers a choice from five different types of contrast. Each one looks at a different possible aspect of the pattern in the repeated measures. The POLYNOMIAL option, used here, extracts orthogonal polynomials, the first of which being regression across the time sequence.

There is a test of whether the absolute value of the contrast is zero and a test of the effect of the covariates on the contrast. In this case, the questions

are "Is there a regression of Log(FEV₁) across the periods?" and "Does this regression of Log(FEV₁) on period (Time) vary less between children within a twin-pair than between unrelated individuals?".

Using MIXED instead of GLM

These standard types of analysis can be readily carried out using the SAS procedure MIXED instead of GLM. However, when they are appropriate it will often be easier to use the GLM procedure as less programming is usually needed. Also, in some of the following equivalent applications, the MIXED procedure uses much more computing time.

The first GLM approach, univariate analyses with one for each period, can be programmed most easily in the MIXED procedure by using a WHERE statement.

```
TITLE 'Analysis for Period 1';
PROC MIXED DATA = CD.D ;
CLASS Twin ;
WHERE Time = 1 ;
MODEL Y = Twin / SOLUTION ;
RUN ;
```

In this example, all the statements used for the MIXED procedure are identical to those which can be used with the GLM procedure. The code could be included in a macro %DO loop to run over the four periods. This MIXED analysis in the first period changes the F value for Twin to 66.4 from 58.1 for GLM. The additional information is coming from the 22 extra children in the data set.

The second GLM approach is where a simple random effect for each subject is added to model. This split-unit type of analysis can be programmed using either the RANDOM statement or the REPEATED statement in the procedure MIXED.

```
TITLE 'Analysis using split-unit analysis' ;
PROC MIXED DATA = CD.D ;
CLASS Sub Time Twin ;
MODEL Y = Time Twin Time*Twin ;
REPEATED Time / TYPE = CS SUBJECT = Sub ;
RUN ;
```

or

```
TITLE 'Analysis using split-unit analysis' ;
PROC MIXED DATA = CD.D ;
```

```
CLASS Sub Time Twin ;
MODEL Y = Time Twin Time*Twin ;
RANDOM Sub / TYPE = SIM ;
RUN ;
```

The covariance matrix R for TYPE=CS has elements $R_{ij} = (\sigma^2 + \sigma_1^2)$ and $R_{ij} = \sigma_1^2$ for $i \neq j$. This is known as *Complex Symmetry*. The covariance matrix G for TYPE=SIM has elements $G_{ij} = \sigma^2$ and $G_{ij} = 0$ for $i \neq j$. This matrix form is known as *Simple*.

The Greenhouse-Geiser and the Huynh-Feldt adjustment to the F tests are not available in the MIXED procedure. However, they are not necessary as it is very simple to fit, and also interpret, a full multivariate model using the MIXED procedure.

The RANDOM statement in this example can be rewritten in an equivalent but computationally more efficient fashion as follows.

```
RANDOM INTERCEPT
/ TYPE = SIM SUBJECT = Sub ;
```

The MIXED analysis gives an F value of 57.8 instead of 51.1 for Twin, 1118 instead of 1284 for Time and 21.8 instead of 21.4 for the Twin*Time interaction.

The third approach in the GLM procedure is a multivariate analysis. The equivalent mixed model allows the observations at each period to have an unstructured variance-covariance matrix within each subject.

```
TITLE 'Equivalent to a Multivariate analysis';
PROC MIXED DATA = CD.D ;
CLASS Sub Time Twin ;
MODEL Y = Time Twin Time*Twin ;
REPEATED Time / TYPE = UN SUBJECT = Sub ;
RUN ;
```

Note how this is similar to the previous program, apart from the replacement of TYPE=CS for *Complex Symmetry* by TYPE=UN for *Unstructured*. The matrix R for TYPE=UN has separate parameters $R_{ij} = R_{ji} = \sigma_{ij}$ for each element.

The multivariate approach used in the GLM procedure produces multivariate tests for the fixed effects based on Wilk's Lambda. The resulting F tests are based on a better approximation to the actual distribution of the test statistic than that for the F tests output by the MIXED procedure.

For this data set the main advantage in using the MIXED procedure is that data for all forty four children can be used. However, care should always be taken before assuming that any missing data values can be assumed to be merely absent. If the process, which controls whether a subject is observed in any period, is dependent upon the potential value in that period, then we cannot proceed by simply excluding this response from the data set. The censoring process itself must be modelled in some way. In this example the lost children can be assumed to be randomly self-selecting. Treating missing values as absent is valid.

Here the MIXED analysis gives an F of 56.6 instead of 51.1 for Twin, 1604 instead of 1516 for Time and 26.7 instead of 12.2 for the Twin*Time interaction.

The fourth approach used in the GLM procedure is to study individual contrasts across the time periods. It is technically possible to extract similar information from the MIXED procedure using the ESTIMATE and/or CONTRAST statements. For instance

```
ESTIMATE 'Helmert2' Time 0 2 -1 -1 /DIVISOR=2;
```

In the MIXED procedure, polynomials are better modelled directly with the MODEL statement.

Models only possible with the MIXED procedure

We should not regard the MIXED procedure as simply a way of handling missing values and covariates that vary across time. It also extends the range of models which can be fitted to repeated measures data. This example illustrates how a full understanding of the data is only possible with the extra facilities available in the MIXED procedure. Initially we are going to disregard the important twin-pair nature of the data and look only at the covariates which vary across the four periods.

Figure 2 shows values of $\text{Log}(\text{FEV}_1)$ plotted against age. Each child is shown as a line connecting the values observed at each age. The eight children who were only recorded once appear as circles on the plot. It is clear from this figure, that FEV_1 increases as the body grows larger. The observation period Time, used in the previous analysis, is masking this facet of the data, as children started the study at different ages. The important influences on FEV_1 are related to the age of the child. It seems logical that the size of the child will be important and perhaps the size of the upper body. Previous studies indicate that height and sitting height are two important

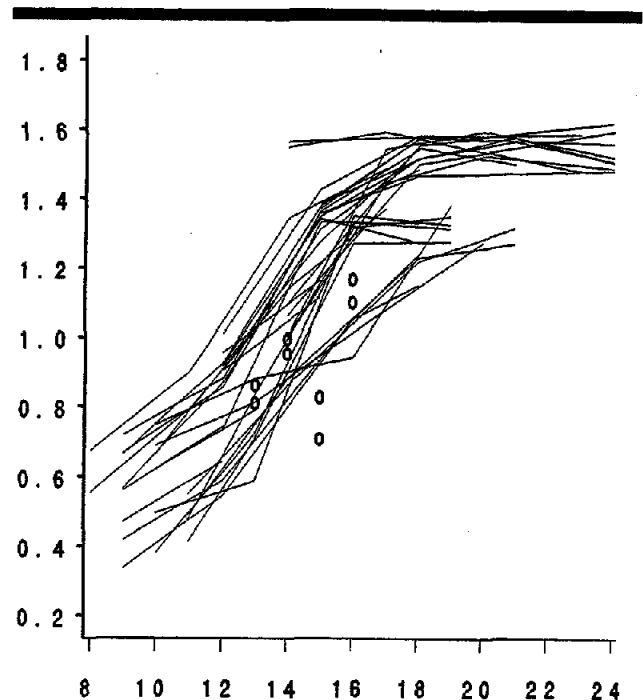


Figure 2. $\text{Log}(\text{FEV}_1)$ against Age in years.

predictors of lung function.

Figure 3 shows the values of $\text{Log}(\text{FEV}_1)$ plotted against the height. In this case we have excluded the eight children who were only observed once in the study. The relationship between $\text{Log}(\text{FEV}_1)$ and height is fairly constant. The relationship within each child appears to be linear, while the slope seems to vary only slightly from child to child.

Here we fit a model with a regression on the height of the child. A random variation is introduced from child to child. This is equivalent to the split-unit assumption discussed earlier.

```
PROC MIXED DATA=CD.D ;
CLASS Sub ;
MODEL Y = Ht / SOLUTION CL ;
RANDOM INTERCEPT
/ TYPE= SIM SUBJECT= Sub CL ;
RUN ;
```

As well as the intercept for the regression changing from child to child we may also want to regard the possibility that the actual slope for an individual child varies from child to child. This is known as a random-coefficient

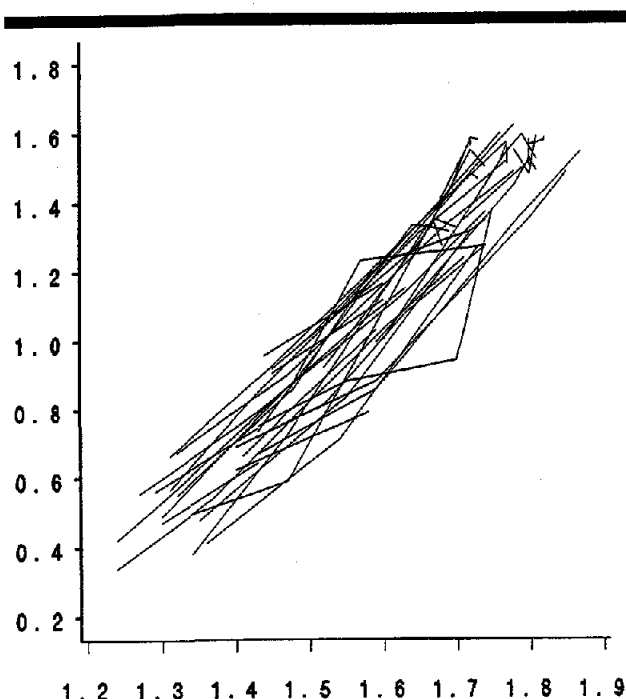


Figure 3. Log(FEV₁) against Height in metres.

regression model. Such a random-coefficient regression model is used in Example 16.5 in the SAS Technical Report P-229.

This is done by including the variable HT in the RANDOM statement as well as the MODEL statement. This is effectively a HT*SUB term in the RANDOM statement as it is nested within the term specified in the SUBJECT= option.

```
PROC MIXED DATA=CD.D ;
CLASS Sub ;
MODEL Y = Ht / SOLUTION CL ;
RANDOM INTERCEPT Ht
/ TYPE=UN SUBJECT= Sub CL ;
RUN ;
```

The covariance matrix G is specified by the TYPE=UN option which chooses an unstructured matrix thus;

$$G = \begin{bmatrix} \sigma_{11}^2 & \sigma_{21} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}$$

Note how we do not only introduce parameters σ_{11}^2 and σ_{22}^2 as variances for the intercept and slope, but also a parameter σ_{21} for the covariance of the intercept and slope across children. We can think of the variation in slope as the pivoting of the regression line about an origin point. The variation in intercept being the random movement of

this pivot point in the vertical (response variate) direction. Lastly, the correlation comes from the position of this pivot point in the horizontal direction relative to the origin. As the origin for the covariate moves further from the pivot point the absolute value of the correlation increases.

The second model in example 16.5 on page 357 of SAS Technical Report P-229 excludes the covariance term, since the default TYPE=SIM is used. In that example it may not be a problem as the origin Month=0 is a control value. Variation in intercept at the start is taken to be independent of the slope across ensuing months. However, in most cases a covariance term will be required. In our example, where the intercept (HT=0) is outside the range of the data, it is clear that we need to estimate a correlation between the intercept and the slope. As the slope increases the intercept will tend to decrease.

Table 1. Covariance parameter estimates without random-coefficient regression.

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > Z
INTERCEPT	1.2467	0.0070	0.0019	3.57	0.0004
Residual	1.0000	0.0056	0.0008	6.68	0.0000

The variance parameter estimates from the first model are shown in Table 1. The same output from the second model (Table 2) is more difficult to read. INTERCEPT UN(1,1) relates to the variation in intercept from child to child, INTERCEPT UN(2,1) is the covariance between the intercept and the slope. INTERCEPT UN(3,3) relates to the variance for the slope. This is because 1 relates to the first parameter (INTERCEPT) in the RANDOM statement and 2 relates to the second parameter (HT).

Table 2. Covariance parameter estimates for random-coefficient regression model.

Cov Parm	Ratio	Estimate	Std Error	Z
INTERCEPT UN(1,1)	23.3259	0.1108	0.0795	1.39
UN(2,1)	-14.5224	-0.0690	0.0511	-1.35
UN(2,2)	9.6178	0.0457	0.0331	1.38
Residual	1.0000	0.0047	0.0008	5.67

As we expect, there is a strong negative correlation (-0.97). The increase in the variance estimate for the intercept is being induced by the variation in the slope. The best way to compare these two models is to look at the REML log-likelihood.

	-2 REML Log-Likelihood
RANDOM INTERCEPT	-233.04
RANDOM INTERCEPT Ht	-236.18

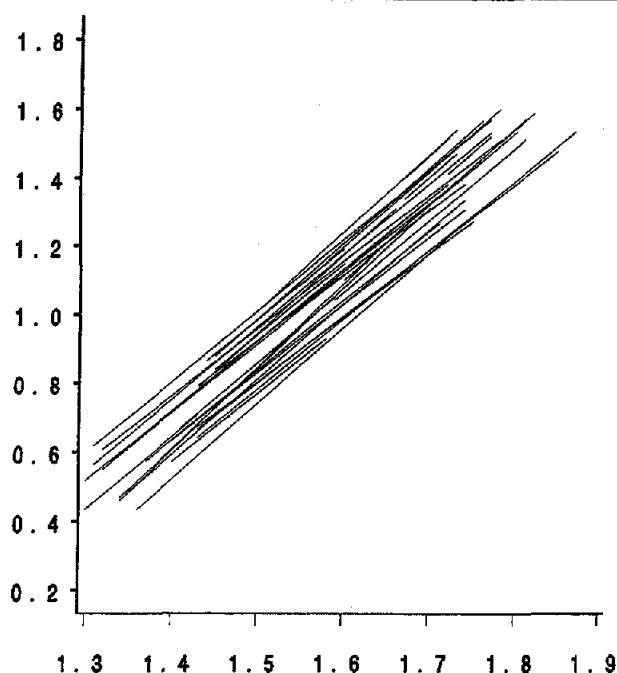


Figure 4. Predicted Log (FEV₁) against Height using random coefficients model.

The change in Deviance (-2 REML Log-likelihood) is 3.14. This is a valid likelihood ratio test as the REML log-likelihoods are marginal likelihoods for the variance parameters. It has an asymptotic χ^2 distribution on the null hypothesis. You can *not* use the equivalent difference to test parameters in the fixed-effects part of the model. The value of 3.14 suggests that the additional random slope is not required in the model. Later we will drop it from the model.

Table 3. Fixed-effect parameter estimates for random coefficient regression model.

Parameter	Est.	Std Err	DDF	T	Pr> T	95% C.I.	
						Lower	Upper
INTERCEPT	-2.239	0.0803	87	-27.85	0.0000	-2.39	-2.08
HT	2.092	0.0495	87	42.21	0.0000	1.99	2.19

The fixed-effects parameter estimates are given in Table 3. The standard deviation for the random variation to the slope is 0.21 ($\sqrt{0.0457}$ from Table 2), so we expect the slope for any individual child to vary from about 1.7 to 2.5 ($2.09 \pm 2 \times 0.21$).

Using the MAKE statement and also the P option on the MODEL statement, we can send the predicted values for

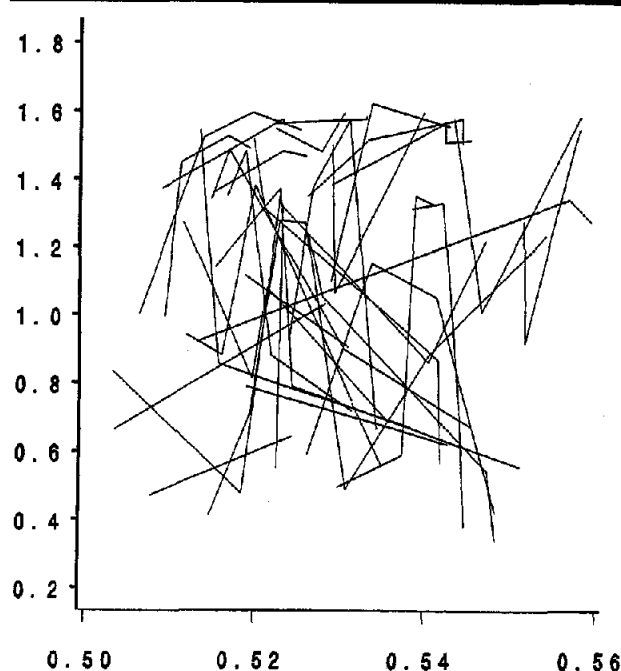


Figure 5. Log(FEV₁) against Sitting height standardised by Height.

each observation to a SAS data set. After merging the resulting data set with the original data, we can draw a graph of the predicted lines (Figure 4).

We know look at the height standardized sitting height (SHT). Figure 5 shows the raw log(FEV₁) values plotted against the standardized sitting height. There does not appear to be much pattern. If we look at the residual from our previous model and plot these against the residuals from fitting height to the standardised sitting height (Figure 6) we can see that there is a tendency for each child to have a separate regression against sitting height with positive slope.

We shall now fit a model with fixed effects for Height and standardized Sitting height. But, rather than include a random-coefficient regression term for Height, we include one for the standardised Sitting height. As before, we fit the model with and without this random slope.

As the INTERCEPT and SHT have a correlation which is very close to unity, we move the origin for the variable SHT. We define and use a new variable NSHT which subtracts 0.53 from SHT. This helps the algorithm in the MIXED procedure to find the maximum REML

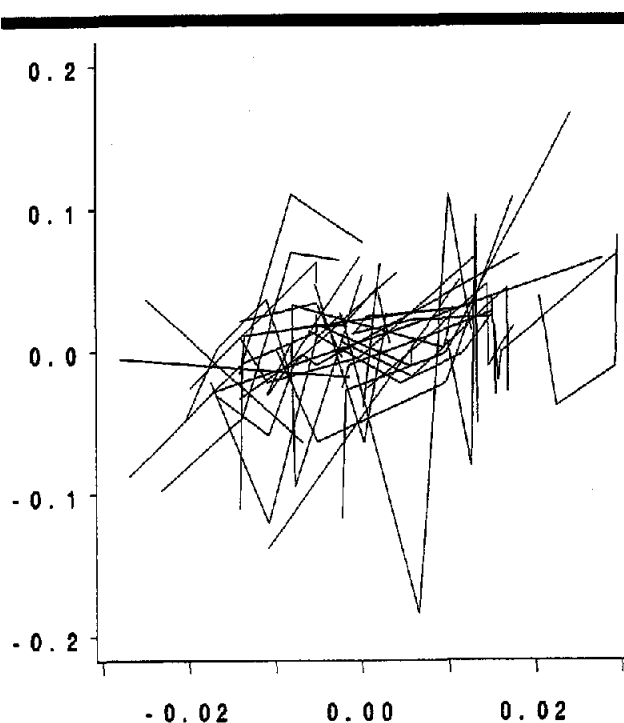


Figure 6. Residual $\text{Log}(\text{FEV}_1)$ plotted against residual from regressing standardised sitting height on height.

log-likelihood solution.

```
DATA CD.D ;
SET CD.D ;
Nsht = Sht - 0.53 ;
```

```
PROC MIXED DATA= CD.D ;
CLASS Sub ;
MODEL Y = Ht Nsht / CL SOLUTION ;
RANDOM INTERCEPT
/ TYPE= UN SUBJECT= Sub CL ;
RUN ;
```

```
PROC MIXED DATA= CD.D ;
CLASS Sub ;
MODEL Y = Ht Nsht / CL SOLUTION ;
RANDOM INTERCEPT Nsht
/ TYPE= UN SUBJECT= Sub CL ;
RUN ;
```

The REML log-likelihoods can be summarised as follows;

	<u>-2 REML Log-Likelihood</u>
RANDOM INTERCEPT	-260.02
RANDOM INTERCEPT Nsht	-269.18

The change in Deviance (-2 REML Log-likelihood) is 9.16. This suggests that the random slope for NSHT is a significant part of the model. So we now look further at the second of these two models.

Table 4. Estimates of covariance parameters and fixed effect parameters for random-coefficient regression model with standardized sitting height.

Cov Parm	Ratio	Estimate	Std Err	Z
INTERCEPT UN(1,1)	1.1961	0.004686	0.001429	3.28
UN(2,1)	-26.79	-0.104960	0.057076	-1.84
UN(2,2)	1824.89	7.148961	4.187761	1.71
Residual	1.000000	0.003917	0.000647	6.05

Parameter	Est.	Std Err	DDF	T	Pr> T	95% C.I.	
INTERCEPT	-2.3341	0.0714	51	-32.69	0.000	-2.47	-2.19
HT	2.1570	0.0441	51	48.84	0.000	2.06	2.24
NSHT	3.5211	0.7113	51	4.95	0.000	2.09	4.94

Table 4 gives the results for the random and fixed effects parameters. The slope for the standardized sitting height is 3.52 with a standard error of 0.71. The random component of the slope has a standard deviation of 2.67 ($\sqrt{7.15}$).

Table 5. Estimates of fixed effect parameters for random-coefficient regression model with standardised sitting height, adding fixed effect Obesity.

Parameter	Est.	Std Err	DDF	T	Pr> T	95% C.I.	
INTERCEPT	-2.3330	0.0718	50	-32.49	0.000	-2.47	-2.18
HT	2.1676	0.0621	50	34.89	0.000	2.04	2.29
NSHT	3.5709	0.7417	50	4.81	0.000	2.08	5.06
OBE	-0.0009	0.0038	50	-0.25	0.807	-0.00	0.00

If we include Obesity as a fixed effect, we find that it has no significant effect in the model (Table 5), even though raw plots of $\text{Log}(\text{FEV}_1)$ against Obesity (Figure 7) show an apparent pattern. This pattern is most likely induced by both variables being associated with age.

Modelling the twins

So far we have ignored the fact that the children are twin-pairs. It is reasonable to assume that there will be less variability from child to child within a twin-pair, as we have seen in the simple analysis using the GLM procedure. One obvious way to handle this is to introduce another strata of variation — twin-pair to twin-pair, as well as child-to-child within twin-pair, and residual.

We can use the MIXED procedure to fit a model where we do not have any random slopes. For instance the

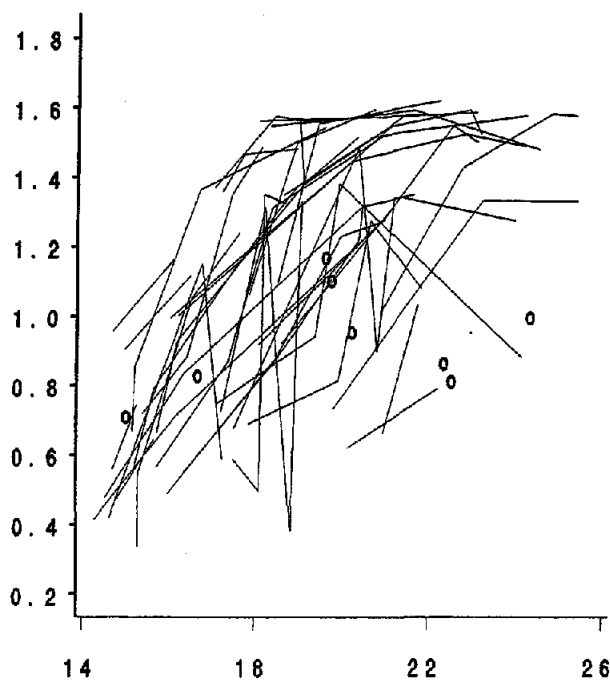


Figure 7. Log(FEV₁) against Obesity (Kgm⁻²)

RANDOM statement would look like;

```
RANDOM INTERCEPT Child
/TYPE=SIM SUBJECT=Twin;
```

where Twin indexes the 22 twin-pairs and the new variable Child has two levels indexing the child within the twin-pair. Note how the default value SIM is used for the TYPE= option. This is very similar to the code for a split-split-unit experiment. For this model, the estimate of the child variance parameter is zero, suggesting that there is very little variation within twins.

Also we could fit TWIN as a fixed effect in our existing random-coefficient regression model.

However, we do not see how to include the random slope at two strata levels within the constraints of the syntax for the MIXED procedure. It would be necessary to constrain off-diagonal elements of the R matrix to zero in ways which are not covered by the current set of TYPE= options. The best model which we have managed to fit to these data is a random-coefficient regression model on NSHT, but at the twin level rather than the child level.

```
PROC MIXED DATA= CD.D ;
```

```
CLASS Sub Twin ;
MODEL Y = Ht Nsht / CL SOLUTION ;
RANDOM INTERCEPT Nsht
/ TYPE= UN SUBJECTS= Twin CL G ;
```

The deviance (-2 REML Log-Likelihood) is -281.58 compared to -269.18 for the random-coefficient regression model at the child level.

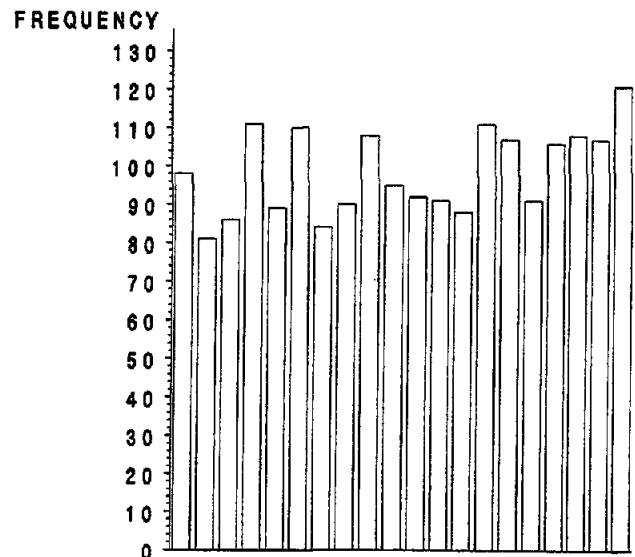


Figure 8. Distribution of P values from test of fixed-effect slope for standardized sitting height (Sht). Each bar represents a 5% interval.

Testing the fixed-effects parameters

The standard errors for the fixed-effects parameters are calculated assuming that the random-effects parameters are known. However they are in fact estimated from the same data set. In the standard general linear model, the use of F statistics rather than the Wald χ^2 , introduces a correction so that the size of any significance tests is exact. The denominator degrees of freedom indicate the precision for the estimate of the residual degrees of freedom. In the case of mixed models, no such simple exact result holds. Indeed the fixed-effect and variance parameter estimates are not independent of each other as they are in the standard general linear model.

In the MIXED procedure a quasi-F statistics is generated by dividing the Wald χ^2 by its degrees of freedom. For simple examples, such as a split-unit experiment, the resulting statistic does have an exact F distribution. In other cases it may approximate an F distribution. The

technical problem is to choose an appropriate value for the denominator degrees of freedom. The MIXED procedure uses a naive algorithm which gives appropriate values when split-unit and other simple designs are specified in the "correct" way (Treatments, rather than structural classifiers of the units, are used to specify levels of variation in the RANDOM statement). In most cases, it uses the residual degrees of freedom. This will be more conservative than the Wald χ^2 which is known to be too liberal. However, in many cases use of the residual degrees of freedom will also lead to too liberal conclusions — the size of the test is too large.

To investigate whether this is a the problem in this example, we simulated data with exactly this same structure and model as that which had been fitted. There were 2000 simulations carried out and the same random-coefficient regression model was fitted each time.

In 26 cases the MIXED procedure would not converge. This was after we had used the following techniques to improve the number of simulations where convergence was complete.

- The PARMS statement was used to start the iterative cycle using the modelled values for the variance parameters.
- The SCORING=20 option was selected to use Fisher scoring.
- The CONVF option was used for checking whether convergence had occurred.

The parameter estimate for the fixed effect NSHT, minus the modelled value, was divided by its standard error and a one-sided P value calculated using the t distribution with degrees of freedom set to those recommended by the MIXED procedure (residual d.f.).

The distribution of the P values is summarised in Figure 8. Figure 9 shows a plot of the P values against the cumulative probability based on their ranks. This graph indicates that there seems to be a slight bias in the estimator, rather than a problem with variance of the test statistic. The results for the INTERCEPT fixed-effect parameter are very similar. We expect the approximation to be less good where the parameter is mostly estimated from between-child rather than within-child information.

Conclusions

The MIXED procedure allows interesting new models to be fitted to repeated-measures data. However, for some types of problem the GLM procedure will remain a better

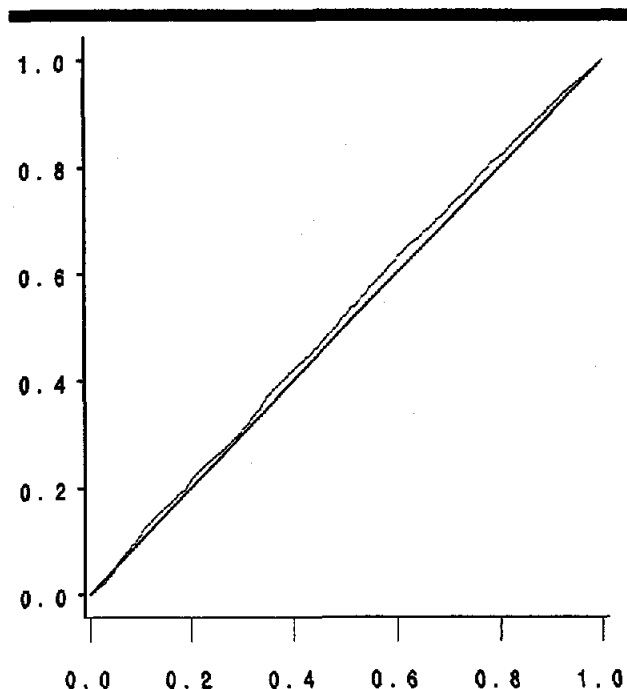


Figure 9. P values from test of fixed-effect slope for standardized sitting height (Sht) against expected probabilities based on ranks.

option. Also it would appear that some important types of mixed model are currently not possible in the MIXED procedure.

Acknowledgements

We thank Dr. J.E Cotes at the Department of Occupational Health, University of Newcastle, for access to the data used in this paper.

References

Ashton, K.M.I. (1984) *Growth and heritability of lung function: Reference value for adolescence*. Ph.D. thesis, University of London, United Kingdom.

SAS Technical Report P-229. *SAS/Stat™ Software: Changes and Enhancements*. SAS Institute Inc, Cary, NC, USA.

Trademark Citations

SAS and SAS/STAT are registered trademarks of SAS Institute Inc, Cary, NC, USA.