Introduction to the analysis of multilevel models with LISREL® 8.30

Stephen and Mathilda du Toit
Scientific Software International, Inc.

Robert Cudeck
University of Minnesota, Minneapolis

October 1999

SSI SCIENTIFIC SOFTWARE INTERNATIONAL

Introduction to the analysis of multilevel models with LISREL 8.30.

1 2 3 4 5 6 7 8 9 0 03 02 01 00 99

Published by:

Scientific Software International, Inc.
7383 North Lincoln Avenue, Suite 100
Lincolnwood, IL 60712–1704
Tel: +1.847.675.0720
Fax: +1.847.675.2140
URL: http://www.ssicentral.com

1. Basic concepts of multilevel modeling

1.1 Introduction

Multilevel models deal with the analysis of data where observations are nested within groups. Social, behavioral and even economic data often have such a hierarchical structure. A frequently cited example is in education, where students are grouped in classes. Classes are grouped in schools, schools in education departments and so on. We thus have variables describing individuals, but the individuals may be grouped into larger or higher-order units.

Traditionally, fixed parameter linear regression models have been used for the analysis of such data. Statistical inference is based on the assumptions of linearity, normality, homoscedascity and independence. It has been shown by Aitkin and Longford (1986) that the aggregation of variables over individual observations may lead to misleading results. Both the aggregation of individual variables to a higher level of observation and the disaggregation of higher order variables to an individual level have been somewhat discredited (Bryk & Raudenbush, 1992). It has been pointed out by Holt, Scott and Ewings (1980) that serious inferential errors may result from the analysis of complex survey data, if it is assumed that the data have been obtained under a simple random sampling scheme.

Multilevel data structures are pervasive. As Kreft and de Leeuw (1998) note: "Once you know that hierarchies exist, you see them everywhere". It was not until the 1980s and 1990s that techniques to properly analyze multilevel data became widely available. These techniques use maximum likelihood procedures to estimate random coefficients and are often referred to as "multilevel random coefficient models" (MRCM) in contrast to "hierarchical linear models" (HLM).

In the following sections, a number of issues pertaining to the analysis of hierarchical data will be addressed.

1.2 Hierarchies, Strata, Subgroups

Many populations of subjects are organized in subgroups. Examples of subgroup distinctions are gender, educational level, or type of job.

Statistical analyses can take advantage of this organization to improve the accuracy of statistical summaries and predictions.

Example 1.1

In education research, a recurrent structure is the subgrouping that exists because students are assigned to particular classrooms.

- The educational environment in one classroom can be quite different from that in others: characteristics of the teacher and the interactions among the students vary.
- When students participate in an experiment, the effect of the treatment may depend on the classroom to which a student is assigned.
- Failure to account for classroom subgroups may distort the effectiveness of the treatment.

1.3 The Interaction Question

Interaction exists when the benefit of an intervention, or the effect of a treatment, depends on particular levels of other variables.

Simple Interaction: Differential performance is attributable to levels of a single variable. For example, students in classes with a female teacher perform better than students with a male teacher, on the same curriculum.

Complex Interaction: A variety of contextual or background variables are associated with performance. Differences in behavior cannot be attributed to any single variable. For example, on a common curriculum, students in certain classrooms perform much differently than in others. However, no specific variable accounts for the differential outcome.

Example 1.2: Can Mathematics Achievement be predicted by Reading Ability?

In the study of cognitive ability, verbal skill is considered by many to be the core feature of intelligence. Individual differences in verbal skill are strongly associated with performance in many different mental tasks. Although this association has been observed in many different settings, the strength of the correlation varies considerably across characteristics of students.

Figure 1.1 shows the correspondence between standardized reading and mathematics scores for 147 high school students (see Tatsuoka, 1988). Subjects are a subsample of students who participated in a large-scale, national longitudinal study known as High School and Beyond. The first 10 observations from the file **tatsu.psf** are shown below.

TATSU.PSF

	Career	ID1	Math	Reading	Intcept
1	1.000	1.000	48.710	15.240	1.000
2	1.000	2.000	43.490	6.330	1.000
3	1.000	3.000	44.080	15.000	1.000
4	1.000	4.000	47.500	23.000	1.000
5	1.000	5.000	63.880	34.670	1.000
6	1.000	6.000	45.620	15.430	1.000
7	1.000	7.000	43.770	12.690	1.000
8	1.000	8.000	49.490	13.200	1.000
9	1.000	9.000	42.890	13.940	1.000
10	1.000	10.000	49.690	8.910	1.000

LISREL spreadsheet presentation of Tatsuoka's data

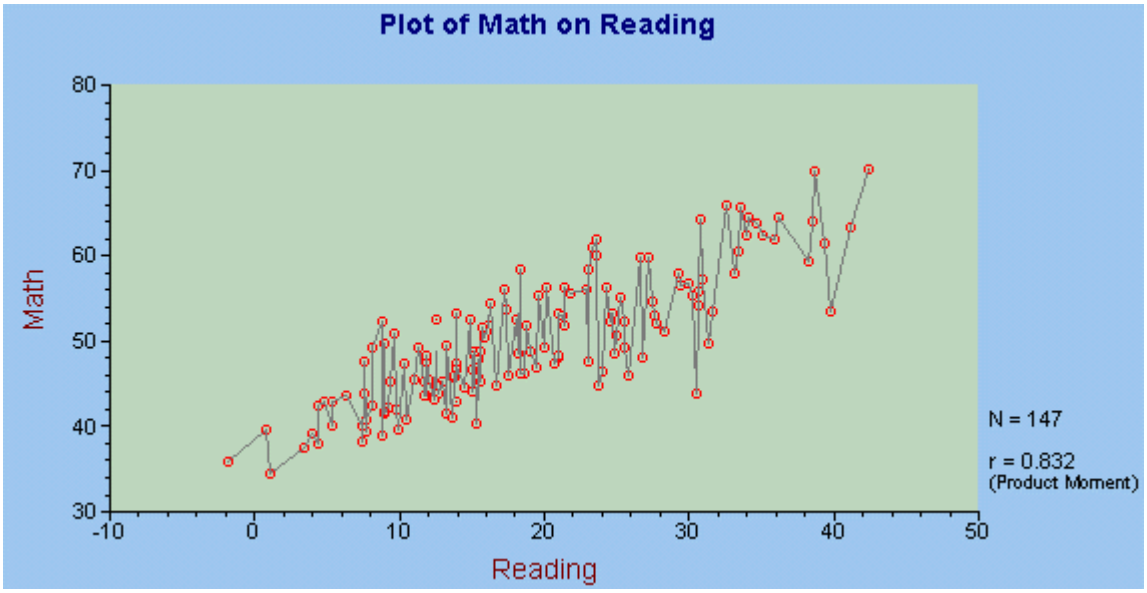


Figure 1.1: Graphical display of Mathematics score versus Reading score

Consider the regression of Mathematics score (M) on Reading score (R) for this sample

$$M = \beta_0 + \beta_1 R + e$$

Estimates of the two regression coefficients and the correlation ρ = corr(M,R) are

$\hat{\beta}_0$	$\hat{\beta}_1$	ρ
38.0	0.640	0.83

For scientific and practical reasons, it is of interest to know whether the prediction of Math from Reading is essentially the same for different types of students.

- Is the association stronger for students with professional goals than for students interested in technical careers?
- Does it differ according to characteristics of the families?
- How does previous educational experience moderate the correlation?

If the regression relationship varies systematically across levels of a covariate, then it is important to take the variable into consideration when making predictions.

Students in this study were pursuing a number of different career options:

- Trades (Tr)
- Police or Security (P)
- Business Management (B)
- Sales (S)
- Military Service (M)
- Teacher Training (T)
- Industrial Operations (I)
- Undecided (U)
- Real Estate Management (R)

Table 1.1 shows the results of nine different regressions, one for each of the nine career subgroups. The sample size available for each regression varies. Of course, and this is the main point, the estimated regression coefficients vary, too. The subgroups have the following estimates:

Table 1.1: Estimates of regression coefficients, independently in each group

	Career Options								
	Tr	B	M	I	R	P	S	T	U
β_{i0}	38.0	36.1	40.5	36.1	39.5	34.4	37.4	41.4	35.3
β_{i1}	0.66	0.71	0.38	0.74	0.63	0.83	0.65	0.68	0.65

In some career options, the subgroup regression is similar to the regression for the total sample, while in other career option subgroups the regression is somewhat different (see Figure 1.2).

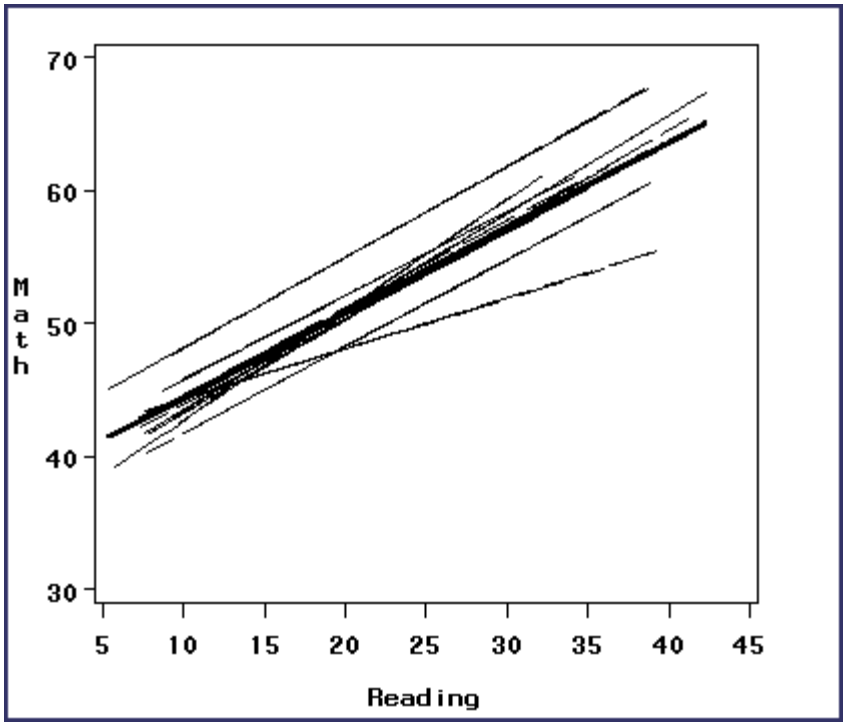


Figure 1.2: Regression lines for the population and 9 career option subgroups

Comparison of Results

The results are consistent with the impression from Figure 1.2 and Table 1.1. Some career options, such as Trades, Real Estate, and Sales, are described adequately by the total sample regression equation. In these subgroups, career option is immaterial to the regression of Math on Reading. On the other hand, for students pursuing other career options, especially the Military Service, the Police or Security, and Teacher Training, and those in the Undecided group, there is a lack of fit by the single common model, and therefore an appreciable loss of prediction efficiency is possible when using the total sample regression.

1.4 The Problem of Structured Populations

Example 1.2 illustrates the kind of problem that can arise when a stratification exists in a population, and a single model is applied to all subjects without regard for potential differences between strata. Parameter estimates obtained on the total sample may perform poorly when applied to particular subgroups. Conversely, differences between subgroups may adversely affect the estimates developed for the total sample.

A method of data analysis is needed that

- produces accurate estimation of the model in the total sample.
- allows submodels for subgroups to differ, perhaps markedly if necessary.
- can relate the overall model of the total sample to the various models of the subgroups in a systematic way.

There is a class of models that exhibits these desirable properties. The methodology is known by a variety of names, including mixed-effect models or random coefficient models. In much of the social science domain, they are known as hierarchical models.

1.5 A Basic Two-level Model

The basic regression approach with a single independent variable requires two coefficients

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij} ,$$

where β_0 denotes the population intercept and β_1 the population slope. In terms of the data described in Example 1.2, the subscript *i* denotes career option while subscript *j* denotes the *j*-th individual within this particular career option subgroup.

The parameters are interpreted as follows.

β_0: Average value of *y* when *x* = 0

β_1: Average change in *y* for every increase of 1 unit in *x*

This interpretation immediately highlights the problem with standard regression when data have a hierarchical structure. As was apparent in Figure 1.2, subjects in some of the subgroups may have different starting points than β_0 at *x* = 0, as well as different rates of increase in β_1. Although β_0 and β_1 may be roughly satisfactory as representations of the average relationship between y and x, they may not describe subgroups effectively.

If the problem is to properly describe the structure in a subgroup, then the obvious approach is to attend to data in one of the strata, not in the overall sample. In a limited sense*, each subgroup is allowed to have its own model*. To accommodate differences in subgroups, and also to relate these differences to an overall model that describes the average relationship between *y* and *x*, a series of regressions is defined at the subgroup level.

- In any single subgroup, the regression may differ from the population model.
- Regression relationships in the collection of subgroups may all differ from another.

Toward that end, we define regression coefficients that are unique to the i-th subgroup as:

b_{i0} : average value of *y* at *x* = 0 for subgroup i

b_{i1} : average change in *y* for every increase of 1 unit in *x* for subgroup i.

The difference between coefficients in a subgroup and the population values are known as random effects:

$$u_{i0} = b_{i0} - \beta_0$$
$$u_{i1} = b_{i1} - \beta_1$$

Random effects are a measure of how different the subgroup's intercept and slope are from those of the population model. Turning the relationship around gives the more common form,

$$b_{i0} = \beta_0 + u_{i0}$$
$$b_{i1} = \beta_1 + u_{i1},$$

where *b* denotes a subgroup coefficient, β a population coefficient and *u* a random effect.

Carrying on with the example of predicting Math scores from Reading scores, the new model has random intercepts and slopes, and includes within it the earlier model with population average parameters

$$\begin{aligned} M_{ij} &= b_{i0} + b_{i1}R_{ij} + e_{ij} \\ &= (\beta_0 + u_{i0}) + (\beta_1 + u_{i1})R_{ij} + e_{ij} \\ &= (\beta_0 + \beta_1 R_{ij}) + (u_{i0} + u_{i1}R_{ij}) + e_{ij} \end{aligned}$$

$(\beta_0 + \beta_1 R_{ij})$ is known as the *fixed* part of the model while $(u_{i0} + u_{i1}R_{ij}) + e_{ij}$ is the *random* part of the model.

1.6 Population and Subgroup Models

The two-level model is more complicated than simple regression, but it has an appealing flexibility that compensates for the extra complexity. For example,

- although the belief is that a linear relationship exists in the population as well as in each of the subgroups, the particular patterns may all be different.
- the population slope, β_1, may be positive, while the slope in a particular group may be zero or negative.
- if in a certain subgroup, u_{i0} is zero but u_{i1} is not, then the group's intercept is the same as that in the population, but with a different slope.

The fact that such wide differences can be represented is an extremely valuable and pleasing feature of the multilevel analysis methodology.

1.7 Levels of a Multilevel Analysis

The overall population of individuals in a multilevel analysis is assumed to be structured in a manner that must be accounted for in order to properly assess the effect of a treatment, or to understand a regression relationship.

The hierarchical framework is based on two kinds of population. These populations are known as the levels of the hierarchy. The interconnection between the levels is the key concept in this kind of analysis.

Level 1: Every subgroup in a structured population is itself a population, or more specifically a subpopulation. Individuals within a subpopulation are known as level-1 units. It is assumed that subjects are members of only one subpopulation. In an actual study, a sample of individuals (or level-1 units) is selected from each subpopulation. Part of the inferential problem is to generalize from the sample to its subpopulation. Continuing with Example 1.2, students following a particular career option are level-1 units within each of the career options.

Level 2: Subpopulations are known as level-2 units. A more general way of stating "students are organized within career options" is to state that "level-1 units are nested within level-2 units". It is assumed that level-2 units are part of a population. For example, it is assumed that there is a population of different careers; from this population, a sample of nine subgroups was selected. The nine career subgroups are not exhaustive -many others could have been obtained.

Another aspect of the inferential problem in a multilevel analysis is generalizing from the sample of career options to the larger population of career options.

Conceptually, the sampling framework has this form:

- A sample of level-2 units is selected. For example, a subset of career options from the larger population of career options is selected.
- From each level-2 unit, a sample of level-1 units is selected. For example, a sample of students pursuing a career option is selected.

Example 1.3: A 2-level model for the Mathematics / Reading Scores

The earlier analysis of these data actually consisted of ten unrelated regressions, one for each of the nine career subgroups, plus the total sample analysis. These analyses are unrelated because each subgroup analysis ignores information from subjects in other subgroups.

Recall that the estimates of the population coefficients (see Example 1.2) were (38.0, 0.640).

The following PRELIS syntax (**tatsu2.pr2**) is used to fit a 2-level model to these data:

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=ALL ;
TITLE=Tatsuoka sample of High School and Beyond ;
SY=TATSU.PSF;
ID1=ID1 ;
ID2=Career ;
RESPONSE=Math ;
FIXED=Intcept  Reading ;
RANDOM1=Intcept ;
RANDOM2=Intcept  Reading ;
```

By specifying OUTPUT=ALL, the usual ***.out** file and two additional files are produced, namely a file containing level-1 residuals (***.res** file) and a file containing estimates of level-2 empirical Bayes residuals (***.ba2** file).

In this analysis, all the information available in the total sample contributes to the accuracy of the estimates in all other subgroups.

Partial output from **tatsu2.out** is given below.

(i) Fixed part of the model

```
 ITERATION NUMBER     5

                       +----------------------+
                       | FIXED PART OF MODEL   |
                       +----------------------+

-------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  Intcept                 37.68531       0.76864      49.02825       0.00000
  Reading                  0.65545       0.04066      16.12127       0.00000
```

From the fixed part of the model it follows that $\hat{\beta}_0$ = 37.68531 and $\hat{\beta}_1$ = 0.64434, where $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimated population's intercept and slope respectively.

(ii) -2 ln L

```
                       +----------------------+
                       |   -2 LOG-LIKELIHOOD   |
                       +----------------------+

 -2 LOG-LIKELIHOOD =     815.327159089409
```

The -2 ln L value may be used to construct test statistics. See Example 3.3 for more details.

(iii) Random part of the model

```
                      +----------------------+
                      | RANDOM PART OF MODEL  |
                      +----------------------+
------------------------------------------------------------------------------
  LEVEL 2                        TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
------------------------------------------------------------------------------
  Intcept /Intcept                0.91092      2.32327      0.39208   0.69500
  Reading /Intcept                0.00319      0.10399      0.03068   0.97552
  Reading /Reading                0.00494      0.00669      0.73802   0.46050

------------------------------------------------------------------------------
  LEVEL 1                        TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
------------------------------------------------------------------------------
  Intcept /Intcept               13.55506      1.67523      8.09146   0.00000

        LEVEL 2 COVARIANCE MATRIX

              Intcept   Reading
 Intcept      0.91092
 Reading      0.00319   0.00494

        LEVEL 2 CORRELATION MATRIX

              Intcept   Reading
 Intcept       1.0000
 Reading       0.0476    1.0000

        LEVEL 1 COVARIANCE MATRIX

                Intcept
 Intcept       13.55506

        LEVEL 1 CORRELATION MATRIX

              Intcept
 Intcept       1.0000
```

The interpretation of the level-2 and level-1 part of the output is given in Section 1.7 and Section 1.8 respectively. The empirical Bayes residuals and variances (saved in **tatsu2.ba2**) for the nine career subgroups are given in Table 1.2.

Table 1.2: Empirical Bayes residuals and variances for the nine career subgroups

Career	Predictor no.	EB residual	Variance of EB residual	Fixed effect
1	1	0.10707	0.61606	Intcept
1	2	0.72738E-02	0.26697E-02	Reading
2	1	-0.35773	0.56775	Intcept
2	2	0.63561E-02	0.13663E-02	Reading
3	1	-0.32494	0.62351	Intcept
3	2	-0.11736	0.16012E-02	Reading
4	1	-0.41321	0.49411	Intcept
4	2	0.18879E-01	0.18271E-02	Reading
5	1	0.52911	0.57515	Intcept
5	2	0.30752E-01	0.20819E-02	Reading
6	1	-0.17362	0.66952	Intcept
6	2	0.19155E-01	0.24943E-02	Reading
7	1	-0.69131E-01	0.68260	Intcept
7	2	-0.68889E-02	0.18396E-02	Reading
8	1	1.2686	0.60906	Intcept
8	2	0.92458E-01	0.17909E-02	Reading
9	1	-0.56618	0.69489	Intcept
9	2	-0.50620E-01	0.19943E-02	Reading

Using the results of the fixed part of the model and those given in Table 1.2, estimates of the individual coefficients summarized in Table 1.3 below are obtained. The computation of these values are illustrated for the Undecided career subgroup.

From Table 1.1, $\hat{u}_{90}$ = -0.56618, and from the fixed part of the model, $\hat{\beta}_0$ = 37.68531, so that

$\hat{b}_{90}$ = 37.68531-0.56618 = 37.119.

Similarly, $\hat{u}_{91}$ = -0.0506 and therefore

$\hat{b}_{91} = \hat{\beta}_1 + \hat{u}_{91}$ = 0.65545 - 0.0506 = 0.604.

Table 1.3: Estimates of regression coefficients from a multilevel model

	Career Options								
	Tr	B	M	I	R	P	S	T	U
$\hat{\beta}_{i0}$	37.792	37.328	37.360	37.272	38.214	37.512	37.616	38.954	37.119
$\hat{\beta}_{i1}$	0.663	0.662	0.538	0.674	0.686	0.675	0.649	0.748	0.604

Trades; Business Management; Military Service; Industrial Operations; Real Estate Management; Police or Security; Sales, Teacher Training; Undecided

1.8 Summary Measures of the Random Effects

Recall that the random slopes and intercepts are

$$b_{i0} = \beta_0 + u_{i0}$$

$$b_{i1} = \beta_1 + u_{i1}.$$

These coefficients vary across level-2 units. Because they differ from unit to unit, they can be summarized in the same way other variables are summarized.

The means of the random effects are both zero, and the means of the individual coefficients equal the parameters

$$\text{mean}(u_{i0}) = 0 \qquad \text{mean}(u_{i1}) = 0$$

$$\text{mean}(b_{i0}) = \beta_0 \qquad \text{mean}(b_{i1}) = \beta_1$$

The variability of the random slopes and intercepts is also useful information. Obviously, from the previous example, the intercepts and slopes vary across career options. This variability is summarized as

$$\Phi_{(2)_{0,0}} = \text{Variance}\,(u_{i0})$$

$$\Phi_{(2)_{1,1}} = \text{Variance}\,(u_{i1})$$

Finally, the covariance between u_{i0} and u_{i1} describes the extent to which intercepts and slope values are correlated.

$\Phi_{(2)_{1,0}}$ = Covariance (u_{i1}, u_{i0})

The variances and covariance are collected into a single matrix known, appropriately enough, as the covariance matrix of the random effects.

$$\Phi_{(2)} = \begin{bmatrix} Var(u_{i0}) & \\ Cov(u_{i1}, u_{i0}) & Var(u_{i1}) \end{bmatrix}$$

The estimate of the covariance matrix of random effects from the Mathematics/Reading data as obtained from **tatsu2.out** is

$$\hat{\Phi}_{(2)} = \begin{bmatrix} 0.91092 & \\ 0.00319 & 0.00494 \end{bmatrix}$$

The correlation between random slopes and intercepts is easier to understand than the associated covariance. This is given by

$$Corr(u_{i1}, u_{i0}) = \frac{Cov(u_{i1}, u_{i0})}{\sqrt{Var(u_{i1})Var(u_{i0})}}$$

In the example, the estimate is 0.0476. This means that across career groups, intercepts and slopes do not appear to be strongly related.

Note on Interpretation:

$\Phi_{(2)}$ is not simply the sample covariance matrix of u_{i0} and u_{i1} over the nine career options in this study. Rather, it is the estimated covariance matrix of random slopes and intercepts in the population. To make the distinction clearer the

- sample variance of the nine intercepts, estimated independently in Table 1.1, is 5.86.
- sample variance of the nine random intercepts from Table 1.2 is 0.33
- estimated population variance of the random intercepts from the multilevel model is 0.911.

1.9 The Variance of Level-1 Residuals

One final concept needs to be reviewed. Repeating from above, the random coefficient model for the example can be written in either of two ways, namely

$$\begin{aligned} M_{ij} &= b_{i0} + b_{i1}R_{ij} + e_{ij} \\ &= (\beta_0 + u_{i0}) + (\beta_1 + u_{i0})R_{ij} + e_{ij} \end{aligned}$$

The residual, e_{ij}, is the difference between actual and predicted Mathematics score for the j-th subject. In other words, e_{ij} is the amount of scatter in the equation for the i-th career option.

The most common assumptions regarding these residuals are that

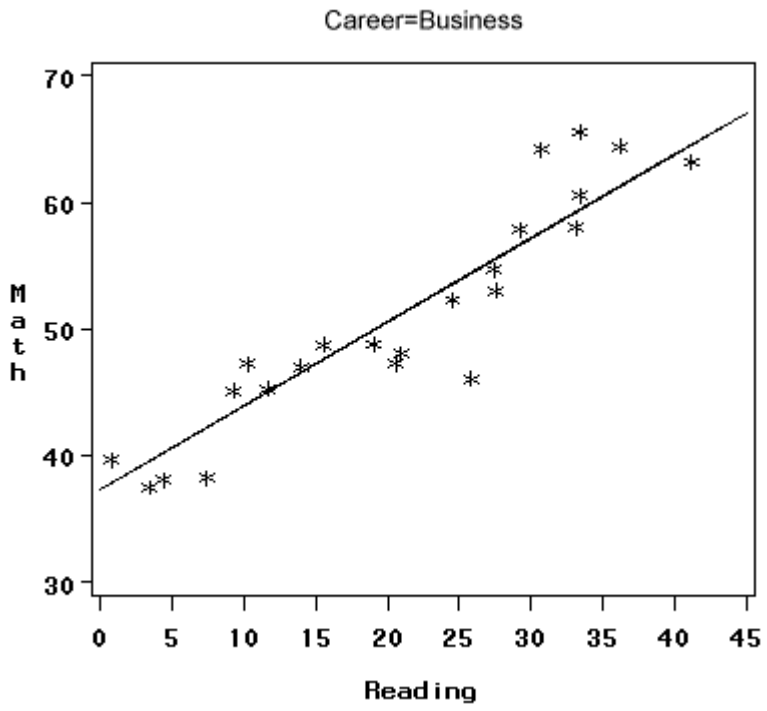
- the scatter is constant everywhere along an individual equation (i.e., lack of fit at reading score R = 12 is the same as at R = 22).
- the scatter is the same across all career options.
- lack of fit for units within one career option is unrelated to lack of fit within other career options.

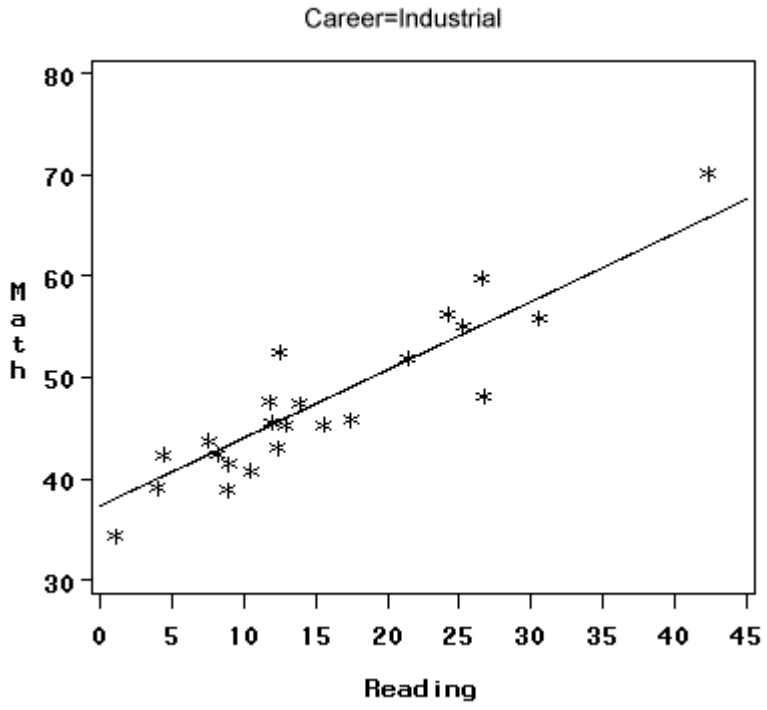
With this background, we need only worry about the variability of e_{ij}. This is written as

$$\Phi_{(1)} = Var(e_{ij}) = \sigma_e^2$$

In the example, the individual regressions fit the data rather well. This can be seen in Figure 1.3. The estimate of σ_e^2 over all subjects and career options is

σ_e^2 = 13.555.


Career=Business
Math
70
60
50
40
30
0
5
10
15
20
25
30
35
40
45
Reading


Career=Industrial
Math
80
70
60
50
40
30
0
5
10
15
20
25
30
35
40
45
Reading

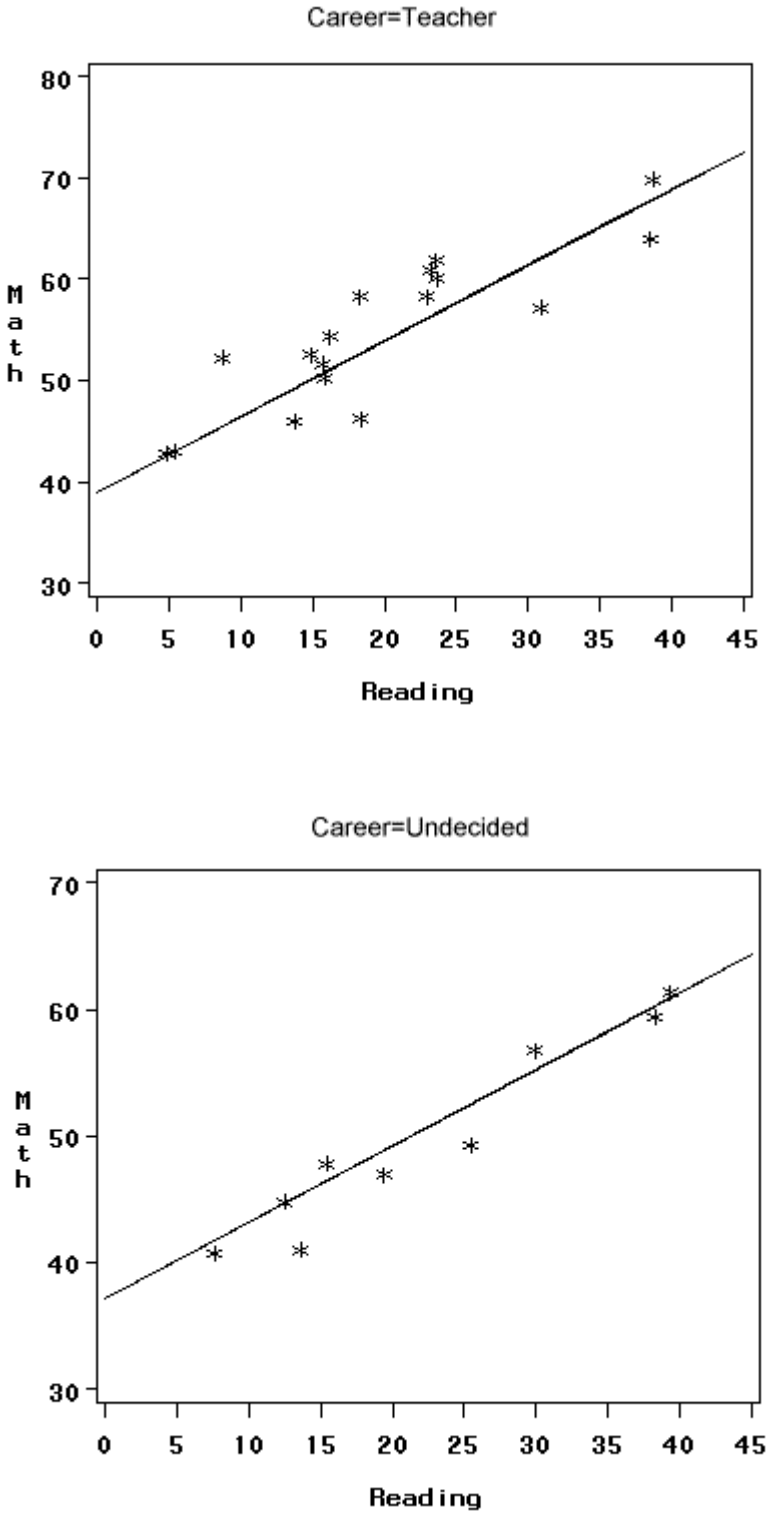


Figure 1.3: Plot of empirical Bayes curves and observed values for four career options

1.10 Grand Mean Centering

An intercept term is generally interpreted as the expected average of the dependent variable, given that all predictors in a model are equal to zero.

Consider the prediction of Mathematics scores from Reading scores using the level-2 model defined and fitted in Examples 1.2 and 1.3. The model formulation is

$$\begin{aligned} M_{ij} &= b_{i0} + b_{i1}R_{ij} + e_{ij} \\ &= (\beta_0 + u_{i0}) + (\beta_1 + u_{i1})R_{ij} + e_{ij} \end{aligned}$$

In this model, the estimated intercept b_{i0} is defined as the expected mean Mathematics score for a student who chose career group *i*, given that Reading score has a value of zero. In the present context, a reading score value of 0 is not meaningful. Consideration should be given to the possible transformation or rescaling of Reading scores that will lead to a more meaningful interpretation of the intercept. The same reasoning may also be used in the case of level-2 predictors, which is not discussed here.

Another example of a situation where rescaling of the predictor may be considered in order to aid interpretation of the results is in the case of a model where a student's test achievement is modeled as dependent on his/her Intelligence quotient. Clearly, the interpretation of the estimated value of b_{i0} in this model as the expected test score for a student from subgroup *i* with IQ quotient equal to zero, is not useful. The same line of reasoning also applies to predictors such as blood pressure, age and weight.

The predictors can be transformed to deviations from the grand mean or to deviations from the subgroup means. However**,** centering in multilevel regression models can have different and unexpected results, depending on the way in which the variables are centered.

The following are two advantages of centering the predictors:

- Obtaining estimates of the intercept, slope and other effects that are easier to interpret, so that the statistical results can be related to the theoretical concerns that motivate the research.

- Removing high correlations between the random intercept and slopes, and high correlations between first-level and second-level variables, and cross-level interactions (for a detailed discussion of this aspect, refer to Kreft and de Leeuw, (1998), pp. 135 to 137).

In the grand mean centered model, the explanatory variable(s) are centered around their overall means. In the model given below, R_M represents the grand mean of all the reading scores, irrespective of career group.

Level-1 model:

$$M_{ij} = b_{i0} + b_{i1}(R_{ij} - R\_M) + e_{ij}$$

Level-1 model:

$$b_{i0} = \beta_0 + u_{i0}$$
$$b_{i1} = \beta_1 + u_{i1}$$

From the equations above it can be seen that the estimated value of b_{i0} is the expected mean value of Math(*M*) , given $R_{ij} = R\_M$, that is, the expected value of a measurement *j* from career option subgroup *i* with Reading score equal to the grand mean of all Reading scores.

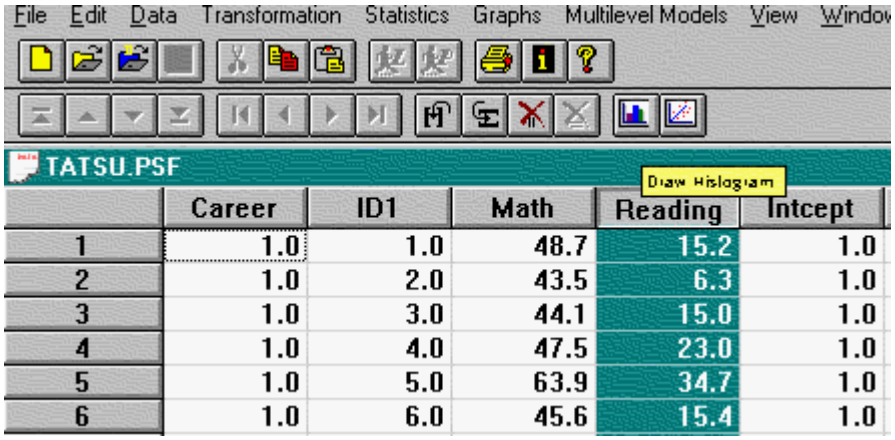
The variance of b_{i0} has a different interpretation, too. It is now the variance between level-2 units in the adjusted means**.**

The parameter β_0 is the adjusted expected mean Mathematics score across career option subgroups, when the Reading score equals its sample mean value.

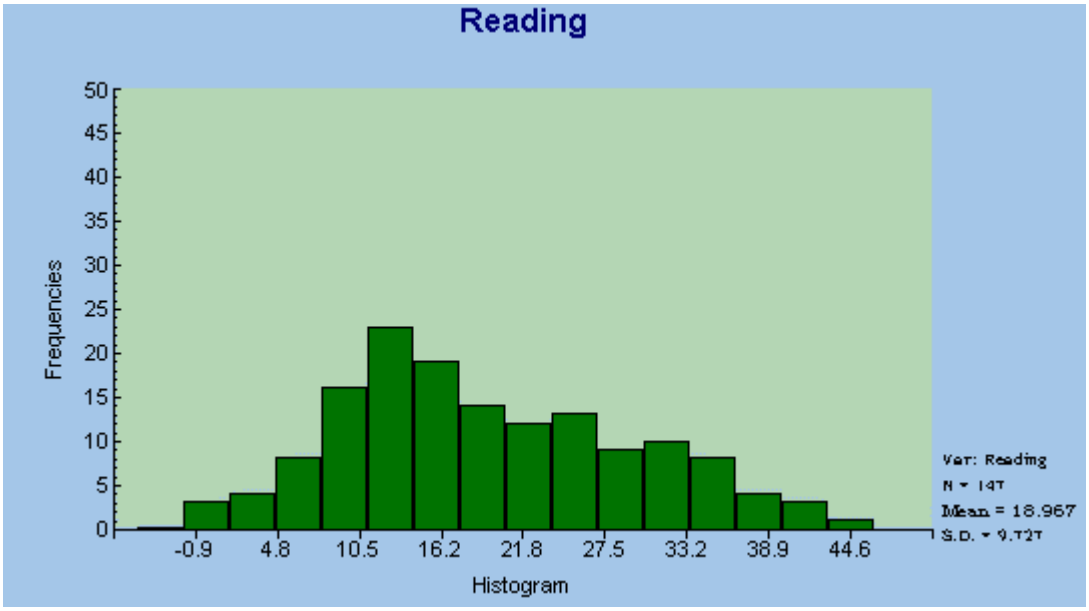
Example 1.4: A 2-level grand mean centered model

It can be shown that the model considered in this section and the model considered in Section 1.5 are equivalent linear models (see Kreft *et al*. (1995) and Kreft & De Leeuw (1998) pp. 106 to 114). Equivalent models will not have the same parameter estimates, but estimates from one can be translated into the estimates from another. They will have the same fit, the same predicted values and the same residuals.

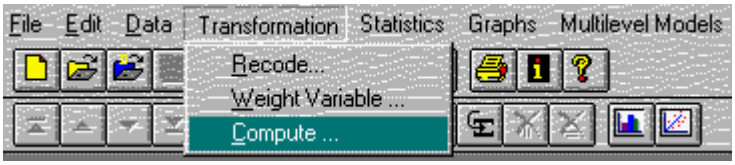
The sample mean of reading scores can be obtained by opening **tatsu.psf**. One could use the **Statistics, Data Screening** option, or, alternatively, obtain the sample mean from the histogram presentation of the variable Reading. To obtain a histogram, left click on the **Reading** label and then left click on the **Draw Histogram** icon.



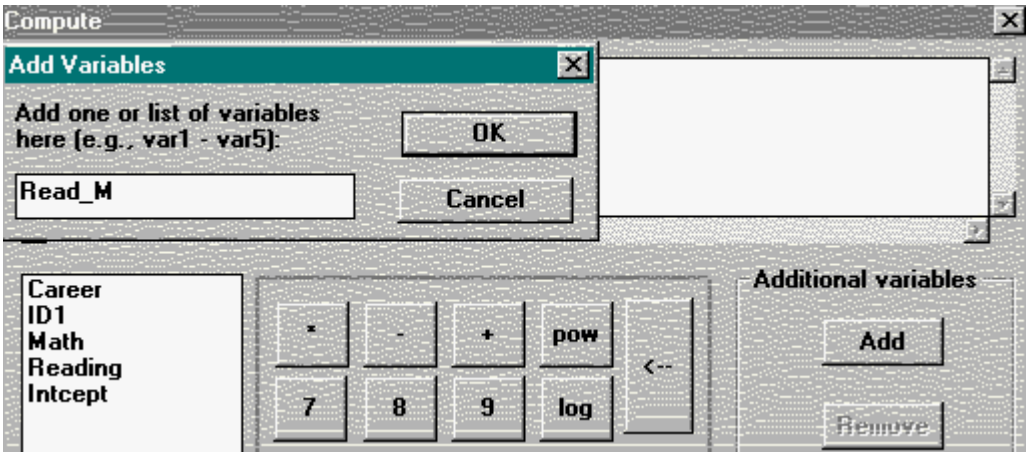
	Career	ID1	Math	Reading	Intcept
1	1.0	1.0	48.7	15.2	1.0
2	1.0	2.0	43.5	6.3	1.0
3	1.0	3.0	44.1	15.0	1.0
4	1.0	4.0	47.5	23.0	1.0
5	1.0	5.0	63.9	34.7	1.0
6	1.0	6.0	45.6	15.4	1.0



From the graph above it is seen that the mean is 18.967. The next step is to add the variable Read_M to the spreadsheet. From the **Transformation** menu, select the **Compute...** option to activate the **Compute** screen.



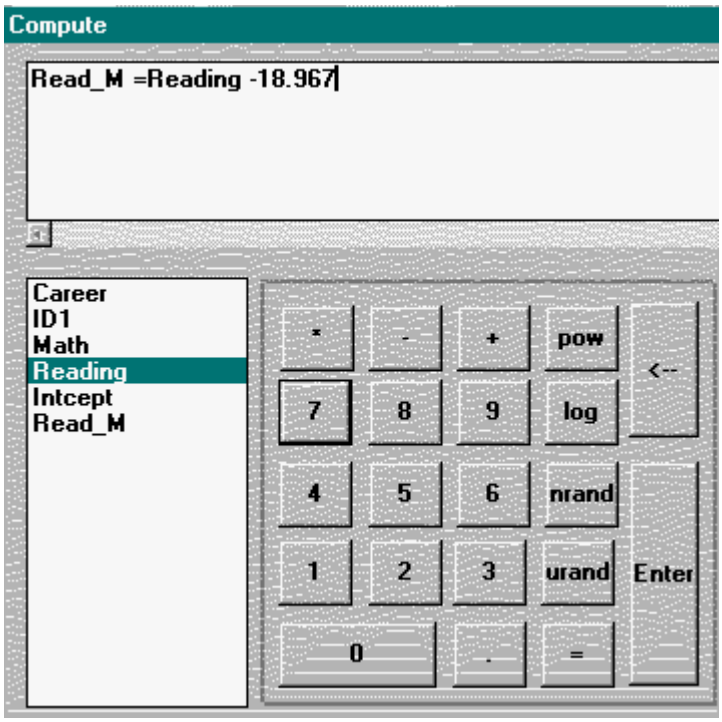
Click the **Add** button in the **Additional variables** field and enter Read_M in the edit box of the **Add variables** screen. Click **OK** when done.



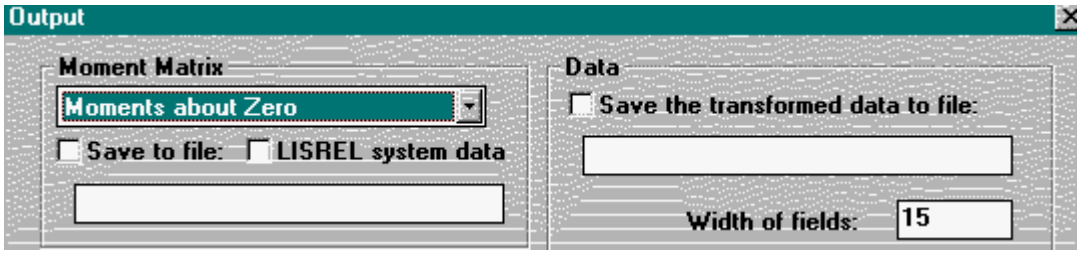
Click on Read_M in the variable list and, with the left mouse button down, drag this label to the **Compute** screen. Click on the equal sign and then drag Reading to the **Compute** screen. Once

Read_M = Reading - 18.967

is displayed, click on the **Output Options** button.



Select the **Moments about Zero** option and click **OK**. This option should be chosen when the data set contains an intercept term. If this is not done, PRELIS may terminate with an error message indicating that the intercept has zero variance.



Make sure that **tatsu.psf** is updated with a new column of values for the variable Read_M. Open **tatsu3.pr2** and run the PRELIS program. The contents of this file are shown below.

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=ALL ;
TITLE=Tatsuoka sample of High School and Beyond - grand mean centering ;
SY=TATSU.PSF;
ID1=ID1 ;
ID2=Career ;
RESPONSE=Math ;
FIXED=Intcept  Read_M ;
RANDOM1=Intcept ;
RANDOM2=Intcept  Read_M ;
```

A portion of the output file is shown below.

(i) Fixed part of the model

```
ITERATION NUMBER     6


                      +-----------------------+
                      | FIXED PART OF MODEL   |
                      +-----------------------+


------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
------------------------------------------------------------------------------
  Intcept                 50.11732       0.64341      77.89362       0.00000
  Read_M                   0.65545       0.04067      16.11523       0.00000
```

The estimated intercept is 50.117. This value is the expected mean Math score for a randomly chosen student with an average Reading score.

(ii) -2 ln L

```
                        +----------------------+
                        |   -2 LOG-LIKELIHOOD  |
                        +----------------------+

 -2 LOG-LIKELIHOOD =     815.327153487148
```

The -2 ln L value is exactly the same as the value obtained for the untransformed Reading scores.

(i) Random part of the model

```
                        +----------------------+
                        | RANDOM PART OF MODEL |
                        +----------------------+

-------------------------------------------------------------------------------
  LEVEL 2                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
  Intcept /Intcept               2.80830      1.75251      1.60245   0.10906
  Read_M  /Intcept               0.09684      0.08387      1.15464   0.24824
  Read_M  /Read_M                0.00494      0.00670      0.73723   0.46099

-------------------------------------------------------------------------------
  LEVEL 1                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
  Intcept /Intcept              13.55518      1.67514      8.09198   0.00000
```

The variance of the intercept term (var(u_{i0})) has increased from 0.91092 (Example 1.3) to 2.80830. It is interesting to note that the correlation between the intercept and the slope has also increased.

1.11 Summary of the Two-level Model

Now we can gather the various pieces together in the whole structure. Denote the dependent and independent variable scores for the j-th individual within the i-th career choice as y_{ij} and x_{ij} respectively.

The linear model:

$$y_{ij} = (\beta_0 + u_{i0}) + (\beta_1 + u_{i1})x_{ij} + e_{ij},$$

where β_0 and β_1 are fixed parameters for the mean curve that applies to all individuals, and u_{i0} and u_{i1} are random effects specific to the j-th career choice.

Level-1 variability (variability for subjects within career choices):

$$\sigma_e^2 = \Phi_{(1)} = Var(e_{ij})$$

Level 2 variability (variability across career choices):

$$\Phi_{(2)} = \begin{bmatrix} Var(u_{i0}) & \\ Cov(u_{i1}, u_{i0}) & Var(u_{i1}) \end{bmatrix}$$

This arrangement means that the variability of y_{ij} is composed of two parts:

variability of y = variability of (u_{i0} , u_{i1}) + variability of e_{ij}

(between careers) (subjects within careers)

2. Hierarchical models for longitudinal data

2.1 Introduction

Hierarchical models have become valuable in many problems in the social sciences because populations are often clustered into subpopulations. It is obviously more realistic to study the effect of a treatment with a hierarchical model in this setting because the subpopulation structure is properly accounted for by the method of analysis.

A second major use of the methodology is in the study of longitudinal or repeated measures data. Instead of students clustered within schools, for example, the basic setup is one in which multiple measurements on the same variable over time are obtained for several individuals. Although some different emphases arise from studies utilizing cluster sampling compared with those arising from repeated measures studies, the methods for analyzing the data exploit many of the same features available with hierarchical models. For a repeated measures application, perhaps the most important feature is the combination of an average profile of change over time that characterizes typical or mean change, plus a collection of individual level models that are potentially different for each case.

2.2 Inclusion of Covariates in the Analysis

The objective of hierarchical models in longitudinal research is to describe a developmental process at the level of the group mean, and especially at the level of individuals.

When the model fits the data adequately, it is desirable to understand how this happens. How do the individual differences represented by this regression approach arise? Which characteristics in the history of the subjects account for variability in initial status, or in rate of change?

A hierarchical model is often expanded to include additional information about the subjects. The way covariates are incorporated into a model is quite interesting. It is based on the following key idea:
Random effects have two interpretations, namely

- they are regression coefficients that quantify the extent to which an independent variable contributes to a dependent variable.
- they are subject-level variables that differ from person to person.

In their capacity as variables, the random effects can be predicted by or correlated with other covariates. If some covariate is correlated with the random effects, then the covariate is an explanation for an important feature (such as initial status or rate of change) of the entire developmental process.

2.3 Dealing with Missing Data

In many studies, data are missing for some subjects. Methods for the analysis of hierarchical models can gracefully accommodate missing data of this kind, as long as the mechanism underlying the missing data is cooperative. If this is true, then even with incomplete data

- estimates of the parameters of the model are obtained that take advantage of whatever information is available.
- arbitrary patterns of missing data are dealt with easily.

There is a sizeable body of literature dealing with missing data (Little & Rubin, 1987, give a convenient overview). A number of comments on this issue follows.

- In a study of cognitive development in children, if dropout occurs in the case of low ability children, then the missing data is not ignorable. The sample becomes biased because low-ability subjects self-select out of the study.
- In a comparison of treatment vs. control group, if treated subjects do not keep their scheduled appointments due to illness, then the missing data is non-ignorable.
- Interestingly, non-response on one variable might be predictable from values on other variables. If non-response is predictable from covariates, but not from the target variable itself, then the missing responses satisfy the missing-at-random assumption.
- Subjects who drop out of a study might be predictable from certain characteristics of their families. As long as the missing data generated by the study drop outs are not directly predicted by the missing values themselves, the non-response is ignorable.
- Judging whether non-response is ignorable is typically decided on a dichotomous basis: yes, it is ignorable, or no, it is not.
- In fact, ignorability is a probability concept based on the conditional distribution of missingness given values of the variable. Strictly speaking, this means there are degrees to which non-response is ignorable. When there are many variables with complicated patterns of missing data, the assessment of ignorability is not an easy task, even in artificial situations.

In practice, it seems that if the missing data are not too obviously non-ignorable, many experienced scientists proceed with methods that can deal with the missing data and are not overly concerned about the consequences. This is not done because they are imprudent or dishonest, but rather because they wish to utilize their valuable data as completely as possible.

Example 2.1: Treatment of prostate cancer

A study was conducted to investigate a treatment for prostate cancer in middle-age men (Cudeck, 1999). The severity of prostate cancer is often assessed by a plasma component known as prostate-specific antigen (PSA), an enzyme that is elevated in the presence of prostate cancer. Lower PSA levels indicate better functioning. PSA levels were collected at baseline and at months 3, 6, 9, and 12 on a sample of 100 men. It is clear from Figure 2.1 that subjects generally improved during the experiment. A small sample of the original data is shown in Table 2.1.

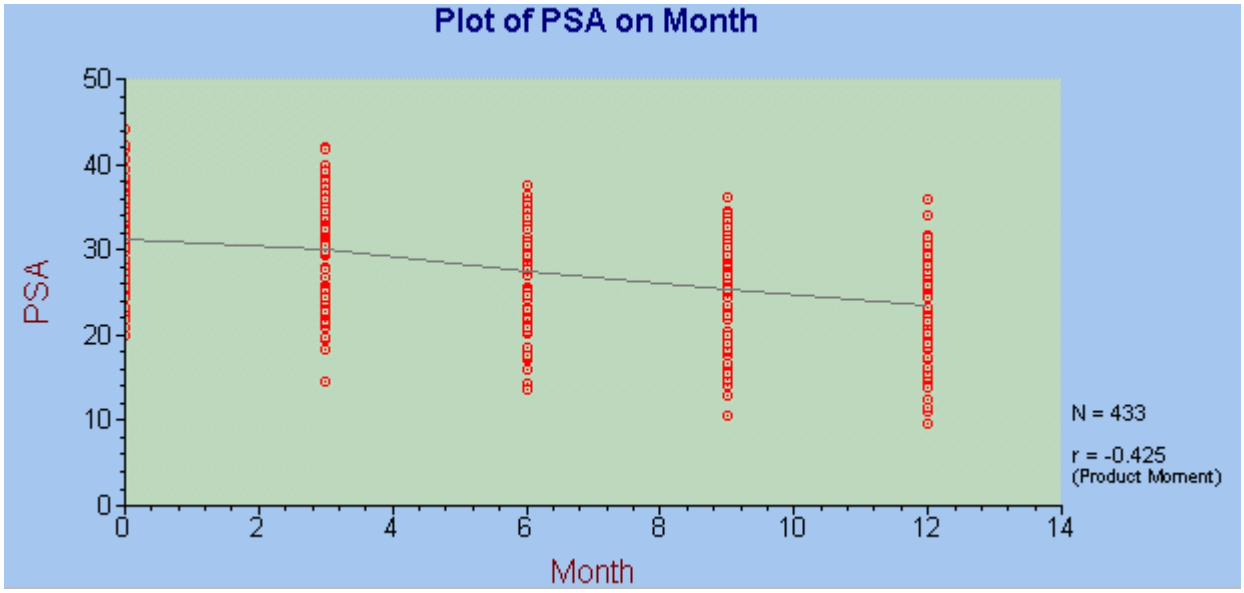


Figure 2.1: Graphical display of PSA levels at months 0, 3, 6, 9 and 12

Table 2.1: Subsample of records from a treatment

Case	Months of experiment					
	Age	0	3	6	9	12
1	69	30.4	28.0	26.9	25.2	19.6
2	58	27.8	26.7	20.5	18.7	18.8
3	53	26.6	21.8	17.8	17.9	14.5
4	61	24.8	24.5	20.2	19.8	18.8
5	63	33.7	30.3	25.4	27.3	20.1
6	49	26.5	24.6	20.9	-9.0	18.9
7	63	26.2	24.4	21.8	22.2	18.4
8	49	24.8	19.5	18.0	16.1	12.5
9	63	28.4	-9.0	22.5	19.4	22.9
10	56	26.1	-9.0	23.3	22.0	14.6
11	68	28.8	31.3	-9.0	23.1	22.8
12	67	29.8	-9.0	25.6	24.5	21.0

Missing data indicated by -9.0.

This example is realistic in that, just as in most studies, data are missing for some subjects. The breakdown is as follows:

	Missing data distribution			
Number missing	0	1	2	3
Number of cases	46	43	9	2

With the prostate cancer data, it seems that the missing data occurs more or less randomly. Lacking evidence that non-response arose on a systematic basis, it can be assumed the missing data is ignorable.

The first 10 cases in the PRELIS speadsheet (**prostate.psf**) are shown below.

PROSTATE.PSF

	ID2	ID1	Month	PSA	Intcept	Age	Mo_Age
1	1.000	1.000	0.000	30.400	1.000	69.000	0.000
2	1.000	2.000	3.000	28.000	1.000	69.000	207.000
3	1.000	3.000	6.000	26.900	1.000	69.000	414.000
4	1.000	4.000	9.000	25.200	1.000	69.000	621.000
5	1.000	5.000	12.000	19.600	1.000	69.000	828.000
6	2.000	1.000	0.000	27.800	1.000	58.000	0.000
7	2.000	2.000	3.000	26.700	1.000	58.000	174.000
8	2.000	3.000	6.000	20.500	1.000	58.000	348.000
9	2.000	4.000	9.000	18.700	1.000	58.000	522.000
10	2.000	5.000	12.000	18.800	1.000	58.000	696.000

From Figure 2.1 it appears that the population regression curve can be adequately described by a linear function.

The corresponding multilevel random coefficient model (MRCM) is

$$PSA_{ij} = b_{i0} + b_{i1}\, Month + e_{ij}.$$

On level-2 of the model, b_{i0} and b_{i1} become outcome variables:

$$b_{i0} = \beta_0 + u_{i0}$$
$$b_{i1} = \beta_1 + u_{i1}$$

As described in Sections 1.7 and 1.8, the covariance matrix of level-2 residuals (u_{i0}, u_{i1}) is denoted by $\Phi_{(2)}$ and the variance of level-1 residuals by $\Phi_{(1)}$ or σ_e^2.

The syntax to fit the above model to the PSA data is contained in the file **prostat2.pr2.**

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
TITLE=;
SY=PROSTATE.PSF;
ID1=ID1 ;
ID2=ID2 ;
RESPONSE=PSA ;
FIXED=Intcept  Month ;
RANDOM1=Intcept ;
```

```
 RANDOM2=Intcept  Month ;
 MISSING_DAT=-9.0;
```

The program took six iterations to achieve convergence. A part of the output file **prostat2.out** is given below.

(i) Fixed part of the model

```
ITERATION NUMBER     6
                       +----------------------+
                       | FIXED PART OF MODEL   |
                       +----------------------+
-------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  Intcept                 31.93378       0.57101      55.92473       0.00000
  Month                   -0.74214       0.01861     -39.86913       0.00000
```

The estimated population intercept and slope are 31.934 and -0.742 respectively. Both these coefficients are highly significant.

(ii) -2 ln L

```
                       +----------------------+
                       |   -2 LOG-LIKELIHOOD   |
                       +----------------------+
 -2 LOG-LIKELIHOOD =     2008.60062393131
```

(iii) Random part of the model

```
                       +----------------------+
                       | RANDOM PART OF MODEL  |
                       +----------------------+

-------------------------------------------------------------------------------
  LEVEL 2                        TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
  Intcept /Intcept              30.89912      4.61240      6.69914   0.00000
  Month   /Intcept               0.30243      0.10758      2.88119   0.00494
  Month   /Month                 0.00392      0.00539      0.72798   0.46663
```

```
------------------------------------------------------------------------------
  LEVEL 1                        TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
------------------------------------------------------------------------------
  Intcept /Intcept               2.28753      0.20830     10.98199   0.00000

        LEVEL 2 COVARIANCE MATRIX

                 Intcept       Month
 Intcept        30.89912
 Month           0.30243     0.00392

        LEVEL 2 CORRELATION MATRIX

               Intcept     Month
 Intcept        1.0000
 Month          0.8685    1.0000

        LEVEL 1 COVARIANCE MATRIX

               Intcept
 Intcept        2.28753

        LEVEL 1 CORRELATION MATRIX

               Intcept
 Intcept        1.0000
```

Example 2.2: Association between age and initial status

The investigators believe that a subject's age is associated with initial status, such that younger subjects have lower overall PSA levels. If this is true, then age should account for some of the differences in intercept that are evident in Figure 2.2.

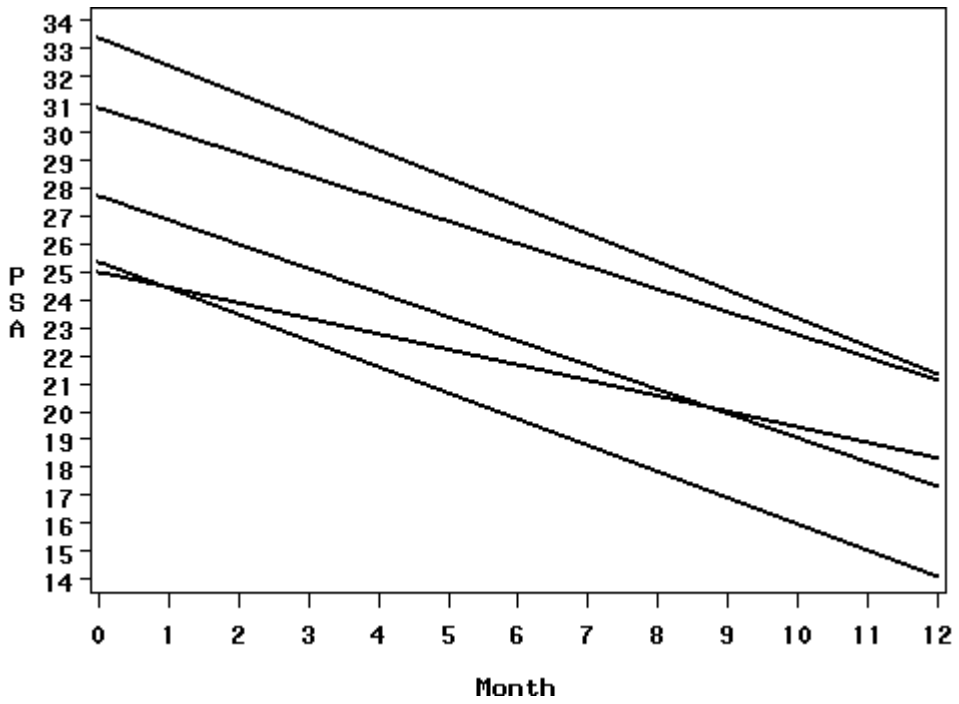


Figure 2.2: Regression lines for cases 1 to 5 of PSA on months

Age was recorded for each subject at the beginning of the experiment. It is known as a time-invariant covariate because, even though subjects get older over the course of the study, age at intake obviously does not change.

Consider the model

$$PSA_{ij} = b_{i0} + b_{i1} Month_j + e_{ij},$$

where

$$b_{i0} = \beta_0 + \gamma\, Age_i + u_{i0}$$
$$b_{i1} = \beta_1.$$

The above model implies that the random intercepts are specified to be a linear function of age.

The question as to whether age affects the initial PSA levels centers on the value of γ :

- If $\gamma = 0$, then there is no association between age and PSA level at the start of the study.
- If $\gamma \neq 0$, then age is associated with initial PSA level.

The value of b_{i0} changes γ units for every increase of one year in age. Substituting the individual intercept into the overall regression, the following is obtained

$$PSA_{ij} = (\beta_0 + \gamma\ AGE_j + u_{0j}) + \beta_1\ Month_j + e_{ij}$$

Syntax to fit the above model to the data is contained in **prostat3.pr2,** which is shown below.

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
TITLE=Treatment of prostate cancer using age as covariate;
SY=PROSTATE.PSF;
ID1=ID1 ;
ID2=ID2 ;
RESPONSE=PSA ;
FIXED=Intcept  Month Age ;
RANDOM1=Intcept ;
RANDOM2=Intcept ;
MISSING_DAT=-9.0;
```

Output is contained in the file **prostat3.out**. It took three iterations for the estimated parameter values to attain convergence. The last part of the output file is shown below.

(i) Fixed part of the model

```
ITERATION NUMBER     3

                        +----------------------+
                        | FIXED PART OF MODEL   |
                        +----------------------+

------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
------------------------------------------------------------------------------
  Intcept                 16.70075       3.94638       4.23192       0.00002
  Month                   -0.74365       0.01785     -41.66932       0.00000
  Age                      0.27513       0.07045       3.90535       0.00009
```

From the fixed part of the model it follows that $\hat{\beta}_0$=16.701, $\hat{\beta}_1$=-0.744 and $\hat{\gamma}$=0.2751.

(ii) -2 ln L

```
                        +----------------------+
                        |   -2 LOG-LIKELIHOOD  |
                        +----------------------+

-2 LOG-LIKELIHOOD =     2003.75379196716
```

(iii) Random part of the model

```
                        +----------------------+
                        | RANDOM PART OF MODEL |
                        +----------------------+

-------------------------------------------------------------------------------
  LEVEL 2                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
  Intcept /Intcept             30.06890      4.33279      6.93984   0.00000

-------------------------------------------------------------------------------
  LEVEL 1                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
  Intcept /Intcept              2.37335      0.18393     12.90344   0.00000
```

From the random part of the model it follows that $\Phi_{(2)} = Var(u_{i0})$ =30.0689, while $\sigma_e^2 = Var(e_{ij})$ =2.37335.

Example 2.3: Correlation between Age and the intercept

Suppose that an estimate of the correlation between b_{i0} and the age of a subject at commencement of the study needs to be obtained. This estimate could indicate to what extent the initial value of PSA is influenced by a subject's age.

The trick to obtaining such an estimate is to add age to the list of dependent variables and to define a third fixed parameter, μ_A, as the mean of age, and the associated random term as the deviation score.

Let

$$Age_i = \mu_A + a_i$$

where a_i has mean zero and variance Var(a_i).

The vector of responses and the design matrix for the fixed effects is coded as follows:

$$y_i = \begin{bmatrix} P_{1i} \\ P_{2i} \\ P_{3i} \\ P_{4i} \\ P_{5i} \\ A_i \end{bmatrix}$$

and

$$X_i = \begin{bmatrix} 1 & M_1 & 0 \\ 1 & M_2 & 0 \\ 1 & M_3 & 0 \\ 1 & M_4 & 0 \\ 1 & M_5 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

where *P* denotes PSA, *M* denotes month and *A* age (in years).

Denote the columns of $\mathbf{X}_i$ by PSA_Icpt, Month and AGE_Icpt respectively. If the coefficients of PSA_Icpt and AGE_Icpt are allowed to vary randomly across individuals, $\Phi_{(2)}$ will have the following structure:

$$\Phi_{(2)} = \begin{bmatrix} Var(u_{i0}) & \\ Cov(a_i, u_{i0}) & Var(a_i) \end{bmatrix}.$$

The revised data set is named **pros_age.psf**, and the first 10 observations are shown below.

PROS_AGE.PSF

	ID2	ID1	PSA	PSA_Icpt	Month	Age_Icpt	Intcept
1	1.000	1.000	30.400	1.000	0.000	0.000	1.000
2	1.000	2.000	28.000	1.000	3.000	0.000	1.000
3	1.000	3.000	26.900	1.000	6.000	0.000	1.000
4	1.000	4.000	25.200	1.000	9.000	0.000	1.000
5	1.000	5.000	19.600	1.000	12.000	0.000	1.000
6	1.000	6.000	69.000	0.000	0.000	1.000	1.000
7	2.000	1.000	27.800	1.000	0.000	0.000	1.000
8	2.000	2.000	26.700	1.000	3.000	0.000	1.000
9	2.000	3.000	20.500	1.000	6.000	0.000	1.000
10	2.000	4.000	18.700	1.000	9.000	0.000	1.000

To run the analysis, open the file **prostat4.pr2** and click the **Run PRELIS** icon. The command file is shown below.

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
TITLE=Prostrate Data: Correlation between Age and Intcept;
SY=PROS_AGE.PSF;
ID1=ID1 ;
ID2=ID2 ;
RESPONSE=PSA ;
FIXED=PSA_Icpt  Month  Age_Icpt ;
RANDOM1=Intcept ;
RANDOM2=PSA_Icpt  Age_Icpt ;
MISSING_DAT=-9.0;
```

Output is written to the file **prostat4.out**. Convergence was attained in four iterations and partial output is given below.

(i) Fixed part of the model

```
Prostrate Data: Correlation between Age and Intcept
 ITERATION NUMBER     4
                       +----------------------+
                       | FIXED PART OF MODEL   |
                       +----------------------+
-------------------------------------------------------------------------------
  COEFFICIENTS              BETA-HAT      STD.ERR.       Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  PSA_Icpt                 31.95664       0.60429       52.88435       0.00000
  Month                    -0.74365       0.01785      -41.66992       0.00000
  Age_Icpt                 55.45000       0.78567       70.57684       0.00000
```

From the output above, it is seen that all the estimated fixed coefficients are highly significant. Note that the value of the age intercept is merely the average age of the subjects, and one would expect this value to be significantly different from zero.

(ii) -2 ln L

```
                      +----------------------+
                      |   -2 LOG-LIKELIHOOD  |
                      +----------------------+
 -2 LOG-LIKELIHOOD =     2699.81445308000
```

(iii) Random part of the model

```
                      +----------------------+
                      | RANDOM PART OF MODEL |
                      +----------------------+
-------------------------------------------------------------------------------
  LEVEL 2                        TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
  PSA_Icpt/PSA_Icpt             34.74141      4.99365      6.95740   0.00000
  Age_Icpt/PSA_Icpt             16.98300      4.96777      3.41864   0.00063
  Age_Icpt/Age_Icpt             59.35415      8.73152      6.79768   0.00000

-------------------------------------------------------------------------------
 LEVEL 1                        TAU-HAT      STD.ERR.     Z-VALUE   PR > Z|
-------------------------------------------------------------------------------
  Intcept /Intcept               2.37335      0.18393     12.90338   0.00000

        LEVEL 2 COVARIANCE MATRIX

               PSA_Icpt    Age_Icpt
 PSA_Icpt      34.74141
 Age_Icpt      16.98300    59.35415

        LEVEL 2 CORRELATION MATRIX

             PSA_Icpt  Age_Icpt
 PSA_Icpt      1.0000
 Age_Icpt      0.3740    1.0000

        LEVEL 1 COVARIANCE MATRIX

               Intcept
 Intcept       2.37335
```

```
        LEVEL 1 CORRELATION MATRIX

              Intcept
 Intcept       1.0000
```

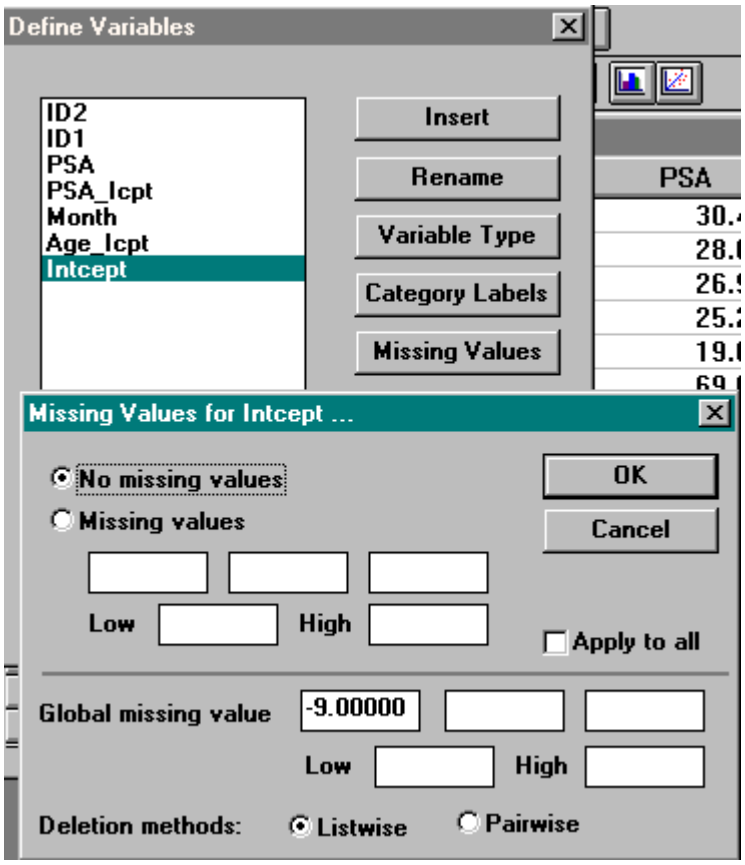
From the level-2 part of the model, the estimated correlation between age and the intercept is found to be 0.37. Since the covariance between the PSA-intercept and the age-intercept is significant, it can be concluded that there is a significant relationship between the age of a subject and the PSA level.

Example 2.4: A structural equation model for finding the correlation between AGE and the PSA intercept

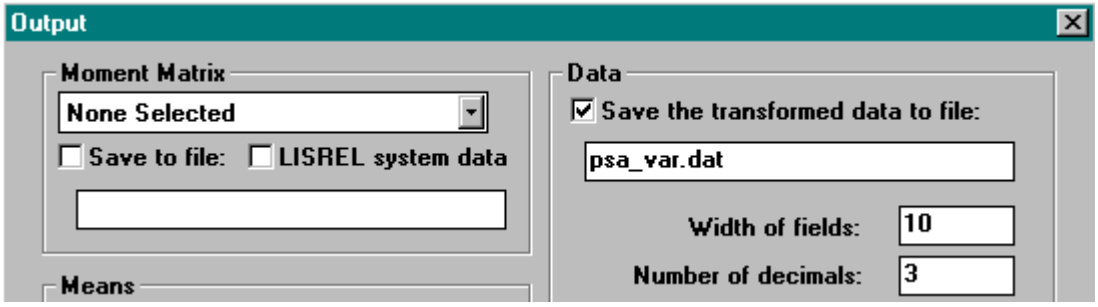
Denote the PSA level measured at occasion 1 by PSA1, etc. To obtain a structural equation model for the prostate cancer data, the sample covariance matrix and the means of the variables PSA1, PSA2, ..., PSA5 and AGE must be calculated. In order to obtain these sample statistics, the following steps are required:

- Creation of a data set with the variables PSA1, PSA2, ..., PSA5 and AGE as columns and the 100 subjects as the rows.
- Use of data imputation to replace missing values where possible.
- Calculation of covariances and means using the listwise deletion option.

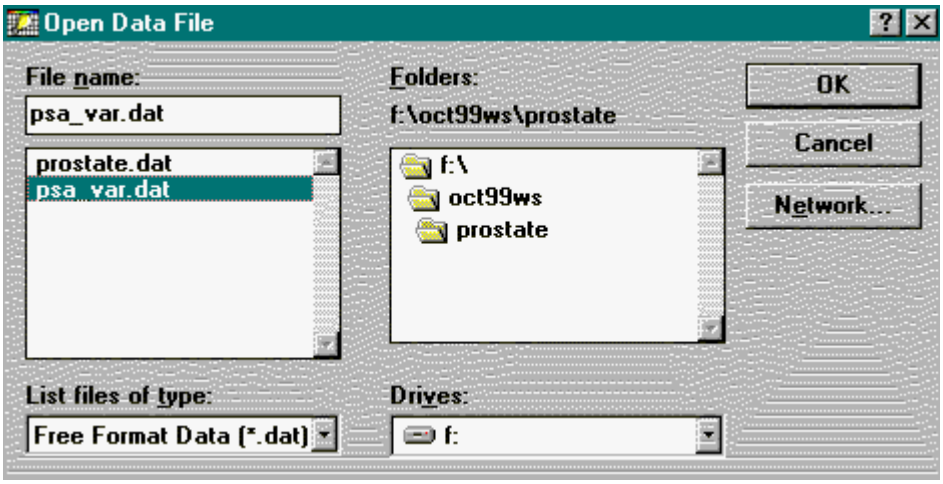
To accomplish these objectives, select **pros_age.psf** from the **File**, **Open** menu. When this data set is displayed in spreadsheet format, click the **Data**, **Define Variables** option and ensure that the **No missing values** option is selected and that no numeric values appear in any of the edit boxes.



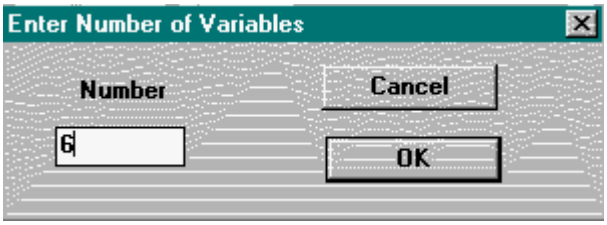
Once this is done, select the **Output Options** item from the **Statistics** menu. Click the **Save the transformed data to file:** check box and enter the name **psa_var.dat**. Click **OK** and return to the **Select Variables** screen. Click **Run**. The data file **psa_var.dat** contains one column and 600 rows of numeric values (-9.0 denotes a missing PSA data value). Each set of six consecutive values is the PSA1 to PSA5 and AGE values for a particular subject.



From the **File** menu select the **Import Data in Free Format** option and from the **Open Data File** menu select the file **psa_var.dat**.



Enter the number of variables (six in this case) and click **OK** when done.

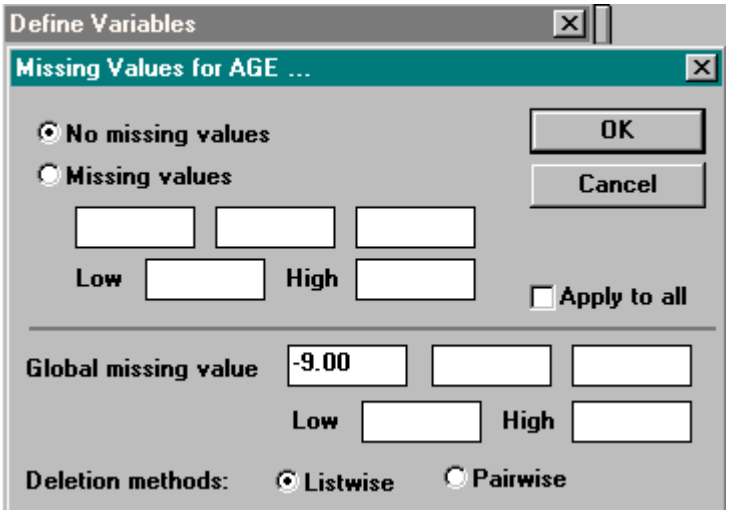


A PRELIS spreadsheet will be displayed. Use the **Data**, **Define Variables** menu to rename the variables from the default VAR1, VAR2, ..., VAR6 names to those shown below. Also, ensure that all variables are defined as continuous.

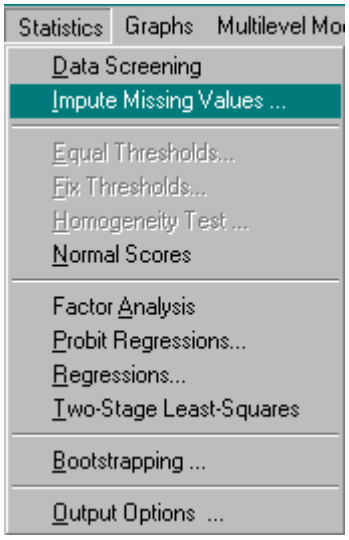
PSA_VAR.PSF

	PSA1	PSA2	PSA3	PSA4	PSA5	AGE
1	30.40	28.00	26.90	25.20	19.60	69.00
2	27.80	26.70	20.50	18.70	18.80	58.00
3	26.60	21.80	17.80	17.90	14.50	53.00
4	24.80	24.50	20.20	19.80	18.80	61.00
5	33.70	30.30	25.40	27.30	20.10	63.00
6	26.50	24.60	20.90	-9.00	18.90	49.00
7	26.20	24.40	21.80	22.20	18.40	63.00
8	24.80	19.50	18.00	16.10	12.50	49.00
9	28.40	-9.00	22.50	19.40	22.90	63.00

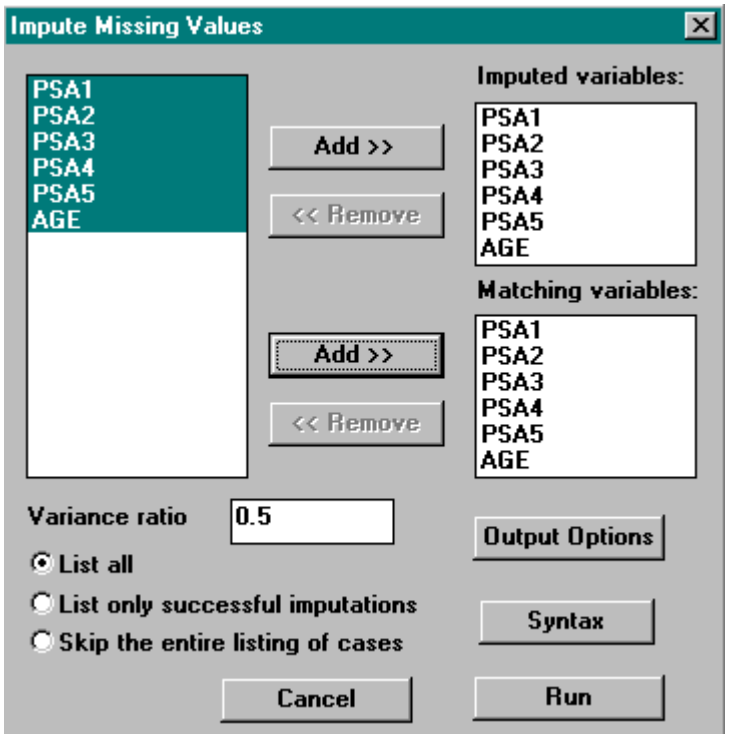
Again, from the **Data**, **Define Variables** menu select the **Missing Values** option and enter -9.0 in the **Global missing value** box. Click **OK** to return to the **Define Variables** screen. Click the **OK** button to return to the main menu.



The next step is the data imputation. From the **Statistics** menu, select **Impute Missing Values**....



Use the **Impute Missing Values** screen to add all the variables to both the **Imputed variables:** and **Matching variables:** list boxes. Once this is done, click the **Output Options** button and select **Covariances** as the **Moment Matrix**. Also, select the **LISREL system data** option.



A portion of the output file is displayed below. After data imputation, the effective sample size is 89.

PSA_VAR.OUT

```
 Total Sample Size =    100

 Number of Missing Values      0     1     2     3
          Number of Cases     89     0     9     2
 Listwise Deletion

 Total Effective Sample Size =     89

 Univariate Summary Statistics for Continuous Variables

 Variable      Mean   St. Dev.   T-Value  Skewness  Kurtosis
 --------      ----   --------   -------  --------  --------
     PSA1    31.515      5.818    51.103     0.049    -0.896
     PSA2    29.469      5.848    47.542    -0.072    -0.721
     PSA3    27.029      5.941    42.921    -0.244    -1.026
     PSA4    24.938      6.277    37.479    -0.252    -1.234
     PSA5    22.751      6.348    33.812    -0.236    -1.076
      AGE    55.360      7.999    65.291    -0.293    -0.283
```

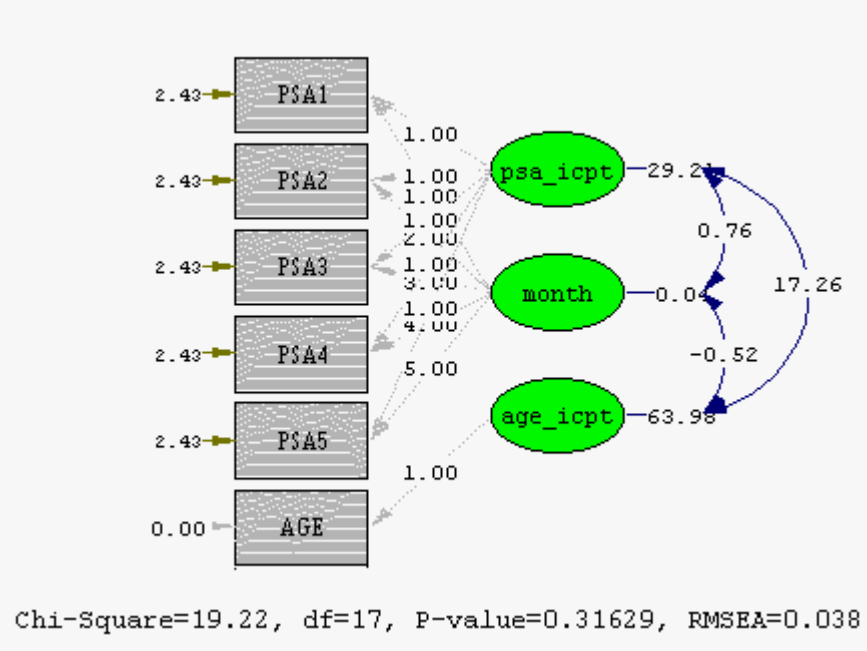
The SIMPLIS input file **pros_age.spl** is shown below. Note that the variances of PSA1 to PSA5 are constrained to be equal. Furthermore, the variance of AGE is set equal to zero to ensure that the variance parameter of age_icpt is identified.

```
Estimation of the correlation between intercept for PSA and intercept for AGE
System File  From File psa_var.dsf
Latent Variables  psa_icpt month  age_icpt
Relationships
PSA1  = 1*psa_icpt  1*month
PSA2  = 1*psa_icpt  2*month
PSA3  = 1*psa_icpt  3*month
PSA4  = 1*psa_icpt  4*month
PSA5  = 1*psa_icpt  5*month
AGE   = 1*age_icpt
psa_icpt month age_icpt  = CONST
Equal Error Variances: PSA1 - PSA5
Set the error Variance of AGE equal to zero
Path Diagram
Iterations = 250
Method of Estimation: Maximum Likelihood
End of Problem
```

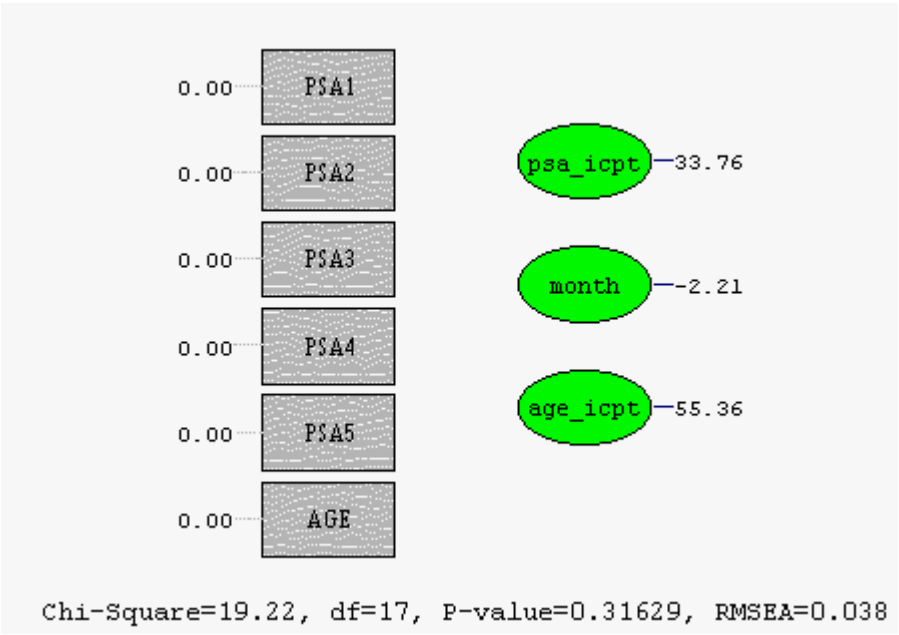
From the path diagram shown below, the estimated correlation between age_icpt and psa_icpt is equal to

$$\frac{17.26}{\sqrt{29.21 \times 63.98}} = 0.40 .$$

The path diagram also displays the goodness of fit chi-square statistic and corresponding p-value. From these values and the value of the RMSEA we conclude that the model fits the data quite well.



By selecting the **Mean Model** from the **Typebar**, **Toolbar** (the **Toolbar** is just below the **run LISREL** and **PRELIS** icons) the estimates of the fixed effects are displayed.



From the path diagram for the **Mean Model** the values of 33.76 and 55.36 are obtained for the estimated PSA and AGE intercepts respectively, and a value of -2.21 for the slope (the month coefficient). Note that the estimated value of the age_icpt coefficient is equal to the sample mean for the variable AGE.

Example 2.5: Use of the FIXVAL, COVnPAT and COVnVAL statements

In this example, the use of the FIXVAL, COVnPAT and COVnVAL statements of the multilevel module to fit the model described in Example 2.4 to the prostate cancer data is illustrated.

The input file **prostat5.pr2** is shown below.

```
OPTIONS OLS=NONE CONVERGE=0.001000 MAXITER=20 OUTPUT=STANDARD ;
TITLE=Prostrate Data: Correlation between Age and Intcept;
SY=PROS_AGE.PSF;
ID1=ID1 ;
ID2=ID2 ;
RESPONSE=PSA ;
FIXED=PSA_Icpt Month  Age_Icpt ;
FIXVAL=
         33.76      -2.21      55.36 ;
RANDOM1= PSA_Icpt Age_Icpt;
COV1VAL= 2.3 0 0.0001;
COV1PAT = 1  0  0;
RANDOM2=PSA_Icpt Month  Age_Icpt;
MISSING_DAT=-9.0;
```

The fixed parameter estimates from the previous analysis are used as initial estimates via the FIXVAL statement. Since these estimates are available, the OLS option was set equal to NONE. In order to set the variation of the Age_Icpt coefficient equal to zero, the COV1VAL statement is used and a value of 0.0001 is entered to indicate that this variance is practically zero. To fix this value at 0.0001, the COV1PAT statement is used. The values 1, 0 and 0 indicate that only the variance of the PSA_Icpt coefficient is allowed to be free.

A portion of the output file is shown below.

(i) Data Summary:

```
                          +--------------+
                          | DATA SUMMARY |
                          +--------------+
NUMBER OF LEVEL 2 UNITS :      100
NUMBER OF LEVEL 1 UNITS :      533

N2  :       1       2       3       4       5       6       7       8
N1  :       6       6       6       6       6       5       6       6
```

```
N2  :       9      10      11      12      13      14      15      16
N1  :       5       5       5       5       5       6       4       3
```

Details are shown for the first 16 subjects. The number of available measurements per subject varies from three to the maximum of six. This implies that the number of available PSA measurements varies from two to five, owing to the fact that AGE records are complete and that AGE is regarded as the sixth measurement.

(ii) Fixed part of the model:

```
                    +----------------------+
                    | FIXED PART OF MODEL   |
                    +----------------------+
------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
------------------------------------------------------------------------------
  PSA_Icpt                31.93422       0.57175      55.85377       0.00000
  Month                   -0.74156       0.01882     -39.41243       0.00000
  Age_Icpt                55.45000       0.78567      70.57678       0.00000
```

The estimated population parameters differ slightly from those reported in Example 2.5. These differences can be attributed to the fact that only the 89 complete records available after imputation were used in the previous example.

(iii) -2 ln L

```
                    +----------------------+
                    |   -2 LOG-LIKELIHOOD   |
                    +----------------------+

 -2 LOG-LIKELIHOOD =     2684.82196668248
```

(iv) Random part of the model:

```
        LEVEL 2 CORRELATION MATRIX

             PSA_Icpt     Month  Age_Icpt
 PSA_Icpt      1.0000
 Month         0.7555    1.0000
 Age_Icpt      0.4081   -0.2849    1.0000
```

```
LEVEL 1 COVARIANCE MATRIX

              PSA_Icpt   Age_Icpt
 PSA_Icpt      2.27349
 Age_Icpt      0.00000    0.00010
```

The estimated correlation coefficient of 0.408 between Age_Icpt and PSA_Icpt is close to the corresponding value obtained in Example 2.4. A similar observation is made concerning the estimated value of the level-1 variance.

3. A growth curve model for Hayashi's Japanese girls data

3.1 OLS Regressions

To illustrate the differences between ordinary least squares regression models (OLS) and multilevel random coefficient models (MRCM), consider the data set shown in Table 3.1.

This data set contains repeated measurements made on 32 Japanese girls over a period of nine years. In Table 3.1, the data for the first girl are shown. The first value in each row is the subject number, the second value the age (in years) at which the measurement was made, and the next four values are various anatomical measurements. A full description of the data are given in Hayashi, 1982.

Table 3.1: Anatomical measurements of 32 Japanese girls over nine years

Girl	Time	Length	Weight	Chest	Crown
1	6	1204	215	560	664
1	7	1262	244	590	694
1	8	1322	282	600	703
1	9	1387	326	625	731
1	10	1454	385	660	758
1	11	1540	415	685	806
1	12	1584	446	730	845
1	13	1610	495	728	863
1	14	1633	516	763	873

In the examples to follow, a number of alternative analyses for this data will be considered. In this section, the change in chest measurements over time is studied.

An ordinary least squares regression line was fitted to the data of each girl. The OLS model can be written as

$$y_j = \beta_0 + \beta_1 t_j + \beta_2 t_j^2 + e_j, \; j = 1, 2, ..., 9,$$

where y_j is the chest measurement at occasion *j*, β_0 and β_1 denote intercept and slope respectively, β_2 the slope of the quadratic term, t_j the occasion of measurement *j*, and e_j the error term.

The results of the OLS regressions are summarized in Table 3.2 and selected regression lines are plotted in Figure 3.1.

Table 3.2: OLS curves for 32 Japanese girls

Girl	Regression coefficients			R^2
	Intercept	Time	Timesq	
1	418.606	22.203	0.17316	0.98423
2	98.212	103.664	-3.79654	0.98288
3	504.955	-0.209	1.61797	0.98318
4	550.439	-8.223	1.94697	0.99332
5	-71.409	129.235	-4.81926	0.98181
6	268.924	47.892	-0.62879	0.96953
7	186.758	68.800	-1.85498	0.95120
8	584.591	-17.798	2.51407	0.98412
9	498.455	-5.402	1.55844	0.98203
10	561.848	-12.358	2.26623	0.96189
11	460.121	15.307	0.36797	0.92232
12	618.152	-24.980	2.49567	0.99549
13	361.682	32.311	-0.03139	0.98028
14	432.939	11.942	1.07792	0.96920
15	489.455	4.726	1.15368	0.98151
16	543.682	-19.032	2.32576	0.91167
17	577.136	-12.786	1.93182	0.96016
18	617.318	-25.978	3.03139	0.97081
19	254.939	90.146	-3.51732	0.89086
20	451.939	10.623	0.95887	0.93388
21	564.667	-7.010	1.35714	0.95570
22	437.470	8.895	1.14610	0.99626
23	504.788	-1.297	1.79654	0.97907
24	640.667	-23.333	2.33333	0.96083
25	391.561	20.361	0.79113	0.96967
26	462.409	16.341	1.46212	0.98481
27	427.879	10.045	1.08442	0.99023

28	655.545	-36.474	2.79879	0.91285
29	626.394	-31.584	2.94589	0.97717
30	395.152	20.220	0.32900	0.96681
31	330.909	30.972	0.25974	0.95828
32	671.606	-44.177	4.29221	0.98861

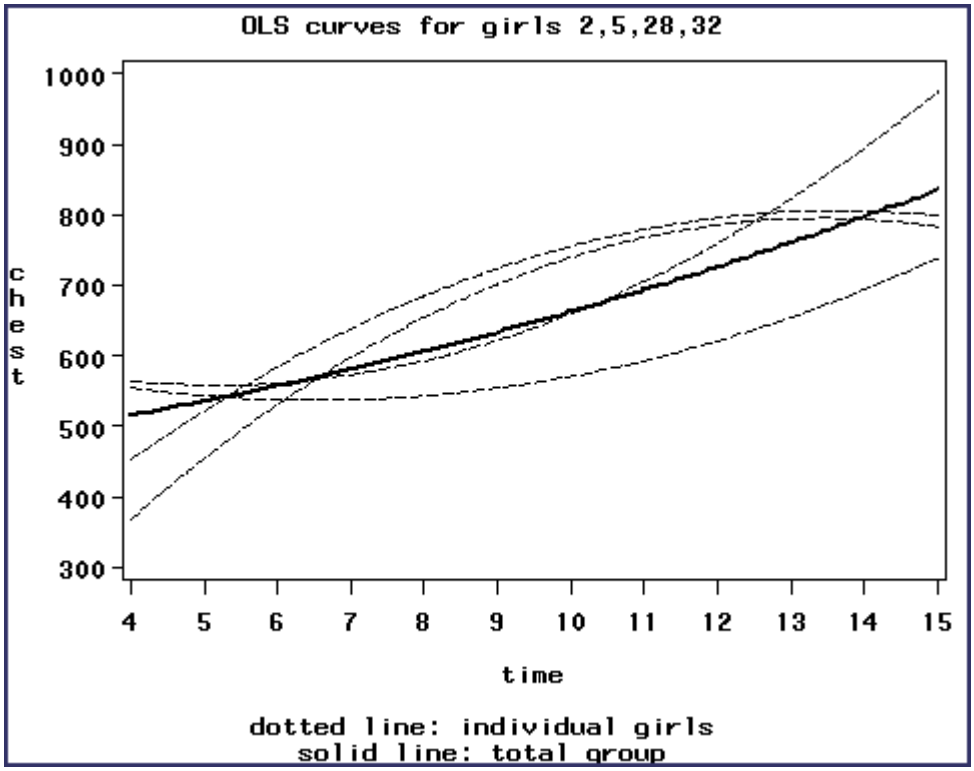


Figure 3.1 Selected OLS curves for growth data

From Table 3.2 and Figure 3.1 it is apparent that there is substantial variation over the intercepts and slopes. Intercepts ranged between -71.409 (girl number 5) and 671.606 (girl number 32), while the coefficient for Time ranged between -44.177 (girl number 32) and 129.235 (girl number 5). The coefficients for Timesq had a minimum value of -4.819 (girl number 5) and a maximum value of 4.292 (for girl number 32). For the total group, the intercept was estimated at 453.681, the coefficient for Time at 11.661 and the coefficient for Timesq at 0.9178. The R^2 for the line fitted to the total group was 0.73036.

3.2. Linear Growth Curve for Hayashi's Data

To distinguish between OLS and MRCM, the intercept-and-slope multilevel random coefficient model is defined as

$$y_{ij} = b_{i0} + b_{i1} t_{ij} + b_{i2} t_{ij}^2 + e_{ij},$$

where the coefficients b_0 and b_1 are assumed to be random variables. The subscript *i* denotes girl *i* and the subscript *j* the *j* -th occasion. It is further assumed that each of the random variables can be written in the form of a regression equation, for example

$$b_{i0} = \beta_0 + u_{i0}$$
$$b_{i1} = \beta_1 + u_{i1}.$$

In the set of equations given above it is assumed that (u_{i0}, u_{i1}) are multivariate normal with mean (0,0) and covariance matrix

$$\Phi_{(2)} = \begin{bmatrix} \text{var}(u_{i0}) & \text{cov}(u_{i0}, u_{i1}) \\ \text{cov}(u_{i1}, u_{i0}) & \text{var}(u_{i1}) \end{bmatrix}$$

and that the (u_{i0}, u_{i1}) are independently distributed from the error term e_{ij}, which has a normal (0, $\Phi_{(1)}$) distribution. β_0 and β_1 are referred to as the fixed part of the model and $\Phi_{(1)}$ and $\Phi_{(2)}$ as the random part of the model.

Example 3.1: A 2-level intercept-and-slopes model

In this example the 2-level MRCM described above is fitted to the Japanese girls data set. It is assumed that an intercepts-and-slopes model will provide an adequate description of the within and between girls variation of the chest measurements. In this example, only linear growth will be considered. The model

$$y_{ij} = b_{i0} + b_{i1} t_{ij} + e_{ij},$$

will be fitted, with

$$b_{i0} = \beta_0 + u_{i0}$$
$$b_{i1} = \beta_1 + u_{i1}.$$

The first 10 lines of the PRELIS data set **japan.psf** are shown below. Note that Time = 1 corresponds to a six-year-old girl, time = 2 to a seven-year-old girl, etc..

JAPAN.PSF

	ident2	ident1	length	weight	chest	crown	time	constant	timesq
1	1.000	1.000	1204.000	215.000	560.000	664.000	1.000	1.000	1.000
2	1.000	2.000	1262.000	244.000	590.000	694.000	2.000	1.000	4.000
3	1.000	3.000	1322.000	282.000	600.000	703.000	3.000	1.000	9.000
4	1.000	4.000	1387.000	326.000	625.000	731.000	4.000	1.000	16.000
5	1.000	5.000	1454.000	385.000	660.000	758.000	5.000	1.000	25.000
6	1.000	6.000	1540.000	415.000	685.000	806.000	6.000	1.000	36.000
7	1.000	7.000	1584.000	446.000	730.000	845.000	7.000	1.000	49.000
8	1.000	8.000	1610.000	495.000	728.000	863.000	8.000	1.000	64.000
9	1.000	9.000	1633.000	516.000	763.000	873.000	9.000	1.000	81.000
10	2.000	1.000	1203.000	218.000	600.000	675.000	1.000	1.000	1.000

The variables Ident2 and Ident1 identify the girl and the observations for each girl. Length, Weight, Chest and Crown represent the measurements at each occasion, Time and Timesq represent the linear and quadratic terms for occasion of measurement, and Constant denotes the intercept.

The input syntax given in the PRELIS command file **japan.pr2** is as follows:

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
TITLE=Growth curve for Hayashi data;
SY=JAPAN.PSF;
ID1=ident1 ;
ID2=ident2 ;
RESPONSE=chest ;
FIXED=constant  time ;
RANDOM1=constant ;
RANDOM2=constant  time ;
```

Partial output from **japan.out** is given below.

```
                          +--------------+
                          | DATA SUMMARY |
                          +--------------+

 NUMBER OF LEVEL 2 UNITS :       32
 NUMBER OF LEVEL 1 UNITS :      288

 N2  :       1       2       3       4       5       6       7       8
 N1  :       9       9       9       9       9       9       9       9

 N2  :       9      10      11      12      13      14      15      16
 N1  :       9       9       9       9       9       9       9       9

 N2  :      17      18      19      20      21      22      23      24
 N1  :       9       9       9       9       9       9       9       9

 N2  :      25      26      27      28      29      30      31      32
 N1  :       9       9       9       9       9       9       9       9
```

The data summary section contains information on the number of units at the different levels of the hierarchy. It can be concluded that there were nine measurements (level-1 units) for each level-2 unit, the level-2 units being the 32 girls. This data set is thus perfectly balanced. One of the advantages of multilevel modeling is that a perfectly balanced design is not a prerequisite for analysis.

This part of the output is followed by details of the fixed part of the model, the -2 log likelihood function value and the estimates of the random components (the variances and covariances of the random coefficients).

(i) Fixed part of the model

```
 ITERATION NUMBER     2

                     +----------------------+
                     | FIXED PART OF MODEL   |
                     +----------------------+
-------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  constant               518.10243       6.81489      76.02503       0.00000
  time                    30.01563       1.03709      28.94216       0.00000
```

From the fixed part of the model it follows that both coefficients are highly significant.

(ii) -2 ln L

```
+-----------------------+
|   -2 LOG-LIKELIHOOD   |
+-----------------------+
```

-2 LOG-LIKELIHOOD = 2736.18244388547

A -2 log likelihood value (the so-called deviance statistic) of 2736.1824 is obtained.

(iii) Random part of the model

```
+-----------------------+
| RANDOM PART OF MODEL  |
+-----------------------+
```

LEVEL 2	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
constant/constant	1243.17948	372.25087	3.33963	0.00084
time /constant	-29.36322	41.89217	-0.70092	0.48335
time /time	26.74444	8.63494	3.09724	0.00195

LEVEL 1	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
constant/constant	460.40010	43.50372	10.58301	0.00000

LEVEL 2 COVARIANCE MATRIX

	constant	time
constant	1243.17948	
time	-29.36322	26.74444

LEVEL 2 CORRELATION MATRIX

	constant	time
constant	1.0000	
time	-0.1610	1.0000

```
       LEVEL 1 COVARIANCE MATRIX

                constant
constant       460.40010

       LEVEL 1 CORRELATION MATRIX

             constant
constant       1.0000
```

There is significant variation in the intercept and in the slope of the linear variable (Time) over the 32 girls. Most of the variation is in the intercept, which indicates the chest measurement of girls at Time = 0, that is, at age five. There is, however, considerable variation in the rate of growth, as indicated by the significant variance component of 26.744 for Time.

Example 3.2: Including a quadratic term in the growth curve

The model fitted in Example 3.1 is now extended to make provision not only for the rate of growth but also for the acceleration in growth, as represented by the variable Timesq in the PRELIS data set **japan.psf**. The model first discussed in Section 3.1 is now fitted to the data:

$$y_{ij} = b_{i0} + b_{i1} t_{ij} + b_{i2} t_{ij}^2 + e_{ij},$$

where $b_{ik} = \beta_k + u_{ik}, k = 0,1,2.$

The contents of the PRELIS command file **japan2.pr2** are shown below.

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=ALL ;
TITLE=Growth curve for Hayashi data non-linear growth curve;
SY=JAPAN.PSF;
ID1=ident1 ;
ID2=ident2 ;
RESPONSE=chest ;
FIXED=constant time timesq;
RANDOM1=constant ;
RANDOM2=constant time timesq;
```

Partial output from the file **japan2.out** is as follows:

(i) Fixed part of the model:

```
                        +-----------------------+
                        | FIXED PART OF MODEL   |
                        +-----------------------+


------------------------------------------------------------------------------
  COEFFICIENTS              BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
------------------------------------------------------------------------------
 constant                534.92783       5.76516      92.78637       0.00000
  time                    20.83814       3.53969       5.88699       0.00000
  timesq                   0.91775       0.35106       2.61421       0.00894
```

(ii) -2 ln L:

```
                        +-----------------------+
                        |   -2 LOG-LIKELIHOOD   |
                        +-----------------------+

 -2 LOG-LIKELIHOOD =     2663.47678201485
```

(i) Random part of the model:

```
                        +-----------------------+
                        | RANDOM PART OF MODEL  |
                        +-----------------------+

------------------------------------------------------------------------------
  LEVEL 2                         TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
------------------------------------------------------------------------------
  constant/constant            591.71638    270.22233      2.18974   0.02854
  time    /constant            -90.69153    127.80375     -0.70962   0.47794
  time    /time                301.45804    100.74831      2.99219   0.00277
  timesq  /constant              7.27857     12.37611      0.58811   0.55646
  timesq  /time                -28.58268      9.77646     -2.92362   0.00346
  timesq  /timesq                2.99756      0.99067      3.02578   0.00248
------------------------------------------------------------------------------
```

```
 LEVEL 1                        TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
 constant/constant            291.44790     29.74578      9.79796   0.00000
```

There is significant variation in the intercept and in the slopes of both the linear and quadratic variables (Time and Timesq) over the 32 girls. Most of the variation is in the intercept, which indicates the chest measurement of girls at Time = 0, that is, at age five. There is a considerable variation in the rate of growth, as indicated by the significant variance component of 301.45804 for Time. The coefficient for growth acceleration, Timesq is also significant but considerably smaller.

Example 3.3: Hypothesis testing

To test the hypothesis that the model discussed in Example 3.2 fits the data better than the intercept-and-slope model of Example 3.1, calculate the difference between the -2 ln L value obtained for the model in Example 3.1 and the - 2 ln L value obtained in Example 3.2. It can be shown that this difference

2736.1824 - 2663.4768 = 72.7056,

has a χ^2 distribution with 10 - 6 = 4 degrees of freedom. The degrees of freedom is the difference in the number of parameters estimated in the two examples. Since the p-value for this test is smaller than 0.01, it is concluded that the model described in Example 3.2 provides a better description of the data than the model fitted in Example 3.1.

Example 3.4: Empirical Bayes estimates and residuals

In the previous example, the model

$$y_{ij} = b_{i0} + b_{i1} t_{ij} + b_{i2} t_{ij}^2 + e_{ij},$$

was considered, where

$$b_{i0} = \beta_0 + u_{i0}$$
$$b_{i1} = \beta_1 + u_{i1}$$
$$b_{i2} = \beta_2 + u_{i2}.$$

Estimates of the fixed coefficients are given in the output file in the previous section. Table 3.3 contains the estimated parameters of these coefficients.

Table 3.3: Estimated parameters for non-linear model fitted to Japanese girls data

Coefficient	Estimate
β_0	534.92783
β_1	20.83814
β_2	0.91775

By specifying OUTPUT = ALL in the OPTIONS paragraph of the syntax file, the program produces a file **japan2.ba2** containing the empirical Bayes residuals (the estimated values of u_0, u_1 and u_2 respectively) and their variances, as well as a file **japan2.res** containing the differences between the observed and predicted chest measurements of the girls.

The empirical Bayes residuals and variances for girls number 2, 5, 28 and 32 are given in Table 3.4 below.

Table 3.4: Empirical Bayes residuals and variances for selected girls

Girl number	Residual	Variance	Predictor
2	1.0760 (u_0)	247.02	intcept
2	36.004 (u_1)	55.634	time
2	-3.7888 (u_2)	0.57238	time_sq
5	-36.734 (u_0)	247.02	intcept
5	39.574 (u_1)	55.634	time
5	-3.8201 (u_2)	0.57238	time_sq
28	-2.3381 (u_0)	247.02	intcept
28	-25.188 (u_1)	55.634	time
28	1.5859 (u_2)	0.57238	time_sq

32	10.872 (u_0)	247.02	intcept
32	-13.839 (u_1)	55.634	time
32	2.4354 (u_2)	0.57238	time_sq

Using the results of Tables 3.3 and 3.4 expected values for girl number 2 are obtained using the following equation:

y = (534.92783+1.0760) + (20.83814+36.004) *time + (0.91775-3.7888)*time_sq .

Similarly, for girl number 5:

y = (534.92783-36.734) + (20.83814+39.574) *time + (0.91775-3.8201)*time_sq .

Shown below are the observed values, predicted values and residuals (e_{ij}) for girl number 2 contained in **japan2.res**. The first column is the observation number, followed by the girl number and the level-1 ID. Columns 4, 5 and 6 are the observed, predicted and residual values respectively.

```
10        2        1      600.00        556.68        43.316
11        2        2      618.00        580.28        37.725
12        2        3      680.00        605.70        74.298
13        2        4      720.00        632.96        87.036
14        2        5      759.00        662.06        96.938
15        2        6      790.00        693.00        97.004
16        2        7      800.00        725.76        74.236
17        2        8      800.00        760.37        39.631
18        2        9      802.00        796.81        5.1913
```

In Figure 3, the residuals for all Japanese girls are plotted by observation number.

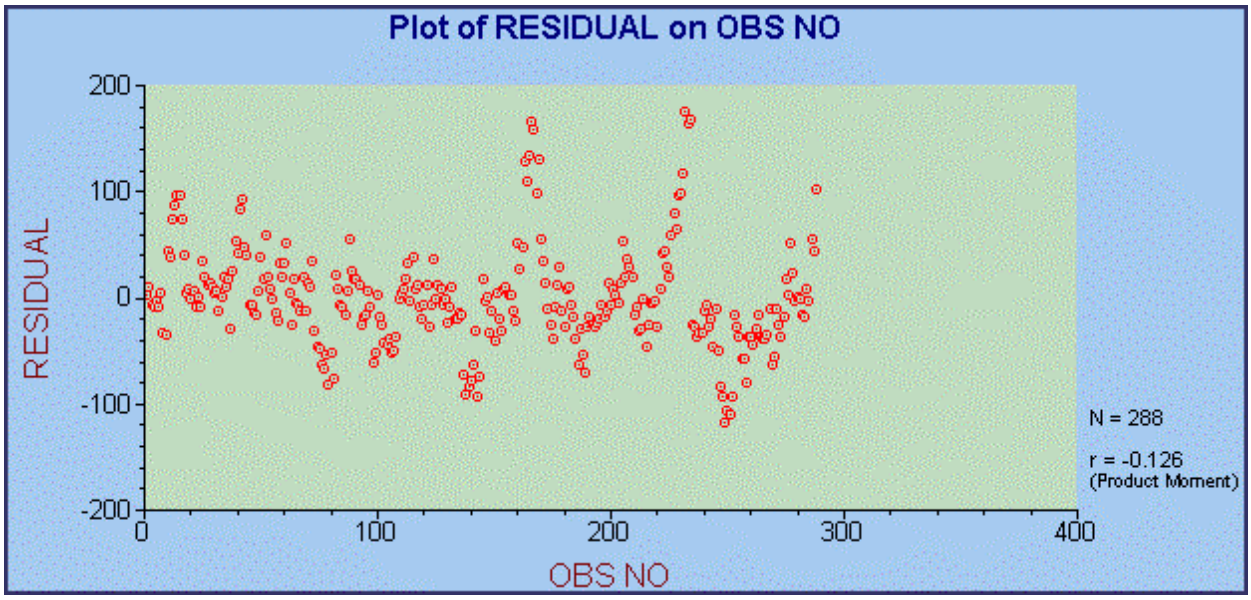


Figure 3.2: Plots of residuals by observation number

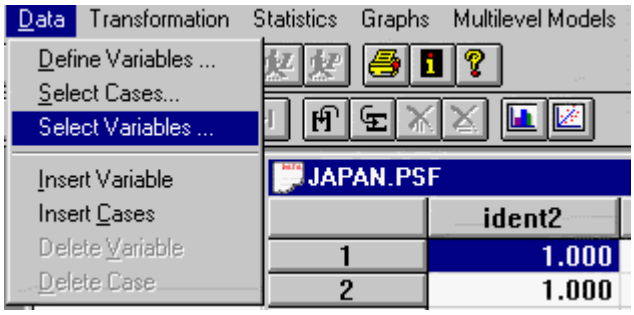
Example 3.5: Structural equation growth curve model

It is relatively straightforward to fit a structural equation growth curve model to the Japanese girls data. This is due to the fact that there is an equal number of repeated measurements on each of the dependent variables for each girl. For illustrative and comparison purposes, consider the dependent variable Chest. To use LISREL or SIMPLIS to fit the structural equation models it is necessary to re-arrange the data in the following manner

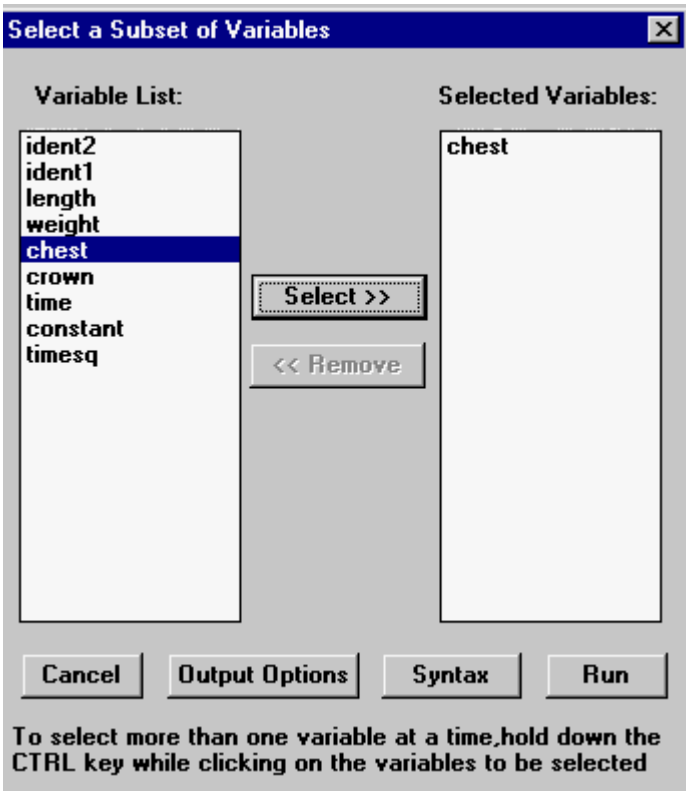
	Chest1	Chest2	Chest3	Chest4	Chest5	Chest6	Chest7	Chest8	Chest9
Girl 1	560	590	600	625	660	685	730	728	763
Girl 2	600	618	680	720	759	790	800	800	802
..	..	..	..	..	..	..	..	..	..
Girl 32	558	580	590	615	670	690	781	805	900

Chest1, Chest2, ... Chest9 denotes the nine consecutive chest measurements for the 32 girls. To obtain a **.psf** file that follows the above layout, proceed as follows:

Use the **File**, **Open** model option to select and display **japan.psf**. From the **Data** menu choose the **Select Variables ...** option.

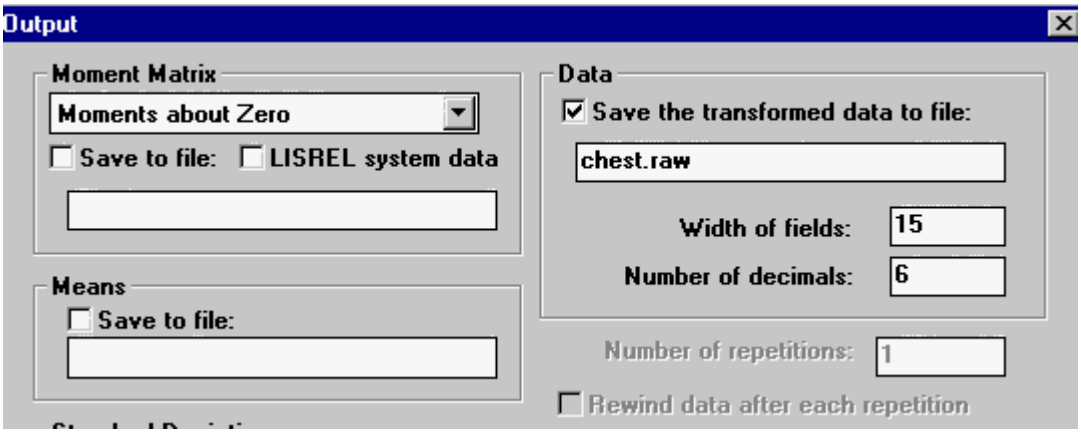


From the **Variable List:**, click on Chest and then click on the **Select >>** button to enter it in the **Selected Variables** list box.

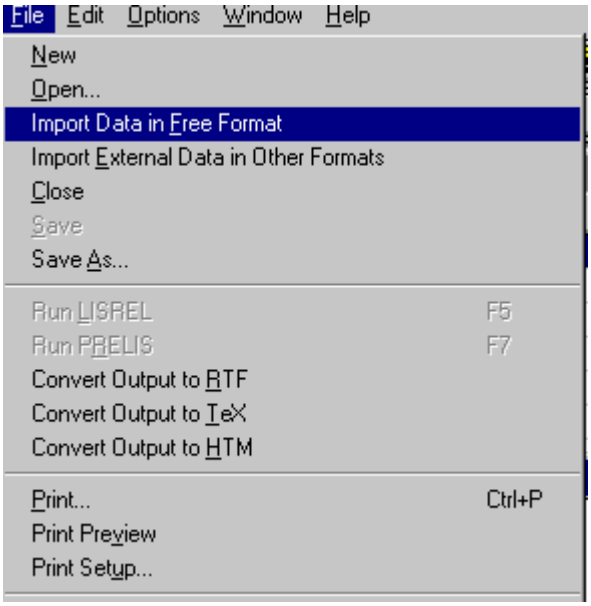


Once this is done, click on **Output Options** to activate the **Output** option screen. From this screen, select **Moments about Zero**, Click the **Save the transformed data to file:** check box and enter **chest.raw** as illustrated below.

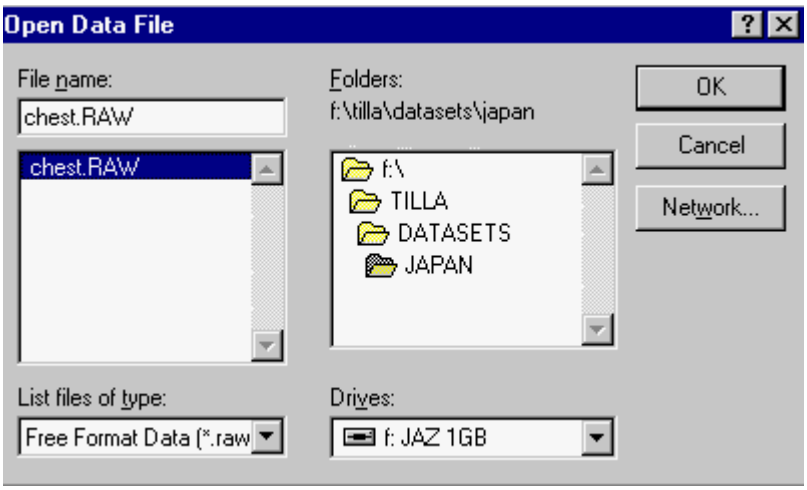
Note: Moments about Zero is selected otherwise PRELIS may terminate with an error message indicating that the variance of the intercept variable is zero.



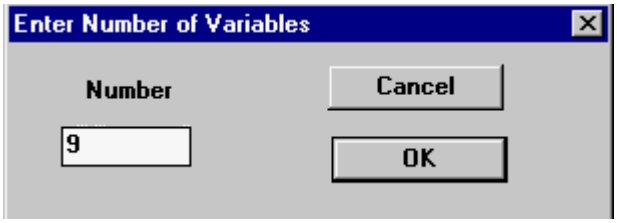
Click the **OK** button on the **Output** options screen to return to the **Select a Subset of Variables** screen. Click the **Run** button to create the file **chest.raw**. Note that **chest.raw** contains the chest measurements stacked in one long column, where the first nine rows are the measurements for girl number 1, and so on. To create a ***.psf** file, select the **Import Data in Free Format** option from the **File** menu.



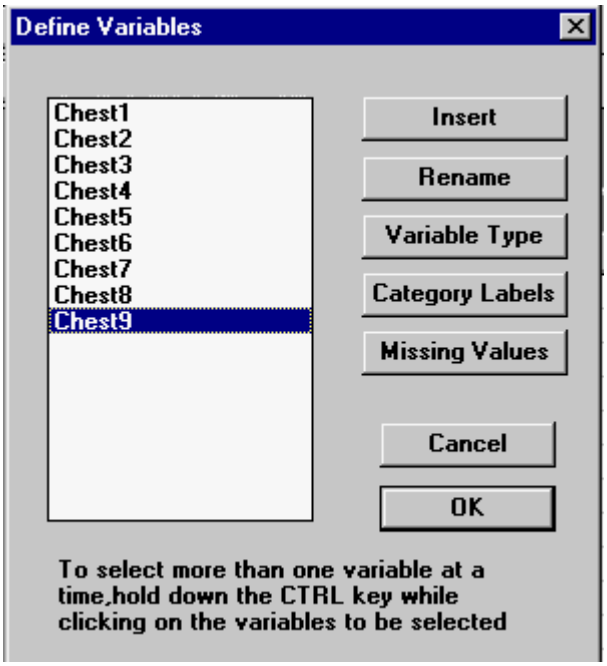
Select the drive and directory where the file **chest.raw** is located and click **OK** when this task is performed.



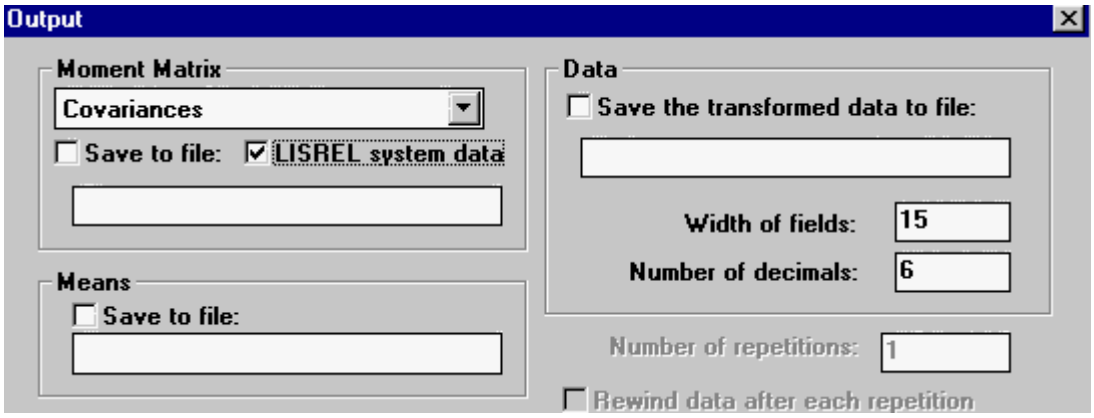
Once the free format data file is selected, the **Enter Number of Variables** screen is activated. Type in 9 and click **OK.**



The file **chest.psf** is created and displayed with the default variable names Var 1, Var 2, . . ., Var 9. Use the **Data**, **Define Variables** option and the **Rename** button to change these names to Chest1, Chest2 . . ., Chest9. The variable type should be changed to continuous. This is achieved by clicking on any variable, then select the **Variable Type** option. Choose **continuous** and the **Apply to All** option.



From the **Statistics** menu, choose **Output Options** and select **Covariances** as the **Moment Matrix** option. Click on the **LISREL system data** check box, and then **OK** to run PRELIS program. In doing so, a data system file, **chest.dsf** is created. This file can be read by LISREL using either SIMPLIS or LISREL syntax.



The observed variables are Chest1, Chest2, . . . , Chest9 and the latent variables are labeled intcept, time and timesq.

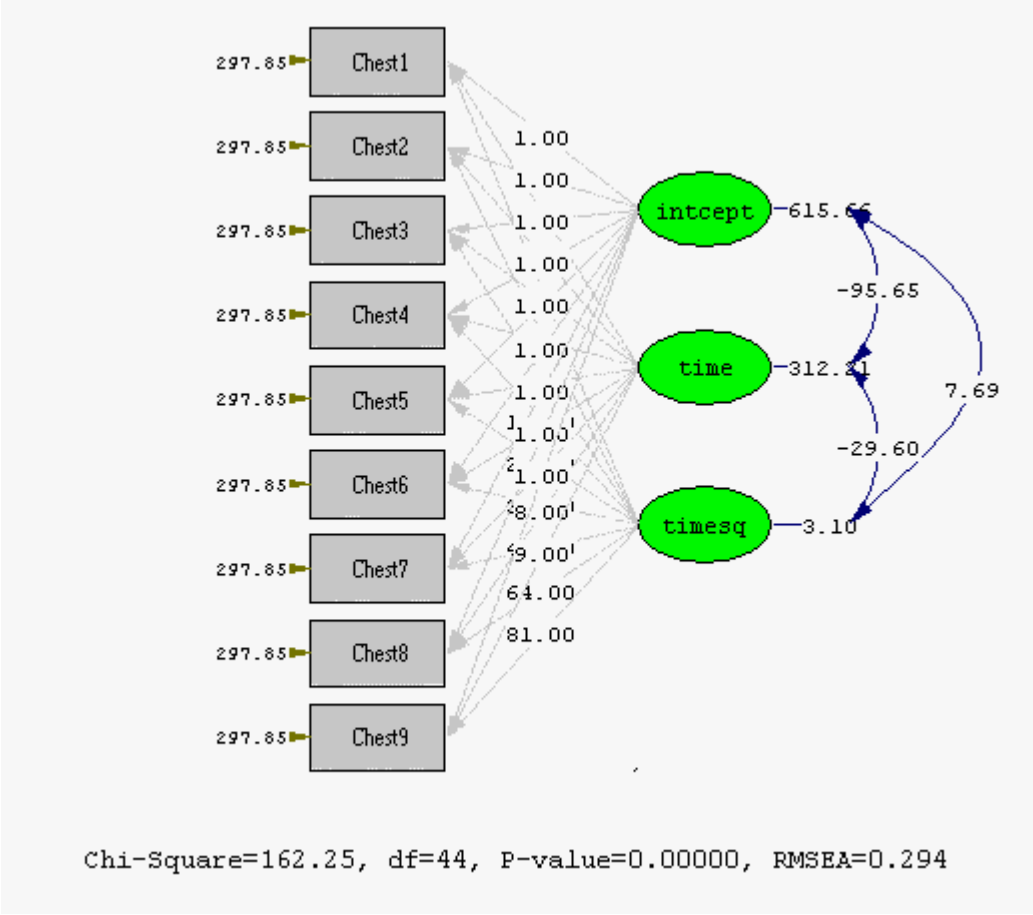
The SIMPLIS input file **japan.spl** is shown below.

```
Non Linear Growth Curve Japanese Girls over 9 years
System File from file CHEST.DSF
Latent Variables  intcept time timesq
```

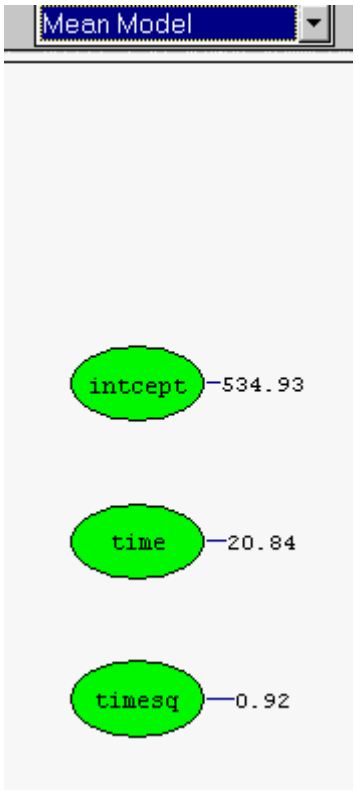
```
Relationships
Chest1  = 1*intcept  1*time 1*timesq
Chest2  = 1*intcept  2*time 4*timesq
Chest3  = 1*intcept  3*time 9*timesq
Chest4  = 1*intcept  4*time 16*timesq
Chest5  = 1*intcept  5*time 25*timesq
Chest6  = 1*intcept  6*time 36*timesq
Chest7  = 1*intcept  7*time 49*timesq
Chest8  = 1*intcept  8*time 64*timesq
Chest9  = 1*intcept  9*time 81*timesq
intcept time timesq = CONST
Equal Error Variances: Chest1 - Chest9
Path Diagram
Iterations = 250
Method of Estimation: Maximum Likelihood
End of Problem
```

The `Equal Error Variances:` statement is used to constrain the variances of the level-1 residuals to be equal.

The X-model part of the path diagram is shown below. From the fit statistics it is apparent that a quadratic term does not describe the change in chest measurement values adequately. Furthermore, the modification indices indicate that the level-1 errors are correlated.

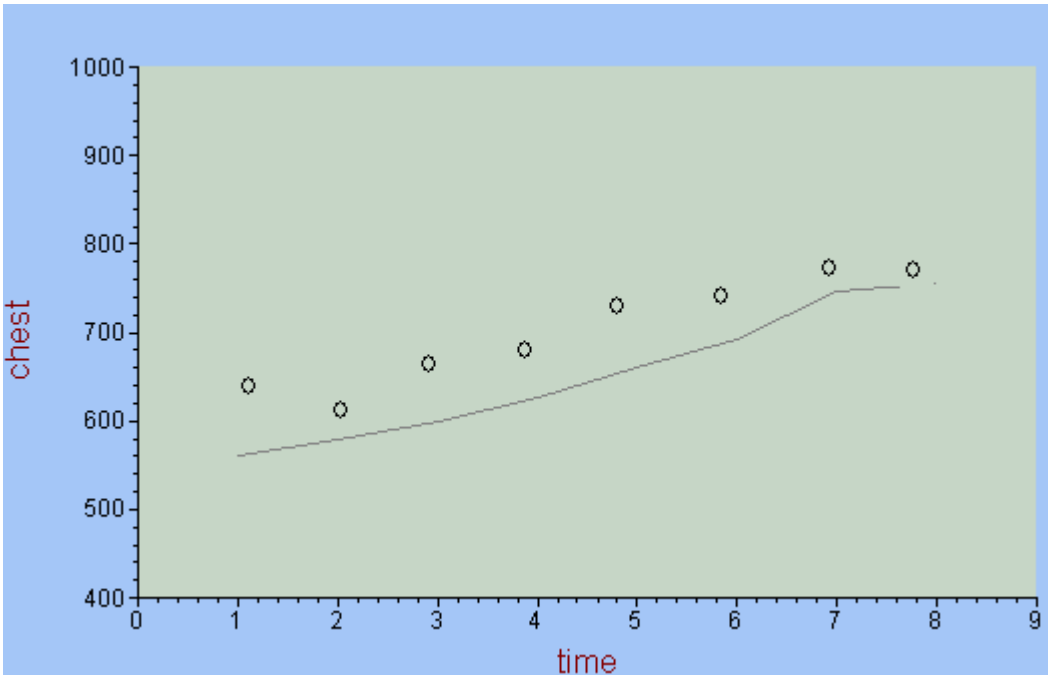


A portion of the mean part of the path diagram is displayed below. The estimated fixed parameters are shown on this diagram. The issue of correlated level-1 residuals is discussed in Example 3.6.



Example 3.6: Correlated level-1 residuals

In the case of longitudinal studies, it may not be possible to fully explain the correlation between successive values of an outcome variable within a standard hierarchical linear modeling framework. The reason for this is that it may be too restrictive to assume that the level-1 error terms have constant variances and are uncorrelated. These assumptions have been used in the results discussed up to now.



The figure above shows the average trend for the chest measurements (solid line) and the observed measurements (circles) of one of the Japanese girls.

From the figure, it appears that if a girl has a higher than average chest measurement at the first occasion of measurement (years, in this case), it is likely that this trend will continue over time. It is also apparent that the magnitude of these differences may change from year to year.

A suitable model to describe this type of behavior is the so-called Autoregressive-Moving Average (ARMA) model, fitted to the level-1 residuals.

Consider the following level-2 model, where *C* denotes chest circumference and *T* denotes the time at which the measurement was taken (*T* = 1 corresponds to age 6, *T* = 2 to age 7, etc.):

$$C_{ij} = b_{i0} + b_{i1}T_j ... + e_{ij},\ j = 1, 2, ..., 9$$

The subscript *i* denotes girl *i* , while the subscript *j* denotes the *j*-th measurement. In this example, however, only the first eight chest measurements will be used.

On level-2 of the model the outcome variables are

$$b_{i0} = \beta_0 + u_{i0}$$
$$b_{i1} = \beta_1.$$

An ARMA model fitted to the residuals $e_{i1}, e_{i2}, ..., e_{i9}$ has the following property:

$$Cov(e_{ij}, e_{ik}) = \sigma^2 \psi_{|j-k|},\ j > k$$

where σ^2 is the so-called white noise variance. The equation above implies that $\Phi_{(1)}$ (the covariance matrix of the residuals) has a Toeplitz structure. A Toeplitz matrix is a matrix in which all the elements on the main diagonal are equal, all the elements below the main diagonal are equal, etc.

$$\begin{bmatrix} \alpha & . & . & . \\ \beta & \alpha & . & . \\ \gamma & \beta & \alpha & . \\ . & \gamma & \beta & \alpha \end{bmatrix}$$

The LISREL multilevel program is based on the assumption that level-1 error terms are uncorrelated, whereas the error terms at higher levels of the hierarchy are allowed to be correlated.

By using dummy variables, it is possible to obtain a correlated structure for level-1 residuals.

Consider the following data:

chest	intcept	time	dummy1	dummy2	dummy3
560	1	1	1	0	0
590	1	2	0	1	0
600	1	3	0	0	1

For each time point, a dummy variable is created, where dummy *j* = 1 if time = *j*, and 0 otherwise.

For the above data, the chest measurement model is

$$C_{i1} = \beta_0 + \beta_1 T_1 + u_{i0} + e_{i1}$$
$$C_{i2} = \beta_0 + \beta_1 T_2 + u_{i0} + e_{i2}$$
$$C_{i3} = \beta_0 + \beta_1 T_3 + u_{i0} + e_{i3}$$

which can be rewritten as

$$\begin{bmatrix} C_{i1} \\ C_{i2} \\ C_{i3} \end{bmatrix} = \begin{bmatrix} 1 & T_1 \\ 1 & T_2 \\ 1 & T_3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_{i0} \\ e_{i1} \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

or

$$C_i = X_{(f)i}\beta + X_{(2)i}u_i^* .$$

Note that this model has no level-1 random part, since the level-1 residuals (via the dummy variables) are now incorporated into the level-2 random part of the model.

We can now use the COVnPAT statement to impose specific structures on the elements of the covariance matrix and, optionally, the COVnVAL statement to assign values to the elements of the covariance matrix.

The input file **chest_ar.pr2** is shown below.

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=30 OUTPUT=STANDARD ;
TITLE= Correlated errors for chest measurements Japanese girls;
SY= CHESTCOR.PSF;
ID2=ident2 ;
ID3=intcept;
RESPONSE=chest;
FIXED=intcept  time;
RANDOM3=intcept;
RANDOM2=intcept  dummy1:dummy8;
COV2PAT = 1
          0  3
          0  5  3
          0  8  5  3
          0 12  8  5  3
          0 17 12  8  5  3
          0 23 17 12  8  5  3
          0 30 23 17 12  8  5   3
          0 38 30 23 17 12  8  5   3 ;
```

Remarks:

- The values in the first column of the COV2PAT matrix are set to zero from rows 2 to 9. This specifies that the covariances between the level-2 intercept and the dummy variables are fixed at the default value of zero, implying that the level-2 and level-1 residuals are uncorrelated.
- The notation dummy1:dummy8 is used to represent dummy1, dummy2, ..., dummy8.
- The LISREL multilevel program requires at least two consecutive levels of IDn and RANDOMn statements. At present, all the random components are included in the level-2 part of the model. To exclude the random-3 part, set ID3 = intcept. Since all values of the variable intcept are equal to 1, there is only one level-3 unit. In this case, the program fixes the random-3 coefficient for the intercept term to a value of zero. Hence the statement RANDOM3 = intcept has no further effect in the estimation procedure.
- The count sequence of the COV2PAT matrix is row-wise. The value of three on the main diagonal instructs the program to constrain all the variances to be equal to the variance of e_{i1}. Similarly, the value of five below the main diagonal is used to constrain all the covariances between e_{ij} and e_{ik}, which are one time unit apart, to be equal.

The program converged in 24 iterations. Partial output for iteration 24 is given below.

(i) Fixed part of the model

```
Correlated errors for chest measurements Japanese girls

 ITERATION NUMBER    24

                      +----------------------+
                      | FIXED PART OF MODEL   |
                      +----------------------+

-------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  intcept                521.60063       8.55373      60.97933       0.00000
  time                    29.34382       1.18088      24.84913       0.00000
```

(ii) -2 ln L

```
                      +----------------------+
                      |   -2 LOG-LIKELIHOOD   |
                      +----------------------+

 -2 LOG-LIKELIHOOD =     2369.47734530539
```

(i) Random part of the model

```
        LEVEL 2 CORRELATION MATRIX

              intcept    dummy1    dummy2    dummy3    dummy4    dummy5
 intcept       1.0000
 dummy1        0.0000    1.0000
 dummy2        0.0000    0.7484    1.0000
 dummy3        0.0000    0.6181    0.7484    1.0000
 dummy4        0.0000    0.4704    0.6181    0.7484    1.0000
 dummy5        0.0000    0.3167    0.4704    0.6181    0.7484    1.0000
 dummy6        0.0000    0.2711    0.3167    0.4704    0.6181    0.7484
 dummy7        0.0000    0.2556    0.2711    0.3167    0.4704    0.6181
 dummy8        0.0000    0.0000    0.2556    0.2711    0.3167    0.4704
```

	dummy6	dummy7	dummy8
dummy6	1.0000		
dummy7	0.7484	1.0000	
dummy8	0.6181	0.7484	1.0000

It is left to the reader as an exercise to show that -2 ln L for the uncorrelated level-1 residual model is 2464.065. In this case, the estimated intercept and slope values are 519.25 and 29.67 respectively, while the estimated level-2 and level-1 variances are 1606.35 and 601.170 respectively.

For the correlated errors model, -2 ln L equals 2369.477 and the number of estimated parameters is 11. The chi-square statistic for testing the null hypothesis that the level-1 residuals are uncorrelated is 2464.065-2369.477 = 94.588, with degrees of freedom equal to 11 - 4 = 7. Since the chi-square value is highly significant, the null hypothesis is rejected in favor of the hypothesis that the residuals are correlated.

References

Aitkin, M.A. & Longford, N.T. (1986). *Statistical modeling issues in school effectiveness studies*. Journal of the Royal Statistical Society A, **149**, 1-43.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Sage Publications.

Cudeck, R. (1999) *Course Notes for Repeated Measurement Models*. Department of Psychology, University of Minnesota, Minneapolis.

Du Toit, S.H.C, du Toit, M., Jöreskog, K.G. & Sörbom, D. (1999). *Interactive LISREL: User's Guide*. Chicago: Scientific Software International.

Hayashi, C. & Hayashi, F. (1982). *A new algorithm to solve PARAFAC-model*. Behaviormetrika, **11**, 49-60.

Holt, D., Scott, A.J. & Ewings, P.D. (1980). *Chi-squared tests with survey data*. Journal of the Royal Statistical Society A, **143**, 303-320.

Jöreskog, K.G., Sörbom, D., du Toit, S.H.C. & du Toit, M. (1999). *LISREL 8: New Statistical Features*. Chicago: Scientific Software International.

Kreft, I.G.G. & de Leeuw, J. (1998). *Introducing Multilevel Modeling*. Sage Publications.

Kreft, I.G.G., de Leeuw, J. & Aiken, L. (1995). *The effect of different forms of centering in hierarchical linear models*. Multivariate Behavioral Research, **30**, 1-22.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Tatsuoka, M. M. (1988). *Multivariate analysis* (2nd ed). New York: Wiley